LATAM Challenge Flight Delay Prediction

Introducción

LATAM airlines pertenece a una industria altamente competitiva, la eficiencia operacional y la sustentabilidad se consideran diferenciadores claves. LATAM tiene muy en claro esto, los valores no solo están integrados en el modelo de negocios, también se alinean con sus compromisos a futuro, crear valor y enfrontar el cambio climático. En este reporte voy a presentar mi implementación o como yo abordaría el problema en cuestión. Enfocado mucho en como mediante el uso de datos podemos obtener insights y realizar predicciones que propongan soluciones a problemas del día a día.

Mi solución se basa en la predicción del retraso de salida de los vuelos (flight departure delay). Flight Delays tienen un efecto cascada en las operaciones de una aerolínea, lo que conlleva un aumento de los costos, decremento de la satisfacción de los clientes y una elevación en las emisiones de carbono a causa de la extensión de fases muy intensivas en el consumo de combustible como el taxiing.

Si logramos predecir correctamente si va a haber o no un delay, se pueden tomar medidas proactivas para mitigar o minimizar los impactos negativos, a su vez, mejorando la eficiencia operacional y reduciendo la huella de carbono

En este reporte, voy a presentar un Análisis exploratorio de los datos utilizados (EDA) y la implementación de algunos modelos de Machine Learning. Si bien no tuve acceso a datos de LATAM utilice datos públicos para demostrar como abordo este tipo de problemas. El objetivo no es entregar un modelo (solución) lista para ser implementada, más **bien ilustrar el potencial que este tipo de soluciones tienen para afrontar los distintos problemas que afecten negativamente a LATAM**.

Los datos utilizados para mi análisis fueron descargados de Kaggle. Tener en cuenta que existen otras fuentes como API's tales como AviationStack API o Rapid API que proveen datos mucho más completos, pero a su vez vienen con un costo.

Con esta propuesta se apunta a demostrar como con los datos podemos contribuir a las estrategias de sustentabilidad y compromisos para el futuro de LATAM. Obviamente con muchas mejoras y con mayor desarrollo considero que este tipo de solución puede ser valioso.

Los resultados y el Código van a ir adjuntos en jupyter notebooks, estos también se van a poder encontrar en un repositorio de GitHub.

Datos

Como mencione previamente los datos utilizados son **simplemente para ilustrar o demostrar como personalmente abordo un problema y el proceso que realizo**. Los datos son de dominio publico y se pueden encontrar en Kaggle. Estos corresponden a datos obtenidos de la Bureau of Transportation Statistics, USA. Contiene información de vuelos en los periodos de enero del 2019 y enero del 2020. Los datos contienen mas de 400,000 vuelos solo del mes de enero.

The Machine Learning Process

El proceso para entrenar modelos de Machine Learning es un proceso iterativo, es decir un vez entrenado un modelo lo evaluamos y vemos que podemos mejorar o cambiar para lograr un modelo que tenga menos errores predictivos. A continuación, detallo el proceso utilizado para lograr a entrenar los distintos modelos de la implementación:

1. Formular una pregunta: ¿qué queremos lograr? Como podemos lograrlo, que criterios utilizo.
2. Encontrar y entender los datos
3. Limpiar los datos y extraer nuevos features
4. Elegir un modelo
5. Tuning y evaluación de los modelos
6. Usar el modelo y los insights y presentar mis resultados.

Objetivo que se quiere lograr

Lo que se quiere lograr es potenciar la sustentabilidad. De acorde a la memoria anual del 2022 estas son algunas de las iniciativas más importantes de sostenibilidad que LATAM está abordando:

- Reducción de emisiones de CO2: LATAM se ha comprometido a reducir sus emisiones de CO2 en un 50% para 2050 en comparación con los niveles de 2005. Esta es una iniciativa crítica dada la creciente preocupación por el cambio climático y el papel de la industria de la aviación en las emisiones de gases de efecto invernadero.

- Uso de combustibles sostenibles: LATAM está explorando el uso de combustibles de aviación sostenibles (SAF) para reducir aún más su huella de carbono. Los SAF son una alternativa a los combustibles de aviación a base de petróleo y pueden reducir significativamente las emisiones de CO2.

