

Predicción de costo para adquirir nuevos clientes

Monografía para optar al título de:

Especialista en Analítica y Ciencia de Datos

Línea de Investigación:

Aprendizaje Supervisado

Juan Pablo Montoya Vallejo

Asesor:
Sebastián Rodríguez Colinas, Msc

Universidad de Antioquia
Facultad de Ingeniería - Departamento de Ingeniería de sistemas
Especialización en Analítica y Ciencia de Datos
Medellín, Colombia
2023

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

Cita	(Montoya Vallejo, 2023)
Referencia **Estilo APA 7 (2020)**	J. P. Montoya Vallejo [2023. *Trabajo de grado Archivo fotográfico de la Universidad de Antioquia: Predicción de costo para adquirir nuevos clientes* [Especialización en Analítica y Ciencia de Datos]. Universidad de Antioquia, Medellín UdeA (A-Z).

creative commons BY NC SA

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería CENDOI

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

TABLA DE CONTENIDOS

RESUMEN EJECUTIVO 6
ABSTRACT 7
I. INTRODUCCIÓN 8
II. PLANTEAMIENTO DEL PROBLEMA 9
III. OBJETIVOS 10
OBJETIVO GENERAL 10
OBJETIVOS ESPECÍFICOS 10
IV. MARCO TEÓRICO 11
MÉTRICAS DE DESEMPEÑO 13
V. DATOS 14
VI. METODOLOGÍA 16
PREPROCESAMIENTO 17
VARIABLE DEPENDIENTE (CAC) 17
VARIABLES NUMÉRICAS 18
VARIABLES CATEGÓRICAS 19
HERRAMIENTAS PARA EL ANÁLISIS SUPERVISADO 20
HERRAMIENTAS PARA CALCULAR MÉTRICA DE DESEMPEÑO 20
APRENDIZAJE SUPERVISADO 21
ITERACIONES y RESULTADOS 21
PRIMERA ITERACIÓN 21
SEGUNDA ITERACIÓN 22
TERCERA ITERACIÓN 23
ENTREGA DEL MODELO 24
VII. RESULTADOS 25
MÉTRICAS 25
EVALUACIÓN CUALITATIVA 25
VIII. CONSIDERACIONES DE POSIBLE PUESTA EN PRODUCCIÓN 26
DISEÑO DEL SISTEMA EN PUESTA DE PRODUCCIÓN 26
OBJETIVOS DE LA IMPLEMENTACIÓN EN PRODUCCIÓN 26
CONTEXTO DEL SOFTWARE A USAR 27
PÚBLICO OBJETIVO DEL PROYECTO 27
DISEÑO INTERNO DE MANEJO DE DATOS 28
ARQUITECTURA DE COMPONENTES Y ESTRUCTURA DEL SISTEMA 29
SEGURIDAD Y ACCESO A LAS BASES DE DATOS 29
DISEÑO DE INTERFAZ DE USUARIO 31
REGLAS Y CONSIDERACIONES DEL DISEÑO DE LA INTERFAZ 32
CONSIDERACIONES PARA TESTEO 32
WHITE BOX TESTING 32
BLACK BOX TESTING 32
SEGUIMIENTO MÉTRICAS DE DESEMPEÑO DEL MODELO 33
ANEXOS 33

INSTALACIÓN DE PAQUETES Y DEPENDENCIAS 33
POSIBLES MÉTRICAS DE NEGOCIO PARA UNA PUESTA EN PRODUCCIÓN 34
IX. CONCLUSIONES 35
X. REFERENCIAS 36

TABLA DE FIGURAS

FIGURA 1 *Distribuciones de la variable dependiente*
FIGURA 2 *Gráfica de correlación de las variables numéricas*
FIGURA 3 *Distribución de valores únicos en las variables categóricas*
FIGURA 4 *Diagrama de flujo del modelo de predicción del C.A.C.*
FIGURA 5 *Resultados de la evaluación del modelo primera validación*
FIGURA 6 *Gráficos de residuos obtenidos de comparar las predicciones y el valor real del C.A.C.*
FIGURA 7 *Gráfica de distribuciones con respecto al C.A.C. de variables numéricas*
FIGURA 8 *Importancia [%] de cada variable en la decisión de predicción del modelo*
FIGURA 9 *Implementación del bucket s3 para almacenar datos de CAC*
FIGURA 10 *Implementación de control de origen en peticiones del backend desde la url*
FIGURA 11 *Implementación de cuenta de servicio en AWS para credenciales e interacción entre aplicaciones*
FIGURA 12 *Implementación de interfaz de usuario en REACT alojada en un servidor en netlify*
FIGURA 13 *Respuesta predicción CAC nuevo cliente*

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

CAC	*Costo de adquisición de un cliente*
CFM.	*Convenient Food Mart*
MAE	*Mean absolute error (error absoluto medio)*
ROI	*Retorno de la inversión*
ML	*Machine Learning*
MSc	*Magister Scientiae*
P.ej.	*Por ejemplo*
UdeA	*Universidad de Antioquia*
Fig	*Figura*

LISTADO DE TABLAS

T*ABLA 1* *Tabla de resultados del del modelo en las iteraciones*

RESUMEN EJECUTIVO

El presente proyecto tiene como objetivo principal predecir el costo de adquisición de nuevos clientes (CAC) en la cadena de supermercados FOODMART. Para lograr esto, se empleó una base de datos real proporcionada por la empresa que contiene información detallada de 60.000 facturas de clientes, incluyendo ingresos, promociones, almacenamiento, ventas y costo de publicidad.

En aras del objetivo principal, se aplicaron técnicas de análisis exploratorio de datos y preprocesamiento para preparar los datos, se entrenaron y validaron un modelo de regresión lineal y dos de random forest. Dichos modelos se evaluaron mediante la métrica de desempeño error absoluto medio.

Finalmente, se compararon los resultados obtenidos y se realizó un análisis crítico del proyecto, incluyendo recomendaciones para futuros trabajos derivados de la puesta en producción del modelo. Además, se implementó una aplicación que permite realizar predicciones en tiempo real del costo de adquisición de nuevos clientes utilizando el modelo entrenado, a través de una interfaz diseñada con React y un backend con Fast API.

Palabras clave — Costo de adquisición de clientes, aprendizaje supervisado, experiencia de usuario.

ABSTRACT

The main objective of this project is to predict the cost of customer acquisition (CAC) in the FOODMART supermarket chain. In order to achieve this, a real database provided by the company was used, which contains detailed information from 60,000 sales bills including incomes, promotions, storage, sales, and advertising costs.

Taking into account the last point, data exploration and preprocessing techniques were applied to prepare the data, regression models such as linear regression and random forest were trained and validated using machine learning concepts. The models were evaluated using performance metrics such as mean squared error.

Therefore, obtained results were compared, and a critical analysis of the project was conducted, including recommendations for future work resulting from the deployment of the best model.

Furthermore, An application that allows real-time predictions of CAC using the trained model was implemented using the trained model, through a user interface designed with React and a backend with FastAPI.

Keywords — Customer adquistion cost, supervised learning, customer experience.

I. INTRODUCCIÓN

El costo de captación de nuevos clientes posee un impacto directo en la rentabilidad y el crecimiento empresarial. Un costo de adquisición reducido permite obtener mayores beneficios económicos por cada cliente incorporado, al tiempo que se pueden asignar más recursos a estrategias de marketing.

En este sentido, el análisis de datos juega un papel crucial, ya que proporciona una comprensión y predicción fundamentales de los costos asociados a este proceso. Así, el análisis de datos brinda beneficios tanto a las marcas como a la compañía, al optimizar los esfuerzos de adquisición de clientes y mejorar la eficacia en las decisiones empresariales.

En este contexto, el presente proyecto se centra en la cadena de supermercados FOODMART y tiene como objetivo desarrollar un modelo de machine learning capaz de predecir el costo de adquisición de nuevos clientes (CAC). Para ello, se cuenta con una base de datos real proporcionada por la empresa, que contiene información detallada de 60.000 facturas de clientes.

El proyecto se divide en tres etapas principales: análisis exploratorio de datos, modelado utilizando técnicas de machine learning y evaluación de los modelos desarrollados. En la etapa de análisis exploratorio, se examinan los datos disponibles y se aplican técnicas de preprocesamiento. A continuación, se procede a entrenar y validar validó un modelo de regresión lineal y dos de random forest, utilizando el conocimiento adquirido del dataset durante el análisis exploratorio. Por último, se evalúan los modelos desarrollados con la métrica de error cuadrático medio.

El proyecto también incluye la implementación de una aplicación que permite utilizar el modelo entrenado para realizar predicciones en tiempo real del costo de adquisición de nuevos clientes. Esto proporciona una herramienta útil para la toma de decisiones y la planificación estratégica en el contexto de costos variables de publicidad y generación de nuevas campañas de marketing.

Aplicación en vivo:

- https://ephemeral-lolly-5a0ff4.netlify.app/
- https://test-api-pr1g.onrender.com/docs

Repositorios:

- https://github.com/jmontoyavallejo/Monografia_frontend
- https://github.com/jmontoyavallejo/Monografia_backend
- https://github.com/jmontoyavallejo/monografia_docker_backend

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

II. PLANTEAMIENTO DEL PROBLEMA

Factores como la falta de información precisa y actualizada, la calidad de los datos, los altos costos asociados con las encuestas, panel de usuarios y actividades de marketing, hacen que calcular el costo de adquisición de clientes sea una tarea desafiante. Para superar esta problemática, las empresas deben trabajar en mejorar la recopilación y calidad de los datos, utilizar herramientas y técnicas de análisis adecuadas, así como considerar los diferentes costos asociados con las estrategias de marketing utilizadas y explorar nuevas alternativas como lo son el aprendizaje supervisado y técnicas de machine learning para tener un conocimiento más profundo de sus clientes sus costos asociados.

