UNIVERSIDAD TECNOLÓGICA METROPOLITANA
UTEM del Estado de Chile

Facultad de Ingeniería

Escuela de Informática

Predicción de comportamiento de clientes en canal web

Diego Oyarce Trejo
doyarce@utem.cl

Marcelo Tapia Riquelme
marcelo.tapiar@utem.cl

Cristobal González Gárate
cristobal.gonzalezg@utem.cl

30 de junio de 2023

Índice general

Índice de figuras

Capítulo 1

Presentación del proyecto

1.1. Resumen

El presente documento de propuesta de Trabajo de Titulación tiene como objetivo mostrar la forma y el plan de trabajo que se utilizan a lo largo del proceso de desarrollo del proyecto propuesto.

El proyecto tiene como objetivo fundamental analizar el comportamiento de los clientes de AFP Capital y sus preferencias de uso en un período igual o inferior a 6 meses, para predecir navegaciones futuras personalizadas.

Este proyecto consta de cuatro fases para su desarrollo, las cuales abarcan la planificación y planteamiento de los antecedentes generales para la realización del proyecto, la investigación de la problemática en estudio en base a la situación actual planteada, el modelamiento y desarrollo del proyecto, que abarca el modelamiento de datos y cómo será afrontado el proceso ETL, hasta el desarrollo del código que soportará y hará funcionar el modelo predictivo, en base a la construcción de bases de datos, APIs y realización de pruebas para mitigar los posibles errores encontrados, y la última fase que dará fin al desarrollo del proyecto, es la fase de las conclusiones y recomendaciones, en la cual se darán a conocer las conclusiones que se fueron recabando a lo largo del desarrollo y elaborando un manual de usuario con las recomendaciones de uso.

Además, este proyecto estará bajo un marco de trabajo de desarrollo ágil, Scrum y metodologías de análisis y minería de datos, las cuales son CRISP-DM y OSEMN. El entorno de desarrollo estará basado en Python, junto a los librerías de análisis y minería de datos (Pandas, Numpy, etc.) y frameworks de desarrollo de APIs (Flask, Django y FastAPI).

El proyecto tiene una duración de dos semestres académicos, los cuales abarcan las asignaturas Título I y Título II, en donde se elaborarán como entregables un

Informe Final de Trabajo de Título y el sistema (MVP) del proyecto propuesto.

1.2. Palabras Clave

- API (Application Programming Interfaces).
- EDA (Exploratory Data Analysis).
- Algoritmos de predicción.
- Algoritmos de clasificación.
- Afiliado
- Administradora de Fondos de Pensiones

1.3. Descripción del trabajo de título

El trabajo de título se basa en un proyecto empresarial que requiere el procesamiento de los registros de navegación del sitio web para afiliados de AFP Capital, con el fin de detectar comportamientos de los clientes y sus preferencias de uso, permitiendo personalizar las futuras experiencias de navegación. La lectura de los registros se realizará extrayendo la información desde Kibana (ElasticSearch), la cual es registrada a través de diversas APIs utilizadas en el sitio web. Los elementos fundamentales del proyecto incluyen el análisis exploratorio de datos, extracciones, transformaciones, cargas, modelos de predicción y detección de preferencias. Todo esto con el objetivo de generar un modelo capaz de predecir el comportamiento de los clientes en el canal web.

1.4. Objetivos

Objetivo general

Analizar el comportamiento de los clientes y sus preferencias de uso en un período igual o inferior a 6 meses, para predecir navegaciones futuras personalizadas.

Objetivos específicos

- Realizar una investigación de las herramientas utilizadas para la predicción de comportamiento de usuarios en un canal web.
- Llevar a cabo un análisis y estudio de los datos entregados por la empresa.

- Realizar un proceso ETL con la información de navegación web de los clientes de AFP Capital, para analizar su comportamiento dentro del sitio web privado.
- Desarrollar un modelo capaz de predecir el comportamiento de los clientes de AFP Capital, para entregar navegaciones personalizadas futuras.
- Establecer recomendaciones de personalización en función de los hallazgos del modelo de predicción para futuras navegaciones dentro del sitio web de AFP Capital.

1.5. Alcances y Limitaciones

Alcances

El proyecto a realizar contempla los siguientes alcances:

- Se analizará el comportamiento de los clientes de AFP Capital en su nuevo sitio web privado.
- El proyecto entregará un modelo capaz de predecir el comportamiento de los clientes de AFP Capital en la web y una API que permita obtener el comportamiento recomendado para un afiliado específico.

Limitaciones

El proyecto contempla las siguientes limitaciones:

- No se tendrá acceso directo a las bases de datos de AFP Capital, por lo que se trabajará con una muestra.
- No se podrá acceder a los ruts e información sensible de los clientes de AFP Capital.
- Solo se trabajará con datos cualitativos de la navegación web de los usuarios.

Capítulo 2

La empresa

2.1. Historia

La historia de AFP Capital se remonta a noviembre de 1980, cuando se implementó en Chile el sistema de pensiones de capitalización individual. El 16 de enero de 1981, se constituyó la sociedad Administradora de Fondos de Pensiones Santa María, que más tarde se transformaría en AFP Capital S.A. Desde sus inicios, la empresa se destacó por su filosofía de servicio, enfocada en satisfacer las necesidades y expectativas de sus afiliados. En 1995, AFP Capital estableció la filial Santa María Internacional S.A., con el propósito de expandir su alcance y ofrecer servicios a personas naturales o jurídicas del extranjero, así como invertir en AFP o sociedades relacionadas con materias previsionales en otros países. Esta iniciativa consolidó la presencia de AFP Capital en el ámbito internacional y fortaleció su posición como una administradora de fondos de pensiones líder en la región. En el año 2000, se produjo una relevante transacción en la historia de AFP Capital. ING Group adquirió Aetna Inc., incluyendo el 96,56 % de las acciones de AFP Capital S.A. Esta adquisición tuvo como objetivo reforzar la posición de liderazgo de AFP Capital en el mercado previsional chileno y contribuir a su crecimiento y desarrollo. Posteriormente, en 2008, AFP Capital llevó a cabo una fusión con AFP Bansander, otra reconocida administradora de fondos de pensiones en Chile. Esta fusión permitió consolidar aún más las operaciones de AFP Capital y fortalecer su presencia en el país. A fines de 2011, Grupo SURA, una empresa líder en el negocio de pensiones en Latinoamérica, adquirió las operaciones de ING en la región. Esta adquisición llevó a AFP Capital a formar parte de Grupo SURA y a beneficiarse de su amplia experiencia y recursos, consolidándose como una compañía destacada en el mercado previsional latinoamericano. En resumen, la historia de AFP Capital está marcada por su constante evolución, consolidación y liderazgo en el mercado de administración de fondos de pensiones en Chile. A lo largo de los años, ha demostrado

su compromiso con la excelencia en la prestación de servicios previsionales y su capacidad de adaptación a los cambios y desafíos del entorno económico y regulatorio.

Figura 2.1: Historia AFP Capital

Fuente: AFP Capital. Recuperado de `https://www.afpcapital.cl/Quienes-Somos/Paginas/Historia.aspx`

2.2. Descripción general

AFP Capital es una destacada compañía chilena dedicada al negocio de pensiones y administración de fondos de pensiones. Forma parte de SURA, una reconocida empresa que ofrece servicios financieros y previsionales en Chile y otros países de América Latina. El principal enfoque de AFP Capital es brindar a sus afiliados una asesoría personalizada y servicios diferenciadores que les permitan alcanzar una mejor pensión al momento de su jubilación. La empresa se distingue por su compromiso con la optimización en la calidad de sus servicios, la entrega de información transparente y relevante a sus afiliados, y su solidez empresarial. Con una trayectoria de más de tres décadas en el mercado, AFP Capital ha logrado posicionarse como una de las principales administradoras de fondos de pensiones en Chile. Esto se debe en gran medida a su administración seria, responsable y eficiente en el manejo de los Fondos de Pensiones, así como a su enfoque en la inversión y gestión de los recursos de manera prudente y estratégica. La compañía cuenta con un equipo de colaboradores altamente capacitados y comprometidos, quienes contribuyen a la excelencia en la atención al cliente y al logro de los objetivos financieros de los afiliados. Además, AFP Capital se distingue por su constante innovación y adaptación a los cambios regulatorios y las necesidades cambiantes de los afiliados, con el fin de brindar soluciones efectivas y satisfactorias en el ámbito de las pensiones.

