EST. 2021 EMC
EDITORIAL MAR CARIBE

CIENCIA DE DATOS EN SISTEMAS DE GESTIÓN DE RIESGOS: ENFOQUE HACIA LA MINERÍA DE DATOS

LIBRO DE INVESTIGACIÓN

RICARDO ANTONIO ARMAS JUÁREZ
MÍA LUCIA GUILLEN GUEVARA
JORSI ERICSON JOEL BALCÁZAR GALLO
MARIELA LIZETY CÓRDOVA ESPINOZA
HUGO LUIS CHUNGA GUTIERREZ
JOSE CARLOS FIESTAS ZEVALLOS

ISBN: 978-9915-9706-9-1
9 789915 970691

Ciencia de datos en sistemas de gestión de riesgos: Enfoque hacia la minería de datos

Ricardo Antonio Armas Juárez, Mía Lucia Guillen Guevara, Jorsi Ericson Joel Balcázar Gallo, Mariela Lizety Córdova Espinoza, Hugo Luis Chunga Gutierrez, Jose Carlos Fiestas Zevallos

Primera edición: Octubre, 2024

Editado por:

Editorial Mar Caribe

www.editorialmarcaribe.es

Av. General Flores 547, Colonia, Colonia-Uruguay.

Diseño de cubierta: Yelitza Sánchez Cáceres

Libro electrónico disponible en https://editorialmarcaribe.es/ciencia-de-datos-en-sistemas-de-gestion-de-riesgos-enfoque-hacia-la-mineria-de-datos/

Formato: electrónico

ISBN: 978-9915-9706-9-1

ARK: ark:/10951/isbn.9789915970691

DOI: 10.70288/emc.9789915970691

EDITORIAL MAR CARIBE

«Ciencia de datos en sistemas de gestión de riesgos: Enfoque hacia la minería de datos»

COLONIA DEL SACRAMENTO, URUGUAY

2024

Sobre los autores y la publicación

Ricardo Antonio Armas Juárez

rarmasj@unp.edu.pe

https://orcid.org/0000-0002-0048-2711

Universidad Nacional de Piura, Perú

Mía Lucia Guillen Guevara

mguillen@unaj.edu.pe

https://orcid.org/0000-0001-8641-0833

Universidad Nacional de San Antonio Abad del Cusco, Perú

Jorsi Ericson Joel Balcázar Gallo

jbalcazarg@unp.edu.pe

https://orcid.org/0000-0002-8378-0609

Universidad Nacional de Piura, Perú

Mariela Lizety Córdova Espinoza

mcordovae@unp.edu.pe

https://orcid.org/0000-0002-7205-461X

Universidad Nacional de Piura, Perú

Hugo Luis Chunga Gutierrez

hchungag@unp.edu.pe

https://orcid.org/0009-0005-7063-9912

Universidad Nacional de Piura, Perú

Jose Carlos Fiestas Zevallos

jfiestasz@unp.edu.pe

https://orcid.org/0009-0008-7860-5911

Universidad Nacional de Piura, Perú

Libro resultado de investigación:

Publicación original e inédita, cuyo contenido es resultado de un proceso de investigación realizado antes de su publicación, ha sido revisada por pares externos a doble ciego, el libro ha sido seleccionado por su calidad científica y porque contribuye significativamente en el área del saber e ilustra una investigación completamente desarrollada y completada. Además, la publicación ha pasado por un proceso editorial que garantiza su estandarización bibliográfica y usabilidad.

Índice

Prólogo

A lo largo de los años, los progresos en la tecnología y los enfoques utilizados en los modelos de riesgo y calificación crediticia han posibilitado la integración de tecnologías avanzadas de inteligencia artificial. Estas técnicas han demostrado significativamente la precisión de los algoritmos, aunque pueden carecer de explicaciones exhaustivas para la toma de decisiones.

Poco se ha explorado objetivamente sobre la ciencia de datos como metodología alternativa, la mayoría de las investigaciones se han enfocado en determinar qué tipos de información deberían utilizar las instituciones financieras para evaluar el peligro crediticio de las pequeñas y medianas empresas (PYME). No obstante, existen escasas pruebas disponibles acerca de los beneficios potenciales de aplicar metodologías de inteligencia artificial explicables para evaluar las calificaciones crediticias de las PYME. Por ende, no existen comparaciones significativas entre los enfoques paramétricos tradicionales y estrategias de aprendizaje supervisado en este contexto de investigación.

Hasta ahora se conoce de enfoques paramétricos tradicionales que ejecutan un modelo probit ordenado, mientras que otros métodos emplean un análisis no paramétrico mediante aprendizaje automatizado, denominado Bosque aleatorio histórico (HRF). Durante el análisis de estos dos enfoques, deseamos obtener una comprensión más detallada de su eficacia y ventajas potenciales al momento de evaluar la rentabilidad crediticia de las PYME. La adopción de métodos alternativos para predecir las calificaciones crediticias de las pequeñas y medianas empresas (PYME) ha tenido un impacto significativo en la investigación científica. Estas estrategias comprende técnicas de minería de datos, enfoque en árboles, inteligencia artificial (IA) y enfoques híbridos. Esta

tendencia ha sido estudiada de manera amplia en estudios pasados, entre ellos, los mapas autoorganizados tipo SOM.

Asimismo, la reciente expansión de la digitalización de los mercados financieros, conocida como Fintech, ha dado lugar a progresos tecnológicos sin precedentes y al surgimiento de diversas metodologías estadísticas en el ámbito financiero.

En consecuencia, los bancos han comenzado a utilizar técnicas de estimación avanzadas para evaluar el peligro crediticio en las PYME. No obstante, los reguladores todavía no han estandarizado el uso de algoritmos de aprendizaje automático y de inteligencia artificial en este ámbito. Esto debido al poco dominio en la usabilidad y corrida de datos a través de las técnicas de aprendizaje automático y que puedan generar sesgos que amenacen la integración financiera y ocasionar cuestiones relacionadas con la protección del consumidor, la ética, la privacidad y la transparencia, que son de suma relevancia para los responsables e impulsores de políticas de mercado.

Los autores del libro disciernen en el campo de la predicción de mercados, finanzas y procesos en general y, el aprendizaje automático que se utiliza comúnmente para crear un modelo de evaluación de estos sistemas de riesgos basados en minería de datos y algoritmos. Este modelo tiene como objetivo clasificar una observación de contexto dado como resultado de "fallo" o "no fallo". Básicamente, el modelo examina variables o indicadores independientes que representan la situación de un mercado financiero durante un período de tiempo específico. Primero aprende de estos datos y luego predice, con cierto nivel de confianza, si el sistema enfrentará la quiebra o no. Esta es la ciencia de datos que se emplea en la actualidad para estimar parámetros, ajustar datos y aplicar propiedades estadísticas a las estimaciones.

Capítulo 1

Machine learning y riesgo financiero

Una de las ventajas clave del machine learning es su capacidad para descubrir patrones complejos dentro de datos que antes eran desconocidos o difíciles de modelar. Esta nueva capacidad ha revolucionado el campo, como lo demuestra un estudio reciente realizado por el Banco de Inglaterra en 2021, que enfatiza la creciente importancia del aprendizaje automático en los procesos de supervisión, particularmente en la detección de prácticas de mercado ilícitas. Estos hallazgos no solo refuerzan el uso del ML en otras tareas de supervisión sino que también resaltan su potencial en los procesos de evaluación de riesgos (Lehnert et al., 2018).

Desde la década de 1990, el sector financiero ha experimentado avances significativos en los sistemas de apoyo a las decisiones. Los métodos estadísticos tradicionales, como las regresiones lineales o logísticas, han desempeñado un papel crucial en estos sistemas y se han utilizado ampliamente en modelos analíticos financieros. Sin embargo, en los últimos años, el aprendizaje automático (ML) ha ganado considerable fuerza como herramienta preferida, principalmente debido a la gran cantidad de datos que se recopilan y la creciente potencia computacional disponible.

1.1 Las medidas de evaluación de riesgos

El tema de la creación de una metodología de evaluación de riesgos que sea universalmente aceptada ha sido durante mucho tiempo un tema de intenso debate entre los investigadores en este campo en particular. A medida que la tecnología continúa avanzando y convirtiéndose en una parte integral de nuestra vida diaria, las metodologías utilizadas para la evaluación de riesgos también

están evolucionando y volviéndose más sofisticadas e intrincadas. Esto se debe principalmente a la necesidad de medir y evaluar con precisión los riesgos que enfrentan los bancos desde diferentes ángulos y puntos de vista. Al mismo tiempo, la aparición de nuevos requisitos regulatorios también sirve como catalizador para el desarrollo de nuevos enfoques y técnicas en el ámbito del análisis de riesgos.

Numerosos investigadores, han propuesto varios enfoques para evaluar la probabilidad de quiebra. Entre estos métodos, el modelo de valoración de opciones de Black-Scholes-Merton, propuesto por Hillegeist et al. (2004), ha demostrado un rendimiento superior en comparación con otras dos medidas confiables y ampliamente reconocidas, a saber, la puntuación Z y O. Los autores enfatizan la importancia de adoptar una metodología estandarizada para garantizar la comparabilidad entre diferentes instituciones.

Mucha de la literatura existente sobre clasificación de riesgos utiliza un enfoque de clasificación binaria, en el que clasifican a un banco como "en quiebra" o "no en quiebra". Esta clasificación se basa en una variedad de ratios financieros, generalmente obtenidos a partir de conjuntos de datos disponibles públicamente o de fuentes indirectas.

En esta sección del libro, examinamos el enfoque de clasificación descrito en una metodología altamente reconocida y universalmente aceptada para evaluar el riesgo conocida como Proceso de Evaluación y Revisión Supervisora (SREP), que fue desarrollado por el Banco Central Europeo (BCE) en colaboración con el Banco Nacional Competente. Autoridades (ANC). Este proceso sirve como un medio para que los supervisores evalúen y midan periódicamente el nivel de riesgo asociado con los bancos individuales, considerando diversos aspectos como liquidez, crédito, mercado, operativo y rentabilidad. Nuestro estudio avala

la utilización del Sistema Automático de Evaluación de Riesgos (ARS) como medio de clasificación de riesgos, que posteriormente se perfecciona mediante la aplicación de juicio de expertos.

Este enfoque particular se basa en datos de supervisión reales obtenidos de la directiva de la Autoridad Bancaria Europea (EBA) para la Aplicación de Normas Técnicas, que cae bajo la jurisdicción del Mecanismo Único de Supervisión (MUS) establecido por la Comisión Europea. Los datos recopilados se utilizan para categorizar cada entidad según su nivel de riesgo, empleando el sistema automatizado de evaluación de riesgos del proceso PRES. Las observaciones abarcan un período comprendido entre 2014 y marzo de 2021. Para garantizar la precisión, los datos utilizados en este estudio han sido objeto de una validación exhaustiva y se ha demostrado que tienen una fuerte correlación con la evaluación de las capacidades de riesgo de liquidez.

1.2 El aprendizaje automático para la evaluación de riesgos

El proceso de evaluación de riesgos implica principalmente un análisis cuantitativo, que con frecuencia se refina y ajusta en función de la opinión de expertos. Recientemente, ha habido un interés creciente en la aplicación de técnicas de aprendizaje automático en la evaluación de riesgos por parte de los bancos centrales. Este interés no solo lo comparten las NCA y otras organizaciones, sino también los académicos que han reconocido los beneficios potenciales de incorporar métodos de aprendizaje automático en este campo.

La evaluación de riesgos ha sido reconocida como un enfoque crucial en la gestión eficaz de los recursos financieros desde principios de la década de 2000. A principios de la década de destacaron los modelos basados en árboles que son más apropiados para tareas de predicción utilizando datos estructurados en

comparación con las redes neuronales artificiales (RNA). Este hallazgo ha sido respaldado por varios estudios a lo largo de los años.

Los avances recientes en tecnología han contribuido en gran medida al desarrollo de modelos más avanzados, como los modelos de aprendizaje profundo (DL), y nuevos métodos de conjunto como el aumento de gradiente extremo (XGBoost). Estos modelos tienen la capacidad de capturar efectivamente la complejidad de diversos fenómenos. Inicialmente, DL ganó prominencia en 2012 con la introducción de ImageNet, pero no fue hasta 2016 que comenzó a utilizarse en el campo de la evaluación de riesgos financieros. DL se muestra muy prometedor como herramienta en la evaluación de riesgos, particularmente en la evaluación del riesgo crediticio. Los autores proponen ampliar este enfoque a otras perspectivas de riesgo, aunque reconocen que la falta de interpretabilidad de los modelos DL sigue siendo un obstáculo importante para su adopción generalizada.

Al mismo tiempo, varios estudios de investigación han demostrado la notable precisión con la que los modelos de aprendizaje profundo pueden adaptarse a datos estructurados. En su estudio, Petropoulos et al. (2018) adoptan un enfoque de supervisión para explorar más a fondo la aplicación de técnicas avanzadas de aprendizaje automático (ML). Los autores desarrollan un sistema de alerta temprana (EWS) para la predicción del riesgo crediticio, utilizando datos de préstamos corporativos proporcionados por los bancos griegos (específicamente, el Banco de Grecia; período: 2005-2015). Si bien XGBoost surgió como el modelo de mayor rendimiento, las redes neuronales profundas (DNN) también mostraron resultados prometedores.

En el campo de la predicción de quiebras, el aprendizaje automático se utiliza comúnmente para crear un modelo de evaluación de riesgos. Este modelo

tiene como objetivo clasificar una observación de contexto dado como resultado de "fallo" o "no fallo". Básicamente, el modelo examina varias variables o indicadores independientes que representan la situación de un banco durante un período de tiempo específico. Primero aprende de estos datos y luego predice, con cierto nivel de confianza, si el banco enfrentará la quiebra o no.

Al considerar las operaciones comerciales, es de suma importancia obtener una comprensión integral de cómo los bancos, así como las autoridades nacionales competentes y otros organismos relevantes, se ajustan y responden a estos avances. Nuestro enfoque se centra particularmente en examinar las estrategias empleadas por los bancos centrales al utilizar tecnologías de vanguardia para reforzar sus capacidades analíticas, con un énfasis específico en el ámbito de la evaluación de riesgos.

Los NCAs desempeñan un papel fundamental en la realización de análisis rigurosos de política económica. Esto implica examinar las condiciones y tendencias económicas actuales, evaluar los impactos potenciales de diversas opciones y sobre el papel de los macroeconometristas dentro de las instituciones políticas, asimismo elaborar recomendaciones y consejos basados en evidencia a los formuladores de políticas, ayudando a guiar la formulación de políticas económicas efectivas, en esencia:

- Resumir y analizar datos.
- Pronosticar las principales variables macroeconómicas.
- Realizar un examen exhaustivo de los riesgos potenciales y sopesar las incertidumbres involucradas.
- Realizar análisis estructurales/causales y análisis de escenarios.
- Tomar decisiones, comunicarlas y justificarlas ante la opinión pública.

Se han realizado estudios limitados sobre la evaluación de riesgos utilizando técnicas de aprendizaje automático y se han centrado principalmente en la utilización de conjuntos de datos públicos o indirectos. Estos primeros estudios han sentado las bases para la aplicación específica de los bancos centrales. Sin embargo, en lo que respecta a la supervisión, los datos utilizados son confidenciales y los procesos se rigen por la legislación de la UE. Por lo tanto, es más probable que este tipo de investigación se realice en colaboración con las autoridades nacionales competentes (NCA).

El objetivo fundamental del aprendizaje automático (Machine Learning, ML) es extraer predicciones de los datos subyacentes (o Big Data). Por lo general, los algoritmos de aprendizaje automático se aplican a los datos para obtener información de ellos (Ban et al., 2018; Evans, 2015). En este caso estamos utilizando datos transversales, que pueden capturarse en cualquier momento. Utilizando información de circunstancias observadas previamente (datos transversales), los algoritmos de ML pueden predecir valores relativos a eventos que aún no se han producido.

La transformación de datos incluye la limpieza de datos, la aplicación de una estrategia para tratar los valores que faltan y el proceso de selección de características. En la fase experimental, comparamos tres enfoques diferentes para evaluar los algoritmos de ML para esta tarea: la clásica división de entrenamiento y prueba, la validación cruzada más precisa y el marco TPOT AutoML (Fernández et al., 2000). A continuación, utilizamos la puntuación f1 y las matrices de confusión para comparar los resultados y, por último, seleccionamos el mejor modelo. En un uso futuro, este modelo puede desplegarse como un Sistema de Alerta Temprana haciendo predicciones para el nivel de riesgo de liquidez.

1.3 El aprendizaje automático y riesgo de crédito

La evaluación de las calificaciones crediticias y el riesgo crediticio de las empresas es un tema crucial que ha provocado un amplio debate en el mundo académico, tanto en contextos teóricos como empíricos. Este tema tiene gran importancia para la industria, así como para los organismos reguladores y de supervisión del sistema financiero. Actúa como una herramienta vital para garantizar la eficacia de la asignación del mercado y la eficiencia de los intermediarios financieros.

Puede resultar particularmente difícil evaluar las calificaciones de las pequeñas y medianas empresas (PYME), ya que normalmente no cotizan en bolsa en el mercado de valores. Esto plantea un obstáculo importante, teniendo en cuenta que las PYME constituyen una parte sustancial del panorama empresarial, especialmente en Europa. Estas empresas a menudo enfrentan desequilibrios de información considerables, que complican aún más la tarea de determinar calificaciones crediticias confiables.