Para abordar este problema, es necesario analizar y utilizar el historial de compras de los productos, la información de los clientes y las instalaciones recopiladas previamente. Estos datos serán fundamentales para predecir el CAC de manera más precisa posible.

El CAC es una métrica de negocio para las empresas que tiene una relevancia importante en estas áreas:

1. Publicidad
2. Promociones
3. Marketing
4. Incentivos
5. Otros gastos relacionados con la adquisición de nuevos clientes.

Conocer el CAC es importante para poder tomar decisiones sobre cómo invertir en la adquisición de nuevos clientes y para comparar el CAC con el valor a largo plazo que un cliente le genera a la empresa y la viabilidad de una campaña publicitaria o un establecimiento comercial.

III. OBJETIVOS

OBJETIVO GENERAL

Desarrollar un modelo de machine learning capaz de predecir el costo de adquisición de nuevos clientes (CAC) en la cadena de supermercados FOODMART, utilizando técnicas de ciencia de datos y regresión.

OBJETIVOS ESPECÍFICOS

- Recopilar y analizar los datos históricos de transacciones hechas por la cadena de supermercados FOODMART, incluyendo información personal, información de los productos e información del establecimiento comercial.
- Aplicar técnicas de preprocesamiento de datos para limpiar y preparar adecuadamente la base de datos, asegurando la calidad y consistencia de los datos utilizados para el modelado.
- Entrenar y validar modelos de machine learning, específicamente regresión lineal y random forest, utilizando la base de datos preparada y los atributos relevantes para la predicción del costo de adquisición de nuevos clientes.
- Evaluar los modelos desarrollados utilizando el error absoluto medio, para determinar su capacidad predictiva y seleccionar el modelo más efectivo.
- Analizar las características y variables más influyentes en la predicción de resultados. Evaluar la importancia de cada variable en el modelo y su contribución al rendimiento predictivo.
- Recomendar el trabajo futuro que podría derivarse de los resultados obtenidos al poner el modelo en producción, identificando posibles despliegues y aplicaciones en relación con la predicción del costo de adquisición de clientes.

IV. MARCO TEÓRICO

Determinar el costo de adquisición de nuevos clientes puede ser una tarea complicada si no se cuentan con las herramientas y metodologías apropiadas para la consecución del objetivo. (Albán, 2016), menciona la dificultad que existe para determinar las variables claves que permitirían otorgar a los clientes una clasificación para la empresa acorde a la prioridad que cada uno suscita, pero se hace complicado debido a que áreas como logística y mercadeo no suelen tener una interrelación efectiva que permitiera obtener información relevante y útil de cara a la segmentación de clientes.

Por lo tanto, es muy importante contar con una buena comunicación entre las diferentes áreas de cualquier empresa, así como también contar con herramientas que consoliden la información en un solo lugar para así para así tener a la mano todas las variables que pudieran afectar el costo de adquisición de nuevos clientes, incluyendo toda la cadena de suministro de la empresa, desde el vendedor que contacta al cliente hasta el operario que despacha los pedidos que el mismo pudiera realizar.

Aplicar métodos como clustering y machine learning puede ser una buena aproximación para así agrupar los diferentes tipos de clientes según sus características comunes e incluso por el costo que cada uno representaría para la empresa o negocio. (Lu, 2016), aborda el tema para realizar pronósticos de ventas minoristas de computadoras. Para ello utilizó un modelo de predicción basado en clustering y machine learning donde inicialmente segmenta los datos según características en común de los diferentes compradores y luego realizó un modelo de predicción con base en las agrupaciones. Pudo determinar que la venta de computadoras no seguía una tendencia lineal, sino que era estacional, por lo que dependiendo de la época del año las ventas se comportan de una u otra manera.

Teniendo en cuenta lo anterior, sería muy útil revisar si la cadena de supermercados Convenient Food Mart presenta ventas estacionales o alguna otra tendencia diferente que pudiera servir como input de estrategias publicitarias y de marketing para así atraer clientes con un costo óptimo haciendo uso de las mejores herramientas en los momentos indicados, evitando así inversiones onerosas en campañas y estrategias que pudieran no llegar al mercado objetivo deseado, que para este caso sería el de adquirir nuevos clientes.

Ahora bien, ser asertivos en determinar el costo óptimo es parte fundamental en la planificación organizacional y la dirección de las instituciones (ACIMED, 1996). Los modelos de rentabilidad de casi cualquier compañía inician desde los costos, pues es desde estos sobre los cuales se determinan los precios de venta y los márgenes de ganancia que se obtendrán en cada una de las ventas. El que los costos sean elevados implica necesariamente que los precios serán más altos o los márgenes menores lo cual afecta negativamente a la empresa perdiendo market share frente a los competidores además de no generar a los clientes una propuesta de valor atractiva (RODRÍGUEZ, 2010).

Por lo tanto, ser eficientes en la administración del costo permite a la empresa tener mayores

palancas de maniobra respecto a las decisiones qué tomar para, en este caso, adquirir nuevos clientes. Ser eficientes con los costos de producción, transporte, publicidad, producción y además, permite tener un mayor margen de maniobra respecto de las estrategias a implementar que vayan en línea con los objetivos de la empresa y de las diferentes áreas que la componen. Cuando se tiene una estructura de costos y gastos muy alta, las posibilidades se reducen en el sentido que cualquier acción repercutirá fuertemente en los márgenes y ganancias de la compañía y finalmente el objetivo de toda empresa es maximizar el valor para los accionistas o propietarios de esta (Madrid, 2005).

Finalmente, la importancia de comprender el valor que el cliente aporta a la empresa y los riesgos asociados a la fuga de estos, son tópicos que han ganado relevancia en los últimos años y las organizaciones se han dado cuenta de ello al percatarse que los indicadores tradicionales no son suficientes para evaluar la rentabilidad de las inversiones destinadas a evitar el abandono (BUADES, 2015). Es por ello que las optimizaciones en el costo deben ir acompañadas de buenas estrategias de mercadeo, fidelización y retención con el fin de que los clientes más importantes de la organización no deserten. Para ello es muy importante estar conectados con las necesidades del cliente día a día y ofrecerles una propuesta de valor que los conecte con la empresa donde al final es un gana-gana tanto para el cliente como para la empresa.

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

MÉTRICAS DE DESEMPEÑO

Para evaluar el rendimiento de los modelos supervisados de machine learning fue el Error absoluto medio (MAE), un MAE más bajo indica un mejor ajuste del modelo a los datos.

El MAE se utiliza para medir qué tan cerca están las predicciones del modelo de los valores reales en promedio. Cuanto menor sea el valor del MAE, mejor será la precisión del modelo, ya que indicará que las predicciones tienen una menor diferencia promedio con los valores reales. Esta métrica se calcula de la siguiente manera:

$$MAE = (1 / n) * \Sigma|y_i - \bar{y}_i|$$

Donde:

MAE: es el Error Absoluto Medio.

n: es el número total de muestras o instancias en el conjunto de datos.

y_i: es el valor real de la i-ésima muestra.

$\bar{y}_i$: es la predicción del modelo para la i-ésima muestra.

Σ: representa la suma de todas las diferencias absolutas entre las predicciones y los valores reales.

Se escoge esta métrica para evaluar el modelo porque el MAE está en las mismas unidades de la variable dependiente y es más entendible para el cliente. en la base de datos el promedio de la variable dependiente CAC de la base de datos es de 99 dólares, para el proyecto es aceptable un error del 5% en las predicciones lo que se traduce en un valor de 5 en el MAE

En la sección 6.3.2, “consideraciones de métricas de negocio para una posible puesta en producción”, se plantea la sugerencia de definir métricas de desempeño que podrían ser relevantes para evaluar la viabilidad de una posible puesta en producción del proyecto. Sin embargo, es importante destacar que estas métricas no serán desarrolladas a lo largo de este documento.

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

V. DATOS

Los datos originales de kaggle fueron compartidos por el usuario "RAM JAS", son los contenidos en el archivo .csv de 60.000 filas y 40 columnas y pesa 2 MB aproximadamente.

Según la descripción de la base datos, incluye datos sobre productos, promociones, clientes, instalaciones, ventas y costo CAC de un conjunto de 60.000 clientes.

Luego de hacer un análisis de la información contenida en el dataset, como la información física que se tiene de los establecimientos comerciales, detalles de los clientes y características del producto que compraron, se sugieren algunas formas en las que los datos podrían haber sido recopilados:

- ventas y costos internos de CFM: CFM podría haber recopilado los datos a partir de sus propios registros internos de las sucursales y contables, que podrían incluir información sobre ventas, costos, promociones, locación, cliente, etc.
- transacciones de tarjetas de crédito: CFM podría haber obtenido los datos a partir de registros de transacciones de sus clientes, que podrían incluir información sobre las compras realizadas, los costos, las promociones, etc.

Cada fila del conjunto de datos representa una transacción en particular, y cada columna proporciona información sobre un aspecto de la transacción ya sea información del cliente, del producto o de las instalaciones del establecimiento comercial.