2.3. Misión y visión

Misión

La misión de AFP Capital es: .ªcompañamos a nuestros clientes, a través de una asesoría experta y diferenciadora en soluciones de ahorro para alcanzar su número, su Pensión, creciendo sustentablemente, desarrollando a nuestros colaboradores e integrándose responsablemente a la comunidad."(*AFP Capital*, 2023)

Visión

La visión de AFP Capital es: "Somos Guías, acompañamos a nuestros clientes a lograr sus sueños a través del ahorro."(*AFP Capital*, 2023)

Capítulo 3

Marco teórico

3.1. Importancia de predecir el comportamiento del cliente en un sitio web

La predicción del comportamiento del cliente dentro de un entorno web se considera a la aplicación de técnicas y modelos analíticos para lograr predecir en cierta manera las posibles necesidades, acciones, preferencias y decisiones que un cliente pueda tomar mientras interactúa en alguna plataforma en línea o sitio web. En los últimos años, ha sido de gran importancia la predicción del comportamiento de los clientes para las empresas, gracias a esto buscan anticipar las necesidades y preferencias de sus clientes, pudiendo adaptar los productos y servicios para entregar una mayor satisfacción al cliente (Zheng, Thompson, Lam, Yoon y Gnanasambandam, 2013). La lealtad de los clientes representa un valor clave para las empresas, ya que un cliente leal seguirá consumiendo los productos y servicios de la empresa, por lo que si se mejora la experiencia del usuario, la satisfacción del cliente aumenta y esto genera un aumento en la ganancia de la empresa. Según Zheng, Thompson, Lam, Yoon y Gnanasambandam (2013), la predicción del comportamiento del cliente ayuda a las empresas a identificar oportunidades de mejora y mercado, además de ayudar a tomar decisiones informadas sobre estrategias de publicidad y marketing. El objetivo fundamental de predecir el comportamiento del cliente en un entorno web es lograr comprender y anticipar las acciones de los clientes con la meta de personalizar, mejorar la experiencia de usuario y poder aumentar la satisfacción y fidelidad de los clientes. Las predicciones pueden abarcar distintos aspectos del comportamiento de un cliente dentro de un canal web, a grandes rasgos existen 4 tipos de predicciones que se pueden realizar, están las predicciones de compras, donde mediante el análisis de patrones de navegación, su historial de compras, preferencias y características demográficas, gracias a esto se busca predecir las compras futuras de un cliente, se encuentra la predicción de clics, esta busca

anticipar los enlaces o elementos con los cuales un cliente va a interactuar dentro de un sitio web, lo que busca mejorar la calidad de contenido que se encuentra desplegado y lograr mejorar la usabilidad del sitio web, también está presente la predicción de abandono de carrito, esta permite tomar acciones de recuperación o retención del cliente, se concentra en identificar aquellos clientes que agregan productos a un carrito de compra pero no finalizan el proceso de compra y por ultimo, esta la predicción de retención de clientes, esta busca predecir qué clientes están más cercanos a abandonar o terminar la relación existente con el sitio web, para poder generar e implementar estrategias para aumentar la fidelización y retención de estos clientes.

3.2. Comportamiento del cliente/afiliado en el canal web

3.2.1. Definición y relevancia del comportamiento del cliente para el negocio

Considerando los modelos de negocios establecidos por las Administradoras de Fondos de Pensiones [AFP], de ahí radica la importancia de la figura del cliente. Según lo que indica la Real Academia Española, el cliente es la persona que realiza una compra o utiliza los servicios que un profesional o empresa pueda ofrecer (Real Academia Española, s.f), no obstante en base al sistema establecido por las Administradoras de Fondos de Pensiones, el cliente obtiene el nombre de afiliado pues estos contribuyen o se encuentran inscritos en un plan de pensiones (Rasekhi, Fard y Kim, 2016). El afiliado es el centro del negocio, cuya gran importancia radica principalmente en la rentabilidad que brinda. Cada trabajador que decida afiliarse se traduce en una ganancia, mientras que cada afiliado que decida desafiliarse genera perdida. Considerando esto es que se puede apreciar la segunda importancia del afiliado, debido a que este promueve la marca si es que la experiencia del servicio de cara al usuario es buena. En tercer lugar, el afiliado, al ser un ganancia para el modelo, este a su vez que obtiene el servicio es capaz de posibilitar el crecimiento de la empresa al tener su preferencia. Por otro lado, la experiencia del cliente y su feedback es valiosa ya que puede brindar conocimiento de los puntos débiles y con posibilidad de mejora que tiene el sistema (Rodriguez, 2023). Dentro de las distintas funciones que el cliente tiene, en primer lugar se puede mencionar al cliente como consumidor. Consiste en unas de las funcionalidades más tradicionales puesto que el objetivo intrínseco del cliente es consumir o contratar servicios. Como consumidor es quien adquiere un producto o servicio y lo aprovecha para un fin o necesidad, por lo que la empresa obtiene su principal fuente de ingresos. En segundo lugar, se tiene al cliente como prosumidor, en otras palabras, consume y produce a la vez (Toffler, 1980). Al momento del consumo, el cliente también deja reseñas o realiza comentarios en lugares especializados, informa-

ción que resulta de utilidad para generar insights que mejoren la experiencia en el servicio. En tercer lugar, se entiende al cliente como crítico, puesto que si la experiencia del cliente es negativa, el feedback y reseñas negativas que este brinde pueden ser de índole constructiva como destructiva. En cuarto lugar, se encuentra el cliente como pieza fundamental en el desarrollo de los productos y servicios. Los comentarios de los clientes pueden conducir al desarrollo de servicios innovadores apegados a las necesidades que los clientes indican. Para poder lograr perfeccionar el servicio y productos ofrecidos, es crucial el aporte de los clientes recurrentes o suscriptores del servicio, en el caso específico de las Administradoras de Fondos de Pensiones se refiere a los afiliados. En quinto lugar, el cliente como evaluador de la experiencia. Relacionado con los puntos anteriores, la mejor forma de mejorar la experiencia del cliente es tomando en consideración los comentarios de los clientes en esta materia, así se puede generar una diferencia de las otras empresas que constituyen la competencia existente en el mercado. Por último, se considera que el cliente puede ser un eventual embajador de la marca, en otras palabras promotores de la misma pudiendo generar recomendaciones, comentarios y reseñas positivas que promuevan el negocio.

3.2.2. Características del comportamiento del cliente en el canal web

Para comprender la experiencia y el comportamiento del cliente dentro de un canal web, es importante reconocer la existencia del consumer journey, el cual describe las distintas etapas por las que un cliente pasa al momento de consumo de un producto o servicio. Según Lemon y Verhoef (2016) las etapas corresponden a conciencia, investigación, consideración, compra, uso y evaluación. La definición de conciencia da cuenta de la necesidad o el problema que debe ser resuelto, mientras que investigación refiere de la búsqueda de información por parte del cliente para posibles soluciones, comparando entre las distintas opciones disponibles (Lemon y Verhoef, 2016). Luego la etapa de consideración donde el cliente puede evaluar entre las opciones disponibles escogiendo la que mejor se adapta a sus necesidades dando paso a la etapa de compra cuando el cliente contrata y/o compra el mejor servicio a su parecer. Posterior viene la etapa de uso donde el cliente puede experimentar y testear la calidad, funcionalidad y experiencia del servicio dando pie a la última etapa que consiste en evaluar la experiencia como satisfactoria o insatisfactoria con la entrega voluntaria de feedback tanto positivo como negativo. Por lo tanto las posibles opciones disponibles para los clientes dentro del canal web buscan hacer del consumer journey una eficiente y grata experiencia. Para poder acceder al canal web de AFP Capital, se debe estar afiliado y tener una cuenta privada personal [Rut y Contraseña] y una vez se hace ingreso al canal web privado, el afiliado tiene disponibles variadas opciones para realizar y que buscan satisfacer sus posibles necesidades, estas corresponden al pago o no de la cotización mensual, la obtención de certificados de cotizaciones, certificado de afiliación, certificado de antecedentes previsionales, certificados de traspaso de fondos, certificado de vacaciones

