Autor(es):

- Daniela Chacón – 201910858
- Esteban Ortiz - 201913613
- Manuel Porras - 201913911
- Angie Rincón - 201114323

Fecha: Diciembre 3 de 2023

PROYECTO FINAL – TERCERA ENTREGA

- **OBJETIVOS**
  - Finalizar las etapas de modelado y evaluación, teniendo en cuenta la retroalimentación brindada durante la sustentación de la segunda entrega.
  - Construir el producto de datos con los diferentes componentes establecidos durante la actividad de ideación.
  - Presentar los resultados del análisis y el producto de datos a los stakeholders de la organización y obtener retroalimentación respecto a los aspectos positivos logrados y elementos a mejorar.
- **ACTIVIDADES DEL SPRINT Y ENTREGABLES:**

1 [35%] CONSTRUCCIÓN DEL PRODUCTO DE DATOS:

1.1 Generación datos finales para la creación de modelo:

```
df_final.shape

(114279, 302)
```

Figura 1. Tamaño DataFrame final información unificada.

Este es el tamaño final de nuestro dataframe unificado el cual contiene la información del clima, vuelos, marcas, tiendas, categorías, ventas, etc.

(El proceso más al detalle se puede encontrar en el notebook Proyecto_OPAINV_Alistamiento_Datos_Entrega3.ipynb en la carpeta de Notebooks entrega final de nuestro repositorio)

1.2 Se escoge el mejor modelo con los mejores resultados obtenidos:

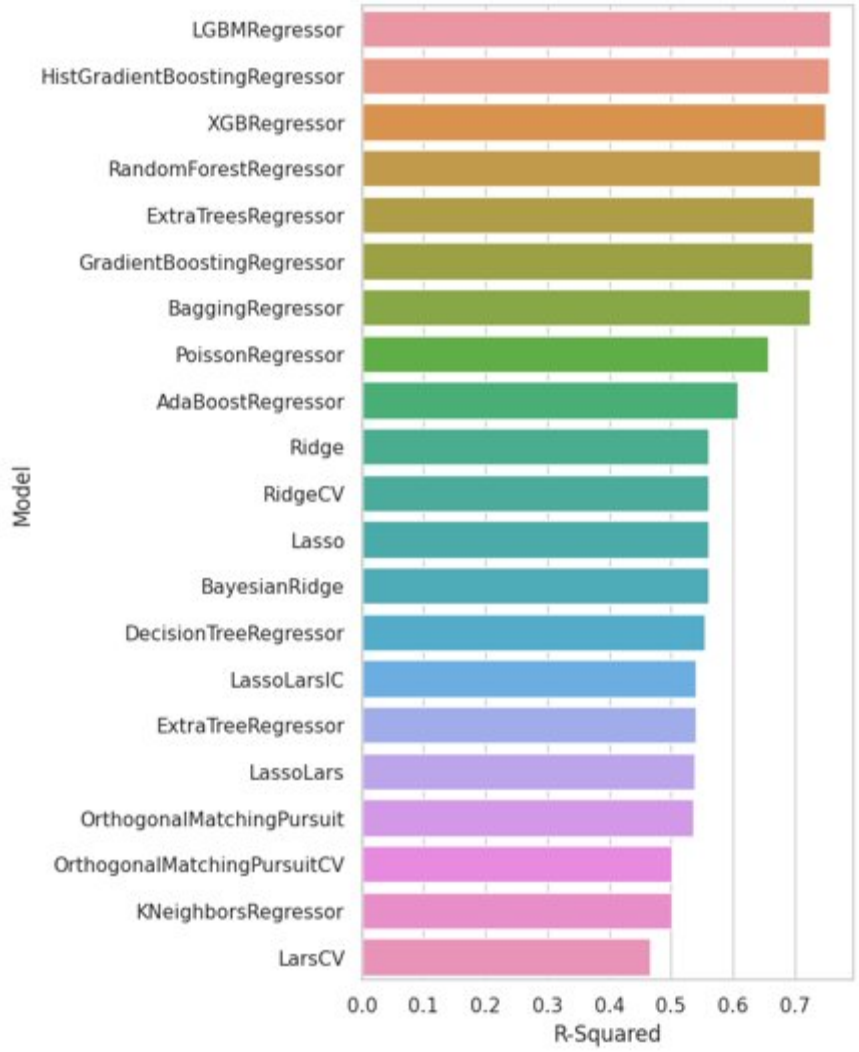

Figura 2. Pruebas diferentes modelos de regresión.