El modo de acceso es a través del post de kaggle en el siguiente link:
https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart

La base de datos en kaggle no tiene información precisa sobre cada columna ni sobre el contexto en el que se recopilaron los datos. Esta aproximación al contenido de las columnas fue inferido a través de los datos que contienen y la distribución que se presenta , a continuación se describen las columnas:

- food_category: categoría del producto comprado
- food_department: departamento del producto comprado
- food_family : familia del producto comprado
- unit_sales(in millions): costo de unidades compradas del producto
- promotion_name: nombre de la promoción aplicada en el descuento
- sales_country: país de la venta
- marital_status: estado civil del cliente
- gender: género del cliente
- total_children: número de hijos del cliente
- education: escolaridad del cliente
- member_card: tipo de tarjeta de asociado del cliente
- occupation: ocupación del cliente
- houseowner: estatus de vivienda del cliente
- avg_cars_at home(approx): número de vehículos del cliente
- avg. yearly_income: ingresos del cliente anuales
- num_children_at_home: número de niños en el hogar del cliente
- brand_name: marca del producto

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

- SRP: precio sugerido de venta por el fabricante del producto
- net_weight: peso neto del producto
- recyclable_package: lleva empaque reciclable
- low_fat: producto bajo en calorías
- units_per_case: unidades del producto
- store_type: tipo de establecimiento comercial
- store_city: ciudad del establecimiento comercial
- store_state: estado del establecimiento comercial
- grocery_sqft: área del establecimiento comercial
- frozen_sqft: presencia de zonas frías en el establecimiento comercial
- coffee_bar: presencia de zonas de café en el establecimiento comercial
- video_store: presencia de zonas de video en el establecimiento comercial
- prepared_food: comida preparada dentro del establecimiento comercial
- florist: presencia de floristería en el establecimiento comercial
- media_type: tipo de medio usado por la publicidad
- cost: costo de adquisición del cliente CAC

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

VI. METODOLOGÍA

En este documento, se desarrollan modelos predictivos para predecir el costo de adquisición de los clientes, con el objetivo de mejorar la experiencia del usuario al encontrar los mejores locales y productos según el público objetivo. En la sección VIII de las posibles consideraciones para una posible puesta en producción se enuncian las posibles métricas de negocio que se deben tener en cuenta en un despliegue del modelo, sin embargo estas no van a ser desarrolladas durante el ejercicio del documento.

A continuación se muestra el diagrama de flujo del modelo, en la que se evidencia el paso a paso desde la entrada de los datos hasta la entrega final del modelo en un archivo .pickle:

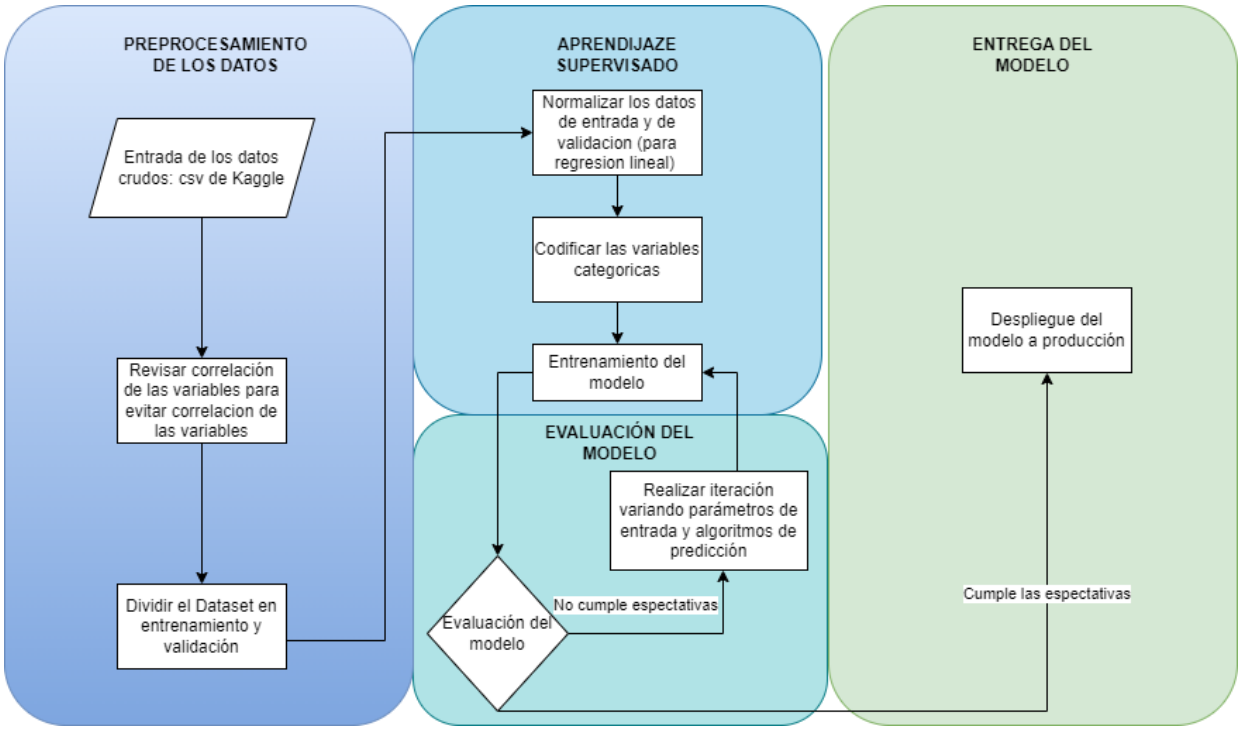

Fig 4. Diagrama de flujo del modelo de predicción del C.A.C.

PREPROCESAMIENTO

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

Para el proyecto de predicción del costo de adquisición de nuevos clientes (CAC) se consideró en el preprocesamiento de datos las siguientes tareas:

- Tratamiento de valores faltantes: En el caso de estudio no se presentaron valores faltantes.
- Tratamiento de valores duplicados: En el caso de estudio no se presentaron registros duplicados.
- Selección de características: Evitar entrar información redundante al modelo para reducir la complejidad del modelo y mejorar la interpretación de los resultados.
- Normalización de los datos: Se ajustaron los valores de las columnas numéricas para que estén en un rango similar, paso necesario para el método de regresión lineal .
- Codificación de variables categóricas: Fue necesario codificar las variables categóricas para poder ser utilizadas en los modelos supervisados.

VARIABLE DEPENDIENTE (CAC)

La variable dependiente está balanceada en cada rango inter cuartil como se observa en la figura 1, esta no posee valores atípicos. También se evidencia que los datos poseen una distribución aleatoria o multimodal ya que no tienen un patrón discernible.

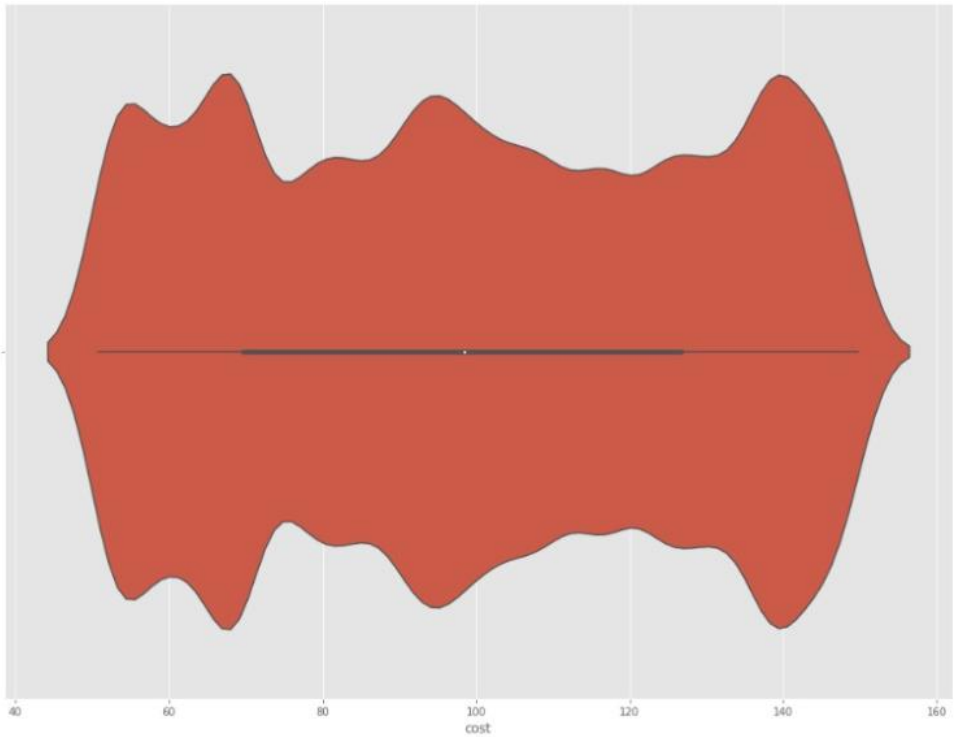

Fig 1. Distribuciones de la variable dependiente

VARIABLES NUMÉRICAS

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

Dentro de las variables numéricas se encuentran algunas variables que tiene una alta correlación entre sí, para evitar enviar información redundante al modelo y posibles errores en la interpretación de los resultados se procede a eliminar del dataset las columnas que tengan una correlación mayor a 0.8 o menor a -0.8.

