Autor: Nicolás Castillo - Sebastián Arango - Camilo Delgado - Johan Rodríguez

PROYECTO FINAL ENTREGA 1 - CIENCIA DE DATOS APLICADA

1 DEFINICIÓN DE LA PROBLEMÁTICA Y ENTENDIMIENTO DEL NEGOCIO

Para el desarrollo del proyecto de ciencia de datos, hemos seleccionado a OPAIN S.A., empresa encargada de la concesión del aeropuerto El Dorado, con responsabilidades de administración, modernización y explotación comercial de este destacado centro aeroportuario. En general, los aeropuertos cuentan con tres fuentes principales de ingresos: ingresos por vuelos comerciales, ingresos por vuelos de carga y las ventas de establecimientos dentro del aeropuerto.

Aunque históricamente las dos primeras fuentes han representado la mayoría de los ingresos, existe un creciente interés en potenciar el segmento comercial. Sobre todo, teniendo en cuenta que según un estudio de la consultora McKinsey que indica que para el año 2024 en la mayoría de las aerolíneas la deuda superará los ingresos lo que llevará a un incremento de los precios de tiquetes (Bower et al, 2022). Esto a su vez, representaría una disminución de los ingresos por parte de vuelos comerciales a nivel global. Actualmente, el aeropuerto El Dorado ha tenido limitadas estrategias para potenciar los ingresos provenientes de los más de 200 comercios situados en sus instalaciones. De ahí surge la necesidad de fusionar datos operativos con datos comerciales, para identificar oportunidades de mejora en esta área.

El propósito principal de este proyecto es descubrir y sugerir tácticas para aumentar las ventas de los establecimientos ubicados en el Aeropuerto Internacional El Dorado, basándonos en información básica de los vuelos. Utilizaremos técnicas de regresión para identificar patrones y comportamientos de los usuarios, lo que facilitará la toma de decisiones y la implementación de medidas más efectivas. Las métricas de negocio claves que servirán para evaluar nuestras propuestas serán:

- Ventas en tiendas cercanas a las puertas de embarque (1.5 horas antes del despegue)
- Número de transacciones en establecimientos cercanos a las puertas de embarque (1.5 horas antes del despegue).

2 IDEACIÓN

A continuación, se describe la propuesta de producto a desarrollar a partir del análisis de los procesos y problemáticas actuales y las necesidades y expectativas del cliente.

Usuarios potenciales del producto de datos

- **OPAIN S.A.:** como entidad administradora del Aeropuerto Internacional El Dorado, tiene interés en aumentar los ingresos comerciales y mejorar la experiencia de los pasajeros. Los resultados y recomendaciones generadas por el producto de datos serán fundamentales para ayudar en la toma de decisiones estratégicas y operativas.
- **Gerentes de tiendas y sus equipos de marketing:** tendrán acceso a la información de cómo la programación de vuelos impacta las ventas en sus tiendas.
- **Analistas de datos:** los profesionales encargados de analizar los resultados generados por el producto de datos serán parte importante del proceso. Sus observaciones y conclusiones contribuirán a la mejora continua de las estrategias operativas y comerciales.

Procesos actuales y problemáticas asociadas

En cuanto a procesos actualmente no se hace un uso de los datos operativos o comerciales para obtener un valor agregado. Como problemáticas se identificaron las siguientes:

- **Operaciones del aeropuerto:** asignación subóptima de vuelos, congestiones o subutilización de algunas áreas.
- **Operaciones de tiendas:** falta de correlación entre los horarios de mayor afluencia de pasajeros y las operaciones de las tiendas.

Propuesta de producto de datos

Se propone el desarrollo de un sistema que brinde al usuario información estimada del aumento de las ventas que pueden llegar a generar determinados vuelos de acuerdo con sus características tales como el destino, la hora de salida y la cantidad de pasajeros a bordo. Esto se traduce en una herramienta de ayuda para la toma de decisiones estratégicas para realizar la asignación de las puertas de abordaje. Este sistema se basará en un modelo de Machine Learning entrenado con datos históricos operacionales y comerciales.

Instrucciones de uso: El usuario deberá ingresar la información de los vuelos internacionales de salida, que incluye detalles como destino, fecha y hora de vuelo, número de pasajeros y tipo de aerolínea, entre otros. Una vez ingresada la información, el sistema proporcionará una estimación del valor de ventas esperado que podría generarse a partir de los pasajeros del vuelo en las tiendas del aeropuerto. Además, se mostrará la importancia de cada atributo en la predicción de dicho valor.