A partir de estos resultados se escoge el modelo LGBMRegressor para realizarle la respectiva búsqueda de hiperparámetros mostrados en la Figura 3.

```
param_grid = {
    'num_leaves': [31, 50, 70],
    'reg_alpha': [0.1, 0.5],
    'min_data_in_leaf': [30, 50, 100],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [-1, 5, 10]
}

lgbm = lgb.LGBMRegressor()

grid_search = GridSearchCV(estimator=lgbm, param_grid=param_grid, cv=5, scoring='r2', verbose=1, n_jobs=-1)

grid_search.fit(X_train, y_train)
```

Figura 3. Búsqueda de hiperparámetros.

Finalmente, mediante tres subcojuntos de datos se obtuvieron los resultados mostrados en la figura 4.

Modelo	MAE Train	MAE Test	RMSE Train	RMSE Test	R2 Train	R2 Test
Todas las variables	**7808426.93**	**11216316.69**	**15335847.97**	**20567723.23**	**0.84**	**0.75**
Sin variables de tienda, categorías no especificadas ni marca	8896343.74	12128998.44	17769446.63	22354600.02	0.79	0.7
Sin variables de tienda, categorías no especificadas, marcas y aerolíneas	8907448.70	12131249.39	17786849.75	22361158.30	0.79	0.7

Figura 4. Métricas resultados pruebas.

Por lo anterior, se decidió escoger el modelo con todas las variables posibles, ya que tiene el mejor comportamiento sobre los datos de prueba.

1.3 Generación de predicciones sobre los datos de prueba:

En este apartado, describimos que a partir del archivo .pkl que se llama *modelo_completo_Entrega3.joblib* el cuál se encuentra en la carpeta llamada *Modelos última entrega* se generaron las predicciones de las ventas del año 2023 y se contrastaron estas predicciones con las ventas reales mediante un tablero de control.

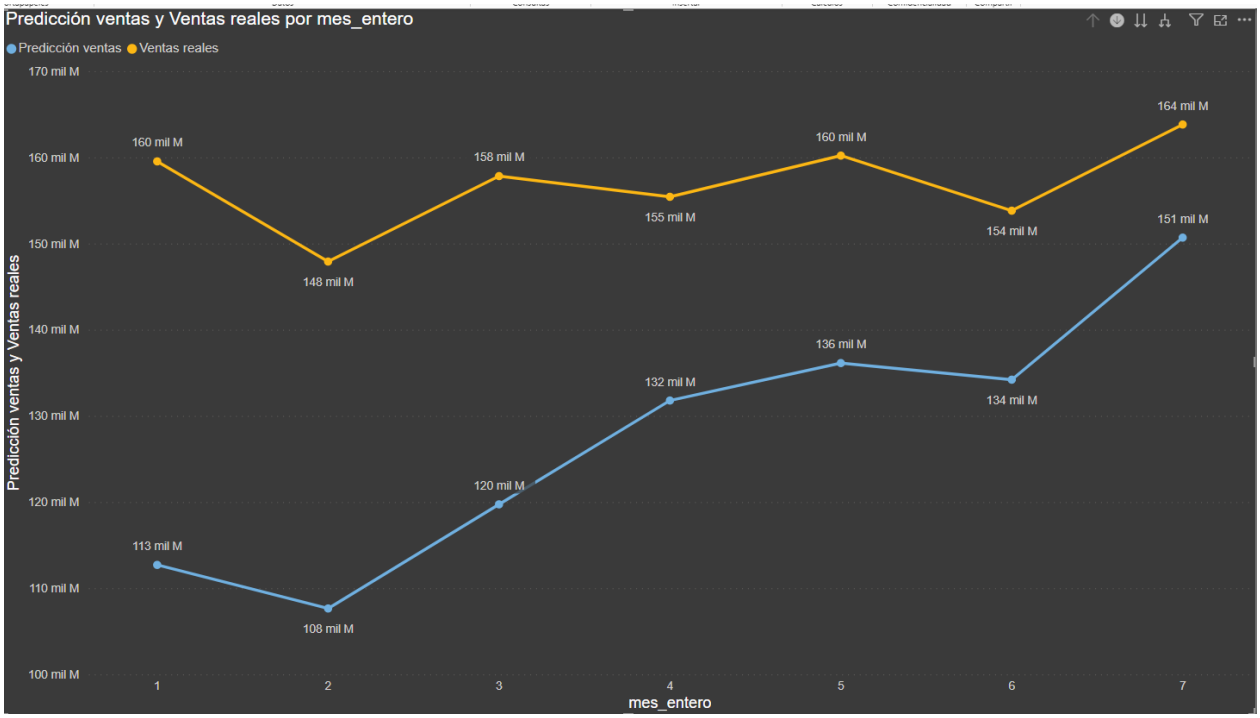

Figura 5. Tablero de control de realidad vs predicción. (Por mes)

Además, este gráfico permite ver la granularidad más baja que sería la de horas como lo muestra la figura 6.