progresivas y certificados tributarios, como también la obtención de certificados generales, como el certificado de residencia, certificado de suscripción de ahorro previsional voluntario [APV], certificado de cuenta 2, certificado de remuneraciones imponibles, certificado de periodos no cotizados y certificado de trabajo pesado, si el afiliado es una persona pensionada puede obtener certificado de asignacion familiar, certificado de calidad pensionado, certificado de pensiones pagadas, certificado de pensión en trámite, certificado de ingreso base y certificado de comprobante de pago de pensión, también poder hacer obtención de la cartola en línea. El canal web privado permite realizar el ahorro obligatorio y ahorrar voluntariamente, dentro de una cuenta de ahorro previsional voluntario [APV] o cuenta 2, realizar inversiones, hacer depósitos directos, tener planillas de pagos y ver las comisión cobrada como afiliado. También le otorga al afiliado la opción de ver su fondo de pensiones, ver los tipos de fondo de pensión, tipo A, tipo B, tipo C, tipo D, tipo E y sus porcentajes de rentabilidad, realizar un cambio de fondo de pensiones y recibir educación previsional. Le otorga al afiliado la opción de realizar giros en sus cuentas personales, acceder a rescates financieros y realizar el trámite de pensión.

3.2.3. Factores que afectan el comportamiento del cliente

Lemon y Verhoef (2016) proponen que los principales factores que influyen en el comportamiento del usuario y su experiencia son sensoriales, afectivos, cognitivos, puntos de contacto y externos. Dentro de la experiencia sensorial se encuentra lo apreciable con alguno de los sentidos del cuerpo, tanto vista, olor, tacto, entre otros. Respecto de la experiencia afectiva, hay que tener en consideración la emocionalidad del cliente producto de la experiencia del producto o del servicio. Al hablar del aspecto cognitivo, este refiere de los pensamientos, creencias y/o actitudes que el cliente pueda tener respecto de la compañía, el producto o el servicio entregado. Sobre los puntos de contacto, estos hacen mención a las distintas maneras en las que el cliente y la compañía entran en contacto, tales como la publicidad, servicio al cliente, redes sociales o interacciones de tipo transaccional (Lemon y Verhoef, 2016). Por último, el factor externo cuya definición hace referencia a considerar el contexto actual, las condiciones socioeconómicas y otros factores que puedan afectar la experiencia del usuario que se encuentren fuera de control de la compañía. Dentro de los factores que pueden influir en el comportamiento de un cliente en el canal web están principalmente, la usabilidad y el diseño. Respecto a la usabilidad, esta depende de 7 características las que garantizan una buena experiencia del usuario. Según Sanchez (2011) la accesibilidad, legibilidad, navegabilidad, facilidad de aprendizaje, velocidad de utilización, eficiencia del usuario y tasas de error del canal web, influyen en la experiencia y posterior feedback que el usuario pueda brindar sobre el uso de los servicios. Por otro lado, el diseño del sitio web depende de 5 características para garantizar un buen contenido y estética para lograr que el usuario encuentre lo que busca en el menor tiempo posible, en otras palabras, eficiencia. El autor Walter Sanchez (2011) indica que el diseño debe de ser en-

tendible, novedoso, comprensible, inteligente y atractivo, consiguiendo acercar los contenidos de mejor manera al usuario y logrando conseguir una navegación más intuitiva. Estos factores son de gran importancia para que el usuario pueda encontrar el contenido que busca en el menor tiempo posible y que la experiencia sea positiva al interactuar con la interfaz del sitio web.

3.3. Herramientas para la predicción del comportamiento del cliente en el canal web

3.3.1. Introducción a las herramientas de análisis de datos

En el entorno empresarial actual, la capacidad de tomar decisiones informadas y basadas en datos se ha vuelto fundamental para el éxito y la competitividad de las organizaciones. El análisis de datos desempeña un papel crucial en este proceso, permitiendo a las empresas obtener información valiosa a partir de grandes volúmenes de información y utilizarla para comprender el comportamiento del cliente de manera más profunda y precisa, esto resulta de suma importancia ya que la calidad de las decisiones tomadas marca la diferencia entre el éxito y el fracaso (Contreras Arteaga & Sánchez Cotrina, 2019, 15). Dentro de las herramientas de análisis de datos, se destacan cuatro conceptos clave que han revolucionado la forma en que se procesan y se obtiene información de los datos: Business Intelligence, Big Data, Machine Learning y Data Mining. Estas herramientas proporcionan a las empresas la capacidad de extraer conocimientos y patrones significativos de los datos, lo que a su vez les permite tomar decisiones estratégicas más acertadas y personalizar sus estrategias de marketing y atención al cliente. El Business Intelligence (BI) se refiere a la recopilación, análisis y presentación de datos empresariales para facilitar la toma de decisiones. Mediante el uso de diversas técnicas y herramientas, el BI permite a las empresas visualizar y comprender mejor los datos de sus operaciones y clientes. Esto incluye la generación de informes, el análisis de tendencias, la monitorización de indicadores clave de rendimiento (KPI) y la creación de tableros de control interactivos. El BI ayuda a las organizaciones a identificar oportunidades, detectar áreas de mejora y optimizar su rendimiento en función de datos históricos y en tiempo real. Sobre la inteligencia de negocios se ha determinado que cada implementación es única para cada proceso empresarial (Garcia-Estrella & Barón Ramírez, 2021, 6). El Big Data se refiere a la gestión y análisis de grandes volúmenes de datos, tanto estructurados como no estructurados, que superan la capacidad de las herramientas tradicionales de almacenamiento y procesamiento. El Big Data se caracteriza por las tres V's: Volumen (gran cantidad de datos), Velocidad (alta velocidad de generación y procesamiento de datos) y Variedad (diversidad de fuentes y formatos de datos). Para aprovechar el potencial del Big Data, las empresas emplean técnicas de procesamiento distribuido y herramientas específicas para el almacenamiento, procesamiento y análisis de

estos datos masivos. El análisis de Big Data permite identificar patrones, tendencias y correlaciones ocultas en los datos, lo que brinda información valiosa para entender y anticipar el comportamiento del cliente. El Machine Learning (aprendizaje automático) es una rama de la inteligencia artificial que permite a los sistemas informáticos aprender y mejorar automáticamente a partir de la experiencia sin ser programados explícitamente. En lugar de basarse en una analítica descriptiva, Machine learning ofrece una analítica predictiva (Sandoval, 2018, 37). Mediante algoritmos y modelos, el Machine Learning permite a las empresas analizar grandes conjuntos de datos y detectar patrones complejos en el comportamiento del cliente. Esto permite realizar predicciones y recomendaciones personalizadas, así como automatizar tareas y procesos, lo que mejora la eficiencia operativa y la experiencia del cliente. El Data Mining (minería de datos) se refiere al proceso de descubrir información valiosa, patrones y relaciones desconocidas en grandes conjuntos de datos. Utilizando técnicas estadísticas y algoritmos avanzados, el Data Mining permite identificar correlaciones y tendencias ocultas en los datos, lo que ayuda a las empresas a comprender mejor el comportamiento del cliente y tomar decisiones más acertadas. Esta herramienta es especialmente útil para la segmentación de clientes, la detección de fraudes, la recomendación de productos y la personalización de ofertas.

3.3.2. Métodos, técnicas y tecnologías de análisis de datos

En la actualidad, el análisis de datos desempeña un papel fundamental en la predicción del comportamiento del cliente. Las empresas y organizaciones buscan comprender y anticiparse a las necesidades y preferencias de sus clientes para mejorar la toma de decisiones y ofrecer productos y servicios más personalizados. Para lograr esto, se han desarrollado diversos métodos, técnicas y tecnologías que permiten analizar grandes volúmenes de datos y extraer información valiosa. A continuación, se listan algunos de los métodos, técnicas y tecnologías más utilizados en el análisis de datos para predecir el comportamiento del cliente.