3 RESPONSIBLE

El presente proyecto plantea diversas implicaciones éticas, de privacidad, confidencialidad, transparencia y aspectos regulatorios que deben ser cuidadosamente considerados.
Es importante destacar que, en la fase inicial del proyecto, llevamos a cabo diversas sesiones de trabajo virtuales en las cuales definimos los aspectos fundamentales relacionados con la privacidad y la confidencialidad de la información. Estas sesiones representaron una etapa

preliminar, donde se acotó la firma de acuerdos de confidencialidad, y posteriormente, *OPAIN S.A* se centró en la gestión de entrega de la data operativa y comercial.
Adicionando, se mencionan principales consideraciones éticas y legales que se deben tener en cuenta a lo largo del desarrollo del proyecto:

- **Transparencia en los Resultados de IA:** En caso de que se decida incorporar modelos de IA en el prototipo que formará parte de la entrega final, es esencial proporcionar una comprensión clara de cómo operan estos modelos y de qué manera impactan en el proceso de toma de decisiones.
- **Confidencialidad de los Datos Comerciales:** Los datos comerciales son valiosos y sensibles. La empresa debe garantizar que estos datos se encuentren protegidos contra el acceso no autorizado y su uso indebido. También deben establecerse políticas claras de acceso a estos datos dentro de la organización.
- **Privacidad de los Pasajeros y Empleados:** *OPAIN S.A.* debe asegurarse de que la recopilación y el uso de datos no violen la privacidad de los pasajeros y empleados del aeropuerto. Esto implica la necesidad de obtener el consentimiento adecuado cuando sea necesario y garantizar la anonimización de los datos personales.
- **Política de Privacidad, Aviso de Privacidad y Consentimiento Informado:**

OPAIN S.A. debe desarrollar una política de privacidad que explique en detalle cómo se recopilan, almacenan y utilizan los datos personales de los pasajeros y empleados del aeropuerto. Además, debe proporcionar un aviso de privacidad fácilmente accesible para todas las partes interesadas que describa claramente los derechos de las personas con respecto a sus datos personales. El consentimiento informado de las personas antes de la recopilación de datos es esencial, especialmente si se utilizan para análisis y toma de decisiones. Este consentimiento debe ser específico, voluntario y permitir a las personas revocarlo en cualquier momento, garantizando así que estén informadas y tengan control sobre el uso de sus datos personales.

4 ENFOQUE ANALÍTICO

Teniendo en cuenta que el producto a desarrollar tiene como objetivo brindar al usuario información estimada de las ventas que puede generar un vuelo para la toma de decisiones, se tiene la siguiente hipótesis a desarrollar durante el semestre:

"A partir de la información categórica y temporal de un vuelo es posible estimar el valor monetario de las ventas que generaran los usuarios que viajen en este"

Para poder comprobar tal hipótesis, se plantea un modelo de regresión, en el cual por medio de características estáticas y categóricas del vuelo como lo son su destino, muelle de salida y aerolínea, así como características temporales cualitativas como la hora de salida se pueda estimar el valor de ventas que generarían sus pasajeros en un momento determinado. Con tal fin se usa una estrategia compuesta de los siguientes pasos, los cuales suceden luego del proceso de recolección de datos:

1. **Análisis univariado:** Análisis individual de las variables o columnas seleccionadas, que se segmenta en los siguientes pasos:
    1.1. **Análisis de variables cualitativas:** Evaluación de la cardinalidad (número de categorías) y distribución de las categorías (análisis de Pareto y diagrama de

barras para hallar valores más frecuentes).

1.2. **Análisis de variables cuantitativas:** Generación de estadísticos de tendencia central (media, mediana), medidas de dispersión (desviación estándar), evaluación de percentiles 5%, 25%, 50% (mediana) 75% y 95% para establecer la centralidad de la distribución, valor mínimo y máximo para establecer si hay datos atípicos, y gráfica de histograma para establecer el sesgo de la distribución y si existen múltiples modas.

2. **Análisis bivariado**: Usando como variable de comparación el valor de ventas estimado para el vuelo se hace un análisis de poder de predicción de cada una de las otras variables contra esta.

   2.1. **Análisis bivariado de variables cualitativas**: Diagramas de cajas donde el eje X es la variable categórica y el eje Y es el valor de ventas estimado. Adicionalmente se aplica el test de Kruskal-Wallis para determinar si los grupos son diferentes respecto al valor de ventas.

   2.2. **Análisis bivariado de variables cuantitativas**: Diagramas de dispersión X es la variable cuantitativa y el eje Y es el valor de ventas estimado. Si esta relación parece ser lineal se aplica la correlación de Pearson y en caso contrario la de Spearman.

   2.3. **Análisis temporal:** Para unas variables categóricas se hace una gráfica de heatmap, donde el eje X es una variable de tiempo (hora o día del año) y el eje Y tiene los valores de la variable categórica para observar si para cada categoría el valor de ventas se acumula en algún periodo de tiempo.

Después de tener estos resultados se plantea como modelo preliminar un modelo de RandomForestRegressor, el cual consiste en un ensamble de árboles que segmentan la información numérica y categórica hasta llegar a un valor estimado de ventas.