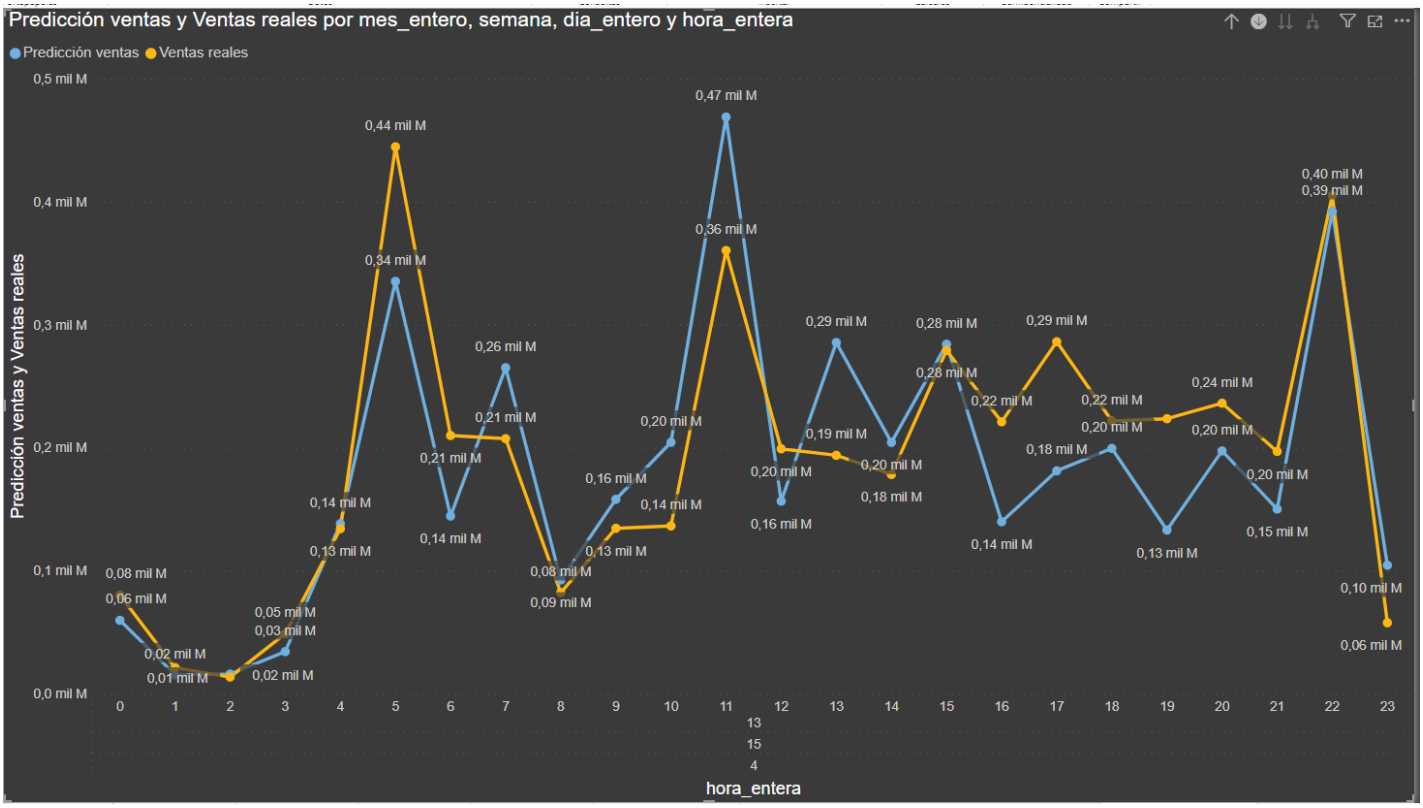

Figura 6. Tablero de control realidad vs predicción. (Por hora entera).

1.4 Creación de tableros de control:

Se diseñó un tablero de control unificado en Power BI como un producto de datos, que permite visualizar los comportamientos de las ventas en diversos aspectos, como las unidades vendidas de ciertas categorías y las ventas promedio según la agrupación de temperaturas, entre otros.