Dentro de este contexto particular, numerosos académicos han dedicado sus esfuerzos a explorar el potencial de utilizar fuentes alternativas de información. Esto incluye la consideración de información blanda que puede derivarse de la práctica de la banca relacional intensiva. Este concepto ha atraído la atención no sólo de los académicos sino también de los organismos reguladores, enfatizando su importancia. Sin embargo, es importante señalar que la eficacia de la información confidencial para mejorar la actividad crediticia de un banco puede no siempre ser consistente. Además, la transferibilidad de dicha información puede resultar difícil dentro de estructuras organizativas complejas. Estos factores resaltan la necesidad de buscar soluciones alternativas que permitan la utilización de información dura. Al hacerlo, se pueden obtener

calificaciones crediticias más precisas, beneficiando así a las pequeñas y medianas empresas (PYME) que carecen de información.

A lo largo de los años, los avances en la tecnología y los métodos utilizados en los modelos de riesgo crediticio y calificación crediticia han permitido la integración de técnicas sofisticadas de inteligencia artificial. Estas técnicas han mejorado enormemente la precisión de la calificación crediticia, aunque pueden carecer de explicaciones claras para sus decisiones. Si bien unos pocos estudios han explorado metodologías alternativas, la mayoría de la literatura se ha centrado en determinar qué tipos de información deberían utilizar las instituciones financieras para evaluar el riesgo crediticio de las pequeñas y medianas empresas (PYME). Sin embargo, hay evidencia limitada disponible sobre los beneficios potenciales de utilizar metodologías de IA explicables para estimar las calificaciones crediticias de las PYME. Además, muy pocos estudios han comparado los enfoques paramétricos tradicionales con técnicas de aprendizaje automático en este contexto.

Para abordar esta brecha en el sector finanzas y proporcionar un análisis más completo, un método es un enfoque paramétrico tradicional que utiliza un modelo probit ordenado, mientras que el otro método utiliza un enfoque no paramétrico a través de una técnica de aprendizaje automático llamada Bosque aleatorio histórico (HRF) (Broby,2022). Al explorar estos dos enfoques, pretendemos obtener una comprensión más profunda de su eficacia y ventajas potenciales a la hora de evaluar la solvencia crediticia de las PYME.

El uso de metodologías alternativas para predecir las calificaciones crediticias de las pequeñas y medianas empresas (PYME) ha ganado un importante impulso en la literatura existente. Estas metodologías incluyen técnicas de minería de datos, metodología basada en árboles, inteligencia

artificial (IA) y métodos híbridos. Esta tendencia ha sido explorada ampliamente en estudios previos. Además, el reciente aumento de la digitalización de los mercados financieros, conocida como Fintech, ha dado lugar a avances tecnológicos sin precedentes y al surgimiento de diversas metodologías estadísticas en el sector financiero (Guerra et al., 2022).

Como resultado, los bancos han comenzado a considerar técnicas de estimación avanzadas para evaluar el riesgo crediticio en las PYME. Sin embargo, los reguladores aún no han autorizado completamente el uso de algoritmos de aprendizaje automático y de IA en este contexto. Esto se debe en parte a la preocupación de que las técnicas de aprendizaje automático puedan introducir sesgos que pongan en peligro la inclusión financiera y den lugar a cuestiones relacionadas con la protección del consumidor, la ética, la privacidad y la transparencia, que son de suma importancia para los supervisores y formuladores de políticas.

Asimismo, los resultados del aprendizaje automático pueden ser difíciles de interpretar y comunicar a las diferentes partes interesadas. En consecuencia, la estimación de las calificaciones crediticias de las PYME ha vuelto a ganar atención recientemente, con la ayuda de la disponibilidad de nuevas técnicas estadísticas y diversas fuentes de datos que complementan la información existente sobre las PYME, lo que resulta en una evaluación más precisa de su riesgo crediticio.

Los resultados demuestran que el enfoque de bosques aleatorios históricos (HRF) supera al modelo probit ordenado tradicional en la evaluación del riesgo crediticio de las pymes. Esto sugiere que las metodologías avanzadas de aprendizaje automático pueden ser adoptadas con éxito por los bancos para predecir el riesgo crediticio de las pymes.

Vale la pena señalar que en este contexto, una posible vía para investigaciones futuras, lo constituye la capacidad predictiva del modelo HRF, podría probarse en condiciones de aumento de precios, aumento de las tasas de interés, shocks externos inesperados (por ejemplo, COVID-19) y otras perturbaciones importantes en el entorno empresarial. incluidos aquellos provocados por fenómenos meteorológicos extremos exacerbados por el cambio climático.

1.4 Aprendizaje automático en predicción de riesgos

La gestión de ganancias es una estrategia frecuentemente empleada por los gerentes de los sectores de contabilidad y seguros. En ambas industrias, la generación de ingresos depende del cobro anticipado de las primas de seguros, mientras que los costos se generan mediante la compensación de reclamaciones futuras. En consecuencia, resulta imperativo que las compañías de seguros establezcan una provisión monetaria suficiente para dar cuenta de reclamaciones inminentes. Este elemento crucial, conocido como "reserva para siniestros", se registra como un pasivo en los balances de las aseguradoras.

La Asociación Nacional de Comisionados de Seguros (NAIC) ha implementado una regulación obligatoria que todas las compañías de seguros deben cumplir, que es la adhesión a los Principios Estatutarios de Contabilidad (SAP). En consecuencia, las aseguradoras están obligadas a informar y modificar consistentemente sus estimaciones de reservas para pérdidas anualmente, ya que estas estimaciones están directamente asociadas con la liquidación de siniestros. Al igual que los Principios de Contabilidad Generalmente Aceptados (GAAP), los Principios de Contabilidad Estatutarios (SAP) son un procedimiento contable que se adapta específicamente a la industria de seguros. Por lo tanto, para garantizar la uniformidad y estandarización dentro del sector de seguros, la

Asociación Nacional de Comisionados de Seguros (NAIC) ha ordenado que todas las compañías de seguros cumplan con los Principios Estatutarios de Contabilidad (SAP).

A medida que se informan y procesan los reclamos de seguros, las aseguradoras ajustan periódicamente sus estimaciones iniciales de los fondos que han reservado para cubrir esos reclamos. Esta práctica permite a las aseguradoras ejercer un cierto nivel de discreción en la gestión de sus cuentas financieras. Comparar la reserva para siniestros revisada con la estimación inicial sirve como una forma sencilla de medir hasta qué punto las aseguradoras manipulan sus ganancias. Esta medida se emplea comúnmente como indicador de las estrategias de las aseguradoras para administrar sus ganancias reportadas.

Normalmente, a la hora de declarar el importe final, los directivos de la empresa tienen la autoridad de elegir en función del rango aceptable sugerido por los actuarios utilizando métodos cuantitativos. A su discreción, los administradores pueden establecer estimaciones de reservas para siniestros más altas o más bajas para diversos fines, como suavizar las ganancias, gestionar los impuestos, evadir controles regulatorios o compensar a los directores. Sin embargo, estas sobreestimaciones o subestimaciones intencionales de pérdidas futuras dan lugar a errores de estimación.

Las compañías de seguros suelen confiar en métodos cuantitativos tradicionales para determinar las reservas para pérdidas. Estos métodos incluyen el método de juicio, el método de pago promedio, el método de escalera de cadena y el método de escalera de cadena estocástica. Este artículo explora específicamente las limitaciones de estos modelos generales al calcular el "error" en las reservas para pérdidas, en lugar de la fórmula matemática real para calcular el monto inicial de la "reserva para pérdidas". Es importante señalar que,

dado que las reservas para pérdidas son estimaciones, ningún método puede predecir con precisión las pérdidas futuras exactas, lo que genera errores en las reservas calculadas. El modelo actual se basa en una estimación lineal, pero ha enfrentado críticas por parte de los investigadores debido a sus limitaciones y desventajas estadísticas. Mack, por ejemplo, sostiene que los estimadores tradicionales dependen en gran medida de unos pocos factores y no son adecuados para todas las situaciones. De manera similar, Meyers demuestra que el modelo de estimación lineal conduce a mayores errores de predicción y requiere superar las dependencias entre variables.

Con base en el conjunto de investigaciones actual, la discrepancia en las reservas para siniestros de las aseguradoras se determina restando las estimaciones iniciales de las reservas para siniestros en el año t del total de siniestros incurridos en el año $t + n$, donde n es igual a cinco. El cálculo del error de reserva se presenta en la ecuación:

$$Reserve\ Error_{i,t} = Incurred\ Losses_{i,t} - Incurred\ Losses_{i,t+n}$$

En la literatura, existen dos enfoques para calcular el error de reserva para pérdidas. Weiss lo mide comparando la pérdida total incurrida por una empresa con las pérdidas acumuladas desarrolladas y pagadas en el futuro. Por otro lado, Samaniego y Mongrut (2014). determinan las diferencias entre las pérdidas totales incurridas y una estimación revisada de las pérdidas que se incurrirán en el futuro. De acuerdo con estudios previos, adoptamos el método KFS ya que no depende del desarrollo de las pérdidas, lo que significa que no depende de cuándo se pagan las pérdidas. Además, las últimas reclamaciones pagadas no representan necesariamente todas las reclamaciones pagadas.

El cuerpo de literatura existente comúnmente emplea el modelo de regresión lineal fundamental presentado como ecuación:

$$RE_{i,t} = \beta_0 + \beta_1^{'} X_{i,t} + \beta_2^{'} I_t + \varepsilon_{i,t}$$

para evaluar las desviaciones en las reservas para pérdidas de las aseguradoras, explorando así una multitud de temas de investigación.

1.5 Mínimos cuadrados ordinarios (OLS)

El modelo actual se basa en métodos cuantitativos y estimación lineal para calcular las reservas para pérdidas de las aseguradoras y su error; Sin embargo, los investigadores a menudo lo critican por sus limitaciones y deficiencias estadísticas. Existe un método simple de escalera de cadenas en el que las estimaciones se basan en algunos factores y que el ratio de siniestralidad final es incierto. De hecho, las estimaciones ya no son suficientes debido a cambios futuros inesperados.

Otros estudios también utilizan otros enfoques, como el modelo de simulación, la técnica bootstrap, el método bayesiano en un entorno estocástico, y muestran evidencia de que la estimación de reservas de las aseguradoras comunes puede conducir a estimaciones sesgadas. Así también existen modelos de simulación para medir la expectativa y la varianza de los errores de predicción de los métodos de reserva de pérdidas, en donde los factores simples conducen a resultados sesgados y errores de predicción mayores.

1.6 El error en la medición

El proceso de estimar el futuro de una empresa basándose en su historial y comportamiento de accidentes pasados es difícil debido a la presencia de errores de medición. Así, analizando el impacto del error de medición en la estimación de las acumulaciones discrecionales, se tiene que la variable

dependiente puede estar sesgada si las variables utilizadas para la división están correlacionadas. Asimismo es cuestionable uso de devengos, ya que pueden llevar a conclusiones engañosas.

La principal preocupación es que estas estimaciones contienen errores en factores relacionados con las variables contribuyentes. Además, si las variables utilizadas para la división están correlacionadas con variables relacionadas con actividades específicas de la empresa, el error de medición en las estimaciones de acumulación puede llevar a conclusiones incorrectas sobre la existencia de gestión de ganancias. Por lo tanto, una prueba insesgada requiere que el error de medición en la estimación acumulada no esté relacionado con las variables utilizadas para dividir en el diseño de la investigación.

1.7 Lo lineal y lo no lineal

El método lineal es el enfoque más comúnmente utilizado para estimar parámetros y se emplea ampliamente para ajustar datos y aplicar propiedades estadísticas a las estimaciones. Por otro lado, la regresión no lineal ocurre cuando las derivadas del modelo dependen de uno o más parámetros que generalmente requieren un proceso iterativo. El procedimiento funciona de la siguiente manera: los usuarios establecen los valores de los parámetros iniciales y luego el software ajusta estos valores para mejorar el modelo. Una vez que el algoritmo alcanza sus mejoras máximas y no se pueden lograr más mejoras, el ajuste se considera convergente. Algunos técnicos utilizan simulaciones iterativas para determinar el vector en la dirección de la mayor reducción del error de la red, que se caracteriza por mínimos locales. Además, un esquema iterativo es beneficioso para modelos multiplicativos no lineales que involucran factores diagonales.

Para comprender mejor la dinámica caótica en situaciones prácticas, se propone la adopción de la ciencia no lineal. Esto se debe a que las complejidades del mundo no pueden representarse con precisión mediante una combinación lineal y la suposición de dimensiones independientes. En apoyo de esta noción, se introdujo un modelo adaptativo no lineal que incorpora términos de interacción para abordar problemas complejos. Este enfoque también está respaldado por otras disciplinas, incluidas la contabilidad y la economía.

En esencia, las regresiones lineales se basan en varios supuestos relacionados con:

- la linealidad de la relación,
- la distribución del término de error,
- la independencia de las variables explicativas,
- la ausencia de autocorrelación y
- la constancia de la varianza del término de error.

Estos supuestos proporcionan una base sólida para estimar los parámetros y obtener la mejor línea de ajuste para los datos. Con estos supuestos establecidos, los modelos de regresión lineal pueden estimar los parámetros minimizando la suma residual de cuadrados, que es la suma de las diferencias al cuadrado entre los valores observados y predichos. El método permite determinar la línea de mejor ajuste que representa la relación entre la respuesta y las variables explicativas.

En relación al cuarto supuesto (que no existe autocorrelación en los términos de error), la autocorrelación se refiere a la correlación entre términos de error en diferentes puntos en el tiempo o el espacio. Si existe autocorrelación,

sugiere que los términos de error no son verdaderamente independientes y pueden conducir a estimaciones de coeficientes sesgadas. Por lo tanto, es importante que los términos de error no estén relacionados entre sí.

Por último, las regresiones lineales suponen que la varianza del término de error es constante. Este supuesto, conocido como homocedasticidad, significa que la variabilidad del término de error es consistente en todos los niveles de las variables explicativas. Si existe heterocedasticidad, es decir, la variabilidad del término de error no es constante, puede afectar la precisión y confiabilidad de las predicciones del modelo. El tercer supuesto es que las variables explicativas no están relacionadas entre sí, lo que significa que no hay multicolinealidad. La multicolinealidad ocurre cuando dos o más variables explicativas están altamente correlacionadas, lo que puede conducir a estimaciones de coeficientes inestables y poco confiables. Para garantizar resultados precisos, es importante que las variables explicativas sean independientes entre sí.

La estimación lineal es efectiva para determinar los efectos incrementales de los cambios en los retrocesos sobre las variables de resultado, pero no es efectiva para determinar el impacto general de estos cambios. En los modelos básicos de OLS con efectos lineales, los coeficientes estimados son equivalentes a los efectos marginales, que miden directamente la influencia de los retrocesos sobre las variables de resultado. Sin embargo, este no es el caso en modelos más complejos cuando se trata de determinar el impacto de los reveses. En ausencia de factores no lineales, todas las variables de impacto se consideran independientes, lo que facilita pasar por alto términos de orden superior y da como resultado una comprensión limitada de los términos de interacción a medida que aumenta la complejidad del modelo.

Debido a la naturaleza desconocida de la forma funcional del modelo de reserva para pérdidas futuras, es imperativo emplear un modelo no lineal. Sin embargo, existe una crítica en torno al uso de la no linealidad, ya que se argumenta que no es un supuesto esencial, ya que la no linealidad ya se considera un término de error en los métodos científicos lineales generales. En este sentido, los modelos lineales con diferentes especificaciones pueden considerarse como un enfoque alternativo, pero en los casos en los que existe un alto grado de complejidad, se hace necesario utilizar modelos no lineales.

1.8 Otros supuestos

En numerosos casos, la estimación de mínimos cuadrados ordinarios no cumple ciertos supuestos. Por ejemplo, la varianza del error puede no exhibir homocedasticidad, lo que significa que la variabilidad de los errores no es constante en todos los niveles de la variable independiente. Además, los errores son propensos a la autocorrelación, lo que implica que existe una correlación entre términos de error consecutivos. En consecuencia, las estimaciones de OLS pueden no ser tan eficientes como se desea, lo que sugiere que podrían existir estimaciones alternativas con discrepancias más pequeñas. No obstante, las estimaciones de Operation SurvivalS siguen siendo imparciales y consistentes, lo que significa que los valores previstos de estas estimaciones son equivalentes a los valores reales de los parámetros.

Los mínimos cuadrados generalizados (GLS) es una versión más avanzada de los mínimos cuadrados ordinarios (OLS) que es particularmente eficaz cuando los errores en un modelo de regresión no están correlacionados y tienen varianzas iguales. GLS se emplea comúnmente para estimar parámetros desconocidos en el análisis de regresión lineal cuando hay heterocedasticidad en el conjunto de datos y/o autocorrelación entre las observaciones. En el contexto de los errores

de reserva de pérdidas de las aseguradoras, se ha observado que estos errores parecen estar distribuidos aleatoriamente y no dependen del año del accidente. En consecuencia, esto indica la presencia de heterocedasticidad. Se propone además que GLS puede ser útil para ajustar las distribuciones de pérdidas agregadas y concluye que el estimador GLS es óptimo, aunque el estimador OLS sigue siendo consistente.

Para analizar las diferentes motivaciones detrás de la gestión de ganancias, se emplean el método del Cuadrado Generalizado Feasible (FGLS) para estimar el error en las reservas para pérdidas de las aseguradoras. Se elige este método en lugar de los mínimos cuadrados ordinarios (OLS) comúnmente utilizados porque FGLS permite la inclusión de términos de error serial heterocedásticos y de panel, que se observan con frecuencia en las estimaciones de reservas de las aseguradoras. Aprovechando el FGLS, los investigadores pueden examinar eficazmente la presencia de correlación serial en estas estimaciones de reservas.