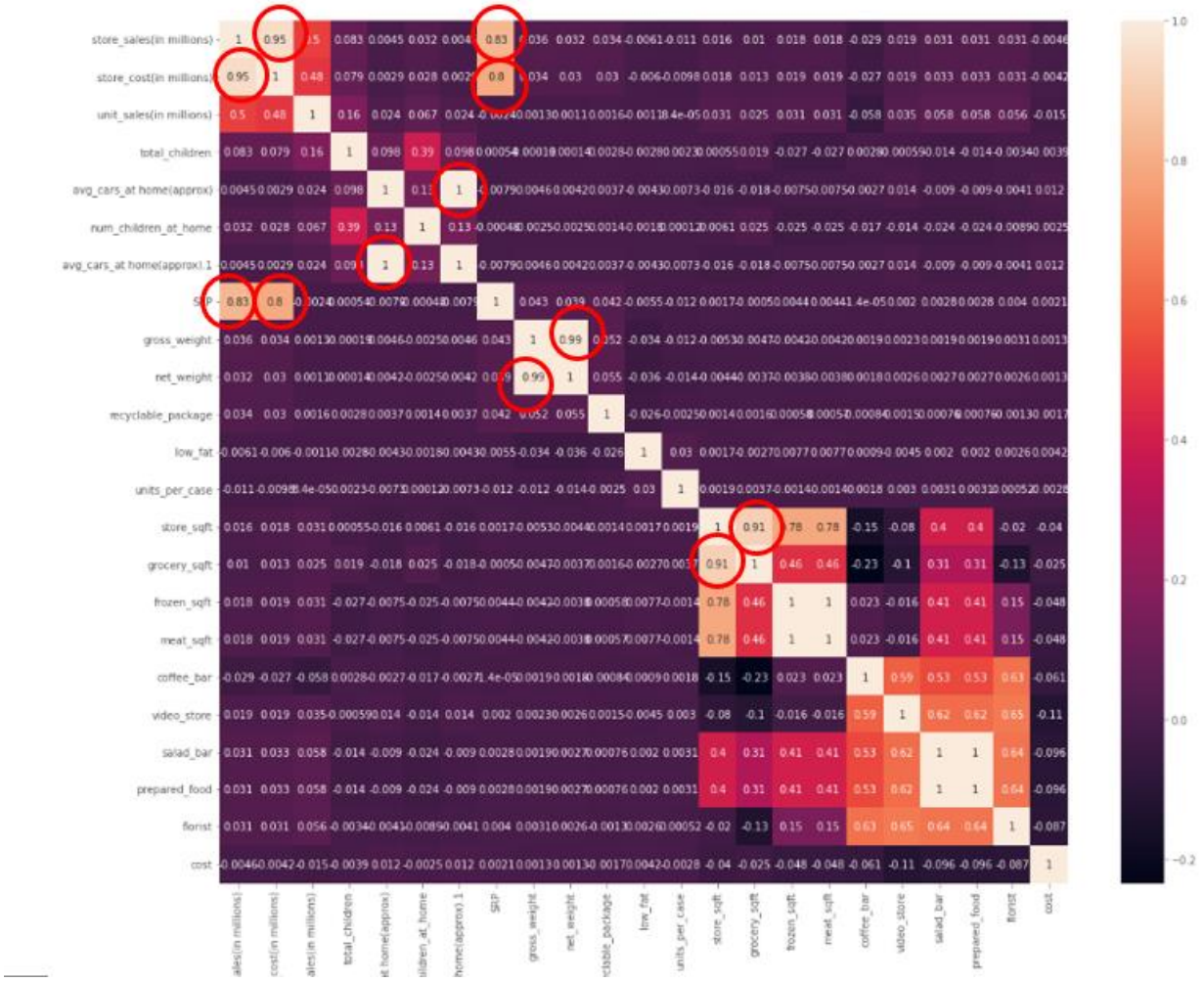

Fig 2. Gráfica de correlación de las variables numéricas

Con este paso las variables que permanecen en el dataset explican en una alta proporción las características que fueron excluidas del mismo y mejora el rendimiento del modelo.

El dataset posee 17 variables categóricas, algunas de ellas con una amplia variación en las opciones como se observa en el gráfico:

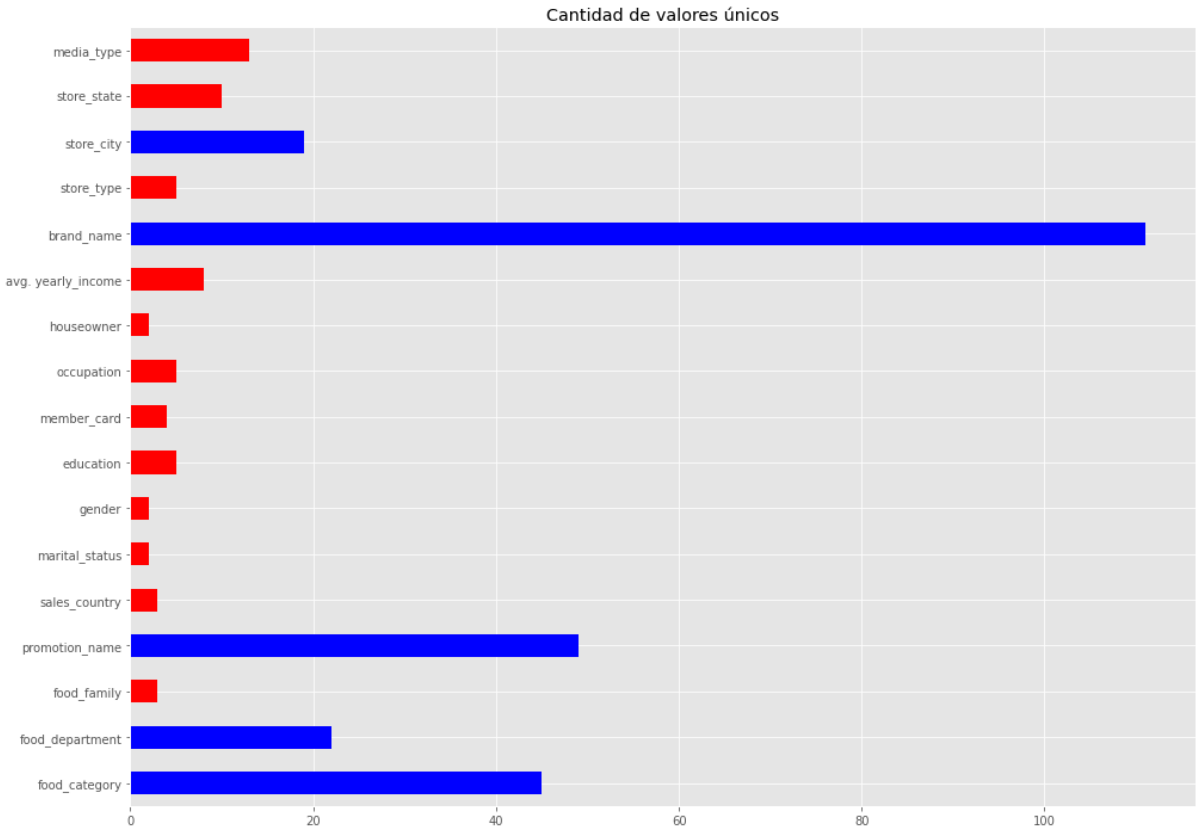

Fig 3. Distribución de valores únicos en las variables categóricas.

Las opciones que tienen más de 20 posibilidades no son tomadas en cuenta en la primera validación ya que esto puede dificultar su uso en algunos modelos de predicción, especialmente si se utilizan métodos de codificación de características del tipo one hot encoding que crea una columna separada para cada valor único. Esto puede aumentar drásticamente la dimensionalidad del conjunto de datos y hacer que el modelo sea más difícil de entrenar y menos eficiente en términos de memoria, en futuras iteraciones se utilizarán estas variables para evidenciar si la suposición fue acertada o afecta negativamente las predicciones.

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

Para el proyecto de predicción del costo de adquisición de nuevos clientes (CAC) con modelos de regresión lineal y random forest en Python, algunas librerías que fueron utilizadas para el desarrollo del ejercicio fueron:

- numpy: librería para realizar operaciones matemáticas y numéricas con arrays y matrices.
- pandas: librería para el análisis y manipulación de datos. También permite leer archivos CSV
- matplotlib: librería base para la generación de gráficos y visualizaciones de datos.
- seaborn: librería para generar gráficos y visualizaciones de datos de forma más accesible utilizando como base matplotlib.
- sklearn: librería de machine learning que proporciona una gran variedad de funciones y modelos para realizar tareas de clasificación, regresión, clustering, etc.
- math: librería soporte para funciones matemáticas dentro del desarrollo del ejercicio

HERRAMIENTAS PARA CALCULAR MÉTRICA DE DESEMPEÑO

Para calcular las métricas de desempeño de un modelo de machine learning y de negocio en Python, se usaron algunas funciones de la librería scikit-learn. Se enumeran de la siguiente manera:

- Error absoluto medio (MAE): La función "mean_absolute_error" se puede usar de la siguiente forma:
    - *from sklearn.metrics import mean_absolute_error*
    - *mae=mean_absolute_error(y_true, y_pred, *, sample_weight=None, multioutput='uniform_average')*

APRENDIZAJE SUPERVISADO

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

En el proyecto de predicción del costo de adquisición de nuevos clientes (CAC), se evaluó la utilización de modelos de aprendizaje supervisado. Se inició con la implementación de un modelo de regresión lineal, Este modelo se basa en la idea de que existe una relación lineal entre las variables independientes (o características) y la variable dependiente (en este caso, el CAC) y que la variable dependiente es continua. Posteriormente, se llevaron a cabo dos iteraciones adicionales utilizando el modelo de random forest.