Métodos y modelos

- Regresión logística
- Clustering
- Árboles de decisión
- Random Forest
- Gradient Boosting Machine

Técnicas

- Redes neuronales artificiales (ANN)

- Support Vector Machine (SVM)

Tecnologías

- Tableau
- Python (con bibliotecas como Pandas, NumPy, Scikit-learn)
- R (con paquetes como dplyr, caret, randomForest)
- Apache Spark
- KNIME
- RapidMiner
- QlikView
- Power BI

3.3.3. Modelos de predicción de comportamiento del cliente

Modelos de regresión

La regresión logística corresponde a un algoritmo de aprendizaje automático supervisado que es empleado para resolver problemas de clasificación. Si bien, su nombre contiene "regresión", en realidad corresponde a un método de clasificación.

Se da uso a la regresión logística cuando la variable de respuesta o variable objetivo es categórica. En lugar de predecir un valor numérico como en la regresión lineal, la regresión logística estima la probabilidad de que una observación pertenezca a una categoría específica.

Los modelos de regresión logística se basan en la función logística, también conocida como función sigmoide, que mapea cualquier valor real a un rango entre 0 y 1. La función sigmoide tiene la siguiente forma matemática:

$$f(z) = \frac{1}{(1 + e^{-z})}$$

En la regresión logística, se ajusta un modelo lineal a los datos de entrada y se aplica la función sigmoide al resultado para obtener la probabilidad de pertenencia a una clase. La ecuación del modelo se expresa como:

$$p(y = 1|x) = \frac{1}{(1 + e^{(-(b0+b1x1+b2x2+...+bn*xn))})}$$

Donde:

p(y=1—x) es la probabilidad condicional de que la variable de respuesta sea igual a 1 dada la entrada x.

b0, b1, b2, ..., bn son los coeficientes del modelo que se ajustan durante el proceso de entrenamiento.

x1, x2, ..., xn son los valores de las variables de entrada.

El proceso de ajuste de la regresión logística implica encontrar los mejores valores para los coeficientes del modelo con la finalidad de maximizar la verosimilitud de los datos observados. Esto se puede hacer mediante métodos numéricos como la maximización de la función de verosimilitud o mediante algoritmos de optimización como el gradiente descendente.

Una vez entrenado el modelo, se puede utilizar para hacer predicciones clasificando nuevas observaciones según la probabilidad estimada. Por ejemplo, si la probabilidad estimada de pertenencia a una clase es superior a un umbral (generalmente 0.5), se clasificará como perteneciente a esa clase.

Para nuestro caso en particular, puede ser utilizado el modelo de regresión logística para predecir el comportamiento de usuarios en un canal web, para ello se necesitaría tener datos históricos que contengan información relevante sobre el comportamiento pasado de los usuarios y las variables predictoras asociadas. Estas variables predictoras pueden incluir características demográficas, patrones de uso del sitio web o aplicación, historial de compras, interacciones anteriores, entre otros.

Una vez que se tienen los datos y las variables predictoras, se puede entrenar un modelo de regresión logística utilizando técnicas de ajuste como la maximización de la verosimilitud o el gradiente descendente. Una vez entrenado el modelo, puede ser utilizado para predecir el comportamiento futuro de los usuarios en función de nuevas observaciones o datos entrantes.

Es importante tener en consideración que la calidad de las predicciones dependerá de la calidad de los datos utilizados para entrenar el modelo y de la selección adecuada de las variables predictoras. Además, es fundamental realizar una validación adecuada del modelo utilizando técnicas como la validación cruzada o la separación de conjuntos de entrenamiento y prueba para evaluar su rendimiento y generalización en datos no vistos.

Ventajas de los modelos de regresión logística

- Interpretación de resultados: La regresión logística proporciona coeficientes que indican la dirección y la magnitud de la relación entre las variables predictoras y la variable de respuesta. Esto permite interpretar el efecto relativo de cada variable en la probabilidad de pertenecer a una clase específica.

- Manejo de variables independientes categóricas: La regresión logística puede manejar tanto variables independientes continuas como categóricas. Incluso puede manejar variables categóricas con más de dos categorías mediante técnicas como la codificación de variables ficticias.
- Estimación de probabilidades: La regresión logística estima la probabilidad de pertenencia a una clase específica en lugar de simplemente clasificar observaciones en categorías. Esto es útil cuando se necesita una medida de certeza o riesgo asociado con la clasificación.
- Buena capacidad de generalización: La regresión logística puede funcionar bien con conjuntos de datos pequeños o moderados, y es menos propensa al sobreajuste en comparación con otros algoritmos más complejos. Esto la hace adecuada para aplicaciones con muestras limitadas.

Desventajas de los modelos de regresión logística

- Linealidad de la relación: La regresión logística asume una relación lineal entre las variables predictoras y la probabilidad logarítmica de la variable de respuesta. Si existe una relación no lineal, la regresión logística puede no ajustarse adecuadamente o requerir transformaciones adicionales de las variables.
- Sensible a valores atípicos y datos faltantes: Los valores atípicos o datos faltantes pueden afectar negativamente el rendimiento de la regresión logística. Es necesario manejarlos adecuadamente para evitar sesgos o imprecisiones en los resultados.
- Suposición de independencia: La regresión logística asume que las observaciones son independientes entre sí. Si hay dependencias o correlaciones entre las observaciones, la precisión de los resultados puede verse comprometida.
- No apto para problemas no lineales: Si existe una relación compleja y no lineal entre las variables predictoras y la variable de respuesta, la regresión logística puede no ser el modelo más adecuado. En tales casos, se pueden requerir técnicas más avanzadas, como modelos no lineales o de aprendizaje profundo.

Modelos de recomendación

Los modelos de recomendación son algoritmos y técnicas utilizados en sistemas de recomendación para ofrecer sugerencias personalizadas a los usuarios. Estos modelos se utilizan en una amplia gama de aplicaciones, como plataformas de comercio electrónico, servicios de streaming de música y video, redes sociales y más.

El objetivo de un modelo de recomendación es predecir o sugerir elementos que sean relevantes o interesantes para un usuario en particular, basándose en su historial de preferencias, comportamiento pasado o en información de usuarios similares. Estos modelos aprovechan el poder del aprendizaje automático y la minería de datos para analizar patrones y relaciones en grandes conjuntos de datos.

Existen varios tipos de modelos de recomendación, entre los más comunes se encuentran:

Filtrado colaborativo: Este enfoque se basa en la idea de que si a un grupo de usuarios con preferencias similares les gusta un conjunto de elementos, entonces a un usuario nuevo con características similares también le podrían gustar esos elementos. El filtrado colaborativo utiliza la información de las interacciones pasadas de los usuarios (por ejemplo, clasificaciones o historial de compras) para generar recomendaciones.

Filtrado basado en contenido: Este enfoque utiliza información sobre las características y atributos de los elementos para recomendar otros elementos similares. Por ejemplo, en un servicio de streaming de música, se pueden recomendar canciones o artistas similares a los que un usuario ha escuchado anteriormente en función de género, estilo o letras.

Modelos híbridos: Estos modelos combinan múltiples enfoques, como filtrado colaborativo y basado en contenido, para aprovechar sus fortalezas y proporcionar recomendaciones más precisas y personalizadas.

Los modelos de recomendación se construyen utilizando técnicas de aprendizaje automático, como regresión logística, árboles de decisión, redes neuronales o algoritmos de factorización matricial. Estos modelos se entrenan utilizando conjuntos de datos históricos que contienen información sobre las preferencias y elecciones de los usuarios, y luego se aplican en tiempo real para generar recomendaciones en función de nuevos datos.

Ventajas de los modelos de recomendación

- Personalización: Los modelos de recomendación ofrecen sugerencias personalizadas a los usuarios, lo que mejora la experiencia del usuario y facilita la búsqueda de productos o contenido relevante.
- Descubrimiento de nuevos elementos: Los modelos de recomendación pueden ayudar a los usuarios a descubrir nuevos elementos que podrían ser de su interés, ampliando así sus opciones y experiencias.
- Mejora de la retención y fidelidad de los usuarios: Al proporcionar recomendaciones precisas y relevantes, los modelos de recomendación pueden aumentar la satisfacción del usuario, mejorar la retención y fomentar la fidelidad a la plataforma o servicio.