Además, las regresiones aparentemente no relacionadas (SUR) abarcan una colección de ecuaciones de regresión que parecen no tener conexión o relación aparente entre sí. Un beneficio notable de este modelo es que los errores en estas ecuaciones lineales muestran correlación dentro de un año específico, mientras que permanecen no correlacionados en diferentes años. En SUR, los predictores son exógenos, lo que significa que no están influenciados por ninguna otra variable. Para mejorar la eficacia del modelo, los parámetros varían entre las diferentes ecuaciones. Esta flexibilidad permite una mayor eficiencia en el análisis.

1.9 Las predicciones de aprendizaje automático

Los algoritmos de aprendizaje automático poseen cinco características clave:

- En primer lugar, implican una multitud de estructuras sistémicas, cada una de las cuales comprende múltiples factores que influyen.
- En segundo lugar, estas estructuras operan sobre la base de reglas deterministas, estableciendo así un marco predecible.
- En tercer lugar, dentro de estas estructuras existen correlaciones interdependientes entre una amplia gama de factores, lo que aumenta aún más su complejidad.
- En cuarto lugar, los fenómenos reales sufren diversas alteraciones que contribuyen al carácter dinámico del sistema.
- Por último, la presencia de características fractales permite predecir estas alteraciones, destacando el potencial para anticipar cambios en el sistema.

1.9.1 RNA

La minería de datos, un concepto que todavía es relativamente nuevo en el campo de la literatura empresarial, implica la utilización de características dinámicas dentro de un intrasistema no lineal para facilitar las interacciones interdependientes. Ye, un destacado investigador, ha empleado eficazmente técnicas de extracción de datos para examinar y analizar diversos patrones de datos, incluidas la clasificación y la predicción. En su investigación, emplearon un método de análisis de agrupamiento jerárquico de enlaces en la habitación con fines de clasificación, mientras que, para fines de predicción, utilizaron un innovador marco analítico dinámico no lineal conocido como red neuronal artificial (Goodell et al., 2021).

El concepto de red neuronal se introdujo inicialmente en la informática en 1954, donde se creó como un programa de software cuyo objetivo era imitar la intrincada estructura del cerebro humano y su sistema neuronal interconectado.

Para obtener una comprensión más profunda de fenómenos sociales y económicos complejos. Este enfoque permite adaptar un modelo para clasificar, controlar y optimizar datos de manera eficiente, así como hacer predicciones sobre patrones futuros en diversos campos de estudio, incluidas las ciencias naturales y sociales.

En su estudio, García y Morales (2016) presentan un enfoque novedoso para predecir la insolvencia de las aseguradoras mediante el empleo de una red neuronal (NN), que es un tipo de modelo de inteligencia artificial. Los autores sostienen que la utilización de NN en este contexto es muy ventajosa por varias razones. En primer lugar, la NN incorpora diversos factores económicos que influyen en el riesgo de insolvencia de las aseguradoras. Al emplear una canalización ponderada como punto de partida, el modelo puede acomodar fácilmente iteraciones y actualizaciones a medida que haya nuevos datos disponibles o se produzcan cambios en el futuro. Esta adaptabilidad de la NN la convierte en una herramienta valiosa como señal de alerta temprana para predecir la insolvencia de las aseguradoras. Los investigadores proporcionan evidencia empírica para respaldar su afirmación, demostrando que la implementación de NN mejora significativamente la precisión de la predicción de la insolvencia para las aseguradoras de responsabilidad civil. Además, los autores destacan que se han observado resultados positivos similares en otros estudios centrados en la industria de los seguros de vida y las reclamaciones por lesiones corporales de automóviles. Estos hallazgos subrayan aún más la efectividad y aplicabilidad de NN en el ámbito de la predicción de la insolvencia de las aseguradoras.

El proceso de cálculo de los errores de reserva para pérdidas para las aseguradoras es una tarea compleja e intrincada. Implica numerosos factores que

están inextricablemente vinculados entre sí. Esta tarea requiere no sólo un enfoque cuantitativo sino también una consideración subjetiva y discreción de gestión. Existen disparidades significativas entre lo que se puede observar y las respuestas resultantes, así como interacciones intrincadas entre diversas variables. Para abordar esta complejidad, hemos optado por emplear en nuestra investigación el primer algoritmo de aprendizaje automático, la red neuronal artificial.

1.9.2 SVM

Support Vector Machine (SVM) es una técnica de aprendizaje automático bien establecida que continúa utilizándose ampliamente en diversos campos, incluidos las ciencias sociales, los negocios, la medicina y los lenguajes naturales. A pesar de ser un método antiguo, SVM sigue siendo valioso debido a su versatilidad para manejar diferentes complejidades, como las lineales, no lineales y lagrangianas. Este uso generalizado es un testimonio de la utilidad de SVM y atrae a numerosos investigadores en el campo.

El principio fundamental de SVM implica encontrar un hiperplano que maximice los márgenes entre clases potenciales. En el caso de SVM lineal, hay dos planos presentes, uno encima y otro debajo del plano lineal. Al utilizar este hiperplano basado en suposiciones lineales, SVM categoriza dimensiones de manera efectiva y hace predicciones. Este enfoque flexible permite que SVM se adapte a supuestos de complejidad como funciones lineales, polinómicas y sigmoideas (Huang et al., 2005).

SVM sigue siendo una herramienta valiosa para hacer predicciones en escenarios financieros, incluido el financiamiento de la cadena de suministro, la evaluación del riesgo crediticio y las calificaciones crediticias. SVM también ha demostrado ser eficaz en la previsión de riesgos en los sectores financiero y

empresarial. Además, los investigadores han descubierto que SVM es un modelo fiable para predecir el movimiento del mercado de valores. Esto ha llevado a la adopción de SVM en la evaluación de riesgos con fines financieros. Dado que la reserva para pérdidas de la aseguradora es un elemento crucial de la gestión de riesgos, SVM puede considerarse como otro enfoque de aprendizaje automático para este estudio.

Capítulo 2

Ciencia de datos, herramientas y bases de datos aplicadas a procesos de logística

Boosting es una técnica de aprendizaje automático que ha demostrado ser eficaz en la industria financiera. Se incluye en la categoría de métodos de conjunto, que implican la combinación de técnicas de clasificación y regresión. El concepto detrás del impulso es mejorar el rendimiento de la máquina amplificando el aprendizaje débil para convertirlo en un aprendizaje fuerte. En otras palabras, si bien la máquina puede identificar patrones débiles en el conjunto de datos de entrenamiento, puede utilizar este conocimiento para hacer predicciones precisas en el conjunto de datos de prueba. Este enfoque es particularmente útil cuando se trata de datos limitados o recursos de datos que no son grandes, como conjuntos de datos de encuestas, donde la resistencia de la máquina al aprendizaje débil se puede superar mediante técnicas de aumento.

En investigaciones financieras recientes, se ha considerado importante la fuerza de la utilización, teniendo en cuenta impulsos como la lealtad financiera del cliente, la solvencia de la empresa y el mercado futuro. Estos impulsos han demostrado ser valiosos para evaluar el error en la reserva de pérdidas entre las aseguradoras. Para evaluar este error, en este estudio se emplearon dos tipos de refuerzos: refuerzo adaptativo y refuerzo de gradiente. Adaptive Boosting, que se introdujo a finales de los años 80, sigue utilizándose en áreas financieras como la predicción del mercado financiero. Por otro lado, Gradient Boosting implica la adición de funciones de parámetros residuales y también ha tenido éxito en contextos financieros.

2.1 Root media de error cuadrado (RMSE)

Para evaluar la precisión de modelos de predicción como redes neuronales artificiales (ANN), aumento de gradiente, aumento adaptativo y máquinas de vectores de soporte (SVM), compararemos el error cuadrático medio (RMSE) entre la estimación lineal (OLS) y el estimaciones proporcionadas por estos algoritmos de aprendizaje automático. Tanto el modelo de estimación OLS como los distintos modelos de aprendizaje automático utilizarán el mismo conjunto de variables. Se emplea RMSE y MAE (error absoluto medio) para evaluar la precisión de las predicciones. Las siguientes ecuaciones ilustran el cálculo de RMSE y MAE, aunque vale la pena señalar que RMSE a veces puede magnificar los errores debido al uso de términos al cuadrado, razón por la cual a menudo se recomienda MAE. para ser utilizado junto con RMSE.

$$RMSE = \sqrt{\frac{\Sigma\left(P-\widehat{P}\right)^2}{n-1}}$$

$$MAE = \frac{\Sigma\left|P-\widehat{P}\right|}{n-1}$$

En este contexto, P significa la probabilidad de que se hayan observado errores de contabilización de pérdidas. Por otro lado, P-hat denota la probabilidad predicha por varios métodos de estimación. Además, la barra P representa el promedio de los errores de pérdida inversa y n representa el número total de observaciones en diferentes grupos. RMSE, o raíz del error cuadrático medio, cuantifica la diferencia entre la probabilidad observada y la probabilidad predicha. En consecuencia, un RMSE más pequeño indica una predicción de probabilidades más precisa y eficiente.

2.2 La segmentación de clientes utilizando la agrupación difusa

La segmentación de clientes se originó en el campo del marketing y la investigación con el fin de analizar y dividir a los clientes potenciales en un mercado de ventas en función de diferentes criterios. Esta división da como resultado la creación de grupos de clientes que son internamente similares pero externamente diversos, lo que permite actividades de marketing específicas. Las instituciones financieras y muchos departamentos comerciales han reconocido la importancia de segmentar a los clientes en diferentes grupos objetivo para mejorar las consultas y los servicios.

Al identificar clientes similares, los productos y servicios se pueden adaptar para satisfacer sus necesidades y expectativas específicas. Sin embargo, implementar un enfoque de este tipo requiere comprender la importancia de la segmentación orientada al cliente y el uso de métodos apropiados para identificar diferentes segmentos de clientes. El análisis de conglomerados tradicional no difuso se ha utilizado ampliamente para la segmentación de mercados y clientes. Este método ayuda a identificar segmentos de clientes específicos y asigna a cada cliente a un grupo determinado. Umbrales claros separan estos grupos, lo que permite una categorización precisa. Sin embargo, el análisis de conglomerados tradicional tiene sus limitaciones.

Las técnicas de agrupamiento difuso, por otro lado, ofrecen una comprensión más matizada de la segmentación de clientes. Al asignar a los clientes a diferentes segmentos con distintos grados de pertenencia, se puede lograr una mejor comprensión de sus necesidades y preferencias. Este enfoque se aplicó con éxito a un banco alemán, que ofrecía productos como cuentas corrientes, tarjetas de crédito y planes de inversión. La agrupación difusa

permitió realizar un perfil más preciso de los clientes y una mejor comprensión de sus requisitos específicos en relación con los productos y servicios del banco.

La experiencia analiza un proyecto que implicó analizar datos de clientes específicos proporcionados por un banco en Alemania con fines de segmentación. El primer paso de este proyecto fue seleccionar y analizar los atributos que se utilizarían para la segmentación, centrándose en identificar cualquier correlación entre estos atributos.

El proceso de selección de atributos relevantes para la segmentación es crucial en el análisis de agrupamiento, ya que determina los criterios que en última instancia conducen a la agrupación de los datos de los clientes. Existe una amplia gama de atributos disponibles, que pueden clasificarse en atributos demográficos como edad, género y situación familiar, así como en atributos socioeconómicos como educación, profesión, ingresos y propiedad. Los atributos seleccionados deben exhibir suficiente potencial discriminatorio sin ninguna correlación. Además, deben ser cuantificables para la aplicación de métodos matemáticos y deben estar en la misma escala.

El propósito del análisis de correlación es garantizar que los atributos elegidos en su mayoría no estén relacionados entre sí y que ningún atributo tenga una influencia excesiva en el análisis de agrupamiento. Este análisis es importante para evitar cualquier sesgo o distorsión en los resultados. Por el contrario, la normalización desempeña un papel crucial a la hora de hacer que los datos sean comparables porque, inicialmente, los datos pueden tener diferentes dimensiones y rangos de escala. Sin normalización, esta variación puede dar como resultado ponderaciones dispares de atributos durante el proceso de agrupación.

Para mejorar la eficacia de las campañas de marketing dirigidas a un grupo particular de clientes, es fundamental analizar exhaustivamente sus patrones de respuesta. Este análisis tiene como objetivo identificar estrategias que pueden emplearse para maximizar la tasa de respuesta de las campañas promocionales, específicamente aquellas dirigidas a promover las tarjetas de crédito a través de correos electrónicos.

El proceso de selección de personas para un envío de correo al banco para este proyecto se ha realizado manualmente hasta ahora. Esto implica crear un prototipo de perfil de cliente y comparar cada registro de la base de datos con este prototipo mediante una consulta a la base de datos. Sólo se incluyen en el envío aquellas personas que cumplen con los requisitos predeterminados. Sin embargo, la composición manual del perfil introduce cierta subjetividad en el proceso. Teniendo en cuenta la gran cantidad de información disponible sobre cada cliente, queda claro que determinar los clientes que cumplen con criterios específicos a través de una consulta es un problema complejo.

Sería muy ventajoso tener la capacidad de configurar automáticamente el perfil del cliente prototipo. Esto se puede lograr mediante la implementación de una red neuronal, que ha demostrado ser eficaz para realizar esta tarea en particular. Específicamente, se ha desarrollado un clasificador neuronal para extraer de una base de datos a las personas que se considera que tienen más probabilidades de obtener una tarjeta de crédito en respuesta a un correo electrónico.

Asimismo, el conjunto de datos se enriquece con información completa sobre la utilización de diversos productos de los clientes. Estos productos abarcan una amplia gama de ofertas, que van desde cuentas corrientes y de ahorro hasta planes de seguros, inversiones y valores. Además, la base de datos

contiene un sistema de clasificación que identifica la ubicación geográfica de residencia de cada individuo. En general, el registro de cada cliente dentro de la base de datos consta de una extensa colección de alrededor de 180 campos de datos distintos, que abarcan una amplia gama de atributos.

Después de realizar un análisis estadístico inicial de los datos, las distribuciones de los campos de datos individuales tanto para entradas como para salidas siempre son notablemente similares. Como resultado, se volvió cada vez más difícil establecer un perfil de comprador típico basado únicamente en estas distribuciones estadísticas. Se hizo evidente que la simple observación de atributos aislados del cliente era insuficiente para comprender su comportamiento de compra. Quedó claro que el perfil del cliente debía tener en cuenta la combinación de diversos datos sobre él. Si bien, la naturaleza específica de esta combinación seguía siendo incierta y no estaba claro qué información se incluiría en última instancia para dar forma al perfil del cliente.

Por este motivo, se determinó que se utilizaría una red neuronal para establecer una conexión entre el perfil del cliente y su inclinación a adquirir un producto. Esta decisión se tomó porque las redes neuronales poseen la capacidad de representar relaciones complejas que no son necesariamente lineales u homogéneas, lo cual es esencial en este caso. Aunque, para agilizar el proceso y centrarse en los atributos más significativos, la siguiente etapa del análisis implicó reducir el número de atributos considerados. Este paso también tuvo como objetivo proporcionar información sobre los atributos que realmente tienen relevancia.

Una descripción concisa del procedimiento de análisis que se describió anteriormente. Para identificar los atributos más importantes, se utilizó una versión modificada del método descrito en Ma y Pohlman (2008). Este

procedimiento modificado permite considerar inicialmente un vasto espacio de atributos y posteriormente seleccionar las combinaciones de atributos que producen las tasas de reclasificación más altas. Como resultado, el procedimiento genera una lista completa de combinaciones de atributos, acompañada de una estimación de sus respectivas tasas de clasificación.

La etapa de selección de atributos implicó examinar combinaciones de dos a cinco atributos con mayor detalle. Para profundizar en el análisis se utilizaron redes neuronales tipo Kohonen, específicamente mapas de atributos autoorganizados. Estas redes se aplicaron a un subconjunto más pequeño de los datos de respuesta y, para cada combinación de atributos seleccionada, se utilizó un conjunto distinto de datos para la evaluación.

En la actualidad, el conocimiento que se ha obtenido a partir de este análisis de datos se está difundiendo y aplicando para alinearse de manera más efectiva con las demandas y especificaciones de diversos productos más allá de las tarjetas de crédito, ampliando el alcance y aplicabilidad del conocimiento adquirido.

2.3 Herramientas

Las investigaciones antes mencionadas se realizaron utilizando la herramienta de software conocida como "DataEngine", que es un paquete de software integral que incorpora técnicas avanzadas como tecnologías difusas y redes neuronales para el análisis inteligente de datos (MIT 2000). Al integrar preprocesamiento, análisis estadístico y sistemas inteligentes para el desarrollo de clasificadores, así como modelado de sistemas, este software demuestra ser un instrumento muy robusto capaz de aplicarse en una amplia gama de aplicaciones.

La estructura de DataEngine tiene una arquitectura abierta, lo que permite a los usuarios mejorar sus capacidades incorporando bloques de funciones definidos por el usuario. Además, los modelos creados con DataEngine pueden integrarse perfectamente con otros paquetes de software y programas de aplicación mediante el uso de la biblioteca ADL de DataEngine.

2.3.1 La minería de datos en el desarrollo de estrategias de mercado

La minería de datos permite a las organizaciones investigar medios para aumentar la eficiencia, estimular la innovación y proporcionar información para la toma de decisiones. Si bien, sorprende que aún no haya ganado mucho impulso en Latinoamérica como herramienta de apoyo a las decisiones de las empresas. Muchas organizaciones en mercados altamente competitivos, incluidos los de telecomunicaciones, comercio minorista, automoción, finanzas y consumo masivo, continúan tomando decisiones a ciegas con respecto a la lealtad del cliente, las ventas adicionales, las ventas cruzadas, la gestión del desempeño y la retención de clientes.