ITERACIONES y RESULTADOS

PRIMERA ITERACIÓN

Durante la primera iteración, se realizó un modelo supervisado de regresión lineal, el cual obtuvo los siguientes resultados:

```
El error MAE   de test es: 23.367551465872463 dolares en promedio
El porcentaje de error respecto a los valores originales es del: 86.95%
```

Fig 5. Resultados de la evaluación del modelo primera validación

Las métricas de desempeño del modelo son bajas ya que el criterio de aceptación es de 5 dólares y está en 23.3 dólares.

Haciendo una revisión de los errores, como se observa en la figura 6, Existen residuos tanto positivos como negativos y la distribución del error no está segmentado en un grupo particular de los datos sino que se produce en todo el dataset. Por tanto el error se puede haber generado por el modelo de regresión lineal que se utilizó o la suposición de que el dataset tiene relaciones lineales con respecto a la variable dependiente CAC.

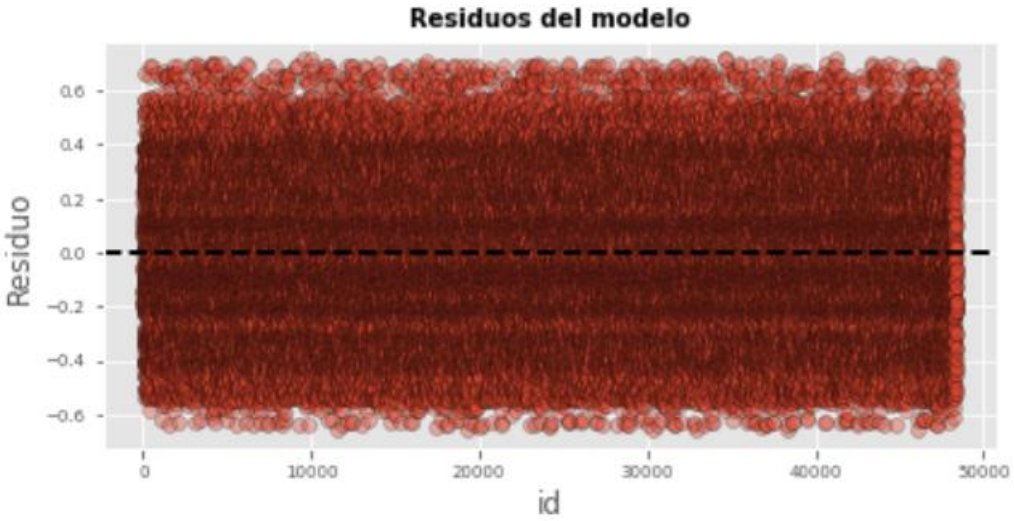

Fig 6. Gráficos de residuos obtenidos de comparar las predicciones y el valor real del C.A.C

.

Durante el preprocesamiento de la primera interacción del modelo se omitió un paso importante del preprocesamiento de datos que es la revisión de las relaciones de las características con

respecto a la variable dependiente. basado en este procedimiento si las relaciones son aproximadamente lineales un método como la regresión lineal es viable para el proyecto.

Como se observa en la figura 7, en la correlación de las variables numéricas con respecto al CAC, ninguna de las variables presenta una correlación lineal con respecto a la variable de salida del modelo, por esta razón se obtuvo un bajo desempeño en la predicción:

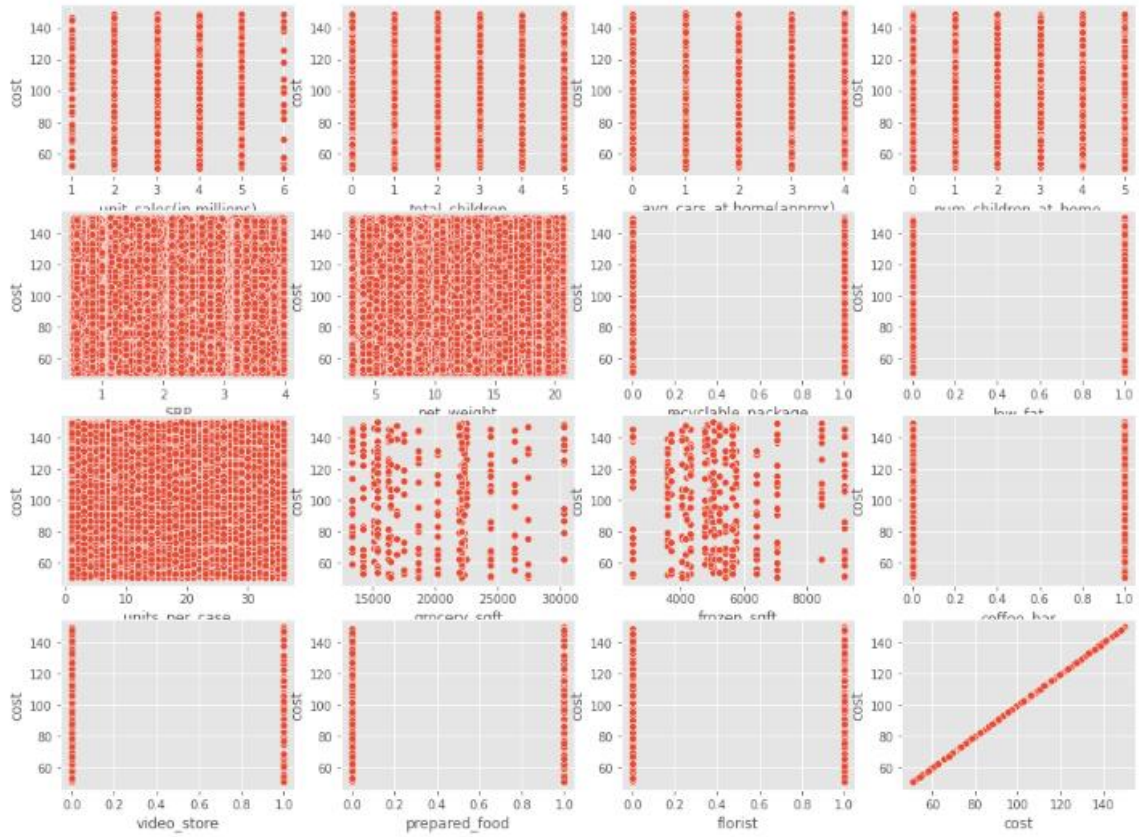

Fig 7. Gráfica de distribuciones con respecto al C.A.C. de variables numéricas

Luego de hacer la primera iteración con el modelo de regresión lineal, se opta por hacer dos iteraciones más con el modelo de RandomForest ya que los resultados no cumplen el criterio de aceptación especificado en la sección 2.4.

SEGUNDA ITERACIÓN

La segunda iteración se realiza un modelo de predicción de random forest con el mismo conjunto de entrenamiento y el mismo conjunto de validación de datos, ya que al evidenciar que la correlación entre las variables y la variable dependiente no es lineal se opta por usar un modelo apropiado para relaciones no lineales

El MAE se reduce sustancialmente de 23,26 a 10,05 solo cambiando el algoritmo de predicción, ya que la distribución de los datos y la correlación de las variables no son lineales, el modelo tiene una mejora sustancial frente a la primera iteración de regresión lineal

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

TERCERA ITERACIÓN

En la tercer iteración del análisis se plantea agregar dos columnas que fueron extraídas durante la exploración de los datos que fueron las columnas promotion name y Media_type, ambas son variables categóricas con un alta cantidad de valores únicos (49 y 19 respectivamente) las cuales fueron extraídas durante el análisis para evitar la alta dimensionalidad del dataset en el modelo.

Se encontró que estás variables tienen una alta importancia en el desempeño del modelo y en el coeficiente de determinación del conjunto de prueba mejorando en una alta medida el desempeño de las predicciones:

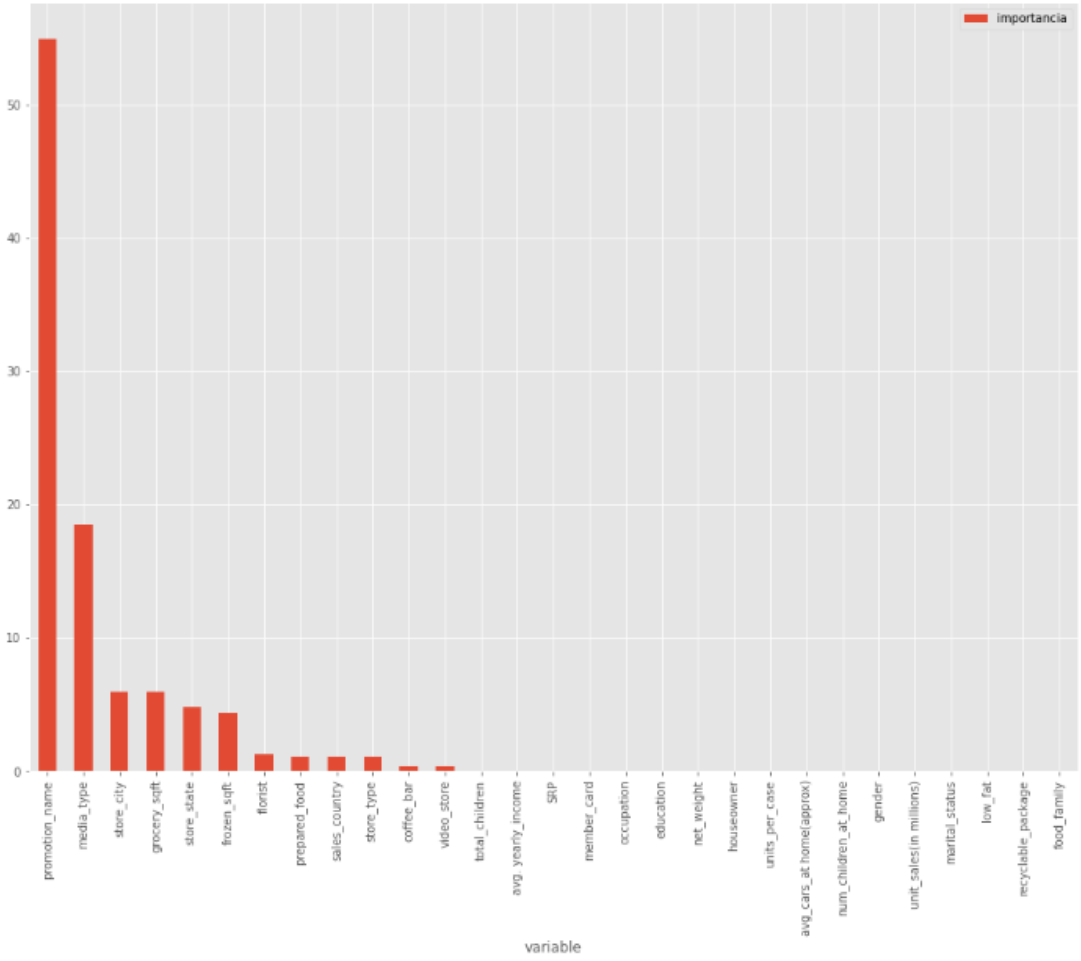

Fig 8. importancia [%] de cada variable en la decisión de predicción del modelo

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

ENTREGA DEL MODELO

Después de completar la tercera iteración, el modelo de aprendizaje supervisado ha logrado satisfacer las expectativas establecidas y ha arrojado resultados prometedores. Estos resultados se presentarán de manera detallada en la sección V del informe. Con el propósito de facilitar su utilización en diversas plataformas, se proporcionará al final del documento un archivo .pickle que contendrá el modelo entrenado. Esto permitirá su implementación en entornos como una API.

En la sección VIII del informe, se explorarán diferentes enfoques para utilizar el modelo entrenado con miras a una posible puesta en producción. Se analizarán opciones que permitan aprovechar al máximo los beneficios del modelo y su integración en sistemas en un posible despliegue integrándose en el diseño del sistema.

VII. RESULTADOS

A continuación, se enuncian los resultados de cada iteración y la evaluación cualitativa del desempeño de los modelos.

MÉTRICAS

Experimento	**MAE [dólares]**	**Cumple métrica del modelo: MAE <= 5**
Primera interacción: regresión lineal	23.26	NO
Segunda interacción: random forest	10.05	NO
Tercera interacción: random forest	0.04	SI

Tabla 1. Tabla de resultados del del modelo en las iteraciones

EVALUACIÓN CUALITATIVA

Para la métrica de machine learning del MAE de la tercera interacción se tuvo una variación de las predicciones del 0.04 dólares en promedio, teniendo en cuenta que se utilizan los hiperparámetros de profundidad máxima, número de árboles y número de características en cada árbol optimizados por el grid search, este es un indicio de que el modelo puede presentar un desempeño alto al inicio de la producción, también se evidencia una clara dependencia con la variable promotion name ya que está representa el 57% de la importancia que toma el modelo para definir una predicción sobre las otras 27 variables que se toman en cuenta, se debe revisar a futuro la dependencia del modelo a esta variable ya que puede estar presentando un caso de overfitting teniendo en cuenta que tiene un acierto de más del 99% en la predicción de los dólares en la tercera iteración que puede reducir el accuracy del modelo cuando se tengan nuevos datos.

También, con base en los resultados del modelo, la cadena de supermercados convenient FoodMart logrará implementar planes de acción acertados e implementar mejores campañas publicitarias en los sitios y productos recomendados por el modelo y con esto lograr reducir costos del CAC sin afectar la fidelización de los clientes al ser las campañas efectivas para el nicho de mercado seleccionado y no sobre la totalidad de tiendas y productos.

VIII. CONSIDERACIONES DE POSIBLE PUESTA EN PRODUCCIÓN

Algunas consideraciones técnicas que son importantes que se mencionan en las siguientes dos subsecciones para el proceso de puesta en producción del modelo de ML para la predicción del costo de adquisición de nuevos clientes (CAC) son:

- Diseño de sistema completo para la puesta en producción
- Posibles métricas de negocio para una puesta en producción

DISEÑO DEL SISTEMA EN PUESTA DE PRODUCCIÓN

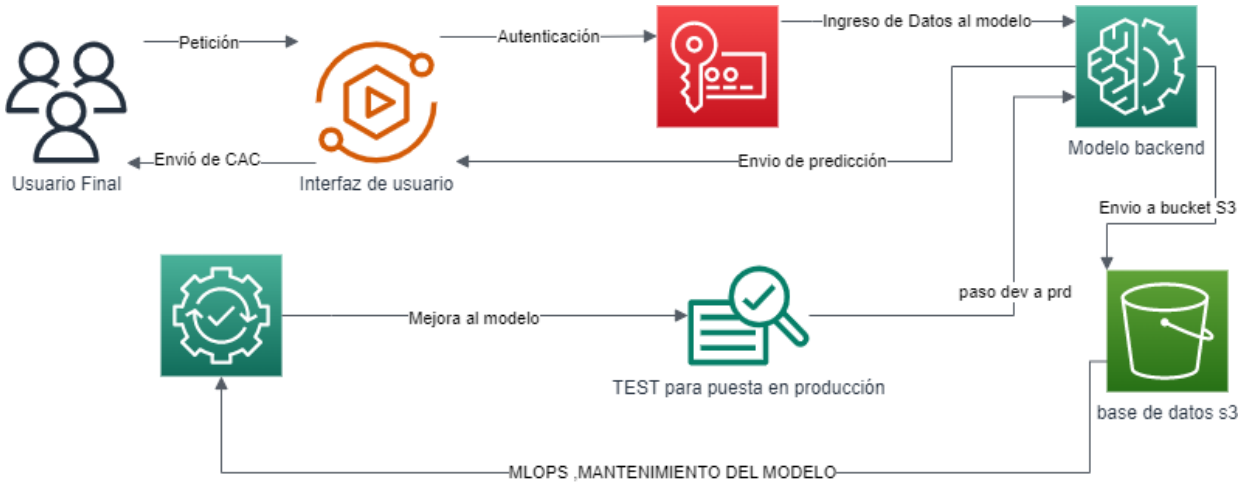

Fig 13. Diagrama de diseño del modelo en producción.

OBJETIVOS DE LA IMPLEMENTACIÓN EN PRODUCCIÓN

El objetivo del diseño de sistema es facilitar el proceso de cálculo de costo de adquisición de cliente, entregando alternativas viables para planes de ahorro dentro de la cadena de supermercados footmart y predicción de budgets para las campañas de la empresa CFM.

De acuerdo a esto, el producto final debe ser ágil, intuitivo, eficiente y fácil de usar para no programadores, debe ofrecer características necesarias para predecir el costo de adquisición de clientes permitiendo que las personas que lo usen puedan tener resultados efectivos y una curva de aprendizaje respecto del manejo de la herramienta muy corta y no requieren de papeleo ni grandes conocimientos previos para interactuar con él.

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

CONTEXTO DEL SOFTWARE A USAR

El sistema de predicción de costos cuenta con dos componentes principales, una interfaz de usuario destinada para el personal administrativo de la compañía que actualiza y realiza peticiones a lo largo de los servidores backend donde se almacena la api con el modelo y debe retornar las predicciones a los usuario.

Por medio de solicitudes http se enviará la información a la api y esta se encargará de generar las funciones necesarias para calcular la predicción, almacenar y medir el acierto que se tenga con el modelo de predicción en el futuro.

Para ello, se planea usar una interfaz de usuario en una aplicación de React en la cual se desplegará la interfaz de usuario en AWS, las conexiones hacia el backend para la predicción del modelo se implementan tanto en un servidor de S3 como en una lambda para generar la predicción, y estas peticiones se almacenarán a su vez en un bucket de S3 destinado para generar un control sobre el desempeño del modelo en el tiempo

PÚBLICO OBJETIVO DEL PROYECTO

Aunque los requerimientos del modelo tienen previsto un público administrativo de mercadeo, el proyecto puede ser escalado para tener más funcionalidades que le permitan a otras áreas como TI, contabilidad, líderes de proyectos y gerentes tener un impacto y uso en la aplicación para generar valor en la compañía.

Este proyecto debe tener un documento anexo en el cual se especifiquen las partes de las aplicaciones y la proyección final de la plataforma de la siguiente manera:

- Introducción: Resumen ejecutivo del modelo
- Estructura de los datos: Cómo la plataforma organiza, almacena y flujos de transporte
- Arquitectura de cada componente: Interfaz, clases, algoritmos, eficiencia y métricas
- Interfaz de usuario: Información preliminar de la aplicación y datos preliminares
- Restricciones y limitaciones: Discusión general de limitaciones impuestas en el inicio
- Testeo: Diseño de pruebas unitarias para testeo y respuestas esperadas
- Anexos: Información adicional que puede ser útil para los lectores acerca de los componentes y la aplicación en detalle

La estructura interna de la data del proyecto se describe en los componentes: interfaz de usuario, servidor backend y almacenamiento de datos S3 de AWS.

En la interfaz de usuario la data es organizada por medio de JSON y será organizada en el formulario donde se envía la información al servidor de python en forma de un diccionario.

Está información viaja a la api por medio de una petición POST en el protocolo HTTP, luego en la api se genera la predicción y se envían los datos completos a un bucket de S3 donde se almacenan en formato .csv.