- Eficiencia en la toma de decisiones: Los usuarios pueden ahorrar tiempo y esfuerzo al recibir sugerencias personalizadas, lo que les ayuda a tomar decisiones más rápidas y eficientes.

Desventajas de los modelos de recomendación

- Sesgo y burbujas de filtro: Los modelos de recomendación pueden verse afectados por el sesgo inherente en los datos de entrenamiento y pueden crear burbujas de filtro, limitando la diversidad y la exposición a nuevas ideas o perspectivas.
- Fracaso en captar preferencias cambiantes: Los modelos de recomendación pueden tener dificultades para captar las preferencias cambiantes de los usuarios a medida que sus gustos y necesidades evolucionan con el tiempo.
- Problemas de inicio en frío: Los modelos de recomendación pueden tener dificultades para ofrecer recomendaciones precisas para nuevos usuarios o elementos que tienen una falta de información histórica.
- Privacidad y preocupaciones éticas: Los modelos de recomendación recopilan y utilizan datos de los usuarios, lo que puede plantear preocupaciones de privacidad y cuestiones éticas relacionadas con el manejo de la información personal.

Modelos de series temporales

Los modelos de series temporales son técnicas utilizadas para analizar y predecir datos secuenciales que están organizados en función del tiempo. En una serie temporal, los datos se registran en intervalos regulares (como horas, días, meses, etc.) y cada punto de datos está asociado con una marca de tiempo.

El objetivo principal de los modelos de series temporales es comprender y capturar los patrones, tendencias y estacionalidad en los datos a lo largo del tiempo, y utilizar esta información para hacer predicciones futuras. Estos modelos son ampliamente utilizados en diversos campos, como la economía, las finanzas, la meteorología, la demanda de productos, la planificación de inventario y más.

Los modelos de series temporales se basan en la suposición de que los datos pasados pueden proporcionar información útil para predecir el futuro. Algunos de los modelos más comunes utilizados en el análisis de series temporales son:

- Media móvil (MA): Este modelo estima el valor futuro de la serie temporal en función de un promedio de los errores pasados. Se utiliza para capturar patrones aleatorios o no sistemáticos en los datos.
- Autoregresión (AR): Este modelo estima el valor futuro de la serie temporal en función de valores pasados de la propia serie. Se utiliza para capturar la dependencia de la serie en sí misma a lo largo del tiempo.

- Autoregresión de media móvil (ARMA): Este modelo combina los enfoques AR y MA para capturar tanto la dependencia de la serie en sí misma como los patrones aleatorios.
- Autoregresión integrada de media móvil (ARIMA): Este modelo amplía el modelo ARMA al considerar también las diferencias entre los valores de la serie temporal. Se utiliza para capturar tendencias y estacionalidad en los datos.

Además de estos modelos clásicos, también se utilizan enfoques más avanzados, como los modelos de espacio de estados, los modelos de suavizado exponencial y los modelos de redes neuronales recurrentes (RNN), que pueden capturar relaciones más complejas y no lineales en los datos de series temporales.

Es importante destacar que el análisis de series temporales requiere un enfoque cuidadoso para la selección del modelo, la identificación de patrones y la evaluación de la precisión de las predicciones. Además, se deben tener en cuenta factores como la estacionalidad, la estacionariedad de la serie y la presencia de datos faltantes o valores atípicos para obtener resultados confiables.

Ventajas de los modelos de series temporales

- Captura de patrones temporales: Los modelos de series temporales pueden capturar patrones, tendencias y estacionalidad en los datos a lo largo del tiempo. Esto permite comprender mejor la dinámica de los datos y hacer predicciones más precisas.
- Predicciones a corto plazo: Los modelos de series temporales son adecuados para hacer predicciones a corto plazo, ya que utilizan la información histórica para predecir los valores futuros. Esto es especialmente útil en aplicaciones donde se necesita anticipar eventos próximos, como demanda de productos o pronóstico del clima.
- Utilización de datos secuenciales: Los modelos de series temporales aprovechan la estructura secuencial de los datos y utilizan la información de los puntos anteriores para hacer predicciones en el siguiente punto. Esto permite tener en cuenta la dependencia temporal en los datos y obtener resultados más precisos.
- Flexibilidad en la elección del modelo: Existen diferentes tipos de modelos de series temporales que se pueden utilizar según la naturaleza de los datos y los patrones presentes. Esto proporciona flexibilidad para seleccionar el modelo más adecuado para el problema específico.

Desventajas de los modelos de series temporales

- Sensibilidad a datos faltantes o valores atípicos: Los modelos de series temporales pueden verse afectados negativamente por la presencia de datos faltantes o valores atípicos. Estos pueden distorsionar los patrones y afectar la precisión de las predicciones.
- Dificultad con tendencias no lineales: Los modelos de series temporales asumen a menudo que las relaciones son lineales o pueden ser capturadas por modelos lineales. Si hay tendencias no lineales en los datos, los modelos lineales pueden no ajustarse adecuadamente y se pueden requerir enfoques más avanzados.
- Necesidad de datos históricos adecuados: Los modelos de series temporales requieren una cantidad suficiente de datos históricos para hacer predicciones precisas. En ausencia de datos suficientes, los modelos pueden tener dificultades para capturar patrones y generar resultados confiables.
- Problemas con cambios estructurales: Si hay cambios estructurales significativos en los datos de series temporales (por ejemplo, cambios en la estacionalidad o en los patrones), los modelos de series temporales pueden tener dificultades para adaptarse y pueden requerir ajustes manuales.

Modelos de atribución

Los modelos de atribución permiten predecir el recorrido que los clientes seguirán al momento de concretar una compra. Este recorrido puede contener las redes sociales, el uso del sitio web del vendedor, el correo electrónico, entre otros. Los modelos de atribución permiten determinar el impacto que tiene el uso de las acciones para el sistema de marketing. Este tipo de modelo permite darle mayor importancia a los canales de marketing y a los puntos de contacto que existen entre el cliente y el vendedor, que llevaron al cliente a realizar una compra.

Al asignar crédito a sus canales de marketing y puntos de contacto, se puede aumentar la posibilidad de que los clientes logren concretar una compra, esto a través de la identificación de las áreas del recorrido del comprador que se puedan mejorar, la determinación del retorno de la inversión para cada canal o punto de contacto, el descubrimiento de las áreas más efectivas para gastar el presupuesto de marketing y la adaptación de las campañas de marketing y muestra de contenido totalmente personalizado por clientes.

Existen variados tipos de modelos de atribución, todos tienen el mismo procedimiento de asignar crédito a los canales y punto de contacto, cada uno de estos tipos de modelo le atribuyen un peso distinto a cada canal y punto de contacto. Los modelos a continuación son los más aptos para lograr la predicción del comportamiento de un cliente:

- Modelo de atribución Multi-Touch: Este modelo demuestra ser poderoso ya que tiene en cuenta todos los canales y puntos de contacto con lo que los clientes interactúan a lo largo de su camino al concretar una compra. Deja en evidencia cuáles de los canales y punto de contacto fueron más influyentes y de cómo estas trabajaron en conjunto para influenciar al cliente.
- Modelo de atribución Lineal: Corresponde a un tipo de modelo de atribución Multi-Touch que le entrega el mismo peso a cada uno de los canales y puntos de contacto con los que el cliente interactúa en su camino al concretar una compra.
- Modelo de atribución Time-Decay: También llamado modelo de atribución de declive en el tiempo, además de considerar todos los puntos de contacto, también considera el tiempo que cada uno de estos puntos de contacto ocurrió, por lo que, los puntos de contacto o interacciones que sucedieron más cercano al momento en que se concretó la compra reciben mayor peso.