Tienen dificultades para pronosticar con precisión la demanda y, a menudo, enfrentan una baja confiabilidad en sus procesos de pronóstico. Sin embargo, a pesar de los beneficios potenciales, las organizaciones a menudo enfrentan desafíos al encontrar nuevas alternativas para mejorar las ventas, la atención al cliente y la gestión de riesgos. También luchan por minimizar las pérdidas y utilizar eficazmente la recopilación de datos. Si bien estos objetivos son de suma importancia, las organizaciones deben superar estos desafíos y aprovechar el poder de la minería de datos para impulsar el crecimiento y el éxito (Dhar, 1998).

La minería de datos, como tecnología de gestión y análisis de información, aprovecha la capacidad existente para el procesamiento, almacenamiento y

transmisión de datos a alta velocidad y bajo costo. Permite a las organizaciones descubrir conocimientos valiosos ocultos en grandes cantidades de información, lo que les permite tomar decisiones mejor informadas para el futuro de su organización. Aunque los algoritmos más utilizados en la minería de datos se crearon hace 30 años, todavía generan resultados altamente confiables que pueden aumentar las ganancias y reducir los costos mediante la toma de decisiones basada en datos.

Actualmente, muchas organizaciones de diversos sectores industriales se enfrentan a retos a la hora de tomar decisiones basadas en estrategias competitivas. Estas estrategias tienen como objetivo incrementar la innovación, la productividad y generar cambios constantes en diversos aspectos como el desarrollo de productos, la gestión publicitaria, las estrategias de marketing, la promoción y la satisfacción del cliente. Desafortunadamente, en Colombia las organizaciones están luchando con el almacenamiento y utilización inadecuada de grandes cantidades de datos e información. Esto conduce a una situación en la que los datos y las especificaciones valiosos no se utilizan de forma eficaz para crear estrategias que puedan contribuir al crecimiento de las industrias.

2.3.2 Negocios y marketing

La necesidad de información está impulsada por dos factores: la incertidumbre y el costo potencial de cometer errores en la toma de decisiones. Estos factores prevalecen actualmente en el mundo empresarial, lo que hace que la información sea crucial para lograr el éxito en el mercado. Varios investigadores y expertos enfatizan la importancia de la información en el panorama empresarial actual. De hecho, algunos académicos incluso sostienen que ahora operamos en una economía basada en la información y el conocimiento.

Para poder recopilar la información necesaria debemos comenzar por la materia prima, que son los datos. Hoy en día, obtener estos datos es más fácil que nunca gracias a los avances en las tecnologías de la información. Las tecnologías facilitan la recopilación, transmisión y gestión de datos. Sin embargo, no basta con tener datos; el verdadero valor radica en transformar estos datos en información significativa y aplicarla a las operaciones comerciales.

El interés del mundo profesional por la minería de datos es evidente en la amplia gama de empresas que han introducido productos de minería de datos en el mercado. Entre estas empresas, las tres más grandes e importantes en términos de herramientas de análisis son ss (Clementina), sas (Enterprise Miner) e IBM (Inteligent Miner)). Desde una perspectiva académica, destacan los algoritmos de minería de datos como un campo de investigación emergente y futuro en marketing. Esto ha dado lugar a la publicación de manuales sobre minería de datos.

2.3.3 Minería de datos en industria de galvanizado

Uno de los usos más intrigantes de la minería de datos en el sector industrial es el desarrollo de modelos de sistemas. Uno de los rasgos frecuentemente observados en los procesos industriales es la expansión continua y acelerada de los datos almacenados. En términos más simples, esto implica que cada día trae un mayor volumen de datos respecto a los datos históricos que engloban información valiosa sobre los procesos productivos.

En consecuencia, a medida que aumenta la cantidad de datos almacenados, disminuye la capacidad de comprenderlos y procesarlos. Así, la utilización de herramientas que permitan la extracción de conocimientos útiles se vuelve esencial. Aquí es precisamente donde entra en juego el importante papel de la minería de datos.

Sólo hay un número limitado de industrias que utilizan herramientas estadísticas para analizar datos. Esto se debe principalmente a los laboriosos cálculos manuales que implican y al tamaño limitado de los conjuntos de datos que pueden manejar. Si bien, en el mundo actual, el sector industrial tiene la oportunidad de aprovechar datos históricos y obtener conocimientos valiosos mediante el uso de técnicas y herramientas avanzadas de análisis de datos, como las redes neuronales. Estas herramientas permiten extraer conocimiento valioso de los datos, proporcionando al sector industrial una ventaja competitiva.

La aplicación de la minería de datos en el modelado de procesos en el sector industrial se puede observar en el proceso de galvanización. El objetivo es predecir las propiedades mecánicas de bobinas de acero galvanizado para mejorar los sistemas de control de una línea de acero galvanizado. Dado que muchas de las características del producto no se pueden medir directamente y requieren pruebas de laboratorio después del proceso de fabricación, no se puede aplicar una estrategia de control tradicional. Sin embargo, al utilizar un estimador en línea que utiliza datos del proceso de fabricación para predecir estas propiedades mecánicas, resulta factible implementar mejoras en los sistemas de control actuales.

Para lograr el mismo objetivo de mejorar los sistemas de control de línea, se ha introducido un nuevo proceso de minería que implica analizar y utilizar datos. Este proceso implica desarrollar un modelo que se centra en la velocidad de la banda de acero en un horno durante el recocido. Al utilizar datos de proceso, el modelo pretende regular la velocidad para garantizar que la temperatura real de la cinta coincida con la temperatura deseada cuando sale de la zona de calentamiento del horno. Este enfoque mejoraría en gran medida el control del proceso de tratamiento térmico al que se someten las bandas de acero

antes de sumergirlas en la olla de zinc. Este tratamiento térmico es crucial para lograr las propiedades y características deseadas de la banda, así como para garantizar una adhesión adecuada del recubrimiento.

2.3.4 Minería de datos en la ingeniería civil

El uso de técnicas heurísticas en ingeniería de datos juega un papel crucial en la búsqueda de patrones complejos dentro de la planificación, operación y gestión de redes de suministro de agua. Este tema de investigación ofrece considerables beneficios prácticos al identificar y analizar de manera eficiente patrones no triviales dentro de los datos disponibles.

El caso de aplicación de minería de datos se centra en las redes de suministro de agua potable y tiene como objetivo predecir la demanda de agua utilizando datos históricos. El proceso implica descubrir reglas de datos que pueden usarse para hacer predicciones precisas basadas en diversos factores, incluidos factores ambientales, sociológicos y de distribución. Además, la aplicación tiene en cuenta información relacionada con el volumen de flujo y otros factores como el día de la semana y las condiciones climáticas, que se sabe que afectan el consumo diario de agua. Al analizar estos extensos datos históricos, la aplicación es capaz de generar rangos de predicción para la demanda de agua potable. Estas predicciones son valiosas para gestionar y suministrar eficazmente la cantidad de agua necesaria.

2.3.5 Minería de datos en distribución

Los distribuidores de refrescos y golosinas, como las populares patatas "Chips", emplean soluciones de información estratégicas para optimizar sus beneficios y garantizar un alto nivel de satisfacción de sus clientes. Lo logran gestionando eficientemente el movimiento de sus productos a través de la red de distribución, basándose en información precisa sobre las ventas en las distintas

tiendas. Estos datos les permiten adaptarse a las variaciones estacionales y entregar productos de alta calidad en el momento oportuno.

Sin embargo, confiar únicamente en esta información es insuficiente, ya que para seguir siendo competitivo, es necesaria información en tiempo real sobre los acontecimientos en curso. Para abordar esta necesidad, se contrata a conductores de camiones equipados con sistemas informáticos conectados por radio para que informen de sus observaciones cada vez que visitan un minorista.

Al recibir rápidamente esta información, las principales empresas pueden realizar rápidamente los ajustes necesarios y optimizar la utilización de su inventario de productos perecederos dentro de la red de distribución. Además, esta información estratégica relativa a ventas y no ventas permite a las empresas modificar su producción en fábrica para alinearse con la demanda predominante.

2.3.6 Minería de datos en industria del turismo

No se puede negar que el panorama de los negocios en el mundo, incluida la industria del turismo, ha sido completamente transformado por la llegada de las tecnologías de la información. Se han logrado importantes avances con la implementación de estos beneficios. Estas ventajas incluyen:

- una comprensión más profunda de las preferencias de los clientes,
- una mayor eficiencia en la prestación de servicios,
- llegar a una clientela más amplia y
- utilizar eficazmente los recursos para mejorar la productividad general.

En la industria del turismo, ejemplos notables en los que se han aplicado estos beneficios incluyen la utilización de sistemas de reservas en línea, la

facilitación de la venta de servicios a través de Internet y el empleo de sistemas de extracción de datos.

La industria del turismo tiene potencial para el desarrollo de tecnologías de la información debido a dos factores principales:

- En primer lugar, el turismo es una actividad que se desarrolla en diferentes territorios, lo que permite promocionar y comercializar actividades que se ofrecen lejos de la ubicación del cliente. Este carácter interterritorial del turismo crea oportunidades para la utilización de las tecnologías de la información.
- En segundo lugar, como parte de la industria del ocio y el entretenimiento, el turismo depende en gran medida de la promoción en los medios utilizando plataformas audiovisuales que sean visualmente atractivas para atraer clientes potenciales. Por lo tanto, las características únicas de la industria del turismo la convierten en una industria adecuada para la integración de tecnologías de la información, particularmente en las áreas de marketing y promoción en medios.

Para comprender mejor la evolución del turismo, es importante profundizar en las primeras aplicaciones que dieron forma a su desarrollo. Los primeros avances notables se produjeron en 1960, cuando las aerolíneas introdujeron sistemas de información con el único fin de reservar billetes de avión. Sin embargo, no fue hasta una década después que estos sistemas se implementaron en las agencias de viajes, permitiendo una mayor accesibilidad al público.

Aun cuando, es importante destacar que durante la década de 1970, las empresas hoteleras tenían un contacto limitado con los sistemas de información,

confiando únicamente en un sistema de reservas computarizado centralizado. Durante este tiempo, sólo unas pocas cadenas hoteleras selectas, como Holiday Inn y Sheraton, junto con un puñado de hoteles independientes, podían ofrecer servicios de reserva computarizados. Estos primeros acontecimientos marcaron los pasos iniciales hacia la modernización de la industria del turismo.

En el año 1980, la industria aérea desarrolló sistemas avanzados que permitían la reserva tanto de vuelos como de hoteles. Estos sistemas innovadores se denominaron entonces Sistemas Computarizados de Reservas. Poco después surgió otro sistema conocido como Lo-Systems. Lo-Systems, también conocido como Global Distributivo Systems (GAS), se convirtió en una herramienta vital de marketing para empresas de hostigoso en aproximadamente 125 países.

Los GAS han demostrado ser muy eficaces a la hora de promocionar sus productos. Los agentes de viajes ahora tienen acceso a una base de datos completa que les proporciona la información más actualizada y confiable sobre varios hoteles y aerolíneas. Para atender a una audiencia global, las principales plataformas GAS incluyen Galileo, Sabre, Amadeus, Worldspan, System One y Book Hotel.

Al utilizar estos diversos sistemas y tecnologías, los hoteles pueden implementar estrategias publicitarias altamente efectivas. Por ejemplo, el sistema Jaguar permite a los agentes de viajes acceder a imágenes electrónicas del hotel, mientras que el sistema Espectro permite a los usuarios identificar y localizar con precisión áreas específicas en un mapa. Además, estos sistemas facilitan un examen exhaustivo y en profundidad de la zona seleccionada, proporcionando una visión detallada y cercana. Como resultado, los hoteles pueden promocionar eficazmente sus ofertas y atraer clientes potenciales.

En la actualidad, aproximadamente el 80% de las reservas hoteleras se realizan mediante este tipo de sistema debido a las numerosas ventajas que ofrece tanto para las empresas hoteleras como para las agencias de viajes. Este sistema demuestra ser una excelente oportunidad de marketing para los hoteles en lo que respecta a la distribución global, al mismo tiempo que sirve como una valiosa herramienta para los agentes de viajes. Les permite acceder a información actualizada en tiempo real, permitiéndoles realizar eficientemente sus operaciones a través del sistema.

Asimismo, este sistema consolida datos de diversas fuentes, que incluyen información sobre hoteles, boletos de avión y alquiler de autos, facilitando así la generación de informes que brindan detalles pertinentes para el buen funcionamiento de sus respectivas empresas. Ante esto, se puede afirmar que la implementación efectiva de tecnologías de la información conduce a una mayor colaboración entre hoteles, restaurantes, agencias de viajes y aerolíneas, beneficiando en última instancia a sus clientes y fomentando el crecimiento mutuo.

La minería de datos es un tema de reciente discusión en el ámbito de los negocios y el marketing. Es una preocupación relativamente nueva que está ganando atención debido a su potencial impacto en la escasez de recursos. Además, el conocimiento y la aplicación de técnicas de minería de datos juegan un papel crucial en el desarrollo y la eficiencia de diversos esfuerzos. Estos factores resaltan la importancia de la minería de datos y su influencia en la medición de la efectividad de los resultados.

En esencia, se puede deducir que uno de los principales beneficios de utilizar herramientas de minería de datos es la conveniencia que ofrecen. Aun cuando, es fundamental poseer suficiente conocimiento sobre los distintos

algoritmos utilizados para maximizar su potencial, ya que no todos los algoritmos producen los mismos resultados ni la misma eficiencia. La eficacia de la minería de datos depende de la adecuada evaluación de los resultados, lo que implica la obtención de indicadores sobre cuatro aspectos: bondad de ajuste, relevancia, novedad y aplicabilidad. Calcular estas medidas permite cumplir las promesas que presenta la minería de datos a través de su definición. Para proporcionar más información, a continuación se presenta un análisis comparativo, que describe las principales ventajas, desventajas, herramientas, contribuciones y logros en cada aplicación específica de la minería de datos.

En el momento actual, los mercados están experimentando un constante estado de cambio y transformación. El consumidor moderno se ha vuelto cada vez más exigente y exige cada día mayores exigencias. Están equipados con una gran cantidad de información al alcance de su mano y buscan activamente productos superiores y experiencias personalizadas. Además, requieren servicios eficientes que puedan abordar eficazmente sus necesidades e inquietudes y al mismo tiempo tener en cuenta la rentabilidad.

Al examinar los casos anteriores, podemos sacar la conclusión de que la utilización de técnicas de minería de datos puede conducir al desarrollo de estrategias competitivas que contribuyan a incrementar las ganancias en los sectores industriales del Departamento del Atlántico. La implementación de estas estrategias también daría como resultado importantes reducciones de costos y mejores servicios auxiliares dentro de los procesos y operaciones de la empresa. La minería de datos sirve como una herramienta valiosa utilizada principalmente para prevenir y diagnosticar situaciones actuales de la empresa, permitiendo tomar decisiones informadas sobre inversiones y la creación de nuevos productos basados en datos reales.

Actualmente se prevé que el suministro de datos mundial se duplica cada 20 meses. Este crecimiento exponencial del volumen de datos plantea importantes desafíos tanto para la comunidad científica como para los sectores productivos de la economía. Se está superando la capacidad de los humanos para analizar, resumir y extraer conocimientos de cantidades tan masivas de datos.

Como resultado, existe una necesidad apremiante de desarrollar herramientas avanzadas capaces de automatizar el análisis de los datos almacenados. Este campo de investigación emergente, conocido como minería de datos, se centra en estudiar y crear estas herramientas que han contribuido a la comprensión y avance de este campo.

La minería de datos se ha convertido en una herramienta y estrategia valiosa en diversos aspectos de las operaciones de una empresa, como el marketing, la producción y la organización. Desempeña un papel crucial en la mejora de la competitividad de la empresa en el mercado. En este artículo profundizamos en el impacto de la minería de datos, una técnica ampliamente utilizada en la investigación de operaciones, en el diseño de estrategias de marketing business-to-business (B2B) dentro del sector industrial (Echeverri et a., 2013).

La minería de datos suele estar vinculada a consultas del departamento de marketing y numerosos ejecutivos la consideran un medio para mejorar su comprensión de la demanda de los consumidores y observar el impacto que tienen las modificaciones en los productos, precios o promociones en las ventas. Si bien, es importante señalar que la minería de datos también posee beneficios sustanciales para otros sectores comerciales.

Por ejemplo, los ingenieros y diseñadores pueden evaluar la efectividad de las modificaciones realizadas a un producto y buscar posibles razones detrás

de su triunfo o fracaso, considerando factores como cómo, cuándo y dónde se utilizan los productos. Además, las operaciones de servicio y reparación pueden optimizar su planificación del inventario de piezas y las necesidades de personal. Además, las organizaciones de servicios profesionales pueden utilizar técnicas de extracción de datos para identificar nuevas oportunidades que surjan de las fluctuaciones en las tendencias económicas y los cambios demográficos.

Así, la minería de datos es cada vez más beneficiosa y valiosa a medida que los conjuntos de datos crecen y los usuarios adquieren más experiencia. Es lógico suponer que más datos contendrán una mayor cantidad de información e inteligencia estratégica. Además, a medida que los usuarios se vuelven más competentes en el uso de las herramientas y obtienen una comprensión más profunda de la base de datos, pueden explorar y analizar los datos de una manera más imaginativa.

La extracción de datos es esencial para diversos fines, como:

- analizar opiniones,
- optimizar precios,
- realizar marketing de bases de datos,
- gestionar riesgos crediticios,
- brindar capacitación y soporte,
- detectar fraude,
- diagnosticar condiciones médicas y de salud,
- evaluar la gestión de riesgos,
- desarrollar sistemas de recomendación y muchas otras aplicaciones.