La respuesta del servidor se envía a la interfaz de usuario en forma de un diccionario de python y es enviada a través de una notificación al usuario.

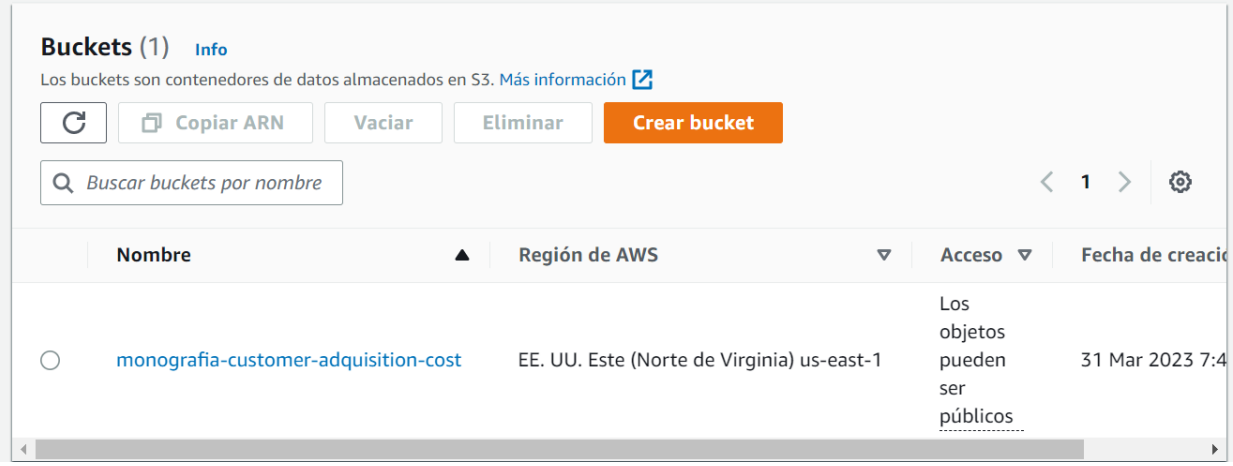

Fig 9. Implementación del bucket s3 para almacenar datos de CAC.

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

La arquitectura de los componentes se divide en la aplicación de interfaz de usuario y el servidor backend.

Para la interfaz de usuario se utiliza el lenguaje Javascript como base y el framework React ,un framework de CSS llamado tailwind en el apartado gráfico y un gestor de alertas llamado Sweet Alerts para crear los componentes necesarios como lo son los contenedores de formularios para información del cliente, de la sucursal y del producto para enviar la información al modelo. También se crea un apartado de helpers que ayuda a manejar las interacciones del usuario con la aplicación, el envío de peticiones a la base de datos y al modelo de predicción y la generación de alertas.

Para el manejo del modelo se hicieron 3 implementaciones en python donde se logra recibir las peticiones y devolver en un diccionario la predicción del modelo, a continuación se dejan los links de las implementaciones:

Implementación de aplicación backend en fast api:
https://github.com/jmontoyavallejo/Monografia_backend

Implementación de docker:
https://github.com/jmontoyavallejo/monografia_docker_backend

Implementación de lambda:
https://us-east-1.console.aws.amazon.com/lambda/home?region=us-east-1#/functions/monografia-function?tab=code

Para la base de datos se utiliza un bucket de S3 donde se almacenan las peticiones enviadas por la interfaz de usuario, estas llegan en un formato json y se almacenan en un archivo .csv para ser revisadas periódicamente y realizar mediciones de desempeño del modelo

SEGURIDAD Y ACCESO A LAS BASES DE DATOS

Para el acceso al servidor donde están almacenados los modelos se implementa un control de orígenes donde se restringen las url donde se pueden recibir las peticiones a solo la url donde está almacenada la interfaz de usuario, esto para tener un control contra ciberataques.

En la figura 10 se muestra que la api sólo recibe solicitudes del servidor de pruebas localhost y de la página web donde está desplegada la interfaz de usuario

```python
from fastapi import FastAPI
from pydantic import BaseModel
from fastapi.middleware.cors import CORSMiddleware
from app.models.model import predict_cost
from app.models.model import __version__ as model_version


origins = [
    "http://localhost",
    "http://localhost:3000",
    "https://ephemeral-lolly-5a0ff4.netlify.app/",
   'https://ephemeral-lolly-5a0ff4.netlify.app',

]
app=FastAPI()

app.add_middleware(
    CORSMiddleware,
    allow_origins=origins,
    allow_credentials=True,
    allow_methods=["*"],
    allow_headers=["*"],
)
```

Fig 10. Implementación de control de origen en peticiones del backend desde la url

Para el acceso a la base de datos y la lambda de AWS se realiza la creación de cuenta de servicio en IAM para tener las credenciales de identidad y conectarse por fuera del entorno de trabajo de AWS sin tener que utilizar la cuenta raíz

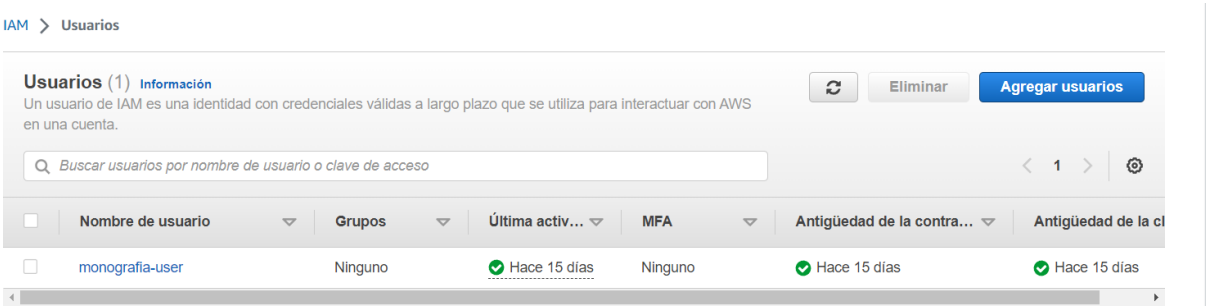

Fig 11. Implementación de cuenta de servicio en AWS para credenciales e interacción entre aplicaciones

Para el acceso a la interfaz de usuario se implementa un control de usuario y una identificación al inicio de la aplicación con el fin de controlar por medio de correo electrónico las personas que tendrán acceso a la interfaz.

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

DISEÑO DE INTERFAZ DE USUARIO

La interfaz de usuario cuenta con 3 contenedores de formulario donde se encuentran las características relevantes con opciones de lista desplegable para el modelo de la información de la transacción, como lo son la información del establecimiento comercial, del cliente y del producto.

En la sección superior de la aplicación se tienen 3 botones, uno para limpiar los datos que han sido ingresados, otro para realizar un ingreso aleatorio de datos con el fin de hacer pruebas de funcionamiento y otro para predecir el costo de adquisición del cliente, este último envía la petición al servidor del modelo y retorna la respuesta por medio de una alerta.

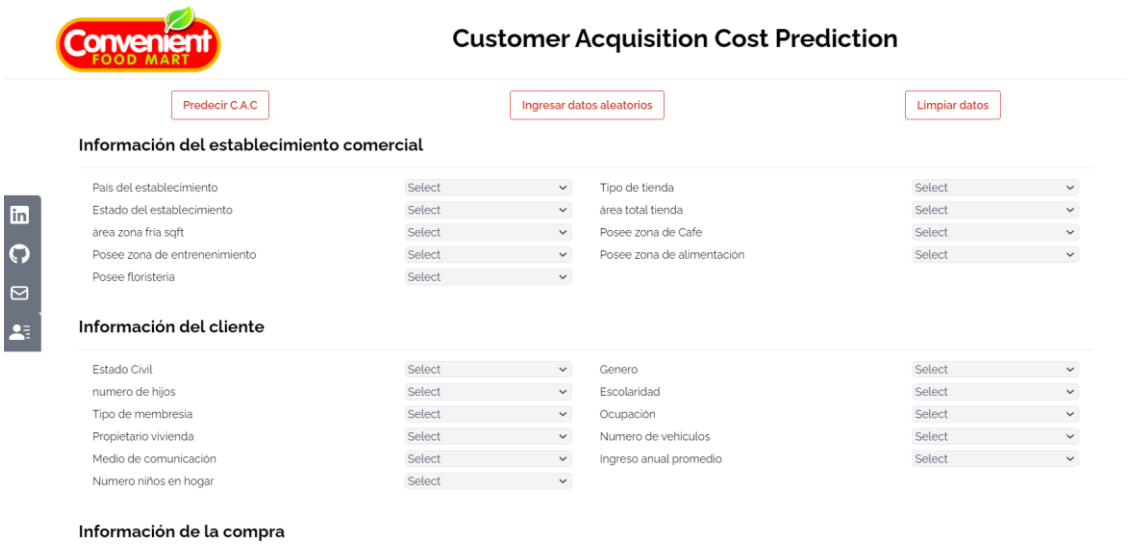

Fig 12. Implementación de interfaz de usuario en REACT alojada en un servidor en netlify