Ventajas de los modelos de atribución

- Facilita el rastrear de mejor manera el paso a paso del cliente: Esto gracias a la atención que se le entrega a cada canal y punto de contacto con el cual el cliente interactúa a la hora de concretar una compra.
- Permiten mayor personalización de rastreo de los clientes: Al saber que canales y punto de contacto tiene cada uno de los clientes, se puede llegar a entregar una experiencia personalizada a cada uno de los clientes.
- Comprender la contribución de cada canal y punto de contacto: Permite comprender como cada canal y punto de contacto contribuye a lograr los objetivos comerciales. Siendo de gran ayuda para identificar como asignar los recursos de manera mas efectiva y lograr optimizar las estrategias.
- Identificar canales y puntos de contacto de alto rendimiento: Un modelo de atribución puede revelar qué canales o puntos de contacto tienen un mayor interacción con los clientes en términos de generación de resultados. Esto permite a las empresas enfocar sus recursos en los canales más efectivos y maximizar su retorno de inversión.

Desventajas de los modelos de atribución

- Poseen una mayor complejidad que los otros modelos: La implementación de un modelo de atribución puede ser compleja y requerir un enfoque personalizado según las necesidades y características de cada empresa. Además, no hay un modelo de atribución único que sea universalmente aceptado, lo que puede generar falta de consenso y confusión en la industria.

- La interpretación de los resultados puede ser subjetiva: La interpretación de los resultados de un modelo de atribución puede estar sujeta a la interpretación y suposiciones del analista. Diferentes personas pueden llegar a conclusiones diferentes basadas en los mismos resultados, lo que puede generar cierta subjetividad en la interpretación de los datos.
- Poseen limitaciones en la medición del seguimiento: El modelo de atribución depende de la disponibilidad y calidad de los datos. Si los datos son limitados o imprecisos, los resultados del modelo pueden no ser confiables o representativos de la realidad.

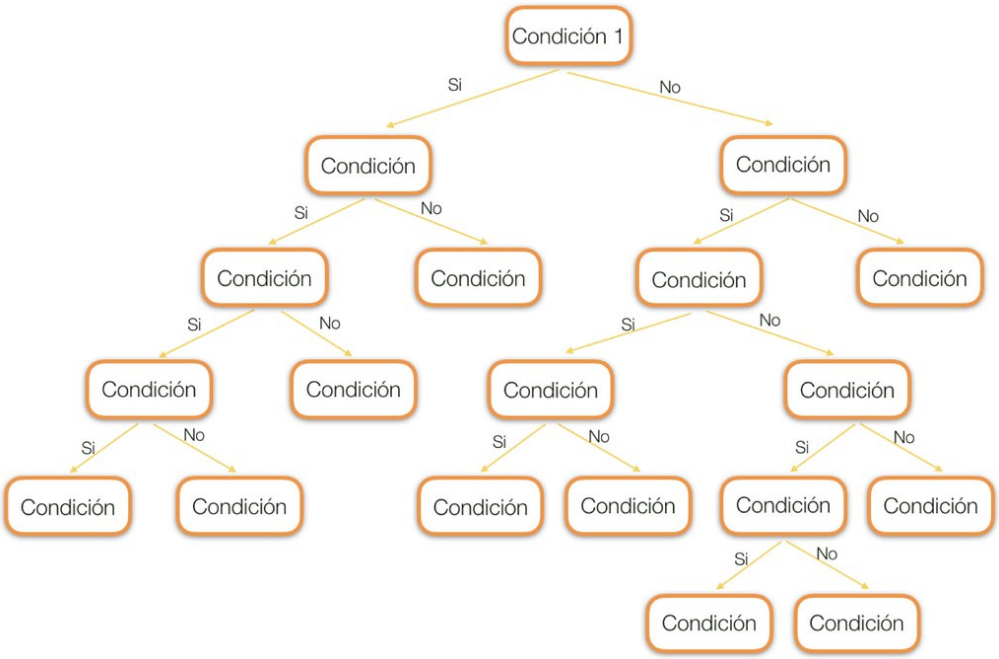
Modelo: Árbol de Decisión y Random Forest

Los árboles de decisión son modelos de aprendizaje supervisado que se utilizan para predecir a qué clase o categoría pertenece un caso conocido mediante uno o más atributos. Estos modelos se construyen utilizando un algoritmo llamado *partición binaria recursiva*. Durante el entrenamiento, el algoritmo realiza divisiones en un subconjunto de los datos basadas en decisiones asociadas a variables conocidas, generando así dos nuevos subconjuntos. Este proceso se repite de manera recursiva hasta alcanzar un punto de terminación predefinido, lo que resulta en la creación del clasificador basado en árbol de decisión. Luego, cada nuevo dato, que posee atributos conocidos, sigue las ramificaciones del árbol siguiendo las reglas y decisiones generadas durante el proceso de entrenamiento.

En la actualidad, los árboles de decisión son unos de los modelos de aprendizaje más utilizados debido a su buen rendimiento (Arana, 2021). Estos algoritmos pueden generar modelos predictivos tanto para variables cuantitativas (regresión) como para variables cualitativas o categóricas (clasificación).

Como se mencionó anteriormente, un árbol de decisión realiza tareas de clasificación. Un clasificador es un algoritmo que nos permite asignar sistemáticamente una clase a cada uno de los casos presentados.

Figura 3.1: Estructura de un árbol de decisión



Fuente: Aprende IA. Recuperado de `https://aprendeia.com/arboles-de-decision-clasificacion-teoria-machine-learning/`

En la figura anterior se puede visualizar la estructura que posee un árbol de decisión, en este se aprecia como actua el algoritmo de partición binaria mencionado al comienzo, tomando un conjunto y separandolo en subconjuntos hasta llegar a un final previamente establecido.

Para estimar la precisión de un clasificador, se calcula la tasa de error de clasificación verdadera. Esta tasa se obtiene evaluando un conjunto de valores X a los que el clasificador asigna una clase incorrecta, y se divide por el total de valores en X. Idealmente, se debería conocer la clase de todos los casos en el universo antes del entrenamiento, o en su defecto, de una muestra de tamaño similar al universo. Sin embargo, en la mayoría de los casos reales, no se dispone de todos los datos del universo, por lo que se trabaja con una muestra y se estima la tasa de error mencionada anteriormente utilizando *estimadores internos*.

Ventajas de los árboles de decisión

- Interpretabilidad: Los árboles de decisión son fácilmente interpretables y comprensibles para los humanos. La estructura del árbol se puede visua-

lizar de manera intuitiva, lo que permite comprender cómo se toman las decisiones y qué atributos son más relevantes para la clasificación.

- Facilidad de uso: La construcción y el uso de un árbol de decisión son relativamente sencillos en comparación con otros algoritmos de aprendizaje automático más complejos. No requieren una preparación exhaustiva de los datos ni un procesamiento previo complicado. Además, los árboles de decisión pueden manejar datos numéricos y categóricos sin requerir transformaciones adicionales, lo que simplifica el flujo de trabajo de modelado.
- Capacidad para manejar datos faltantes y variables irrelevantes: Los árboles de decisión tienen la capacidad de manejar datos faltantes en los atributos de forma natural. Durante la construcción del árbol, si un atributo tiene valores faltantes, el modelo puede utilizar otros atributos para tomar decisiones sin requerir imputación de datos. Además, los árboles de decisión son resistentes a variables irrelevantes, lo que significa que pueden ignorar atributos que no aportan información útil para la clasificación.
- Flexibilidad y robustez: Los árboles de decisión pueden manejar tanto problemas de clasificación como de regresión. Además, son capaces de capturar relaciones no lineales entre los atributos y la variable objetivo. Aunque cada árbol individual puede ser susceptible al sobreajuste, se pueden aplicar técnicas de regularización, como la poda, para mejorar la generalización y evitar el sobreajuste.
- Eficiencia en tiempo de entrenamiento y predicción: Los árboles de decisión tienen tiempos de entrenamiento y predicción rápidos, ya que solo implican la evaluación de una serie de reglas de decisión. Aunque el tiempo de construcción puede ser mayor para conjuntos de datos grandes, una vez construido, el árbol puede ser utilizado eficientemente para hacer predicciones en tiempo real.

