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

26 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: Figura: 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, información 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 entendible, 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

Existen diferentes modelos empleados para realizar predicciones del comportamiento de un usuario en un canal web, el empleo de dichos modelos se encuentran explicados a continuación:

Modelos de regresión

El modelo de regresión es empleado para la predicción de variables, la regresión logística estima la probabilidad de que suceda un evento basándose en un conjunto de datos. Existen 3 tipos de modelos de regresión, los cuales corresponden a los modelos de regresión logística binaria, regresión logística multinomial y regresión logística ordinal.

Siendo el Modelo de regresión lógica binaria es empleado para predecir comportamientos de variables que tienen un comportamiento dicotómico, es decir, que solo cuentan con dos resultados posibles, como ejemplo podemos mencionar a la clasificación de correo electrónico si es spam o no lo es, si una opción es verdadero o falso, entre otros ejemplos. Dentro de la regresión logística es el más utilizado y en general, corresponde a uno de los clasificadores más comunes para la clasificación binaria.

El Modelo de regresión logística multinomial es empleado para predecir cuando la variable dependiente cuenta con tres o más resultados posibles, cabe recalcar que los valores a predecir no se encuentran ordenados.

Finalmente, el Modelo de regresión logística ordinal es utilizado cuando la variable de respuesta tiene tres o más resultados posibles, pero a diferencia del modelo de regresión logística multinomial, los valores empleados si poseen un orden definido.

Ventajas de los modelos de regresión

- La implementación del modelo es más sencillo comparado con otros modelos.
- Interpretación de los resultados relativamente sencilla.
- Simplificación de problemas.
- Existencia de documentación respecto a los modelos.

Desventajas de los modelos de regresión

- Propensos a sobreentrenarse.
- Alta sensibilidad a valores atípicos.
- Son propensos a realizar suposiciones.

Modelos de recomendación

Los modelos de recomendación corresponden a una subclase de aprendizaje automático que es utilizado para clasificar o valorar productos y usuarios. En modo de resumen, es un sistema de recomendación en un sistema que predice las valoraciones que un usuario puede dar a un producto y/o servicio. Este modelo es empleado en grandes empresas como Google, Amazon, Instagram, Spotify, entre otros.

Este modelo se puede clasificar en sistemas de filtrado colaborativo, sistemas basados en contenido o sistema de recomendación híbrido.

Los sistemas de filtrado colaborativo corresponden al proceso en el cual se predicen los intereses de un usuario, esto se realiza al identificar las preferencias e información de varios usuarios, realizando una búsqueda de patrones utilizando de filtrado de información.

Los sistemas basados en contenido se enfocan en generar recomendaciones basadas en las preferencias y los perfiles de usuario. El modelo trata de hacer coincidir a los usuarios con elementos o productos que les hayan gustado anteriormente. Un punto a destacar de este sistema es que se basa en los productos destacados por el usuario objetivo, en cambio, los otros modelos se basan en el usuario y en otros usuarios que utilizan la plataforma.

Para terminar con los modelos de recomendación, los sistemas de recomendación híbridos se encuentran diseñados para utilizar diferentes fuentes de información para generar las recomendaciones.

Ventajas de los modelos de recomendación

- Aumenta la participación de los usuarios.
- Recomendación de elementos acorde a los gustos de los usuarios.
- Poseen la capacidad de automatizar sus sugerencias.

Desventajas de los modelos de recomendación

- Propenso a presentar sesgo y falta de diversificación.
- Autorización de terceros por privacidad (dado que utilizan datos personales).
- Presentan problemas de arranque en frío.
- Difíciles de interpretar.

Modelos de series temporales

Los modelos de series temporales corresponden a un proceso en el cual son utilizados datos pasados con la finalidad de predecir acontecimientos futuros. Estos modelos analizan tendencias y patrones en los datos para extraer información para realizar predicciones sobre valores futuros, este tipo de modelos son utilizados en el ámbito financiero para predecir ventas o cotizaciones, otro de los campos en los que es empleado este modelo es en el científico, enfocándose en predecir los patrones meteorológicos.

Ventajas de los modelos de series temporales

- Capaces de capturar patrones.
- Útiles para predecir a corto plazo.
- Proporcionan facilidades para la interpretación de los resultados.

Desventajas de los modelos de series temporales

- No son eficientes para predecir a largo plazo.
- Son muy sensibles a los datos atípicos.
- Requieren información consistente.

Modelos de aprendizaje automático

Los modelos de aprendizaje automático corresponden a programas capaces de identificar patrones o tomar decisiones. Estos pueden ser entrenados para realizar predicciones, identificar objetos e imágenes, también puede ser empleado para predecir comportamientos de usuarios.

Ventajas de los modelos de aprendizaje automático

- Permiten trabajar con grandes volúmenes de datos.
- Permite la automatización de tareas.
- Mejoran a medida que son utilizados.

Desventajas de los modelos de aprendizaje automático

- Vulnerables a ataques de terceros.
- Requieren de una capacidad significativa de recursos.
- Algunos modelos entregan resultados complicados de interpretar.