Link al archivo compartido en la nube:

https://app.powerbi.com/reportEmbed?reportId=ac2e0b60-2eab-4269-81ee-6416d55ecb99&autoAuth=true&ctid=fabd047c-ff48-492a-8bbb-8f98b9fb9cca

Archivo original del tablero de control (.pbix):

https://drive.google.com/file/d/1kiPRSB6vfhazjDihq1ZgPlCevhNnexV1/view?usp=sharing

Nota: Es de aclarar que para poder ver el informe en línea se tiene que pedir permiso primero, ya que, las políticas de la DSIT (Departamento de la seguridad y la información) no permite generar un link para que cualquier usuario lo pueda ver, si no que toca compartirle directamente a cada usuario el power BI. Por otro lado, se dejó enlazado también el archivo .pbix para que se pueda modificar, se pueda utilizar para mayor facilidad si es el caso o se pueda subir al power bi de su respectiva organización u espacio de trabajo propio.

2 [20%] RETROALIMENTACIÓN POR PARTE DE LA ORGANIZACIÓN:

2.1 Cronograma establecido por el cliente:

El cliente en este caso especifico OPAIN y CAOBA, nos entregaron y estuvimos de acuerdo con el siguiente cronograma con fecha de actualización al 15 de octubre del año 2023:

A continuación, se presentan las actividades que se realizaron hasta la fecha indicada:

Fecha	Actividad
11/08/2023	Primera reunión Caoba, profesor de la universidad los Andes y Opain para revisar alcance del proyecto
25/08/2023	Primer acercamiento entre Opain y grupo 1 de los estudiantes de la universidad de los Andes
29/08/2023	Formalización de aceptación por parte de Caoba
01/09/2023	Primer acercamiento entre Opain y grupo 2 de los estudiantes de la universidad de los Andes
17/09/2023	Los estudiantes hicieron la primera entrega que tenían en el cronograma del curso al profesor
29/09/2023	Presentación a Opain de la primera entrega
06/10/2023	Opain hace entrega a estudiantes del cronograma propuesto por la Dirección Proyectos Aeroportuarios e Innovación Tecnológica

Programa de actividades

Se presentan las entregas que deben realizar los estudiantes a Opain de acuerdo con las entregas que deben hacer a la universidad de los Andes:

Fecha de Entrega	Meta	Entregable
20/10/2023	Comprensión de datos y elección de modelo.	• Verificar significancia del modelo seleccionado. • Presentación de hallazgos o tendencias encontradas en el análisis exploratorio de datos. • Código abierto comentado en formato ipynb. • Sustentar que modelo se usó y ¿por qué?

24/11/2023	Prototipo inicial del proyecto	• Construcción del prototipo inicial donde se detallen los procesos de desarrollo (así sean de overview o experimentales) y tengan un grado de significancia optimo, así como una relevancia constante. • Presentación de hallazgos y análisis encontrados del desarrollo del prototipo inicial. • Código abierto comentado en formato ipynb.
15/12/2023	Modelo finalizado	• El modelo debe ser capaz de sugerir tácticas para aumentar las ventas de los establecimientos ubicados en el Aeropuerto Internacional El Dorado, basándose en información histórica comercial y de operaciones. • Presentación final a alto nivel donde se detalle el funcionamiento del modelo. • Código abierto con comentado en formato ipynb.

2.2 Observaciones generadas a partir de cada entrega:

A continuación, se presenta un resumen de los comentarios proporcionados por el cliente durante las reuniones realizadas casi semanalmente desde el inicio del semestre. Es importante señalar que no contamos con observaciones específicas escritas por ellos, ya que estas fueron recomendaciones expresadas durante las reuniones, las cuales también han sido grabadas en nuestro espacio en Microsoft Teams.

Entrega	Observaciones hechas por el cliente:
Entrega 1 Universidad	En esta primera entrega, se presentaron al cliente las primeras visualizaciones de las construcciones de varios modelos, incluyendo un total de 43 modelos de regresiones. El objetivo era determinar cuál de ellos se ajustaba mejor a los datos. Se proporcionó una visión general de todos los datos disponibles, abarcando todas las marcas disponibles. A raíz de esta presentación, el cliente nos aconsejó evitar posibles problemas derivados de la creación de tantas columnas dummy en los modelos. Además, se enfatizó la falta de información con respecto a una marca que denomina sus categorías como "categoría 1", "categoría 2", etc.

