PROYECTO FINAL

Airline Passenger Satisfaction

Integrantes

- Lourdes Aparicio
- Luciano Ghidella
- Martin Rodriguez Valiente
- Mariana Moreyra
- Jonathan Carrasco

Curso

- Data Science - Comisión 23050 – Coderhouse

Tabla de Contenidos

Descripción del caso de negocio

La industria aérea se caracteriza por una intensa competencia por lo que resulta de suma importancia conocer a los clientes, sus intereses y preferencias. Este trabajo se centra en analizar los datos de encuestas realizadas a pasajeros de una cierta aerolínea para evaluar distintos aspectos y cómo afectan en el nivel de satisfacción, para esto se utilizarán técnicas de Data Science.

Objetivos del modelo

El objetivo general del presente trabajo es predecir a través de un modelo de Machine Learning la satisfacción con la mayor asertividad posible, según ciertos contextos y analizar qué variables impactan con mayor correlación en la satisfacción, realizando el entrenamiento de un modelo predictivo

Se plantean como objetivos específicos:

- Conocer las características y preferencias de los clientes de acuerdo con el género, edad, tipo de cliente y clase de vuelo que utilizan.
- Identificar cuáles son los servicios que deben mejorarse, y ver si se asocian a las características generales de los clientes
- Analizar las características generales de los vuelos que tienen mayores inconvenientes (demoras en partida/arribo por ej.), y ver si la información recolectada es de utilidad para proponer soluciones a los mismos.
- Desarrollar un modelo predictivo que permita identificar el nivel de satisfacción de los pasajeros respecto a los servicios brindados.

Descripción de los datos

Los datos se obtuvieron del sitio web Kaggle. Se trata de una base de datos estructurados, generada a partir de encuestas realizadas a más de 100k clientes. La misma cuenta con campos que permiten describir las características generales del cliente, como género, edad, tipo de viaje, categoría de pasajero, distancia del vuelo; como así también cuáles son las opiniones del mismo con relación a distintos aspectos del viaje. En este punto, utilizando escalas de Likert, se les consultó sobre distintos aspectos con el grado de satisfacción donde : 0 correspondía a variables donde la respuesta No Aplica, y los puntajes de 1 a 5 indican el nivel de satisfacción de los pasajeros.

- Gender: Género de los pasajeros (variable categórica, “Female /Male”)
- Customer Type: tipo de cliente, categorizado como: “cliente leal” / “cliente desleal”) (Loyal customer, disloyal customer)
- Age: Edad actual de los pasajeros (variable numérica, en años)

- Type of Travel: Motivo del viaje de los pasajeros (variable categórica: Personal Travel / Business Travel)
- Class: Tipo de clase en la que viajaban los pasajeros (variable categórica: Business / Eco / Eco Plus)
- Flight distance: Distancia recorrida en el viaje (variable numérica, en kilómetros)
- Inflight wifi service: Nivel de satisfacción respecto al servicio de wifi durante el vuelo (escala de likert)
- Departure/Arrival time convenient: Nivel de satisfacción con relación a la conveniencia entre el tiempo de partida/arribo (escala de likert)
- Ease of Online booking: Nivel de satisfacción respecto a la reserva online (escala de likert)
- Gate location: Nivel de satisfacción respecto a la ubicación de la puerta de embarque en el aeropuerto (escala de likert)
- Food and drink: Nivel de satisfacción respecto a la comida y bebida (escala de likert)
- Online boarding: Nivel de satisfacción respecto a Satisfaction level of online boarding (escala de likert)
- Seat comfort: Nivel de satisfacción respecto a la comodidad de los asientos (escala de likert)
- Inflight entertainment: Nivel de satisfacción respecto al entretenimiento durante el vuelo (escala de likert)
- On-board service: Nivel de satisfacción respecto al servicio durante el vuelo (escala de likert)
- Leg room service: Nivel de satisfacción respecto al servicio de espacios para piernas Satisfaction level of Leg room service (escala de likert)
- Baggage handling: Nivel de satisfacción respecto al manejo del equipaje (escala de likert)
- Check-in service: Nivel de satisfacción respecto al servicio de check-in (escala de likert)
- Inflight service: Nivel de satisfacción respecto al servicio durante el vuelo (escala de likert)
- Cleanliness: Nivel de satisfacción respecto a la limpieza (escala de likert)
- Departure Delay in Minutes: minutos de demora en la partida (variable numérica, en minutos)
- Arrival Delay in Minutes: minutos de demora en el arribo (variable numérica, en minutos)
- Satisfaction: Nivel de satisfacción respecto a la aerolínea en general, medido como “satisfactorio” o “neutral / no satisfactorio”.