5 RECOLECCIÓN DE DATOS

A medida que se avanzaba en las sesiones de trabajo con *OPAIN S.A.* durante las últimas semanas, se ha llevado a cabo un proceso de análisis que ha culminado en la identificación y consenso en torno a los datos de mayor relevancia. Dentro de este proceso, se ha determinado que los datos relacionados con los vuelos de salida son de particular interés y poseen un valor estratégico significativo. Esto se debe a que, durante el período previo a la salida de un vuelo, los pasajeros suelen pasar una considerable cantidad de tiempo en las áreas de espera del aeropuerto. Durante este intervalo, es más probable que los viajeros se involucren en actividades como compras, lo que incluye la adquisición de souvenirs, productos de moda, accesorios, y la experiencia gastronómica en los restaurantes y locales de comida disponibles en el aeropuerto.

Teniendo en cuenta los anteriormente mencionado, se puede comentar que el enfoque en los vuelos de salida como punto focal de estudio se fundamenta en la idea de que, al comprender y adaptarse a las necesidades y preferencias de los pasajeros en este momento crítico, se puede mejorar la satisfacción del cliente, aumentar los ingresos generados en el aeropuerto y fomentar una experiencia de viaje más placentera en general. A través de la recopilación y análisis inteligente de estos datos, estaremos en posición de tomar decisiones informadas que

beneficiarán tanto a los pasajeros como a las partes interesadas involucradas en la operación del aeropuerto.

Descripción

El Dataset proporcionado de vuelos de salida comprende un período de tres años, desde 2020 hasta 2023, y consta de 344,146 filas y 31 columnas. Entre los atributos más destacados se encuentra información relevante como destino, tipo de vuelo, muelle, fecha, tipo de aerolínea y el número de pasajeros que salen. Estos datos son fundamentales para analizar patrones de vuelo, eficiencia operativa y preferencias de los pasajeros en el aeropuerto durante ese período. El desarrollo técnico de este punto se presenta en la sección **5. Recolección de datos** del notebook adjunto: **ENTREGA_UNO_PUNTOS_5_Y_6.**

6 ENTENDIMIENTO DE LOS DATOS

El reporte del análisis exploratorio de datos y calidad de los datos se presenta en la sección **6. Entendimiento de los datos** del notebook adjunto: **ENTREGA_UNO_PUNTOS_5_Y_6.**

7 CONCLUSIONES E INSIGHTS

- Este proyecto subraya el valor de los datos operativos y comerciales como activos estratégicos. Utilizar análisis de datos avanzados y técnicas de ciencia de datos puede proporcionar a OPAIN S.A. información vital para tomar decisiones más informadas y eficientes.
- Del análisis exploratorio inicial podemos observar que variables como la hora del día, la cantidad de pasajeros en el muelle y el tipo de aerolínea tienen un impacto significativo sobre las transacciones y las ventas. En los próximos avances esperaríamos evidenciar que estas variables tengan un impacto mayor en comparación al resto de los features observados. Adicional a esto, el año 2020 al ser impactado por la pandemia del Covid-19 tuvo unas ventas extremadamente irregulares. Por esto, lo mejor será no tomar datos de este año. Además, al enfocarse en los vuelos de salida, se busca mejorar la experiencia del pasajero. Durante este período previo a la salida, los pasajeros son más propensos a involucrarse en actividades como compras, lo que brinda oportunidades significativas para aumentar las ventas comerciales.

- **Optimización de Asignación de Puertas:** La asignación estratégica de puertas de embarque, considerando la coordinación con los horarios de mayor afluencia de pasajeros, puede prevenir congestiones y mejorar la experiencia del cliente. Esto probablemente aumentaría las ventas en las tiendas cercanas a estas puertas.

- **Sincronización de Horarios:** Sincronizar las operaciones de las tiendas con los horarios de mayor afluencia de pasajeros puede aumentar significativamente las oportunidades de venta y mejorar la eficiencia comercial. Esto podría lograrse mediante la programación de promociones y eventos especiales en momentos estratégicos.

8 PREPARACIÓN DE DATOS

Para la preparación de los datos, se llevaron a cabo una serie de actividades para garantizar que estos estén en un formato adecuado y listos para su procesamiento y análisis. A continuación, se presenta un diagrama de bloques que ilustra el proceso completo que se realizó hasta la creación y prueba de un primero modelo de regresión. Las actividades realizadas a la preparación de datos y que serán detalladas en esta sección van de la actividad 1. **Obtención de fuentes de datos** hasta la actividad **5. Transformación de datos.** El diagrama de bloques se encuentra como anexo:

- ***Anexo 2. Figura 1. Preparación de datos, generación y prueba del modelo base.***

1. Obtención de fuentes de datos: El proceso se inició con la adquisición de los datos. Para el proyecto, se dispuso de dos fuentes de datos: datos operativos que contenían información sobre los vuelos y datos comerciales que registran las ventas por marca. Este último conjunto de datos abarcaba información de ventas de 10 marcas distintas.