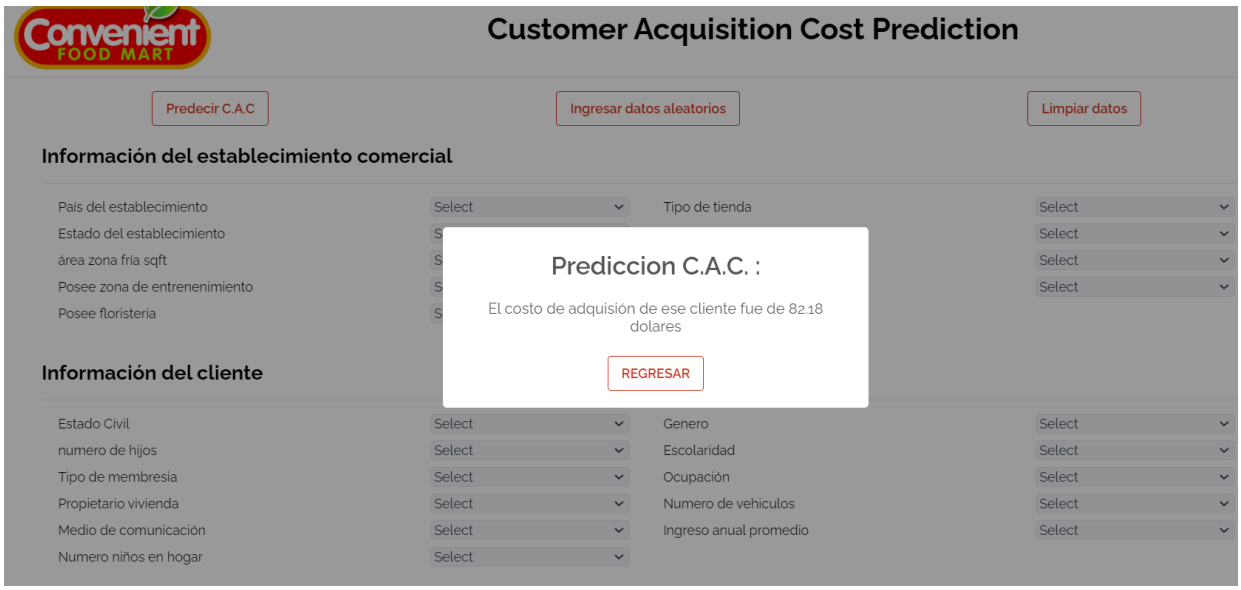

fig.13 Respuesta predicción de CAC nuevo cliente

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

REGLAS Y CONSIDERACIONES DEL DISEÑO DE LA INTERFAZ

En la interfaz de usuario se tienen las siguientes reglas para el manejo de las interacciones con los usuarios:

- En cada menú desplegable se debe elegir una opción, esto con el fin de tener una consistencia en los datos y reducir los datos nulos desde la entrada
- Las variables del modelo de predicción han sido modificadas al idioma español para tener una información.
- Hay un tiempo entre peticiones de 5 segundos que debe respetarse para no sobrecargar las peticiones del modelo.
- En la parte inferior de la aplicación se encuentra un abstract explicando la interfaz y el manejo de la misma para los usuarios que la utilizan por primera vez
- El sistema no almacena las peticiones ni las respuestas en caché con el fin de no llenar al usuario en memoria en su interfaz web

CONSIDERACIONES PARA TESTEO

WHITE BOX TESTING

Cuando cada aplicación sea dispuesta para el ambiente de desarrollo, los desarrolladores asignados a las pruebas tendrán la responsabilidad de debugear el código e implementar pruebas unitarias para las funcionalidades iniciales y futuras funcionales.

El código debe ser versionado y documentado a través de github como se puede evidenciar en los repositorios que se mencionan en apartados anteriores.

BLACK BOX TESTING

Black Box Testing contiene una gran parte del proceso de prueba. Esto se hará después de que todos los componentes hayan sido evaluados/dirigidos y consistirá en ejecutar todas las situaciones posibles que puedan ocurrir en el uso de la interfaz web, el bucket de s3 y el servidor donde se almacena el modelo.

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

SEGUIMIENTO MÉTRICAS DE DESEMPEÑO DEL MODELO

Para las métricas de desempeño se plantea una revisión mensual del modelo, en la cual se medirá con respecto al MAE la variación del CAC que entrega el modelo vs los datos resultantes de las futuras campañas publicitarias objetivo que se tengan en el plan de ahorros, con base en ellos se harán ajustes en el modelo con los nuevos datos para afinar las métricas de desempeño de las nuevas campañas y no reducir las mismas durante el proyecto.

ANEXOS

Como se mencionó en el apartado de público objetivo se tendrá un manual de usuario con generalidades completas del proyecto tanto para desarrolladores como para público general de la compañía.

INSTALACIÓN DE PAQUETES Y DEPENDENCIAS

Para la interacción de los usuarios no es necesario la instalación de aplicaciones ni paquetes, basta con conectarse a la aplicación web e interactuar con la interfaz gráfica para recibir las predicciones.

Dependiendo de la aplicación donde se decida alojar el modelo se tienen dos metodologías en un servidor o serverless (lambda) , se tienen los siguientes pasos para la instalación de paquetes y dependencias:

- Servidor
    - Crear un ambiente virtual para almacenar las dependencias
    - Si se cuenta con la aplicación de docker seguir los pasos descritos del dockerfile para hacer el build de la aplicación
    - Si no se cuenta con la aplicación de docker se debe instalar los requerimientos del archivo requirements.txt
    - Buildear la aplicación de fast api con el comando uvicorn app.main:app –reload
- Serverless
    - Almacenar la función del modelo dentro de una lambda
    - Revisar las dependencias que se alojan en el archivo requirements.txt
    - Disponer la interfaz de usuario con las librerías de AWS para permitir que el handle Submit del formulario active la lambda y haga la producción, esto ingresando las credenciales de la cuenta de servicio

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

POSIBLES MÉTRICAS DE NEGOCIO PARA UNA PUESTA EN PRODUCCIÓN

Al analizar los datos históricos de las transacciones y la información de los clientes, productos y sucursales, es posible identificar patrones y tendencias que ayuden a la empresa a comprender mejor a sus clientes y ofrecerles productos personalizados y relevantes.

Entender el histórico permite saber cuáles promociones y productos han tenido más éxito sobre el público, aumentar la fidelidad de los clientes en las marcas y reducir la probabilidad de que se vayan a la competencia.

Las métricas de desempeño de negocio que se podrían considerar para evaluar el rendimiento de los modelos de predicción son:

- Retorno sobre la inversión: mide el beneficio obtenido en relación con la inversión que puede suponer implementar en producción el modelo de predicción del CAC. El método por el cual se puede monetizar en primera instancia el proyecto es ejecutar campañas publicitarias según el público objetivo en las sucursales donde se genere el menor costo de adquisición predicho por el modelo y comparar con campañas publicitarias hechos sin tener en cuenta el modelo, para encontrar el % de reducción de costos en publicidad y tener un retorno de la inversión por medio de este plan de ahorros.
- Tiempo de amortización: mide el tiempo que se tarda en recuperar la inversión inicial del ejercicio del modelo en producción a través de los beneficios sobre las campañas, se espera un tiempo de amortización de 7 años basado en los 14,2% de retorno por año en 7 años se tendrá el retorno completo de la inversión.

PREDICCIÓN DE COSTO PARA ADQUIRIR NUEVOS CLIENTES

IX. CONCLUSIONES

El proyecto de predicción del costo de adquisición de nuevos clientes (CAC) en una cadena de supermercados podría ser una tarea de alto interés para las áreas de ventas y mercadeo que permitiría a la empresa tomar decisiones más informadas sobre cómo se están desempeñando sus estrategias de marketing, a qué tipo de público están impactando realmente, la efectividad de las campañas de publicidad y los medios en los que más se informan los clientes y los costos finales de adquisición de clientes.

A partir de los resultados satisfactorios obtenidos por el modelo desde las métricas de desempeño de ml y de negocio, la empresa convenient Food Mart podría tener una mejor comprensión de los factores influyen más en el CAC, como impactar más a cada tipo de usuario y puntos de partida para optimizar sus esfuerzos de marketing para minimizar el costo.

Si se decide implementar el modelo en producción, se podrían realizar trabajos futuros para mejorar y optimizar su desempeño para incluir más promociones y mejorar los puntos de venta. Para esto se podría incluir la adición de nuevos datos y revisión continua del desempeño del modelo en el tiempo. También podría ser útil explorar la utilización de técnicas de aprendizaje automático no supervisado, como el clustering, para obtener una comprensión más profunda del público objetivo de cada promoción y campaña publicitaria para identificar patrones o tendencias que podrían no ser evidentes a simple vista y mejorar el impacto que se tienen en los usuarios finales desde los productos que se comercializan y los establecimientos comerciales.

X. REFERENCIAS

- ACIMED (Ed.). (05 de 1996). Scielo. Obtenido de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94351996000200006
- Albán, H. M. (2016). APPLICATION OF KOHONEN MAPS FOR THE PRIORITIZATION. EIA, Antioquia. Envigado: Revista EIA, ISSN 1794-1237. Obtenido de https://bit.ly/45RvIpd
- BUADES, J. C. (2015). rio.upo. Obtenido de https://bit.ly/3NgYEju
- GONZÁLEZ, Á. A. (04 de 12 de 2021). Fundación Universitaria Los Libertadores. Obtenido de https://bit.ly/3MNBsYB
- Lu, I.-F. C.-J. (03 de 02 de 2016). Springerlink. Obtenido de Neural Computing and Applications: https://link.springer.com/article/10.1007/s00521-016-2215-x
- Madrid, C. d. (Ed.). (2005). ECONOMISTAS, COLEGIO DE MADRID. Obtenido de https://bit.ly/3qpFuPg
- RODRIGUEZ, O. L. (2010). repository.javeriana.edu.co. Obtenido de https://bit.ly/3CgLcWp