Desventajas de los árboles de decisión

- Sensibilidad a cambios pequeños en los datos: Los árboles de decisión son muy sensibles a cambios pequeños en los datos de entrenamiento. Una modificación mínima en los datos de entrada puede dar lugar a un árbol de decisión completamente diferente. Esto puede hacer que el modelo sea inestable y su rendimiento pueda variar significativamente.
- Tendencia al sobreajuste: Los árboles de decisión tienen la capacidad de adaptarse demasiado a los datos de entrenamiento. Si no se controla adecuadamente, el árbol puede memorizar el ruido o las fluctuaciones aleatorias en los datos de entrenamiento, lo que puede resultar en un mal rendimiento en datos nuevos y no vistos. La poda y otras técnicas de regularización se utilizan para mitigar este problema.

- Limitaciones en la representación de relaciones complejas: Aunque los árboles de decisión pueden capturar relaciones no lineales entre atributos y la variable objetivo, pueden tener dificultades para representar relaciones complejas que requieren una combinación de múltiples atributos. Las decisiones tomadas en cada nodo se basan en un solo atributo, lo que puede limitar su capacidad para modelar interacciones más sofisticadas.
- Propensión a sesgos en los datos de entrenamiento: Los árboles de decisión pueden verse afectados por sesgos en los datos de entrenamiento, especialmente cuando hay desequilibrios en las clases o falta representación de ciertas categorías. Esto puede resultar en una clasificación desigual o inexacta en casos minoritarios o poco representados.

3.3.4. Metodología del proyecto

Para llevar a cabo el desarrollo del proyecto, se definieron cuatro fases que corresponden a la totalidad del proyecto, las cuales corresponden a:

Fase 1: Planteamiento y planificación

Para la primera fase del proyecto, se llevará a cabo una planificación de la manera en la que será abordada la problemática, para desarrollar un anteproyecto que será utilizado para evaluar y planificar las actividades correspondientes al desarrollo del proyecto. Entre ellas se encuentran:

- Planteamiento del proyecto y sus objetivos.
- Definición de alcances y limitaciones.
- Creación de un cronograma de actividades.

Fase 2: Investigación

Para la segunda fase, se realizará una investigación de herramientas y recursos necesarios para llevar a cabo un diseño de la solución para la problemática del proyecto planteado, sumado a un análisis de las bases de datos brindadas por la empresa AFP Capital. Una vez realizado lo anterior, se llevará a cabo una propuesta de diseño para la problemática, siendo entregada y analizada por la empresa, con la finalidad de pasar a desarrollo. Algunas de las actividades de esta fase corresponden a:

- Investigación del problema.
- Toma de requerimientos.
- Investigación de tecnologías de análisis de datos.

Fase 3: Modelamiento y desarrollo

Para la tercera fase, se llevará a cabo el diseño y desarrollo del sistema propuesto, además de realizar pruebas para verificar el correcto funcionamiento. Algunas de las actividades de esta fase corresponden a:

- Modelado del sistema ETL.
- Modelado de la API.
- Implementación del modelo propuesto.
- Pruebas y validaciones.
- Correcciones de errores.

Fase 4: Conclusiones y recomendaciones

Para la última fase, se dará fin al desarrollo del proyecto, elaborando un manual de usuario el cual indicaría algunas funcionalidades del sistema. Algunas de las actividades de esta fase corresponden a:

- Desarrollo de manual de usuario.
- Redacción de conclusiones y recomendaciones.
- Cierre del proyecto.

3.3.5. Metodología del sistema

CRISP-DM

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un proceso estándar utilizado para realizar proyectos de minería de datos. La metodología CRISP-DM se divide en seis fases distintas que se describen a continuación:

1. **Comprensión del problema:** En esta fase se define el problema a resolver y se establecen los objetivos del proyecto. También se recopilan los datos necesarios para el proyecto.
2. **Comprensión de los datos:** En esta fase se realiza una exploración de los datos para comprender su calidad, estructura y relevancia para el problema en cuestión.
3. **Preparación de los datos:** En esta fase se limpian y procesan los datos para que puedan ser utilizados en la etapa de modelado.

4. **Modelado:** En esta fase se aplican técnicas de modelado para desarrollar un modelo predictivo. Se prueban diferentes modelos y se selecciona el que mejor se ajuste a los datos.
5. **Evaluación:** En esta fase se evalúa el modelo desarrollado en la fase anterior. Se verifica que el modelo funcione correctamente y se ajuste adecuadamente a los datos.
6. **Implementación:** En esta fase se implementa el modelo desarrollado en la fase de modelado en un entorno de producción. También se establecen planes para monitorear el rendimiento del modelo y actualizarlo según sea necesario.

Las fases de la metodología CRISP-DM son iterativas, lo que significa que es posible volver a una fase anterior si es necesario.

OSEMN

La metodología OSEMN (acrónimo de las palabras en inglés: Obtain, Scrub, Explore, Model, Interpret) es un proceso utilizado en la minería de datos y el análisis de datos para trabajar con grandes conjuntos de datos de manera efectiva.

1. **Obtener (Obtain):** En esta etapa, se recopilan los datos necesarios para el análisis. Los datos pueden provenir de diferentes fuentes, como bases de datos, archivos en línea o registros de sensores. La calidad y la cantidad de los datos obtenidos son cruciales para el éxito del análisis.
2. **Limpieza (Scrub):** Una vez que se han obtenido los datos, es necesario realizar una limpieza para eliminar datos innecesarios o incorrectos. Esta etapa puede implicar la eliminación de duplicados, la corrección de errores y la eliminación de valores atípicos. El objetivo de esta etapa es obtener datos limpios y coherentes para el análisis.
3. **Exploración (Explore):** En esta etapa, se utilizan técnicas de visualización y estadísticas para explorar los datos y obtener información sobre ellos. Se pueden identificar patrones, tendencias y relaciones entre diferentes variables. El objetivo es obtener una comprensión más profunda de los datos y de cómo se relacionan entre sí.
4. **Modelado (Model):** En esta etapa, se utilizan técnicas de modelado estadístico o de aprendizaje automático para crear modelos que puedan predecir resultados futuros o identificar patrones en los datos. El objetivo es utilizar los datos para crear un modelo que pueda utilizarse para tomar decisiones informadas.
5. **Interpretación (Interpret):** En esta etapa, se interpretan los resultados obtenidos en la etapa de modelado. Los resultados pueden ser utilizados

para tomar decisiones o para generar nuevas hipótesis que puedan ser exploradas en futuros análisis.

Se propone el uso de la metodología OSEMN, ya que se enfoca en el análisis de datos y la creación de modelos predictivos. OSEMN también es una metodología más flexible que CRISP-DM, lo que puede ser útil en un proyecto de SCRUM donde se busca una mayor adaptabilidad.

Por otro lado, también se propone el uso de la metodología CRISP-DM, ya que el proyecto incluye una etapa de exploración y análisis de datos, seguida por una fase de construcción de modelos. CRISP-DM se enfoca en el proceso completo de minería de datos, desde la comprensión del problema hasta la implementación del modelo, lo que puede servir para realizar un trabajo más estructurado.

Ya que este proyecto se encuentra bajo el marco de trabajo SCRUM, ambas metodologías pueden ser utilizadas de manera complementaria, utilizando OSEMN para las fases de creación de modelos y CRISP-DM para la etapa de exploración y análisis de datos.

Capítulo 4

Proceso ETL

4.1. Diseño Proceso ETL

El diseño de un proceso ETL (Extracción, Transformación y Carga) implica seguir distintos pasos para asegurar que este proceso y el flujo de datos sean eficientes, precisos y cumplan con los requisitos del proyecto. Los pasos que se acordaron seguir son los siguientes:

- Requisitos ETL
- Identificicación fuente de datos
- Diseño modelo de datos objetivo
- Planificación de las transformaciones
- Selección herramientas
- Construcción y prueba proceso ETL
- Monitoreo proceso ETL

4.1.1. Requisitos ETL

En esta etapa se definen los requisitos del proyecto, las fuentes de datos, los objetivos comerciales y del proceso ETL, las necesidades de análisis y los plazos para realizar el proceso. Estableciendo una base solida para el diseño y buen funcionamiento del proceso ETL.

- **Fuente de datos:** La fuente de datos corresponde a un archivo .CSV que contiene información de la navegación web de los clientes en forma de Web Logs.