2. Análisis de dimensiones de calidad y limpieza de datos: Se llevó a cabo un análisis exhaustivo de la calidad de los datos y se realizó una limpieza minuciosa. Este proceso se ejecutó de forma independiente para las dos áreas de datos. Como resultado de estas actividades, se obtuvieron los conjuntos de datos procesados. A continuación, se detallan los pasos específicos llevados a cabo en esta etapa. Los diagramas de bloques se encuentran como anexos:

- ***Anexo 3. Figura 2. Análisis de dimensiones de calidad y limpieza de datos operativos.***
- ***Anexo 4. Figura 3. Análisis de dimensiones de calidad y limpieza de datos comerciales.***

3. Integración de datos comerciales y operativos: esta integración de datos comerciales y operativos fue un componente crucial en el proyecto, ya que se buscaba unir información operativa y transaccional para desarrollar un modelo de regresión efectivo. La integración se llevó a cabo a través del atributo "**MUELLE**", que permitió vincular los conjuntos de datos operativos y comerciales. Este proceso de unificación permitió obtener un único conjunto de datos que contiene la información necesaria para el modelo propuesto.

Un aspecto importante a considerar fue el intervalo de tiempo de dos horas antes de la salida del vuelo. Se identificó que este lapso era el momento clave en el cual los pasajeros potenciales realizaban compras en los establecimientos comerciales del aeropuerto. Por lo tanto, se enfocó en recopilar y combinar los datos relevantes dentro de este intervalo de tiempo para capturar con precisión el comportamiento de compra de los pasajeros y su relación con la información operativa del vuelo.

La integración de estos conjuntos de datos permitió obtener una visión más completa y detallada de los patrones y tendencias de compra de los pasajeros en relación con los vuelos, lo que resultó en un conjunto de datos más enriquecido y listo para su análisis y modelado posterior. El proceso completo de las actividades 1, 2 y 3 se encuentra detallado en los notebooks adjuntos:

- 1_Limpieza_datos_comerciales,
- 2_Limpieza_datos_operativos,
- 3_Union_datos_operativos_y_comerciales_por_muelle
- 4_Union_de_datos_operativos_y_comerciales_por_marca

Estos notebooks proporcionan una descripción exhaustiva de las diferentes etapas de limpieza de datos, incluidas las estrategias específicas utilizadas para abordar los problemas de calidad de datos y la estrategia de fusión de los conjuntos de datos operativos y comerciales.

4. Análisis y visualización exploratoria de datos: Se realizaron análisis exhaustivos y visualizaciones detalladas de la variable objetivo (**valor_venta**) y las relaciones entre distintas características para comprender la naturaleza de los datos y detectar posibles tendencias.

5. Transformación de datos: Se agregaron nuevas características, como información geográfica y socioeconómica, con el fin de enriquecer los conjuntos de datos y mejorar la capacidad predictiva del modelo. Estas variables son: **País Destino**, **Continente Destino PIB** y **HDI** (Índice de Desarrollo Humano) estas dos últimas son indicadores por país.

El proceso completo de las actividades 4, y 5 se encuentra detallado en el notebook adjunto:

- 5_Baselined_de_modelo_de_regresion.

9 ESTRATEGIA DE VALIDACIÓN Y SELECCIÓN DE MODELO

Para esta etapa iniciaremos con un modelo de regresión lineal como modelo base, ya que esto nos brindara una comprensión inicial de cómo las variables independientes influyen en la variable dependiente y qué tan bien el modelo puede adaptarse a los datos.

El proceso de entrenamiento del modelo de regresión lineal se realizará únicamente utilizando el conjunto de datos de entrenamiento (80% de los datos), lo que permitirá al modelo aprender los patrones y relaciones presentes en los datos de entrenamiento. Posteriormente, el modelo se evaluará utilizando el conjunto de prueba (20% de los datos). Los resultados obtenidos a partir de esta evaluación proporcionarán información valiosa sobre la idoneidad del modelo de regresión lineal en la resolución del problema. Los resultados los utilizaremos como referencia para comparar y evaluar la eficacia de otros modelos más complejos.

Después de evaluar el modelo base, nuestro siguiente paso implica el entrenamiento de modelos más sofisticados, utilizando un conjunto de validación para seleccionar los parámetros óptimos. Este enfoque nos permitirá identificar tanto el mejor modelo en términos de resultados de predicción como los atributos más y menos influyentes en la predicción del modelo. Utilizaremos técnicas más avanzadas, como la regularización y el ajuste de hiperparametros, para mejorar la precisión y la generalización del modelo.

Finalmente se analizará cómo se conserva la distribución de los subconjuntos de datos en relación con el conjunto original. Esto implicará un análisis de las estadísticas descriptivas de

cada subconjunto, incluyendo medidas como la media, la mediana, la desviación estándar y los percentiles, buscando identificar la similitud en las distribuciones de las características clave entre el conjunto original y los subconjuntos derivados.

10 CONSTRUCCIÓN DEL MODELO

El proceso de construcción de modelo parte de que los supuestos de regresión se cumplen en el modelo baseline. Por ende, se genera una serie de scripts que hacen provecho de la validez de estos supuestos para explorar la mejor configuración en un proceso de dos etapas: Exploración de configuración de transformaciones y exploración de configuración de modelo.