Su utilidad se extiende a una amplia gama de industrias, incluidas:

- la venta minorista,
- la distribución mayorista,
- los sectores de servicios,
- las telecomunicaciones,
- las comunicaciones,
- los seguros,
- la educación,
- la manufactura,
- la atención médica,
- la banca,
- la ciencia,
- la ingeniería y
- el marketing en línea o las redes sociales.

Las empresas que se dedican al diseño, producción o distribución de bienes físicos tienen el potencial de mejorar sus estrategias de orientación de productos analizando cuidadosamente los patrones de compra junto con datos económicos y demográficos. Además, al realizar un examen exhaustivo de los comentarios de los clientes y usuarios, los registros de reparación y otros datos relevantes, sus diseñadores e ingenieros pueden descubrir información valiosa que puede conducir a oportunidades de mejora del producto.

En las industrias orientadas a los servicios, se pueden encontrar oportunidades similares para mejorar los productos a través del análisis cuidadoso de los comentarios de los clientes, ya sea obtenidos directamente o de fuentes como las redes sociales. Al cruzar esta retroalimentación con información relacionada con servicios, canales, desempeño de pares, regiones geográficas, precios, datos demográficos y económicos, las empresas del sector de servicios pueden identificar áreas donde se pueden realizar mejoras para atender mejor las necesidades y preferencias de sus clientes.

Los fabricantes pueden monitorear y rastrear de manera efectiva las tendencias de calidad, los datos de reparación, las tasas de producción y la información sobre el rendimiento del producto obtenida en el campo. A través de este proceso, pueden identificar cualquier inquietud o problema en el proceso de producción. Además, tienen la oportunidad de identificar posibles actualizaciones o mejoras en sus procesos de fabricación que pueden dar como resultado una mayor calidad, ahorro de tiempo y costos, un mejor rendimiento del producto e incluso indicar la necesidad de equipos de fábrica nuevos o mejorados.

En conclusión, es crucial incorporar estos descubrimientos en los procesos de previsión y planificación para garantizar que toda la empresa esté consciente y preparada para los cambios esperados en la demanda de los consumidores. Al obtener una visión más profunda de las necesidades y preferencias de los clientes, la organización puede alinear sus estrategias y capitalizar los prospectos recientemente reconocidos. Este enfoque integral mejorará en última instancia la capacidad de la empresa para adaptarse y prosperar en un entorno de mercado dinámico.

Capítulo 3

Minería de datos en finanzas

Numerosos expertos en teoría organizacional han señalado que las organizaciones tienden a adquirir más conocimiento rechazando alternativas deficientes en lugar de descubrir alternativas exitosas. Esto se debe a su capacidad para observar los resultados negativos de las malas decisiones, mientras que carecen de información sobre decisiones que no son fácilmente observables. En consecuencia, las organizaciones tienden a tener una inclinación natural a corregir errores (errores de tipo I) con el tiempo a medida que los reconocen.

Si bien, a menudo no abordan los errores de tipo II, que implican el rechazo o la falta de consideración de alternativas potencialmente buenas, ya que no se dispone de la información necesaria sobre sus resultados. Esta exploración limitada de opciones potencialmente beneficiosas limita en última instancia la capacidad de una organización para aprender.

Si ampliamos el concepto representado por Dhar (1998), la decisión de aceptar o rechazar hipótesis en el sector finanzas, con base en la oferta y la demanda global, se basa en valores dentro de un rango y no en un simple sí o no, los diferentes resultados no se concentran en regiones específicas sino que aparecen dispersos por todo el espacio multidimensional. Determinar las regiones específicas donde ocurren estos resultados se convierte en una tarea desafiante, y la escasez de información previa exacerba la dificultad de estimar con precisión la distribución de los resultados correspondientes a las diversas condiciones de entrada. A medida que aumenta el número de factores o insumos, este problema se vuelve aún más complejo.

Este sección explora cómo se puede utilizar la combinación de contrafactuales y algoritmos específicos de aprendizaje automático para generar distribuciones condicionales de resultados de grandes sistemas de bases de datos. Estas distribuciones son valiosas para estimar correlaciones entre acciones y resultados. Además, demuestran ser efectivas para abordar la cuestión de los errores de tipo II.

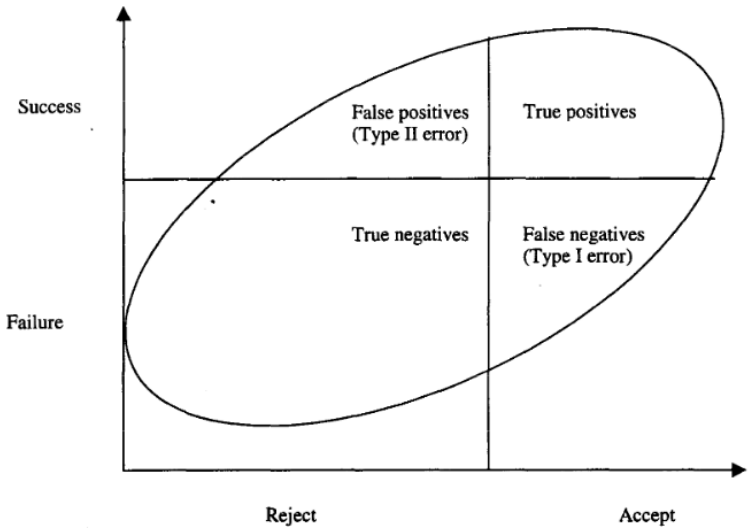
Los contrafactuales tienen una rica historia en lógica, particularmente en el razonamiento sobre causalidad y escenarios hipotéticos. Representan eventos que en realidad nunca sucedieron, pero que si hubieran sucedido, habrían tenido resultados específicos. Al evaluar con precisión estos contrafactuales y emplear herramientas apropiadas para su generación y evaluación, brindan una solución al problema de los errores de Tipo II. Básicamente, un generador de hipótesis enfocado y una función de evaluación definida con precisión son componentes necesarios en este enfoque.

Una función de evaluación desempeña un papel fundamental a la hora de establecer un vínculo fuerte entre diversas acciones potenciales y los resultados correspondientes que pueden producir. Aprovechar el poder de una base de datos completa ayuda a guiar su exploración de acciones que contienen un elemento de entusiasmo, ya que poseen la capacidad de generar resultados positivos y ventajosos.

Para proporcionar una comprensión integral de las complejidades involucradas en el aprendizaje a partir de datos, comenzaré presentando dos casos de la vida real que ocurren en el ámbito de la industria financiera. Estos casos se derivan de aplicaciones exitosas implementadas actualmente en una destacada organización financiera. La Figura 1 gira en torno a la intrincada tarea

de comprender las relaciones con los clientes basándose en una gran cantidad de datos transaccionales.

Figura 1. Toma de decisiones y dispersión de datos en el espacio multidimensional



Lo que implica la identificación de clientes que poseen un valor financiero más alto en comparación con otros. En este contexto particular, una transacción denota una operación en la que un cliente participa en la compra o venta de un volumen específico de un producto particular en un momento específico, facilitado por un vendedor designado. Con una cantidad monumental de transacciones de este tipo que se producen diariamente, una gran reserva de datos queda disponible, interconectando a los clientes, los productos y todo el proceso de ventas.

El objetivo principal de la organización es extraer información valiosa sobre cómo interactuar eficazmente con varios segmentos de clientes y, en última instancia, mejorar su experiencia general. La segunda cuestión que nos ocupa tiene que ver con la exploración de las complejidades de los mercados de valores financieros. En concreto, profundiza en las formas en que los precios de las

acciones están determinados por una multitud de datos que inundan continuamente el mercado. Estos datos abarcan una amplia gama de factores, incluidos anuncios de ganancias (incluidos resultados inesperados), pronósticos o revisiones de analistas con respecto a las ganancias de empresas o sectores industriales, divulgaciones continuas de los balances y estados de resultados de las empresas, tendencias de precios y volúmenes, informes de noticias y más.

Es ampliamente reconocido en la industria financiera que dichos datos ejercen una influencia significativa en el desempeño del mercado de valores. Aun cuando, muchas conexiones simplistas entre estas variables han demostrado ser engañosas, y las complejas a menudo presentan desafíos en términos de comprensión y alineación con la intuición. En consecuencia, esta cuestión se vuelve bastante exigente ya que las relaciones identificadas deben ser lo más sencillas posible, permitiendo a los tomadores de decisiones formular un marco de comprensión plausible y al mismo tiempo lograr un nivel satisfactorio de precisión en la predicción de resultados.

Entre los aportes que aquí se pretenden, se encuentran:

- En primer lugar, mostrar la aplicación exitosa de métodos de descubrimiento de conocimiento para resolver problemas industriales a gran escala. Estas aplicaciones, que llevan casi tres años implementadas, sirven como prueba anecdótica de la eficacia de la minería de datos en el ámbito de las Finanzas.
- En segundo lugar, se enfatiza la importancia de utilizar contrafactuales para superar un obstáculo común en las organizaciones: la consideración inadecuada de los errores de Tipo II. Al utilizar datos históricos para generar distribuciones condicionales de resultados, los tomadores de

decisiones pueden obtener una mejor comprensión del dominio del problema y tomar decisiones más informadas.

- En tercer lugar, presenta un marco para determinar las circunstancias bajo las cuales los patrones descubiertos mediante la minería de datos pueden usarse para respaldar las decisiones y cuándo deberían reemplazar la toma de decisiones humana. Este marco enfatiza que la automatización de la toma de decisiones depende de factores más allá de la propia estructura del problema.
- Por último, se muestra cómo el dominio de la aplicación influye en el proceso de minería de datos. Contrariamente a la idea errónea de que la minería de datos es una búsqueda aleatoria de patrones interesantes, se caracteriza mejor como una parametrización iterativa del conocimiento existente, que se asemeja a reglas más que a una regresión estadística.

El uso de reglas en la minería de datos ofrece beneficios prácticos, como la facilidad de comprensión y la capacidad de razonar sobre eventos y causalidad.

3.1 La generación de conocimiento

A continuación para el desarrollo de este punto se presentan dos ejemplos:

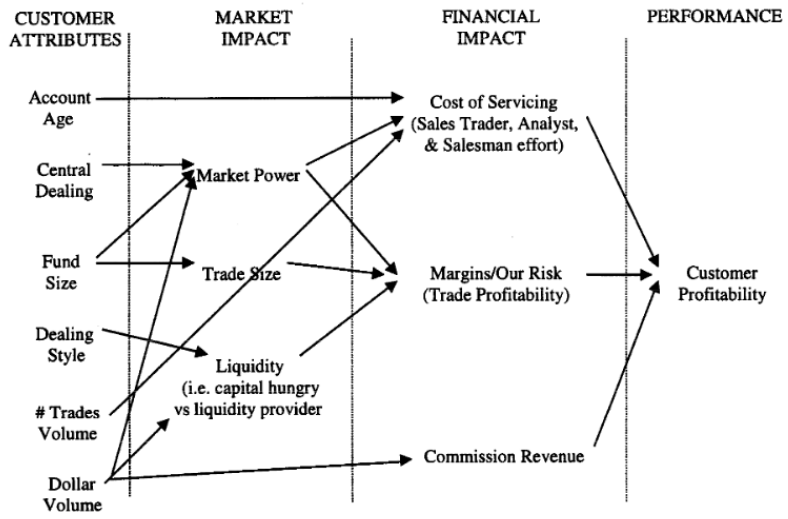
Ejemplo A:

Dentro de una importante firma de valores, existía una divergencia en los puntos de vista entre dos estimados altos directivos con respecto a cuál era la clientela más rentable. Un gerente abogó por el cultivo de relaciones más sólidas con clientes más grandes, afirmando que ofrecían oportunidades comerciales superiores y más abundantes, particularmente en términos de tarifas de transacción. Por el contrario, el otro gerente cuestionó este punto de vista afirmando que los clientes más importantes eran, en realidad, aquellos que

ejercían dominio en el mercado y manipulaban su posición para obtener costosas concesiones. A partir de sus astutas observaciones de las actividades de la mesa de operaciones, este director argumentó que los clientes más grandes a menudo poseían la capacidad de negociar tasas de comisión más bajas y frecuentemente participaban en transacciones "hambrientas de capital", lo que implicaba asumir riesgos sustanciales.

Después de múltiples rondas de discusión con los gerentes para generar ideas y conceptualizar el problema en cuestión, se representan y describen visualmente las variables relevantes relacionadas con el problema en la Figura 2 (Dhar, 1998). Esta representación visual sirve como modelo inicial del campo específico, proporcionando una base de comprensión de causa y efecto.

Figura 2. Diagrama de causa-efecto de un modelo de negocios de inversión de capital



La Figura 2 ilustra varios atributos de los clientes, como su volumen de operaciones en términos tanto de operaciones como de valor monetario, así como su estilo de negociación preferido, que indica si están dispuestos a poner en riesgo su capital. Estos atributos se pueden analizar y agregar desde la base de datos de transacciones mensualmente, lo que permite clasificar a los clientes en función de estos atributos. Asimismo, los atributos demográficos como el tipo de cuenta, el tamaño del fondo y la propiedad de una operación comercial centralizada se pueden obtener de la base de datos de la cuenta maestra.

Así, los atributos de los clientes juegan un papel importante en la configuración de la dinámica del mercado. Factores como el poder de mercado, el tamaño del comercio y su contribución o agotamiento de la liquidez del mercado contribuyen a su impacto en el mercado. Comprender las implicaciones de estos atributos es crucial para que los participantes del mercado tomen decisiones informadas sobre sus estrategias comerciales y, en última instancia, maximicen sus beneficios financieros.

Es fundamental reconocer que el impacto de proporcionar liquidez en el mercado es relativamente menor en comparación con el agotamiento de la liquidez. En términos simples, ofrecer productos o servicios a la venta tiene un efecto menor en el mercado en comparación con realizar compras. Esta distinción se vuelve particularmente importante cuando se consideran las ventajas financieras de cada función. Suponiendo que todos los demás factores permanezcan constantes, generalmente es más ventajoso desde el punto de vista financiero ser un proveedor de liquidez que un consumidor de ella.

Para comprender la importancia de estos atributos, consideremos un ejemplo en el que el mercado está dominado por los compradores. En tal escenario, un vendedor que ingrese al mercado contribuiría a la liquidez general

del mercado. Al ofrecer sus productos o servicios a la venta, brindan a los compradores la oportunidad de satisfacer su demanda y mantener un nivel saludable de actividad comercial. Por otro lado, un comprador en esta situación agotaría la liquidez del mercado, ya que consume recursos y reduce la disponibilidad de bienes o servicios para otros compradores. Existen numerosos factores que influyen en el mercado y uno de los factores clave son los atributos de los clientes involucrados.

Estos atributos tienen un impacto significativo en la dinámica del mercado y pueden influir en gran medida en su funcionamiento general. Algunos de los atributos cruciales de los clientes que determinan su impacto en el mercado incluyen su poder de mercado, el tamaño de sus operaciones y si sus operaciones contribuyen a la liquidez del mercado o la agotan.

Las variables del mercado tienen un impacto significativo en la situación financiera de una empresa, afectando tanto los ingresos que genera como los costos en los que incurre. Una de las variables más sencillas de medir son los ingresos por comisiones, ya que se registran por cada transacción realizada. Sin embargo, hay otros factores a considerar, como el riesgo que implica la ejecución de operaciones para los clientes. Esto abarca el riesgo de mantener una posición que de otro modo no se habría tomado y la posibilidad de incurrir en altos costos de ejecución. Para evaluar estos factores, se deben analizar los datos comerciales y de mercado en el momento de la transacción y durante el período posterior. Además, existen gastos asociados con la gestión de las relaciones con los clientes, incluida una parte de los costos fijos y los costos variables relacionados con la satisfacción de sus necesidades. Determinar el costo de brindar servicios a los clientes puede ser complejo, ya que requiere asignar costos con precisión entre los diferentes clientes.

El objetivo principal de este estudio es examinar los factores que afectan la rentabilidad del cliente. Los ingresos por comisiones, el riesgo y el costo del servicio son variables importantes que influyen en la rentabilidad del cliente. Sin embargo, la dinámica específica y la interacción entre estos factores siguen en gran medida inexploradas y requieren más investigación.

Resulta evidente que determinar la rentabilidad del cliente es una tarea compleja que implica recopilar y analizar datos relevantes de bases de datos. Es interesante observar que los altos directivos tenían opiniones diferentes sobre qué tipos de clientes eran realmente rentables. Esta discrepancia puede deberse a las diferentes prioridades que asignaron a cada flecha del diagrama. Por ejemplo, el primer gerente minimizó la importancia del poder de mercado de los clientes en relación con la rentabilidad, mientras que el segundo gerente reconoció la influencia de comisiones más altas pero se centró más en los gastos y riesgos potenciales involucrados al tratar con clientes más grandes.

Una vez recopilados los datos esenciales, resulta fácil identificar a los clientes que generan las mayores ganancias. Aunque, determinar los tipos exactos de clientes que son más lucrativos es una tarea mucho más difícil. Por ejemplo, si poseemos una base de datos que contiene 20 atributos diferentes, cada uno con 10 valores únicos, el gran número de combinaciones potenciales suma la asombrosa cifra de 102. Esta amplia gama de posibilidades puede resultar bastante abrumadora para un problema de escala relativamente modesta. Por lo tanto, es crucial desarrollar un enfoque preciso para generar hipótesis que puedan acelerar el proceso de descubrir relaciones pertinentes.