Entrega 2 Universidad	En esta segunda entrega, se presentó al cliente un modelo de regresión seleccionado a partir de la búsqueda de hiperparámetros realizada en tres modelos distintos. El HistGradientBoostingRegressor fue identificado como el mejor modelo, demostrando un rendimiento equitativo tanto en los datos de entrenamiento como en los de prueba. Además, se llevaron a cabo dos pruebas: una considerando solo las categorías con información de las marcas que tenían información de las categorías y otra exclusivamente con información de ventas. Es importante destacar que en esta entrega se incorporó información climática con la misma granularidad horaria que las ventas. A raíz de esta adición, el cliente expresó su agrado, ya que la inclusión de datos climáticos proporciona mayor explicabilidad al modelo. Esto resulta especialmente relevante para obtener conclusiones sobre las ventas, dado que la principal importancia de este modelo no radica tanto en su capacidad predictiva, sino en su habilidad para explicar de manera efectiva las ventas ocurridas en el año 2022.
Entrega 3 Universidad	En esta entrega final, se presentó al cliente la primera versión del tablero de control, que permite analizar diversos comportamientos de las ventas basándonos en los datos procesados y en las predicciones generadas por el modelo. Es importante señalar que en esta última entrega se incorporó un modelo con nuevas variables relacionadas con vuelos, con el objetivo de mejorar la explicabilidad del modelo. A partir de esta adición, al cliente le complació que fuéramos más allá de simplemente proporcionar un modelo. Se destacó la importancia de presentar de manera gráfica la información, permitiendo tanto a ellos como a nosotros extraer conclusiones sobre las ventas específicas de cada marca o tienda.

3 [35%] PRESENTACIÓN FINAL:

Realizamos la presentación con unas ciertas mejoras respecto a lo mostrado en la sustentación final en el siguiente link:
https://drive.google.com/file/d/1hK-8R1WPm76uhIev_Lz2Z5wJARRGkpoa/view?usp=sharing

Nota: Dejamos los archivos subidos en Google drive para evitar restricciones por temas de organización que impone la universidad.

4 [15%] CONCLUSIONES Y TRABAJO FUTURO:

- Las predicciones generadas por el modelo desarrollado resultan ser de gran utilidad para Opain, ya que permiten optimizar la cantidad de ventas al basarse en las tendencias de compra de los viajeros en el Aeropuerto El Dorado.
- El modelo presentado puede desempeñar un papel crucial como sistema de toma de decisiones para Opain, proporcionando seguridad al momento de reubicar tiendas, ajustar productos y/o establecer nuevas alianzas con compañías.
- Nuestro modelo final tiene la capacidad de informar sobre los factores que inciden en el comportamiento de las ventas. Permite indagar sobre qué variables y en qué medida influyen en las predicciones respectivas.
- El producto de datos entregado superó las expectativas del cliente, quienes esperaban únicamente un modelo funcional en términos de predicciones.
- Aunque los años 2022 y 2023 son comparables hasta cierto punto, debido a limitaciones temporales, no tendremos todos los datos disponibles para el año 2023.
- Las marcas que representan aproximadamente el 76% de las ventas corresponden a las marcas 3 y 6, siendo las tiendas Duty Free las que generan las mayores ventas.
- Las categorías con mayor influencia en las ventas incluyen tabaco y alcohol, juegos y juguetes, productos de lujo, regalos y obsequios, así como accesorios de mano, belleza y cuidado.
- Las zonas/ubicaciones en el muelle internacional que más influyen positivamente son A6-A8 y A10-A12, mientras que la zona A2-A4 muestra una relación inversa con las ventas.
- Los vuelos en horas tempranas, especialmente entre las 0-3 de la mañana, presentan una relación negativa con las ventas. Además, se observan picos de venta entre las 5-8, 10-15 y 19-22 horas.
- Los días de la semana con mayores ventas son el domingo, lunes y sábado.
- De cara al futuro, sería interesante incorporar aún más variables externas al propio aeropuerto. Al explicar el comportamiento de los pasajeros dentro del aeropuerto con estos datos adicionales, sería crucial para generar conclusiones más acertadas y desarrollar modelos de predicción de ventas más efectivos.
- La falta de información en relación con las categorías no clasificadas generó un problema en la construcción del modelo. Aunque es posible explicarlas, carecen de contexto, a diferencia de las marcas que proporcionaron descripciones detalladas de sus productos.
- Los datos se limitaron a las ventas en el muelle internacional. Sería interesante realizar también un análisis de estas ventas en vuelos nacionales, dado que se comprende que estos constituyen la mayor parte de los vuelos que atraviesan el aeropuerto.