En la primera etapa se considera la categorización de variables hallada para el modelo baseline (numéricas, categóricas de alta cardinalidad y categóricas de baja cardinalidad) para aplicar transformaciones numéricas (Scalers y Power transformers) y de categoría a valor numérico (One Hot Encoder y Feature Hashing). Esto se hace para cada dataset de marca por separado y todas en una sola, lo que se conoce como “muelle”, para llegar a un total de 40 combinaciones de transformaciones por cada dataset, donde también se evalúa el efecto de incluir variables exógenas, como **HDI**.

Para cada configuración de dataset se entrenan tres flujos de Validación Cruzada con Random Search, teniendo cada flujo un modelo sofisticado de los anteriores como estimador. Este proceso se puede revisar en los scripts de configuración (**config.py**, **config_extended.py** y **exploration.py**), así como el notebook que los invoca**, 7_Exploracion_de_hparams.ipynb**. Como resultado se tiene el valor de las métricas de evaluación por cada configuración para cada dataset, desagregadas en train y test, que se almacenan en las carpetas **results** y **extended_data_results**.

Partiendo de todas las configuraciones se usan las métricas de evaluación seleccionadas, R2 y MAE, para seleccionar las mejores configuraciones de transformación y modelo, se eliminan variables de alta colinealidad de nuevo y se seleccionan las mejores 20 features para cada marca y todas en una, con motivo de poder presentar mayor explicabilidad del modelo en su despliegue productivo.

11 EVALUACIÓN DEL MODELO

Se llevó a cabo una búsqueda de hiperparámetros en los tres modelos examinados para este proyecto. En total, se evaluaron 540 modelos, 180 para cada modelo: LGBM, Regresión Lineal y Random Forest. Se exploraron diversos hiperparámetros, incluyendo métodos de normalización para variables categóricas (comparando One Hot con Feature Hasher) y técnicas de normalización para variables numéricas, como Min-Max Scaler y Standard Scaler, además de otros parámetros relevantes para cada modelo. A continuación, se detallan las métricas de evaluación de estos modelos.

Tabla 1- Resultados Modelos

Modelos	TEST		TRAIN	
	R2	MAE	R2	MAE
Regresión Lineal Modelo Base	0.16	188.8M	0.16	188.7M
Light GBM	**0.90**	**58.37M**	**0.99**	**18.9M**
Regresión Lineal	0.33	162.7M	0.34	163.4M
Random Forest	0.78	81.28M	0.98	17.5M

De los tres modelos evaluados, Light GBM destaca con un R2 de 0.90 en pruebas y 0.99 en entrenamiento, aunque muestra signos de sobreajuste al tener una diferencia notoria entre estas métricas. Por otro lado, la Regresión Lineal, con un R2 de solo 0.33 en pruebas y 0.34 en entrenamiento, indica un ajuste pobre al conjunto de datos, lo que sugiere que la relación entre las variables no es estrictamente lineal. Random Forest logra un R2 de 0.78 en pruebas y 0.98 en entrenamiento, mostrando también señales de sobreajuste. Para perfeccionar estos modelos, es crucial considerar técnicas como la regularización, feature engineering y explorar otros modelos de aprendizaje automático. Sin embargo, es importante recalcar que es una muy buena primera aproximación en especial el prometedor modelo encontrado con el LGBM.

12 CONCLUSIONES

- **¿Cuáles son las mayores dificultades que se han tenido en el proyecto?**

Durante el desarrollo del proyecto, nos enfrentamos a varios desafíos importantes, especialmente la gestión de una gran cantidad de variables categóricas en los datos proporcionados, como el destino, tipo de vuelo, sala, aerolínea, tipo de aerolínea y fecha del día. Además, nos encontramos con otros obstáculos significativos. En el aeropuerto existen tiendas con 200 marcas, pero actualmente solo estamos utilizando 10 de ellas debido a que solo para estas 10 existe información comercial. Lamentablemente, estas marcas carecen de información sobre la puerta de embarque cercana a las tiendas, lo que nos llevó a tomar la decisión de utilizar el campo de muelle como alternativa. Sin embargo, la información disponible se limita a un solo muelle, concretamente el "A internacional oriente" con vuelos de salida que es donde se encuentran las 10 marcas para las cuales se cuenta con información comercial.

Durante la limpieza y el análisis exploratorio de los datos, observamos que 6 de las 10 marcas analizadas presentaban valores negativos, lo que resalta la importancia de contar con datos de ventas brutas para evitar inconsistencias. También notamos que algunas marcas carecen de datos completos. Por ejemplo, la marca 4 solo tiene registros hasta fines de 2022, mientras que la marca 9 es la única que cuenta con datos hasta septiembre de 2023. En términos de gestión de datos o valores nulos, no encontramos problemas significativos en este aspecto.