Una estrategia alternativa y más eficaz para abordar el problema implica estimar distribuciones condicionales de resultados. En lugar de explorar todas las combinaciones posibles de atributos, nos concentramos en las distribuciones

más interesantes, considerando que construir una distribución multivariada completa de resultados no es práctico, especialmente cuando se trata de numerosos insumos. Además, es importante tener en cuenta que es posible que determinadas combinaciones de atributo y valor no sean significativas o no estén disponibles en la base de datos.

Por lo tanto, es más ventajoso ver el proceso de aprendizaje como la generación de distribuciones condicionales respaldadas estadísticamente, donde el aspecto condicional puede representarse en diversas formas, como cláusulas, proposiciones lógicas, restricciones algebraicas lineales o reglas. Estas reglas pueden ser evaluadas por expertos, lo que permite el desarrollo de una teoría de dominio basada en los datos. Si bien un método de aprendizaje automatizado consiste en determinar los parámetros del modelo representado en sí, no siempre es apropiado utilizar un modelo lineal ya que las relaciones entre variables suelen ser condicionales. Por ejemplo, el costo del servicio de una cuenta institucional puede ser solo alto cuando el volumen de operaciones es bajo o cuando las cuentas tienen menos de cierto tamaño. La simple asignación de pesos numéricos a las variables también puede oscurecer el problema, particularmente cuando la relación entre dos variables no es consistente en todos los valores o cuando está influenciada por otras variables.

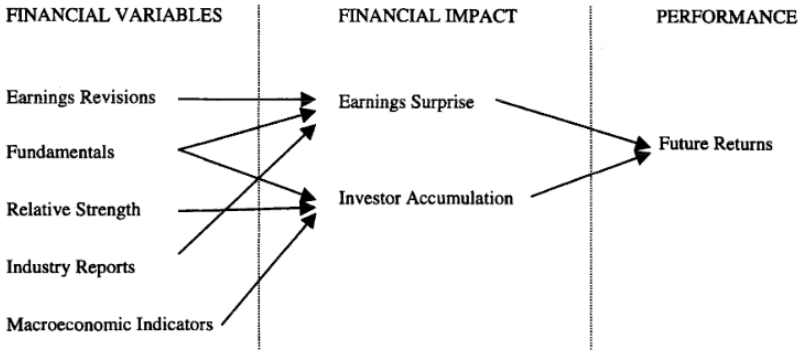
Ejemplo B:

Las empresas de valores están muy interesadas en comprender el impacto de un flujo constante de información diversa sobre los precios de las acciones en el mercado. Esta información incluye anuncios de ganancias, pronósticos de analistas o evaluaciones de ganancias para empresas o sectores industriales específicos, estados financieros periódicos proporcionados por empresas (denominados fundamentos), indicadores técnicos como "fuerza relativa" que

indican el impulso de los precios, informes de la industria y datos macroeconómicos. El objetivo es establecer las relaciones entre estas variables y utilizar estos conocimientos para participar en actividades comerciales lucrativas y una gestión de riesgos eficiente. La capacidad de negociar de manera inteligente plantea un desafío importante para las empresas de valores.

Cuando se trata de analizar la rentabilidad del cliente, el primer paso crucial implica identificar las variables relevantes y formular hipótesis que tengan el potencial de generar conocimientos significativos. En la Figura 3 se muestra un modelo inicial que esquematiza los elementos esenciales de esta materia (Dhar, 1998).

Figura 3. Gestión de riesgos en acciones de mercado según hipótesisfinancieras



Como en el problema anterior, el desafío es encontrar distribuciones condicionales de resultados interesantes, en otras palabras, reglas que expresen relaciones sólidas entre las variables del problema. Por ejemplo, una regla descubierta para este problema podría verse así: "Las sorpresas de rendimiento positivo/negativo están asociadas con rendimientos futuros positivos/negativos".

Una iteración más avanzada de esta regla implicaría identificar áreas particulares de sorpresa de ganancias positivas y negativas que exhiban la conexión más significativa con los rendimientos futuros. Por ejemplo, un enfoque más matizado podría proponer que las sorpresas positivas o negativas en las ganancias que excedan 2 desviaciones estándar del monto acordado tengan un impacto notable en los rendimientos futuros, o que los sólidos fundamentos subyacentes combinados con una sorpresa positiva/negativa en las ganancias estén asociados con resultados positivos/negativos.

3.2 Los contrafactuales e el aprendizaje automático

Los algoritmos de aprendizaje automático tienen la capacidad de generar contrafactuales, lo que puede ser extremadamente valioso para descubrir conexiones o patrones interesantes dentro de partes específicas de una base de datos. Estos contrafactuales juegan un papel crucial al ayudarnos a crear distribuciones condicionales que sean cautivadoras e intrigantes.

La creencia subyacente en este enfoque es que no es necesario ni rentable generar la distribución multivariada completa de resultados, ya que se prevé que una porción sustancial del espacio del problema no tendrá nada de especial.

Para establecer la jerarquía de contrafácticos, existen varias técnicas disponibles en teoría de la información y estadística que pueden emplearse para cuantificar su grado de importancia. Cuando se intenta identificar patrones que se parecen a reglas, es racional emplear representaciones y algoritmos diseñados específicamente para manejar dichas estructuras.

Un enfoque comúnmente utilizado implica representar cada elemento de una regla como una expresión booleana, que se determina en función de las variables relevantes para el problema en cuestión. Si las reglas se componen de

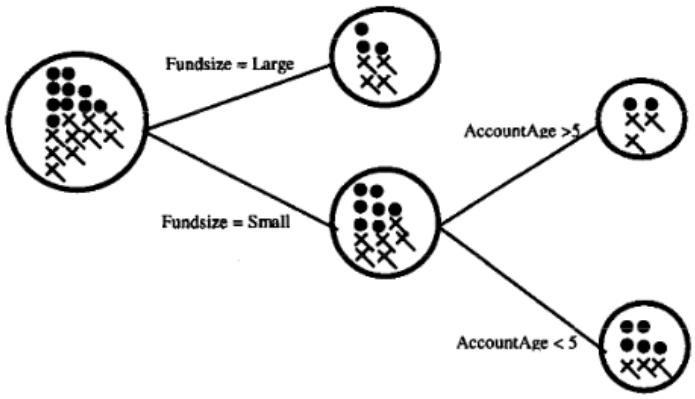
conjunciones de estas expresiones, reflejando el ejemplo mencionado anteriormente, se dice que el problema está en forma normal disyuntiva.

Cuando se trata de generar contrafactuales, existen dos alternativas principales disponibles: algoritmos de inducción de reglas y algoritmos genéticos. Los algoritmos de inducción de reglas se utilizan ampliamente en el ámbito del aprendizaje automático, mientras que los algoritmos genéticos no han obtenido el mismo nivel de popularidad principalmente debido a su velocidad de procesamiento más lenta. Sin embargo, cabe señalar que los algoritmos genéticos poseen la ventaja de ser capaces de realizar búsquedas más exhaustivas y exhaustivas.

La Figura 4, es una representación visual que muestra la división de los datos en el ejemplo de rentabilidad del cliente en varios grupos (Dhar, 1998). Esta división se lleva a cabo paso a paso, por lo que cada grupo abarca un segmento o "porción" distinto de los datos. A medida que avanzamos en la jerarquía, cada grupo se vuelve cada vez más refinado o "puro" en términos de la similitud de los valores de las variables dependientes entre los casos que comprende.

Para facilitar el análisis, hemos asumido que la variable dependiente y el tamaño del fondo se clasifican como altos o bajos, mientras que la antigüedad de la cuenta se trata como una variable continua.

Figura 4. Conglomerado de datos jerárquicos de consumo en un modelo de mercado



El grupo del extremo izquierdo representa la colección completa de datos. Este grupo está formado tanto por consumidores de bajos ingresos, indicados por cruces, como por consumidores de altos ingresos, indicados por círculos. La proporción de consumidores de bajos ingresos y consumidores de altos ingresos en la base de datos general es igual, con una proporción de 50:50.

La separación inicial basada en el tamaño del fondo conduce a una proporción ligeramente mayor de clientes que se consideran "rentables". Sin embargo, la razón para seleccionar esta variable como base para la división es sencilla. Provoca la mejora más significativa al reducir la imprevisibilidad del grupo original, que está determinado por la teoría de la información. Vale la pena mencionar que cualquier proporción daría como resultado una menor imprevisibilidad en comparación con una distribución igual como 50:50.

Es importante entender que el antecedente y el consecuente de esta regla se refieren a las partes del enunciado que vienen antes y después de la palabra clave "Entonces". Si aceptamos la validez y amplitud de la regla antes mencionada, inicialmente parece corroborar la hipótesis planteada por el

segundo gerente de que las cuentas más pequeñas son más lucrativas. Sin embargo, el grado de importancia atribuido a la antigüedad de la cuenta en relación con su rentabilidad merece una mayor exploración.

Al principio, el resultado fue desconcertante, lo que llevó al desarrollo de múltiples explicaciones alternativas. Algunas de estas explicaciones se centraron en las características de las últimas cuentas, mientras que otras se concentraron en la noción de que el equipo de ventas había experimentado transformaciones sustanciales en los últimos cinco años, lo que potencialmente podría haber afectado la dinámica de sus interacciones. Aunque ciertas explicaciones podían validarse o refutarse mediante el examen de los datos, también había explicaciones que no podían verificarse con los datos existentes.

La importancia del resultado se ve subrayada por el hecho de que embarcarse en la minería de datos pone en marcha una serie de contemplación introspectiva sobre puntos de vista descuidados y posibilidades sin explotar que normalmente permanecerían inexploradas en las actividades organizacionales rutinarias. A través de la realización de un ejercicio de minería de datos, las organizaciones se ven obligadas a profundizar en el ámbito de las potencialidades y cuestionar nociones preconcebidas, lo que conduce en consecuencia a una comprensión más profunda del tema que se aborda actualmente. La lección que se puede aprender del problema de predicción del mercado es similar a la mencionada anteriormente.

Cuando se observan valores extremos en la distribución de las ganancias sorpresivas, existe una relación más precisa entre una mayor sorpresa y mayores rendimientos. Esto significa que cuando la sorpresa es significativamente mayor o menor que la estimación de consenso de los ingresos medios en al menos una desviación estándar, el efecto es verdadero.

Por otro lado, si la sorpresa cae dentro de una desviación estándar, no se observa el efecto. Esto sugiere que la mayoría de los casos dentro de una desviación estándar pueden considerarse "ruido" aleatorio, mientras que la "señal" significativa se encuentra en áreas específicas, específicamente en los extremos de la distribución.

Hay casos en los que podemos generar fácilmente ilustraciones en las que la distribución prominente de la variable dependiente puede existir en la mayoría de las regiones, excluyendo los extremos y similares. El elemento crucial a considerar es que el patrón o "señal" significativo sólo puede identificarse en ciertas áreas designadas, lo que resulta en una relación no lineal entre las variables independientes y dependientes. El objetivo es descubrir estas conexiones no lineales.

El impacto sobre una variable que está influenciada por otros factores a menudo se produce debido a la combinación de dos o más factores que no tienen una dependencia directa entre sí. Los sistemas complejos, se caracterizan por que los efectos de las interacciones entre variables son más significativos que los efectos de las variables individuales. Por ejemplo, el efecto sorpresa sobre las ganancias es más notable cuando los fundamentos subyacentes de una empresa, como la relación precio-beneficio, son sólidos. Estos efectos de interacción no son necesariamente de naturaleza lineal. Se hace evidente cómo un algoritmo de inducción de árbol generaría un árbol de decisión, cuando se trabaja con dichos datos.

Un algoritmo genético es un método que trabaja con una colección de cromosomas, donde cada cromosoma funciona como una hipótesis que se compara con una base de datos. En la figura anterior se muestra que un

cromosoma representa una ruta completa de un árbol de decisión, comenzando desde la raíz y terminando en un nodo de hoja.

Básicamente, el algoritmo genético genera un escenario hipotético al producir un cromosoma de una sola vez. Debido a la posibilidad de paralelización en el proceso de evaluación, el algoritmo puede evaluar toda la población de cromosomas al mismo tiempo. La métrica de aptitud se emplea para ordenar los cromosomas dentro de la población según su clasificación.

Una ventaja clave de este algoritmo es su capacidad para mejorar y ajustar hipótesis previas a medida que continúa evolucionando. Al aprovechar las ocurrencias aleatorias, el algoritmo genético puede mejorar en gran medida sus capacidades de búsqueda. Además, el algoritmo posee inherentemente una naturaleza paralela, lo que lo hace altamente compatible para abordar problemas combinatorios complejos. En situaciones donde no es posible utilizar métodos numéricos o descenso de gradientes, el algoritmo genético surge como una estrategia extremadamente eficiente.

Capítulo 4

Árboles, algoritmos, entropías y minería de datos

La complejidad de la función de evaluación utilizada en la inducción de árboles y los algoritmos genéticos puede diferir significativamente. Puede abarcar pruebas estadísticas básicas o teóricas de la información o programas más complejos. Un ejemplo, implica realizar divisiones basadas en una sola variable y utilizar la reducción de entropía como criterio. Para determinar la entropía de un grupo *i*, se emplea la fórmula estándar:

$$H_i = -\sum_{ji} P_i \, log_2(P_i)$$

En donde, la entropía, conocida como *Hi*, sirve como métrica para determinar la cantidad promedio de información necesaria para clasificar un ejemplo dentro de un conjunto de datos. Este cálculo se basa en la probabilidad *(pi)* de que un ejemplo pertenezca a un grupo específico (irhcluster). Cuando todos los miembros de un grupo comparten la misma clase, la entropía está en su punto más bajo, lo que indica que se necesita información mínima para identificar la clase. Por el contrario, en situaciones en las que un ejemplo tiene la misma probabilidad de pertenecer a cualquier clase, la entropía está en su punto más alto, lo que significa un mayor requisito de información para clasificar el ejemplo.

La ventaja de una división se determina comparando la entropía del grupo original con la entropía combinada de los subconjuntos resultantes después de la división. En otras palabras, si un grupo (denominado grupo *i*) se divide en varios subconjuntos (denominado grupo *j*), el cálculo tiene en cuenta la diferencia entre la entropía del grupo *i* y la entropía total del grupo *j*.

$$gain_{ij} = H_i - \sum_j H_j * R_j$$

En donde, *Rj* representa la correlación entre el número de casos en el grupo j y los del grupo i. Esta métrica sirve como evaluación teórica de la importancia de la información adquirida de la división. Cuantifica el grado en que la división influye en la distribución de la variable dependiente, indicando así el nivel de discriminación alcanzado.

Al evaluar la efectividad de una división, es importante normalizar el valor de ganancia para garantizar que las divisiones más pequeñas y los grupos más grandes tengan preferencia sobre los más pequeños. Sin esta normalización, el algoritmo tendería a generar grupos muy pequeños, potencialmente tan pequeños como el tamaño 1, lo que minimizaría la entropía pero los volvería inútiles para fines predictivos, ya que probablemente estarían sobreajustados a los datos.

¿La generación de contrafactuales es simplemente la ejecución de un algoritmo de agrupamiento? De hecho, es mucho más amplio. Se puede utilizar cualquier herramienta o método que pueda producir declaraciones similares a reglas y evaluarlas. En los ejemplos que hemos examinado, los resultados estaban predeterminados y se incluyeron en los datos para simplificar. Sin embargo, la evaluación también puede ser dinámica, lo que requiere la ejecución de un programa como una simulación de Monte Carlo o el entrenamiento de una red neuronal para generar una puntuación de idoneidad para la hipótesis alternativa. Incluso si el vector de resultados puede calcularse previamente para cada estado de la naturaleza registrado, evaluar el contrafactual todavía conlleva un costo significativo.

Recuerde que en el ejemplo de la predicción de acciones, el rendimiento depende de varios factores, incluidos los rendimientos y la volatilidad de los rendimientos asociados con la "ejecución" de una estrategia comercial contrafactual o hipotética. Evaluar esto implica generar un resultado que no existe en una base de datos existente. En otras palabras, "aplicar" la relación hipotética es una tarea más extensa y costosa desde el punto de vista computacional en comparación con simplemente analizar los datos. Requiere anticipar las consecuencias de una acción.

4.1 La evaluación de contrafactuales

El poder de los contrafactuales radica en su capacidad de revertir el proceso típico de análisis de datos basado en consultas. En lugar de preguntar qué datos se ajustan a un patrón determinado, los contrafactuales nos incitan a considerar qué patrones se alinean con los datos disponibles. Esta inversión nos permite automatizar el proceso creativo de generar, evaluar y refinar hipótesis, que es un aspecto crucial en el desarrollo de teorías.

Si bien, es importante reconocer un problema importante que surge con el proceso de extracción de datos. Si se ejerce suficiente esfuerzo y se dedica una cantidad sustancial de tiempo a la tarea, es muy probable que un modelo eventualmente produzca resultados favorables cuando se emplee en la presa. En consecuencia, se vuelve imperativo determinar si el modelo generado simplemente se ajusta a los datos existentes o si realmente representa una correlación resiliente entre las variables del problema que probablemente persistirá en el futuro.

Independientemente del método utilizado para la evaluación, es crucial reconocer patrones que no sean simplemente coincidentes desde un punto de vista estadístico. Cuando se trata de predecir resultados, es esencial simular una

hipótesis durante un período prolongado y potencialmente en escenarios alternativos. En el caso de evaluar la rentabilidad del cliente, la evaluación a lo largo de un cronograma puede no ser particularmente relevante, mientras que la categorización basada en factores geográficos podría ofrecer un enfoque más significativo y revelador para segmentar la población.