Tabla 1. Resumen de las variables incluidas en el informe

Variables que caracterizan a pasajeras/os	Variables que caracterizan los vuelos	Variables con escala de satisfacción
Age	Flight distance	Inflight wifi service
Gender	Departure Delay in Minutes	Departure/Arrival time convenient
Customer Type	Arrival Delay in Minutes	Ease of Online booking
Type of travel		Gate location
Class		Food and drink
		Online boarding
		Seat comfort
		Inflight entertainment
		On-board service
		Leg room service
		Baggage handling
		Check-in service
		Inflight service
		Cleanliness

Variable Target → **Satisfaction**

Hallazgos encontrados por el EDA

- En cuanto a satisfacción (nuestra variable a predecir) En un (56%) la opinión en satisfacción fue Neutral o negativa, versus 44% donde los clientes quedaron satisfechos. No hay desbalance de la variable target.
- Variable género: La proporción de hombres y mujeres encuestados es similar.
- Según histograma por edad, el rango de más concentración es entre 25-60 años.
- Respecto a las clases: Es menor al 10% las personas que viajan en eco plus, el resto se dividide entre business y eco en partes prácticamente iguales. Es probable que la oferta de eco plus no esté disponible en muchos vuelos, y no que sea por una preferencia del cliente.
- Evaluando las encuestas, vemos que hay más exigencia (o menos conformidad) en wifi en vuelo y en facilidad de reserva on line.
- Los clientes de business tienen un promedio de edad más elevado (+33 a 50) que quienes viajan en las otras dos (+25 a 50)
- Los clientes que realizan viajes cortos (menores a 1000 km) tienden a tener un nivel de satisfacción neutral o negativa. Esta No Satisfacción, puede tener que ver con otra variable que se relacione con viajes cortos más que con la distancia en sí misma. (Por ejemplo, la clase, edad o tipo de viaje)
- Los que viajan en clase eco en su mayoría son No satisfechos y los que viajan en clase business en mayoría son Satisfechos. También los viajes cortos poseen mayor concentración en clase eco, lo que explica una de las causas porque en viajes cortos hay menor grado de satisfacción. Siguiendo con la misma idea, vemos que los que viajan por motivos personales en su gran mayoría son No satisfechos, y viajan distancias más cortas que los que viajan por trabajo.
- Un cliente no leal es más probable que sea No satisfecho.
- Los No satisfechos poseen entre 20 a 40 años (los conformes poseen entre 40 a 60).
- Vemos que los satisfechos que viajaron en business, les dieron importancia a los asientos y al servicio en vuelo. Mientras que los que los satisfechos que viajaron en económica le dieron importancia al wifi y a los entretenimientos. (Dato importante para el cliente, donde fortalecer en cada clase) Se observa que los No Satisfechos de todas las clases puntuaron en promedio bajo al servicio de wifi en vuelo.
- Observamos que aquellas variables que tienen mayor correlación lineal con la satisfacción de los pasajeros corresponden a: Online Boarding, Class y Type of Travel. Mientras aquellas que peor correlación lineal tienen con la variable de satisfacción son: Gate Location, Gender y Departure/Arrival Time Convenient,

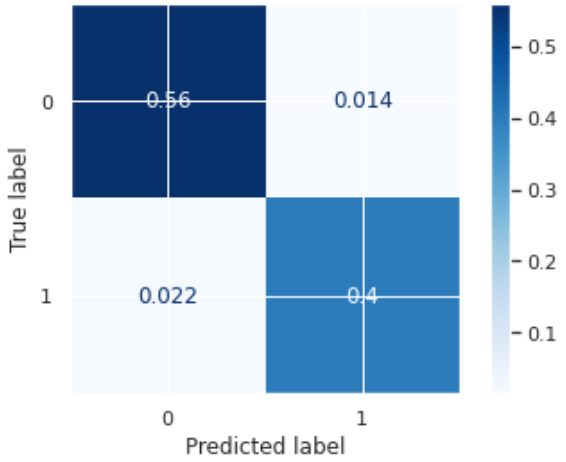
Algoritmo Elegido

Para lograr nuestro objetivo seleccionamos el algoritmo LightGBM. Realizamos pruebas con otros modelos como Decision Tree, Random Forest y Regresión Logística, sin embargo, comparando las métricas de Accuracy y Area Under the Curve decidimos seleccionar el modelo LightGBM.