- **¿Qué estrategias se plantean para mitigarlas?**

Considerando los desafíos mencionados anteriormente, se plantearon diversas estrategias para mitigarlos. Inicialmente, se desarrolló un modelo Baseline simple, que incorporaba variables numéricas relacionadas con el tiempo, como la hora del día, la semana, el mes, el día y el día del año, junto con variables categóricas que representaban las características estáticas de los vuelos. Para manejar las variables categóricas de baja cardinalidad, se generaron columnas binarias, mientras que para aquellas de alta cardinalidad se aplicaron transformaciones utilizando el método hash. Dado que las variables numéricas eran limitadas, solo fue posible aplicar transformaciones a algunas de ellas, especialmente aquellas disponibles antes de la generación de la variable de salida "valor de venta". Asimismo, se implementó el proceso de Clipping para estas variables numéricas, estableciendo límites de valores sin considerar los valores atípicos.

Además, se identificó el problema de la heterocedasticidad durante el análisis. Para abordar este desafío, se optó por emplear modelos de ensamble, que no requerían cálculos computacionalmente intensivos que consumieran una gran cantidad de memoria, en contraste con otros métodos como la regresión por Kernel o modelos más complejos que requerirían una búsqueda exhaustiva de hiperparámetros, como las redes neuronales profundas. A lo largo del proceso, se aplicaron varias transformaciones a los datos, incluyendo el escalado estándar y el escalado robusto, mediante enfoques tradicionales. Por último, se utilizó un Framework de exploración de hiperparámetros para ajustar el modelo y mejorar las métricas de rendimiento.

- **¿Qué condiciones considera que deberían tener los datos para obtener mejores resultados? Más datos, diferentes características, menor sesgo, etc.**

Para mejorar los resultados del modelo desarrollado, es esencial llevar a cabo una comprensión exhaustiva de los datos y garantizar su calidad. Esto implica la necesidad de disponer de una cantidad adecuada de datos que sean representativos y variados, al mismo tiempo que se minimiza cualquier sesgo existente. Además, sería altamente beneficioso contar con un mayor número de variables numéricas, dado que la mayoría de las disponibles eran de naturaleza categórica. No obstante, es importante destacar que, debido a las limitaciones discutidas con el grupo de OPAIN, no fue factible obtener más datos comerciales ni utilizar muelles distintos al "A internacional oriente". A pesar de estas restricciones, comprender a fondo y garantizar la calidad de los datos disponibles sigue siendo fundamental para simplificar la fase de modelado y lograr resultados analíticos más sólidos.

- **¿El mejor modelo obtenido hasta el momento es suficiente para dar solución al problema u oportunidad de negocio abordado?**

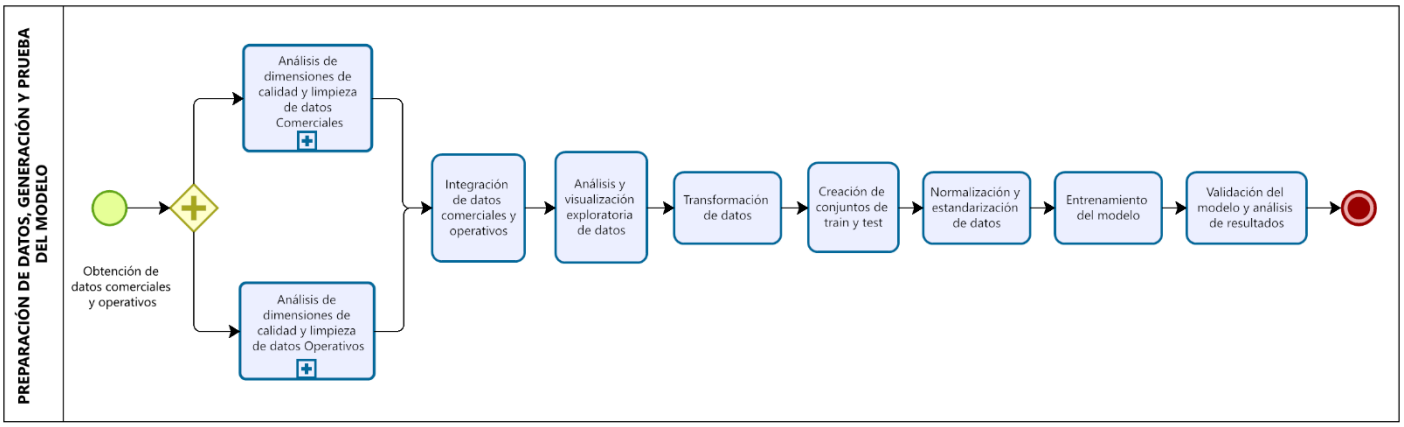
Basándonos en las métricas validadas hasta la fecha, el modelo Light GBM se presenta como una solución sólida para abordar el problema u oportunidad de negocio planteada con OPAIN, gracias a su capacidad para explicar la variabilidad de los datos y realizar predicciones precisas en las ventas. No obstante, es crucial tener en cuenta otros factores, como la interpretabilidad del modelo y el costo computacional, antes de tomar una decisión definitiva.