En los problemas en los que el tiempo desempeña un papel importante, las reglas derivadas de los datos tienen muchas más probabilidades de ser sólidas si dividimos los datos por universo y tiempo. Cuando los datos son abundantes, esto garantiza que el algoritmo de descubrimiento sólo perseguirá aquellas hipótesis que funcionen de forma coherente en los distintos conjuntos de datos, en lugar de aquellas que sean muy buenas en algunos subconjuntos y malas en otros.

Para ello, antes del ejercicio de extracción de datos, hay que pensar detenidamente por qué se espera que un modelo funcione de manera uniforme en los distintos subconjuntos de datos. En la práctica, estos subconjuntos pueden corresponder a distintos segmentos de clientes, distintos periodos de tiempo, etcétera.

Otra forma de ver lo anterior es que la minería de datos no es un ejercicio de pesca ascendente no dirigido. Los experimentos de minería de datos llevan tiempo de preparación e interpretación. Estos experimentos no son gratuitos; de hecho, pueden ser bastante caros. Por esta razón, es importante formular el problema cuidadosamente con hipótesis a priori sobre los tipos de relaciones que deberíamos esperar descubrir y por qué. En otras palabras, aunque el ejercicio de extracción de datos es "exploratorio", como señalaba Tukey, la exploración debe plantearse lo más cuidadosamente posible al inicio del ejercicio.

4.2 Uso de los patrones descubiertos

En la sección anterior he enfatizado la importancia de examinar a fondo los contrafactuales. Después de considerar esto, una pregunta lógica que surge naturalmente es: si los patrones identificados son genuinamente sólidos, ¿por qué no eliminar completamente de la ecuación al que interviene?

La respuesta a esta pregunta no está determinada únicamente por la medida en que se desglose el problema. Numerosos marcos en el dominio de la literatura sobre sistemas de soporte a la decisión (DSS), surgen del concepto de Simon de decisiones programables versus no programables. El concepto fundamental gira en torno a la noción de que los problemas no programables requieren el juicio humano, por lo que los modelos sirven para ayudar y facilitar el juicio humano en lugar de reemplazarlo con la automatización.

Aun cuando, el debate en torno a la automatización versus el soporte no está determinado únicamente por un factor. Hay otros elementos cruciales que entran en juego. Uno de esos elementos es el nivel de claridad en la función de pago. Los diferentes marcos de sistemas de apoyo a la decisión (DSS) suponen que los problemas con funciones de resultados ambiguas son los más desafiantes. Sin embargo, el ejemplo de la predicción del mercado demuestra que incluso cuando la función de recompensa es precisa, el problema en sí puede seguir estando estructurado de manera inadecuada.

¿Es lógico automatizar la toma de decisiones basándose únicamente en el hecho de que la función de recompensa está bien definida y puede calcularse mediante un modelo? La respuesta a esta pregunta depende de otro factor: el objetivo de la extracción de datos. Si el objetivo es desarrollar una teoría sobre el dominio del problema, es crucial sacar del proceso a quien toma las decisiones. El modelo debe probarse sin ninguna intervención humana; de lo contrario,

resulta imposible diferenciar la contribución de los humanos de la del modelo basado en el aprendizaje automático.

Lo anterior plantea una interesante paradoja. Cuando utilizamos el aprendizaje automático y métodos contrafactuales para encontrar patrones en situaciones complejas no lineales donde el resultado del modelo es una decisión, debemos tener cuidado de asegurarnos de que sea el modelo aprendido el que distinga entre ruido y señal, y no el responsable de la decisión en sí. De lo contrario, no podremos determinar si el modelo aprendido captura con precisión la verdadera estructura del problema subyacente o si el juicio humano está compensando un modelo defectuoso.

Cuando el modelo es eficaz y realmente captura la estructura de datos subyacente, anticipamos que funcionará bien incluso con datos invisibles. Sin embargo, si los resultados se lograron mediante intervención humana, resulta difícil determinar si nuestro modelo aprendido realmente capturó un efecto genuino en los datos. No podemos determinar si quien tomó las decisiones lo dirigió eficazmente cuando tomó malas decisiones o mal cuando tomó buenas decisiones.

El punto señalado anteriormente es que puede parecer contradictorio, pero incluso en situaciones en las que las decisiones están mal programadas, es necesario utilizar soluciones totalmente automatizadas para probar el modelo. En el ejemplo proporcionado, donde se está probando una teoría sobre el comportamiento del mercado, el nivel de precisión en la definición de la función de recompensa y el objetivo del ejercicio de modelado tienen una mayor influencia en si la decisión es automatizada o asistida, que la complejidad de la decisión. problema en sí.

A pesar de la complejidad de la decisión, la función de recompensa está bien definida, ya que se basa en la ganancia o pérdida de una operación y los datos necesarios para calcularla están fácilmente disponibles. La automatización proporciona un método confiable y completo para probar teorías.

Por otro lado, es posible que los problemas que están más organizados y bien definidos no se presten fácilmente a la automatización. Tomemos, por ejemplo, el escenario de un gerente de ventas que intenta determinar estrategias de ventas efectivas que generen ganancias. Resulta una tarea desafiante identificar una ecuación exacta que relacione las acciones realizadas por los vendedores con las ganancias resultantes.

Digamos que se establece una correlación entre el tamaño de una operación y su rentabilidad; no puede utilizarse simplemente como modelo para automatizar el comportamiento de los vendedores. Aparte del hecho de que esta correlación no puede transformarse en un conjunto preciso de instrucciones, también sería extremadamente difícil determinar si los vendedores están realmente intentando generar acuerdos más importantes. Además, puede que no exista una relación precisa entre sus acciones y los resultados financieros que logran.

La clasificación se basa en el nivel de precisión de la función de pago y los objetivos teóricos que se abordan. Cuando consideramos el cuadrante superior izquierdo, que puede ejemplificarse con el ejemplo del comercio, resulta evidente que la automatización es una opción lógica. Este cuadrante se caracteriza por la disponibilidad de datos relevantes necesarios para la formación de la teoría y una función de resultados bien definida.

Cuando se trata de problemas de gestión, la presencia de intangibles y riesgos asociados con la automatización dificulta la definición precisa de las

funciones de compensación. Sin embargo, el sector financiero destaca como una excepción porque muchas tareas de toma de decisiones en áreas como el comercio, la gestión de riesgos y la cobertura tienen funciones de resultados bien definidas que pueden especificarse con precisión.

En consecuencia, es posible que en el futuro seamos testigos de una mayor automatización de los procesos de toma de decisiones en la industria financiera. Esta tendencia ya se ha observado hasta cierto punto con la introducción de tecnologías que reemplazan a los humanos en áreas como el comercio programado y el arbitraje de riesgos.

4.3 Minería de datos en riesgos financieros

Las técnicas de minería de datos se han vuelto cada vez más frecuentes en el campo de las finanzas, con aplicaciones que van desde la gestión del riesgo crediticio hasta la detección de fraude. Recientemente, ha habido una tendencia creciente a utilizar técnicas de extracción de datos para la detección de riesgos financieros empresariales. Esto es particularmente importante en países en desarrollo, donde el riesgo financiero de las empresas tiene una importancia significativa para los administradores.

Varios factores, incluidas las condiciones macroeconómicas, el desempeño de la industria y el entorno empresarial general, pueden contribuir a las crisis financieras dentro de las empresas. Estas crisis pueden provocar importantes pérdidas económicas empresariales e incluso quiebras, lo que genera pérdidas sustanciales para los acreedores, accionistas, inversores y empleados que se enfrentan al desempleo. En consecuencia, la financiación empresarial tiene un profundo impacto en los intereses de diversas partes interesadas. Asimismo, las crisis financieras que enfrentan empresas individuales pueden tener efectos dominó en industrias enteras e incluso en la economía en general.

Muchas personas están trabajando para identificar los factores clave que contribuyen al fracaso financiero dentro de las empresas y tomar las medidas preventivas necesarias para evitar crisis futuras. Los sistemas de alerta temprana de crisis financieras son de vital importancia para los administradores y con este fin se emplean numerosos métodos de previsión. El objetivo central es predecir con precisión el riesgo financiero determinando los factores críticos e implementando medidas preventivas.

Las técnicas de minería de datos se han utilizado ampliamente para detectar crisis financieras, siendo el enfoque predominante actualmente el uso de una única técnica. Sin embargo, las técnicas individuales a menudo carecen de precisión suficiente para crear modelos de predicción altamente precisos, lo que expone a las empresas a pérdidas significativas cuando se materializan posibles riesgos financieros.

Históricamente, los investigadores se han basado en modelos estadísticos como el modelo de puntuación Z de Edward Altman, que empleaba regresión lineal múltiple, para predecir crisis financieras. Altman, Haldeman y Narayanan desarrollaron posteriormente un modelo mejorado llamado ZETA, que superó la precisión del modelo de puntuación Z. Ohlson también introdujo el modelo de regresión logística, logrando una precisión aún mayor que los métodos anteriores (Zhang, et al., 2013). En los últimos años, se han aplicado cada vez más técnicas de minería de datos en este campo, incluidos árboles de decisión, clasificación bayesiana, redes neuronales y máquinas de vectores de soporte. Sin embargo, es posible que una sola técnica no satisfaga las necesidades de toma de decisiones, lo que nos lleva a combinar dos técnicas en este artículo para mejorar los resultados.

La transformación digital es el proceso de integración de las empresas en la economía digital en constante cambio mediante la utilización de tecnologías de la información. Esta transformación desafía constantemente a las empresas a tener un conocimiento profundo de sus procesos y operaciones para poder adaptarse y evolucionar. Para ayudar en esto, se puede utilizar un marco orientado a procesos para identificar puntos de contacto con el cliente, determinar flujos de trabajo y especificar requisitos de datos durante todo el proceso de creación de valor (Najem et al., 2022).

La gestión de procesos de negocio (BPM) es un concepto que se ha vuelto cada vez más significativo para lograr operaciones eficientes, reducciones de costos y mejoras en la calidad y la productividad. A lo largo de los años, la investigación y la práctica han introducido varios enfoques de BPM con diferentes objetivos y resultados.

Los principios subyacentes de estos conceptos giran en torno a la idea de que el valor de un proceso, así como su necesidad y capacidad de mejora continua, dependen de varios atributos, incluida su importancia, solidez y capacidad de prosperar. Si bien, la viabilidad económica de las iniciativas de mejora está limitada por factores como la creciente complejidad de los proyectos y la necesidad de cualificaciones especializadas, que restringen el número de procesos que pueden gestionarse eficazmente en un momento dado. En consecuencia, los procesos se priorizan meticulosamente y se eligen en consecuencia. Vale la pena señalar que no se ha cuantificado el alcance del potencial de mejora no aprovechado.

Según la teoría de la economía de cola larga de Fisher et al. (2021), proponen que el potencial de mejora de procesos en una empresa u organización sigue una distribución conocida como cola larga. Esto significa que hay un

pequeño número de procesos que tienen un alto potencial de mejora, formando la “cabeza corta” de la distribución.

Por otro lado, existen numerosos procesos de menor valor que a menudo se pasan por alto y no se consideran para su optimización, lo que constituye la "cola larga". Es importante señalar que la teoría sugiere que el potencial acumulado de mejora en la cola larga no es insignificante, sino que representa una cantidad sustancial de valor sin explotar, de acuerdo con el principio de Pareto.

4.4 La teoría de la larga cola de los procesos empresariales

La Gestión de Procesos de Negocio (BPM) es un enfoque integral que abarca varios métodos, técnicas y herramientas para identificar, descubrir, analizar, rediseñar, implementar y monitorear de manera efectiva los procesos de negocio a lo largo de todo su ciclo de vida. Implica el uso de sistemas de información conscientes de los procesos, que son sistemas automatizados que ejecutan procesos basados en modelos de procesos predefinidos. Estos sistemas están diseñados no sólo para manejar la ejecución de procesos sino también para gestionar las aplicaciones, las personas y la información que son parte integral del proceso general.

Muchas empresas adoptan un enfoque centralizado para la gestión de procesos de negocio (BPM) mediante el establecimiento de un centro de excelencia dedicado a BPM. Esta iniciativa central aúna recursos, conocimientos y competencias, permitiendo la especialización y la generación de efectos de aprendizaje. El propósito de este centro BPM es obtener beneficios de estos esfuerzos colectivos. Sin embargo, debido a la limitación de recursos, las empresas necesitan priorizar ciertos procesos sobre otros. Es esencial centrarse en los procesos que tienen el potencial de generar los mayores beneficios

esperados, ignorando aquellos que pueden tener poco impacto directo en el éxito general del negocio.

Para comprender mejor este concepto, Fisher et al. (2021), se basó en la economía de cola larga de Anderson y desarrolló la teoría de procesos de negocio de cola larga. Según esta teoría, BPM puede verse como un problema de maximizar los beneficios de las empresas. En otras palabras, las empresas apuntan a maximizar sus ganancias considerando los excedentes y costos asociados con BPM. El superávit esperado representa los beneficios que se obtienen al implementar mejoras estructuradas en los procesos. Por otro lado, la complejidad de las operaciones de una empresa, las inconsistencias en la toma de decisiones y la disminución general de la eficiencia en grandes proyectos contribuyen a reducir el excedente obtenido con BPM.

Además, existen costos de instalación para capacitación de empleados y soporte informático, así como costos de gestión directa asociados con un proceso específico. Estos costos están influenciados por el nivel de estandarización y la difusión del conocimiento dentro de la organización, lo que puede generar efectos de aprendizaje.

En muchos casos, las empresas optan por manejar procesos que ofrecen un equilibrio favorable entre los excedentes previstos y los costos debido a consideraciones económicas. La línea de manejabilidad representa el punto en el que ambos factores alcanzan un equilibrio, y este concepto tiene asociaciones históricas con el principio de Pareto.

4.5 El minado y descubrimiento de los procesos

La minería de procesos se centra en la recopilación, el análisis y la interpretación de datos procedentes de registros de eventos de procesos. Los registros de eventos comprenden información sobre ejecuciones de procesos

extraída de sistemas de información conscientes de los procesos, que conservan mensajes, transacciones o modificaciones. Entre otras cosas, los registros de eventos especifican los nombres o ID de los objetos de negocio procesados, las partes interesadas de la empresa (es decir, los recursos utilizados), el momento en que se produjeron los eventos y, posiblemente, los datos de negocio. La minería de procesos se utiliza para las tareas de descubrimiento de procesos, comprobación de la conformidad de los procesos y mejora de los procesos.

La investigación ha desarrollado varias técnicas para implementar el descubrimiento de procesos en la práctica. Aunque algunos procesos persiguen el mismo objetivo, pueden implicar diferentes secuencias de actividades, que se fundamentan en el concepto de agrupación de trazas, que es adecuado para descomponer los registros de eventos en subconjuntos homogéneos, por ejemplo procesos o variantes de procesos. Entre otros, los modelos *bag-of-activities, hamming distance, generic edit distance y n-gram* describen los enfoques más populares para la agrupación de trazas.

El enfoque de *bag-of-activities* no tiene en cuenta la información contextual ni el orden en que se ejecutan las actividades. Frente a este inconveniente, los modelos de *n-gramas* no sólo analizan las actividades aisladas, sino que también incorporan las actividades anteriores y posteriores como contexto de una actividad. La distancia de Hamming o *hamming distance* operativiza el número de posiciones de caracteres en las que dos secuencias difieren y se limita a trazos con la misma longitud.

Este problema puede atenuarse con enfoques de alineación de secuencias. La distancia de edición o *edit distance* penaliza mucho las secuencias con longitudes diferentes y no es adecuada para analizar registros de eventos de la vida real. En este contexto, Bose y van der Aalst introdujeron una distancia de

edición generalizada que tiene en cuenta las propiedades de los registros de procesos. Sin embargo, el cálculo de estas puntuaciones resultó inviable para los conjuntos de datos más grandes. Además, muchas medidas de clúster interno no funcionan bien con medidas de distancia personalizadas, ya que los centros de clúster no se pueden calcular trivialmente.

4.6 La medición del rendimiento de los procesos

La medición del desempeño de los procesos es un aspecto esencial de BPM y ofrece a las empresas la oportunidad de medir sus capacidades de gestión. Más allá de simplemente recopilar datos pertinentes, este proceso implica un análisis exhaustivo de qué tan bien se alinea un proceso particular con objetivos y criterios predeterminados. A través de este análisis, las empresas pueden establecer una base sólida para tomar decisiones de priorización y asignar de manera efectiva los recursos organizacionales a las iniciativas de mejora más ventajosas.

La medición del desempeño ha sido ampliamente investigada y explorada como un tema multidisciplinario en el pasado. Diferentes autores han propuesto varios métodos, marcos y conceptos, entre los que destacan el cuadro de mando integral y EFQM[1]. A pesar de sus supuestos y objetivos compartidos, estos métodos a menudo difieren en términos de su enfoque, como si operan a nivel empresarial o a nivel de proceso, así como los tipos de medidas que emplean para evaluar el desempeño del proceso, como las basadas en el desempeño. o medidas no basadas en el desempeño (Martínez, 2008).

[1] La abreviatura EFQM significa Fundación Europea para la Gestión de la Calidad, que es una organización sin fines de lucro fundada en el año 1988 por un colectivo de 14 empresas europeas. El objetivo principal de esta organización es emprender la tarea de formular un modelo de excelencia diseñado específicamente para la región europea.

Los métodos que no dependen del desempeño implican la evaluación del desempeño del proceso utilizando un conjunto predeterminado de criterios. Estos criterios pueden ser establecidos por las empresas basándose en factores críticos de éxito o recopilando aportaciones de los empleados involucrados. Luego se pide a las partes interesadas que evalúen individual o colectivamente el desempeño del proceso, utilizando una escala que capture percepciones tanto positivas como negativas.