Métricas de Desempeño del Modelo

La performance del modelo es la siguiente:
Accuracy = 0.9633
AUC = 0.9950

Matriz de confusión:



Curva ROC:

Iteraciones de Optimización

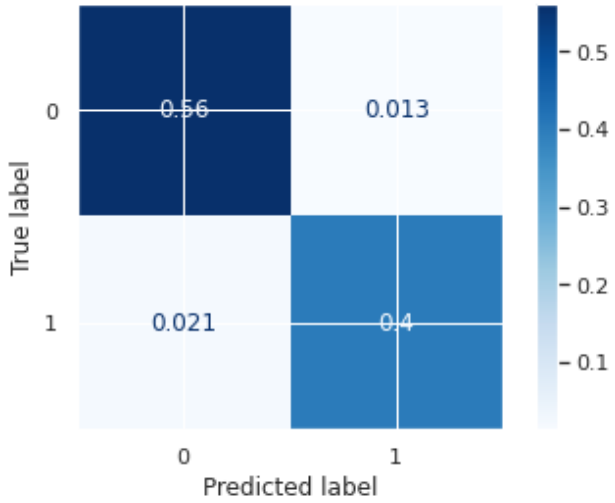
Para mejorar la performance del modelo realizamos un tuneo de hiperparámetros utilizando el método GridSearchCV. A si mismo, para encontrar los hiperparámetros óptimos realizamos un análisis de sensibilidad, observando cómo cambia las métricas de nuestro

modelo ante un cambio del número de estimadores del algoritmo o del hiperparámetro “max depth” de LightGBM.
También realizamos feature selection utilizando tres técnicas, Variance Threshold, SelectKBest y SelectFromModel.

Métricas finales del Modelo Optimizado.

La performance del modelo optimizado es la siguiente:
Accuracy = 0.9653
AUC = 0.9954

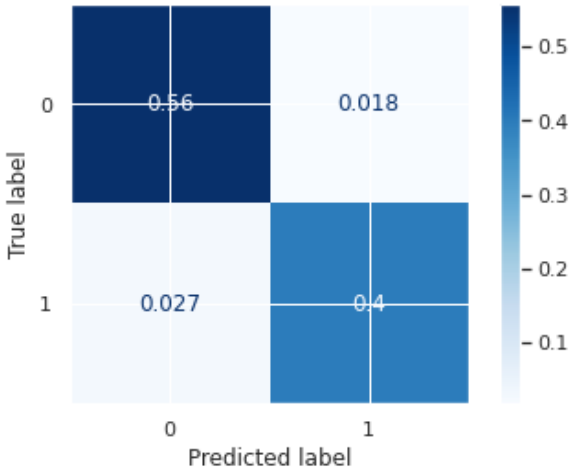
Matriz de confusión:



Curva ROC:

Resultados de Feature Selection:
Se observa que utilizando el método SelectFromModel pasamos de 22 features a un dataset con **10 features** y nuestro Accuracy disminuyo solamente de 0.9653 a 0.9555 y el AUC de 0.9954 a 0.9925. (Utilizando el mejor lightGBM)

Matriz de confusión:



Curva ROC

Conclusiones

El algoritmo LightGBM nos permitió predecir la satisfacción del cliente con un 96.5% de aciertos y el AUC (relación entre el ratio de Falsos positivos y el ratio de verdaderos positivos) es 99.5%. Nuestro error de predecir que un cliente está satisfecho cuando en realidad no lo está es: 1.3%. Y el error de predecir que un cliente no está satisfecho cuando en realidad sí lo está es: 2.1%.
Consideramos muy apropiado utilizar lo aprendido en feature selection reduciendo más del 50% la cantidad de variables de entrada (en vez de 22 se analizan 10) bajando apenas el 1% la asertividad, que en nuestra target (satisfacción del cliente) es completamente aceptable.

Futuras líneas

Durante el trabajo encontramos la dificultad de mejorar tanto el Accuracy como el AUC, se intentaron distintos métodos, pero no fueron satisfactorios. No se supera el nivel de Accuracy de 0.965 y AUC 0.995. Creemos que un análisis más exhaustivo de feature engineering quizás pueda superar esa barrera.