Modelos de análisis de sentimientos

Como dice el nombre, el modelo de análisis de sentimientos permite procesar y analizar información a tiempo real, es utilizado principalmente en las redes sociales para analizar cómo ve la gente un producto nuevo y lo que no les gusta a sus clientes.

Ventajas de los modelos de análisis de sentimientos

- El modelo es escalable.
- Permite la automatización de los procesos del modelo.
- Permite segmentar y personalizar los métodos de análisis.

Desventajas de los modelos de análisis de sentimientos

- Al modelo se le dificulta identificar el sarcasmo.
- Posee limitaciones al momento de identificar sentimientos complejos.
- Los resultados del modelo pueden verse afectados por sesgos en los datos de entrenamiento.

Modelos de detección de anomalías

Los modelos de detección de anomalías corresponden a modelos de machine learning enfocados en detectar actividades extrañas en la información que está analizando, con esto queremos decir que detecta anomalías.

Existen diferentes tipos de métodos para detectar anomalías con machine learning, entre ellos se encuentran los modelos supervisados, sin supervisión y semi supervisado.

Para el método de modelos supervisados, el entrenamiento del modelo consta de dos variables de entrenamiento: normal y anormal. El modelo utiliza los ejemplos para detectar patrones y detectar anomalías en los datos proporcionados. Es importante destacar que, en el entrenamiento supervisado es fundamental asegurar la calidad de los datos con el cual será entrenado el modelo.

Para el método de modelos sin supervisión, es común emplear redes neuronales para implementar el modelo, dado que al utilizar redes neuronales se disminuye la carga de trabajo manual, con esto queremos decir que no es necesario preparar los datos de antemano. Un punto negativo es que la dificultad de implementar una red neuronal es alta.

Finalmente, para el modelo semi supervisado es una combinación de los dos métodos anteriores, permitiendo contar con las ventajas de ambos métodos. Al ser supervisado durante su entrenamiento, es posible tener un mayor control del tipo de patrones que aprender el modelo.

Ventajas de los modelos de detección de anomalías

- El modelo permite automatizar la detección de anomalías.
- Posee gran adaptabilidad a datos y entornos.
- Permite detectar de manera temprana las anomalías.

Desventajas de los modelos de detección de anomalías

- Es difícil definir una anomalía.
- Requiere de un entrenamiento adecuado al modelo.
- Es muy propenso a entregar posibles falsos positivos y falsos negativos.

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.

Existen varios modelos adicionales para el uso del marketing y la predicción de comportamiento, sin embargo, para este proyecto se limitó la búsqueda a los modelos mencionados anteriormente, los cuales se adaptan mejor a nuestras necesidades.

Ventajas de los modelos de atribución

- Facilita el rastrear de mejor manera el paso a paso del cliente.
- Permiten mayor personalización de rastreo de los clientes.

Desventajas de los modelos de atribución

- Poseen una mayor complejidad que los otros modelos.
- La interpretación de los resultados puede ser subjetiva.
- Poseen limitaciones en la medición del seguimiento.

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 sea eficiente, preciso y cumpla con los requisitos del proyecto, los pasos a 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.

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 y archivos .CSV y API´s. Esto además comprende la estructura la definición de la estructura, el formato y ubicación de cada fuente de datos dentro del proyecto.

4.1.3. Diseño modelo de datos objetivo

Dentro de esta etapa se diseña el modelo de datos que se utilizara y soportara el proyecto, como las bases de datos. Esto implica identificar las entidades, sus atributos y relaciones necesarias para lograr satisfacer los requisitos del proyecto antes definido. Se opto por ocupar un modelo dimensional del tipo estrella.

4.1.4. Planificación de las transformaciones

Dentro de esta etapa se determinan las transformaciones necesarias para tener una base sólida para desarrollar el proyecto, estas transformaciones conllevan limpiar, filtar, combinar los datos y el enriquecimiento de estos. Por esto es que se diseño el siguiente plan detallado con las transformaciones a aplicar.

4.1.5. Selección herramientas

En esta etapa se eligen las herramientas de software para poder realizar el ETL que se ajusten a las necesidades y requisitos antes mencionados, es por esto que se seleccionaron las siguientes herramientas:

4.1.6. Construcción y prueba proceso ETL

Es en esta etapa en la cual se implementa el diseño del proceso ETL ya definido en puntos anteriores utilizando las herramientas seleccionadas, desarrollando los flujos de extracción, transformación y carga de los datos según lo establecido. Luego se realizan distintas pruebas para asegurar el correcto funcionamiento del proceso y que se obtengan los resultados esperados.

4.1.7. Monitoreo proceso ETL

Se establece un sistema de monitoreo para poder supervisar el rendimiento del proceso ETL, logrando identificar posibles problemas y garantizar la calidad de los datos. Es en esta etapa donde se hace un mantenimiento del proceso, pudiendo tener actualizaciones de las transformaciones, resolución de problemas y optimizar el proceso.

Referencias

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