13 ANEXOS

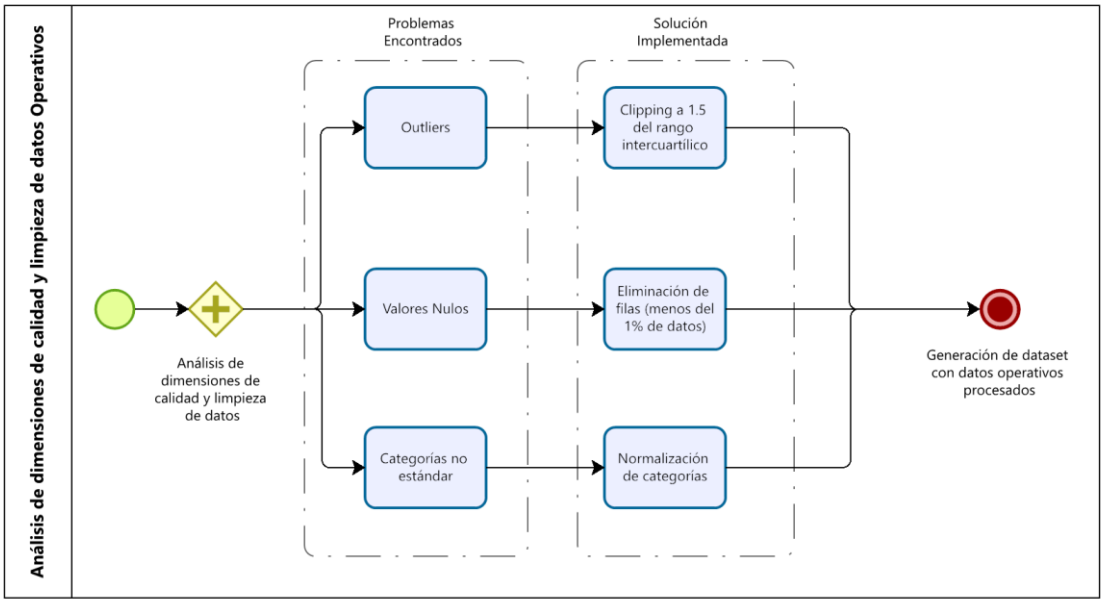
Anexo 1. Mockup conceptual de la herramienta a desarrollar (*Es una representación conceptual y no refleja la interfaz final del producto)*

https://www.canva.com/design/DAFuoRRkrc4/XBKCnspP7GvuYoYQk9oa_w/view?utm_content=DAFuoRRkrc4&utm_campaign=designshare&utm_medium=link&utm_source=publishsharelink#1

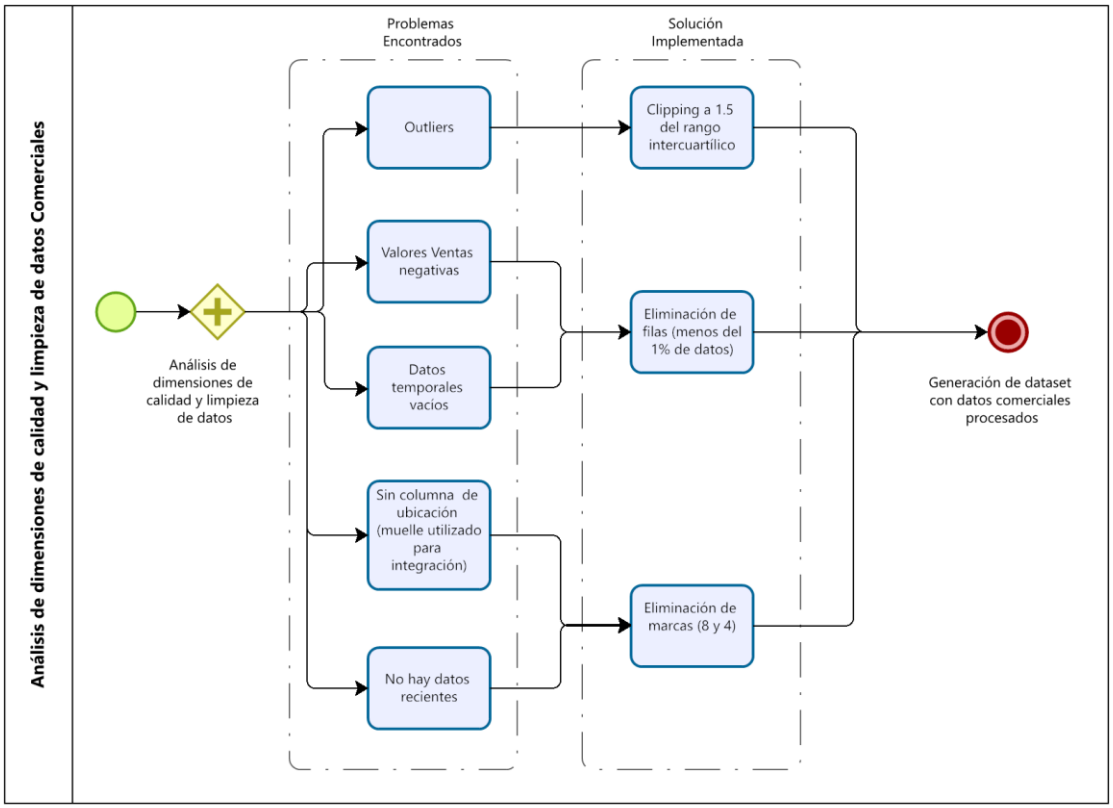
Anexo 2. Figura 1. Preparación de datos, generación y prueba del modelo base.