Al combinar estas evaluaciones individuales, las empresas pueden obtener información valiosa sobre los factores que contribuyen o dificultan el éxito empresarial. Sin embargo, es importante señalar que los métodos no basados en el desempeño tienen algunos inconvenientes, como ser lentos, complejos y susceptibles a sesgos de evaluación. Si bien estos métodos proporcionan evidencia cualitativa para medir el desempeño del proceso, pueden ser costosos e inadecuados para un análisis frecuente debido a su complejidad y al desafío de obtener resultados claros. Fischer y sus colegas han utilizado métodos no basados en el desempeño para examinar la distribución de procesos en una pequeña y mediana empresa.

Por otro lado, los métodos basados en el desempeño brindan a las empresas la capacidad de evaluar automáticamente el estado actual de un proceso utilizando los datos recopilados durante su ejecución. Estos métodos permiten comparar los indicadores de desempeño con puntos de referencia establecidos o rangos de valores aceptables. Además, ayudan a evaluar los riesgos asociados con las modificaciones del proceso y a monitorear la progresión a largo plazo de un proceso de manera ágil y precisa.

La eficacia de los métodos basados en el desempeño depende en gran medida de la presencia de sistemas de información conscientes de los procesos.

Si bien estos métodos comparten similitudes con la minería de procesos, su enfoque no radica en el descubrimiento de procesos, las pruebas de cumplimiento y la mejora, sino más bien en el análisis de los datos de ejecución para identificar patrones y relaciones a través de un procedimiento de extracción-transformación-carga. Este enfoque promueve la precisión y la reproducibilidad, facilitando así la evaluación de los indicadores de desempeño, la complejidad del proceso, los defectos de diseño existentes y las áreas potenciales de mejora.

4.7 Priorización de procesos

El tema de la mejora de procesos es complejo y diverso, ya que depende en gran medida de la selección de procesos apropiados. Por lo tanto, al priorizar los procesos, es crucial considerar no sólo su desempeño en costos sino también otros factores. Investigaciones anteriores y aplicaciones prácticas han introducido varios conceptos para medir el desempeño del proceso. El enfoque de Rosales et al. (2003) y Paiva et al. (2019), presentan los criterios como de importancia estratégica, saludable o disfuncional y viables, que pueden usarse para evaluar el potencial de mejora en un proceso y guiar la toma de decisiones durante su priorización.

- La importancia se refiere a la transcendencia estratégica de un proceso, determinada por su influencia en los objetivos estratégicos. En consecuencia, los procesos que se alinean estrechamente con las actividades comerciales clave y la estrategia corporativa tienen mayor valor. Evaluar la importancia de un proceso requiere una comprensión integral de la propia trayectoria estratégica.
- La salud es una medida integral que evalúa la condición actual de un proceso, proporcionando información valiosa sobre qué tan bien se alinea con estándares predeterminados de calidad y desempeño. No sólo llama

la atención sobre los procesos que son particularmente susceptibles a errores, sino que también identifica aquellos que muestran una falta de satisfacción de los empleados y clientes. Derivado del ámbito de la ciencia médica, el término implica inherentemente una perspectiva positiva, asumiendo que prevalece un estado de normalidad en lugar de disfunción.

- Al final, la viabilidad se refiere al grado en que un proceso puede ajustarse y su susceptibilidad a limitaciones culturales y políticas. Cuando se trata de mejorar los procesos, se debe dar prioridad a aquellos que encuentran barreras mínimas, no sólo en relación con la resistencia de los empleados al cambio, sino también considerando las limitaciones impuestas por la tecnología.

Aunque las empresas tienen acceso a varios indicadores de desempeño que pueden aplicarse a diferentes dimensiones, a menudo necesitan ajustarse para adaptarse a sus contextos organizacionales específicos. Por ejemplo, se han identificado un conjunto completo de indicadores que pueden utilizarse fácilmente en empresas minoristas, mientras que existen conceptos similares para las industrias manufactureras y la administración pública. Sin embargo, estos conceptos se adaptan a sectores específicos y se centran en medir factores específicos de la empresa, como el inventario mínimo, la depreciación de las existencias o el índice de pedidos perfecto. Como resultado, carecen de la capacidad de analizar datos de ejecución de procesos independientes de la industria y, a menudo, requieren información comercial adicional que puede no estar disponible en las configuraciones de registro de eventos estándar. Por lo tanto, existe la necesidad de indicadores más generalizados que puedan

derivarse de un número limitado de atributos proporcionados por los sistemas de información sensibles a los procesos.

Kratsch y sus colegas proponen un enfoque novedoso que va más allá de centrarse en indicadores individuales al priorizar procesos. En cambio, utilizan un método basado en datos que incorpora varias características, como tiempos de ejecución y uso de materiales, para predecir con precisión el desempeño futuro de un proceso. Luego, este enfoque genera una lista completa de proyectos de mejora con prioridades programadas. De manera similar, Lehnert y su equipo presentan ProcessPageRank, que evalúa los procesos en función de sus requisitos específicos para una mejora ajustada a la red. Por otro lado, Ohlsson y su equipo emplean dos componentes, a saber, un mapa de calor del proceso y un mapa de categorización. Mientras que el mapa de calor actúa como un modelo detallado para el análisis de procesos, el mapa de categorización ayuda a determinar la brecha de rendimiento existente de un proceso.

4.8 Modelo de alerta temprana de financiación de empresas con minería de datos

4.8.1 El modelo SVM

La SVM, también conocida como máquina de vectores de soporte, es una técnica de aprendizaje supervisado ampliamente utilizada para tareas de clasificación y regresión de datos. Se considera un modelo de aprendizaje automático increíblemente robusto y flexible que ha ganado considerable popularidad en diversos campos, como los sistemas financieros de alerta temprana (Huang et al., 2005). Su notable capacidad para procesar datos con numerosas dimensiones y lidiar con límites de clases complejos lo ha convertido en una opción extremadamente favorecida para una amplia gama de operaciones de minería de datos. Su objetivo principal es identificar el hiperplano más

adecuado que separe efectivamente el conjunto de datos dado en distintas clases. Por lo tanto, descubre el hiperplano que maximiza la brecha o margen entre puntos de datos que pertenecen a diferentes clases, a menudo denominado margen máximo.

Cuando se trata de datos que no pueden separarse por una línea recta, SVM emplea un método conocido como "truco del núcleo" para convertir los datos en un espacio de dimensiones superiores donde se pueden separar por una línea recta. SVM se utiliza ampliamente en la minería de datos y ha encontrado aplicaciones en varios dominios, como la identificación de spam, la clasificación de imágenes y la predicción de riesgos financieros.

Es importante tener en cuenta que SVM funciona como un clasificador binario, lo que significa que únicamente puede categorizar datos en dos clases distintas. No obstante, su funcionalidad se puede ampliar para manejar tareas de clasificación de múltiples clases mediante la utilización de diversas técnicas, incluidas metodologías uno contra uno o uno contra todos.

Una de las características notables de SVM es su eficacia para tratar datos que tienen una gran cantidad de dimensiones. Esto lo hace particularmente adecuado para tareas como clasificar texto o reconocer imágenes. Otro aspecto clave de SVM es la inclusión de un parámetro de regularización, denominado C, que desempeña un papel en el equilibrio de la importancia de maximizar el margen y minimizar los errores de clasificación. Cuando C se establece en un valor más bajo, el margen resultante será más amplio, lo que permitirá una mayor flexibilidad en la clasificación, pero podría dar lugar a algunas clasificaciones erróneas. Por el contrario, cuando C se establece en un valor más alto, el margen se vuelve más estrecho, lo que potencialmente resulta en una clasificación más precisa, pero también corre el riesgo de sobreajustar el modelo.

Dentro del ámbito de los sistemas de alerta temprana financiera, las máquinas de vectores de soporte (SVM) se han convertido en una herramienta confiable para pronosticar posibles dificultades financieras y quiebras dentro de las empresas. El modelo SVM, en este caso, se entrena utilizando una amplia gama de ratios financieros y otros indicadores financieros pertinentes. El resultado resultante es una clasificación binaria que determina efectivamente si la empresa en cuestión es susceptible a dificultades financieras o permanece segura.

4.8.2 Modelo logístico de alerta financiera

El modelo logístico de alerta financiera, es una técnica de minería de datos ampliamente utilizada para la predicción de riesgos financieros, que se basa en la regresión logística, una técnica estadística utilizada para modelar la probabilidad de un evento binario, como la quiebra de una empresa. Utiliza una serie de variables financieras, como el flujo de caja, el endeudamiento, la rentabilidad y la liquidez, para predecir la probabilidad de que una empresa experimente dificultades financieras en el futuro.

Estas variables se utilizan para construir un modelo de regresión logística que puede predecir la probabilidad de que una empresa experimente dificultades financieras en el futuro. Este modelo se ha utilizado ampliamente en la industria financiera para la evaluación del riesgo crediticio y la toma de decisiones de inversión. También se ha utilizado en la investigación académica para la predicción de la quiebra de empresas y la identificación de factores de riesgo financieros. Sin embargo, tiene algunas limitaciones, en particular, puede ser difícil de interpretar y puede ser sensible a la selección de variables y la calidad de los datos. Además, el modelo logístico no tiene en cuenta la interacción entre las variables y puede no ser adecuado para la predicción de eventos raros.

4.8.3 Modelo de alerta temprana financiera con fusión de información

Este modelo integra las fortalezas respectivas de diferentes métodos de minería de datos, como SVM y regresión logística, para mejorar la tasa de precisión de predicción. El enfoque fusiona los diferentes resultados de minería de datos para obtener resultados de predicción confiables para la toma de decisiones. Los resultados muestran que el enfoque propuesto supera a los modelos SVM y Logístico individuales en términos de tasa de precisión de predicción.

En general, el modelo de alerta temprana financiera con fusión de información es un enfoque prometedor para mejorar la precisión de la predicción de la angustia financiera utilizando métodos de minería de datos. Al integrar las fortalezas de diferentes modelos, el enfoque puede proporcionar resultados de predicción más confiables para la toma de decisiones en la alerta temprana financiera.

La mayor parte de la literatura desarrollada se concentró en evaluar el riesgo de instituciones financieras específicas basadas en indicadores de performance del banco en particular. La poca capacidad anticipativa de estos índices frente a los fracasos bancarios observados en la década pasada han generado preocupación sobre cuáles pudieron ser los determinantes sistémicos más que idiosincrásicos de estas crisis.

En este sentido, hay un reconocimiento creciente de la relevancia del ambiente macroeconómico y la salud del sistema financiero en el desempeño de los indicadores de performance bancario individuales. La literatura de alerta temprana y de predicción de crisis bancarias puede ser clasificada en dos, según sea el alcance de la predicción: crisis bancaria individual o crisis sistémica; o según la metodología empleada: indicadores de performance bancarios

cualitativos, enfoque de extracción de señales, modelos de estimación de variables dependientes dicotómica, modelos de duración y de redes neuronales, entre otros.

Conclusiones

En cuanto a desafíos de los sistemas de gestión de riesgos, la existencia de elementos intangibles y sesgos vinculados a la automatización, complica la determinación exacta de las funciones de compensación. No obstante, el sector financiero sobresale debido a que numerosas tareas de toma de decisiones en campos como las cuentas nacionales, la administración pública y la estratificación de la oferta y la demanda global, poseen funciones de resultados claramente establecidas que pueden ser detalladas con exactitud, empleando algoritmos de Redes Neuronales Artificiales.

Por lo tanto, podríamos observar en el futuro un incremento en la automatización de los procesos de decisión en el sector financiero. Esta tendencia ya ha sido detectada hasta cierto grado con la implementación de tecnologías que sustituyen a los humanos en campos como el comercio automatizado y el arbitraje de riesgos.

Las técnicas de minería de datos se han vuelto cada vez más frecuentes en el campo de las finanzas, con aplicaciones que van desde la gestión del riesgo crediticio hasta la detección de fraude. Recientemente, ha habido una tendencia creciente a utilizar técnicas de extracción de datos para la detección de riesgos financieros empresariales. Esto es particularmente importante en países en desarrollo, donde el riesgo financiero de las empresas tiene una importancia significativa para el Estado.

Varios factores, incluidas las condiciones macroeconómicas, el desempeño de la industria y el entorno empresarial general, pueden contribuir a las crisis financieras dentro de las empresas. Estas crisis puede provocar importantes pérdidas económicas empresariales e incluso quiebras, lo que genera pérdidas sustanciales para los acreedores, accionistas, inversores y empleados que se

enfrentan al desempleo. En consecuencia, la financiación empresarial tiene un profundo impacto en los intereses de diversas partes interesadas. Asimismo, las crisis financieras que enfrentan empresas individuales pueden tener efectos dominó en industrias enteras e incluso en la economía en general.

Las empresas del sector finanzas están trabajando para identificar los factores clave que contribuyen al fracaso financiero dentro de las empresas y tomar las medidas preventivas necesarias a través de métodos estadísticos paramétricos o no, asistido por software con entrenamiento de datos, esto es lo que se conoce como "sistemas de alerta temprana de crisis financieras", es decir, métodos de previsión.

El resultado clave de la usabilidad de estos algoritmos es predecir con precisión; el riesgo financiero, que organizan datos multidimensionales en una representación bidimensional o tridimensional, de manera que los datos similares se agrupen cercanamente. Esto facilita la visualización y comprensión de patrones en grandes conjuntos de datos y la detección de relaciones entre ellos.

Bibliografía

Ban, G. Y., El Karoui, N., & Lim, A. E. (2018). Machine learning and portfolio optimization. *Management Science*, *64*(3), 1136-1154.

Broby, D. (2022). The use of predictive analytics in finance. *The Journal of Finance and Data Science*, *8*, 145-161.

Dhar, V. (1998). Data mining in finance: using counterfactuals to generate knowledge from organizational information systems. *Information Systems*, *23*(7), 423-437.

Echeverri, L.A., Retamoza, A.M.P., de la Rosa, M.O., Barros, I.V., Alvarez, D.D. O., y Guerrero, E.C. (2013). Minería de datos como herramienta para el desarrollo de estrategias de mercadeo B2B en sectores productivos, afines a los colombianos: una revisión de casos. *Sotavento MBA*, (22), 126-136.

Evans, P. (2015). De la deconstrucción a los big data: cómo la tecnología está transformando a las empresas. En BBVA (Eds.), *Reinventar la empresa en la era digital* (pp. 17-36).

Fernández, E., Menchero, A., Olmeda, I., & Insua, D. R. (2000). Redes neuronales artificiales en finanzas. *Bolsa y estadística bursátil*, 273-290.

Fischer, M., Hofmann, A., Imgrund, F., Janiesch, C., & Winkelmann, A. (2021). On the composition of the long tail of business processes: Implications from a process mining study. *Information Systems*, *97*.

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of accounting studies*, *9*, 5-34.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & operations research*, *32*(10), 2513-2522.

García Salgado, O. y Morales Castro, A. (2016). Desempeño financiero de las empresas: Una propuesta de clasificación por RNA. *Dimensión Empresarial*, *14*(2), 11-23.

Goodell, J.W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and

research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance, 32.*

Guerra, P., Castelli, M., & Côrte-Real, N. (2022). Machine learning for liquidity risk modelling: A supervisory perspective. *Economic Analysis and Policy*, *74*, 175-187.

Lehnert, M., Röglinger, M., & Seyfried, J. (2018). Prioritization of interconnected processes. *Business & Information Systems Engineering, 60,* 95-114.mexiucanas

Ma, L., & Pohlman, L. (2008). Return forecasts and optimal portfolio construction: a quantile regression approach. *The European Journal of Finance*, *14*(5), 409-425.

Martínez, B. (2008). Calidad.¿ Qué es el modelo EFQM (European Foundation for Quality Management)?. *Anales de pediatría continuada*, *6*(5), 313-318.

Najem, R., Amr, M. F., Bahnasse, A., & Talea, M. (2022). Artificial Intelligence for Digital Finance, Axes and Techniques. *Procedia Computer Science*, *203*, 633-638.

Paiva, F. D., Cardoso, R. T. N., Hanaoka, G. P., & Duarte, W. M. (2019). Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications, 115*, 635-655.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *IFC Bulletins chapters*, *49*.

Rosales, S.G., Ramos, I. R., y Cuervo, C.M. (2003). Revisión del estado del arte de la aplicación de redes neuronales artificiales en economía y finanzas. In *Emergent solutions for the information and knowledge economy: proceedings of the Tenth International Association for Fuzzy-Set Management and Economy Congress. León, October 9-11, 2003* (pp. 167-184). Secretariado de Publicaciones y Medios Audiovisuales.

Samaniego Alcántar, Á., y Mongrut, S. (2014). Relación entre la creación de valor y la inversión en I+D: una aproximación mediante redes neuronales artificiales. *Innovar*, *24*(51), 19-30.

Zhang, L., Zhang, L., Teng, W., & Chen, Y. (2013). Based on information fusion technique with data mining in the application of finance early-warning. *Procedia Computer Science*, *17*, 695-703.

De esta edición de *"Ciencia de datos en sistemas de gestión de riesgos: Enfoque hacia la minería de datos"*, se terminó de editar en la ciudad de Colonia del Sacramento en la República Oriental del Uruguay el 24 de octubre de 2024

EST. 2021 EMC
EDITORIAL MAR CARIBE

CIENCIA DE DATOS EN SISTEMAS DE GESTIÓN DE RIESGOS: ENFOQUE HACIA LA MINERÍA DE DATOS

LIBRO DE INVESTIGACIÓN

ISBN: 978-9915-9706-9-1

2024