- **Objetivos comerciales:** Analizar el comportamiento de los clientes y sus preferencias de uso en un período igual o inferior a 6 meses, para poder predecir navegaciones futuras, y a partir de esto proporcionar atenciones personalizadas.
- **Objetivos proceso ETL:** Realizar las transformaciones necesarias para asegurar que el flujo de datos sea eficiente y preciso, a través de la limpieza de los datos, la normalización, la agregación, el filtrado, el enriquecimiento de datos, así como los cálculos y derivaciones necesarios.
- **Necesidades de análisis:** Realizar un análisis exploratorio de los datos entregados.

4.1.2. Identificicación fuente de datos

En esta etapa se determinan las fuentes de datos a ser usadas para el proyecto, incluyendo bases de datos, archivos .CSV y APIs. Esto además comprende la definición de la estructura, el formato y ubicación de cada fuente de datos dentro del proyecto.

La fuente de datos corresponde a un archivo .CSV que contiene información de la navegación web de los clientes en forma de Web Logs. Los Web Logs registrados vienen con 4 atributos, especificados a continuación:

- **rut cliente**: Este atributo representa un identificador único por cliente.
- **fecha evento**: Representa la fecha y hora de la interacción del cliente con el sitio web.
- **metodo**: Este atributo representa cuál fue el método al cual el cliente llamó al interactuar con el sitio web.
- **canal**: Corresponde al canal web con el cual el cliente realizó la interacción en el sitio web.

Para almacenar la fuente de datos se utiliza una estructura de carpetas, siendo las siguientes:

- **Input**: Dentro de esta carpeta se encontrará el archivo .CSV tal cual es entregado.
- **Intermediate**: Aquí se almacenará el archivo con la información preprocesada.
- **Output**: Dentro de esta última carpeta se almacenará la información ya procesada y lista para ser usada.

4.1.3. Diseño del modelo de datos objetivo

Dentro de esta etapa se realiza el diseño del modelo de datos objetivo, que se basará en un modelo dimensional del tipo estrella. Este modelo es ampliamente utilizado en el diseño de bases de datos para data warehousing y análisis de datos.

El diseño del modelo dimensional se centra en organizar los datos de manera que sean óptimos para el análisis y la generación de informes. En este enfoque, se identifican las dimensiones clave del negocio, que representan las categorías principales de interés, y se establecen relaciones con una tabla central conocida como tabla de hechos.

En el diseño del modelo dimensional, las entidades se convierten en dimensiones, que contienen atributos descriptivos que permiten realizar análisis en función de estas características.

La tabla de hechos es el núcleo del modelo y contiene las medidas numéricas o factores que se analizarán, como ventas totales, cantidad de productos vendidos o ingresos generados. Estas medidas se vinculan con las dimensiones a través de claves externas.

Al utilizar un modelo dimensional del tipo estrella, se logra un diseño optimizado para consultas y análisis de datos. La estructura simplificada y desnormalizada permite realizar operaciones de agregación y filtrado de manera eficiente, lo que facilita la generación de informes y análisis de datos complejos.

Durante el diseño del modelo de datos objetivo, se deben considerar los requisitos específicos del proyecto y las necesidades de análisis del negocio. Es importante realizar una cuidadosa identificación de las dimensiones clave, seleccionar los atributos relevantes y establecer las relaciones adecuadas entre las dimensiones y la tabla de hechos.

4.1.4. Planificación de las transformaciones

Dentro de esta etapa se realiza la planificación detallada de las transformaciones necesarias para construir una base sólida y consistente para el desarrollo del proyecto. Estas transformaciones implican una serie de pasos que permiten limpiar, filtrar, combinar y enriquecer los datos de manera adecuada.

La planificación de las transformaciones es fundamental para garantizar la calidad y la integridad de los datos que serán utilizados en el proyecto. Durante esta etapa, se identifican las tareas específicas que deben llevarse a cabo para lograr los objetivos establecidos, teniendo en cuenta los requisitos del proyecto y las necesidades del negocio.

Algunas de las transformaciones comunes incluyen:

- **Limpieza de datos:** Se realizan tareas de limpieza para corregir errores, eliminar valores duplicados o inconsistentes, y garantizar la coherencia de los datos. Esto puede incluir la corrección de formatos incorrectos, la normalización de datos, el manejo de valores faltantes o la estandarización de la información.
- **Filtrado de datos:** Se aplican filtros para seleccionar y extraer los datos relevantes para el proyecto, descartando aquellos que no cumplen con ciertos criterios o condiciones específicas. Esto ayuda a reducir el volumen de datos y a enfocarse en la información más relevante y útil.
- **Combinación de datos:** Se integran datos provenientes de diferentes fuentes o fuentes de datos diversas. Esto implica fusionar conjuntos de datos relacionados, realizar uniones o cruces de tablas, y establecer relaciones entre los datos para generar una visión global y coherente.
- **Enriquecimiento de datos:** Se agregan atributos o información adicional a los datos existentes para enriquecer su contexto y mejorar su valor. Esto puede implicar la incorporación de datos externos, la realización de cálculos derivados, la normalización de datos o la aplicación de reglas específicas.

Es importante tener en cuenta que la planificación de las transformaciones considera el orden y la secuencia adecuada de ejecución, así como la documentación de cada paso y los criterios de validación y verificación para garantizar la calidad de los datos transformados.

4.1.5. Selección herramientas

En esta etapa, se realiza la selección de herramientas de software que se ajusten a las necesidades y requisitos del proyecto para llevar a cabo el proceso ETL de manera eficiente. Se evalúan diferentes opciones disponibles en función de su capacidad, compatibilidad y facilidad de uso, para garantizar una elección adecuada.

4.1.6. Construcción y prueba proceso ETL

En esta etapa, se lleva a cabo la implementación del diseño del proceso ETL previamente definido, utilizando las herramientas seleccionadas. Se desarrollan los flujos de extracción, transformación y carga de los datos según lo establecido en el diseño.

Una vez implementado, se procede a realizar pruebas exhaustivas para garantizar el correcto funcionamiento del proceso. Estas pruebas incluyen la verificación de la extracción de datos de las fuentes, la correcta aplicación de las transformaciones definidas y la carga exitosa de los datos en el destino final.

El objetivo de las pruebas es asegurar que el proceso ETL cumpla con los requisitos establecidos y que los resultados obtenidos sean los esperados. Esto implica validar la integridad y coherencia de los datos transformados, así como verificar el rendimiento y la escalabilidad del proceso.

En caso de encontrar inconvenientes o desviaciones durante las pruebas, se realizan los ajustes necesarios en el diseño o en la configuración de las herramientas utilizadas. Es fundamental realizar iteraciones y pruebas adicionales hasta obtener resultados consistentes y satisfactorios.

4.1.7. Monitoreo proceso ETL

Se establece un sistema de monitoreo para supervisar de manera continua el rendimiento del proceso ETL, permitiendo identificar y abordar posibles problemas a tiempo y garantizar la calidad de los datos. En esta etapa, también se realiza el mantenimiento del proceso, lo cual implica actualizaciones de las transformaciones, resolución de problemas y optimización del proceso.

El sistema de monitoreo juega un papel fundamental en la detección temprana de cualquier anomalía o disrupción en el flujo de datos. Mediante la implementación de métricas y alertas, se pueden supervisar aspectos clave como el tiempo de ejecución, el uso de recursos, los volúmenes de datos y la integridad de los resultados.

Además, el mantenimiento del proceso ETL implica la capacidad de adaptación a medida que evolucionan las necesidades del proyecto. Esto puede implicar la actualización de las transformaciones para reflejar cambios en las fuentes de datos o requerimientos del negocio, así como la solución de problemas que puedan surgir durante la ejecución del proceso.

Asimismo, se busca optimizar el proceso ETL a través de la identificación de posibles cuellos de botella o ineficiencias. Esto puede implicar ajustes en el diseño de las transformaciones, mejoras en la selección de herramientas o la optimización de los recursos utilizados.

Referencias

Afp capital. (2023). Sitio web. Descargado de `https://www.afpcapital.cl/Paginas/default.aspx`

Arana, C. (2021). *Modelos de aprendizaje automático mediante árboles de decisión.*