Anexo 3. Figura 2. Análisis de dimensiones de calidad y limpieza de datos operativos.



Anexo 4. **Figura 3. Análisis de dimensiones de calidad y limpieza de datos comerciales.**



Anexo 5. Distribución de los subconjuntos de datos respecto al conjunto original.

	Pasajeros saliendo	Año	Día	Hora entera	Hdi2021	Pasajeros por Muelle y Hora	valor_venta
count	57477.000000	57477.000000	57477.000000	57477.000000	57036.000000	57477.000000	5.747700e+04
mean	157.874941	2022.428467	15.710632	12.320563	0.826511	202.982202	4.180949e+08
std	46.973945	0.494861	8.814630	6.653599	0.079482	111.057170	2.573630e+08
min	0.000000	2022.000000	1.000000	0.000000	0.627000	0.000000	0.000000e+00
25%	126.000000	2022.000000	8.000000	7.000000	0.758000	133.000000	2.486026e+08
50%	154.000000	2022.000000	16.000000	13.000000	0.805000	165.000000	4.072462e+08
75%	179.000000	2023.000000	23.000000	17.000000	0.921000	248.000000	5.904153e+08
max	261.500000	2023.000000	31.000000	23.000000	0.942000	909.000000	1.783547e+09

Tabla 2. Estadísticas descriptivas del conjunto de datos original

	Pasajeros saliendo	Año	Día	Hora entera	Hdi2021	Pasajeros por Muelle y Hora	valor_venta
count	45981.000000	45981.000000	45981.000000	45981.000000	45644.000000	45981.000000	4.598100e+04
mean	157.837824	2022.427655	15.713229	12.315261	0.826473	202.887464	4.186520e+08
std	47.036316	0.494744	8.825197	6.642735	0.079574	111.038802	2.570508e+08
min	0.000000	2022.000000	1.000000	0.000000	0.627000	0.000000	0.000000e+00
25%	126.000000	2022.000000	8.000000	7.000000	0.758000	133.000000	2.496512e+08
50%	154.000000	2022.000000	16.000000	13.000000	0.805000	165.000000	4.076016e+08
75%	179.000000	2023.000000	23.000000	17.000000	0.921000	248.000000	5.910822e+08
max	261.500000	2023.000000	31.000000	23.000000	0.942000	909.000000	1.783547e+09

Tabla 3. Estadísticas descriptivas del conjunto de datos de entrenamiento

	Pasajeros saliendo	Año	Día	Hora entera	Hdi2021	Pasajeros por Muelle y Hora	valor_venta
count	11496.000000	11496.000000	11496.000000	11496.000000	11392.000000	11496.000000	1.149600e+04
mean	158.023399	2022.431715	15.700244	12.341771	0.826665	203.361126	4.158667e+08
std	46.725387	0.495337	8.772615	6.697127	0.079116	111.134629	2.586074e+08
min	0.000000	2022.000000	1.000000	0.000000	0.627000	0.000000	0.000000e+00
25%	126.000000	2022.000000	8.000000	7.000000	0.758000	133.000000	2.441407e+08
50%	154.000000	2022.000000	16.000000	13.000000	0.805000	165.000000	4.046871e+08
75%	180.000000	2023.000000	23.000000	18.000000	0.921000	249.000000	5.885711e+08
max	261.500000	2023.000000	31.000000	23.000000	0.942000	909.000000	1.695835e+09

Tabla 4. Estadísticas descriptivas del conjunto de datos de prueba

14 BIBLIOGRAFÍA

- Bower, et al. (2022). Back to the future? Airline sector poised for change post-COVID-19. McKinsey & Company
- Krishnan, K. (2013). Data Warehousing in the Age of Big Data. Newnes.
- Pinto, J. K. (2015). Gerencia de proyectos (Tercera edición). Pearson.
- Nicholas, J. M. (2017). Project Management for Engineering, Business and Technology (5a Ed.). Partes 1, 2 y 3.
- Project Management Institute. (2013). A Guide to the Project Management Body of Knowledge (PMBOK® Guide) (5a Ed.). PMI.org. (6a ed. a partir de septiembre de 2017).