- Compensación de carbono: LATAM también está trabajando en programas de compensación de carbono, que permiten a los pasajeros compensar las emisiones de sus vuelos comprando créditos de carbono.

- Eficiencia operacional: LATAM está trabajando para mejorar la eficiencia operacional de sus vuelos, lo que puede reducir las emisiones de CO2 y otros gases de efecto invernadero.

En esta oportunidad mi implementación va de la mano del Machine Learning en el cual se intenta predecir si la salida de un vuelo se va a atrasar o no. En mi opinión esta propuesta puede tener un impacto en varias formas:

1. Mejora de la eficiencia operacional: Al predecir los retrasos, LATAM puede tomar medidas preventivas para minimizarlos o evitarlos. Esto puede implicar ajustar la programación de los vuelos, reasignar los recursos, o tomar medidas para acelerar los procedimientos de embarque y desembarque. Al reducir los retrasos, LATAM puede mejorar la eficiencia de sus operaciones, lo que puede resultar en costos operativos más bajos.
2. Mejora de la satisfacción del cliente: Los retrasos en los vuelos pueden ser una fuente importante de insatisfacción para los clientes. Al predecir y minimizar los retrasos, LATAM puede mejorar la experiencia de viaje de sus clientes, lo que puede resultar en una mayor satisfacción del cliente, una mayor lealtad y, en última instancia, más clientes.
3. **Reducción de la huella de carbono:** Los retrasos en los vuelos pueden resultar en un consumo innecesario de combustible, lo que puede aumentar las emisiones de CO2 de LATAM. Al predecir y minimizar los retrasos, LATAM puede reducir su consumo de combustible y, por lo tanto, sus emisiones de CO2. Esto puede ayudar a LATAM a cumplir sus objetivos de sostenibilidad y a reducir su impacto en el cambio climático.

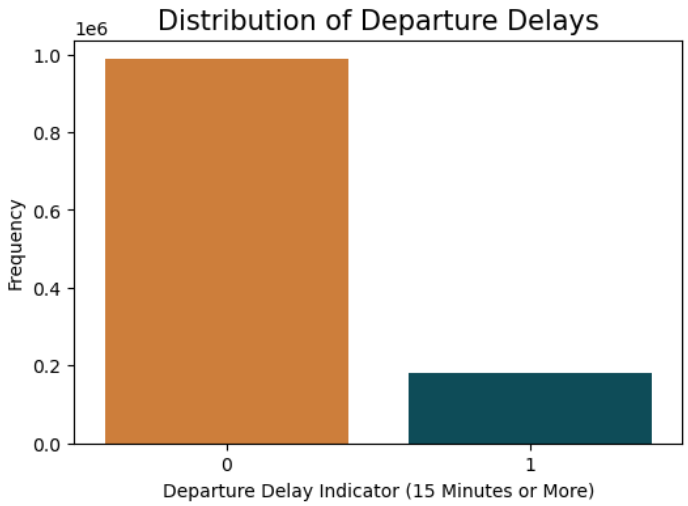
En resumen, un sistema de predicción de retrasos en los vuelos puede proporcionar a LATAM una valiosa herramienta para mejorar la eficiencia de sus operaciones, mejorar la satisfacción del cliente, reducir su huella de carbono y mejorar su planificación y toma de decisiones.

Análisis exploratorio

Por la extensión del reporte solo voy a entregar algunos ejemplos, para un análisis más completo ver EDA.ipynb notebook.

Mediante este proceso logramos entender algunas características de mis datos, se detectan anomalías, sirve para comprobar suposiciones y entender de una mejor manera la estructura e información que mi conjunto de datos tiene. Los hallazgos que se logran en este proceso sirven como una base para la elección de mi modelo y el enfoque que se va a tomar. Algunas preguntas que intente resolver durante este proceso.

¿Cómo está distribuida la variable objetivo (Retraso de Salida)? ¿Qué porcentaje de vuelos están retrasados?



¿Cuáles son las aerolíneas con mas vuelos, mas delays en promedio?

¿Cuáles aeropuertos son los mas visitados y cuales aeropuertos de origen son más utilizados?

En nuestros datos pude concluir que Soutwest Airlines, Delta Air y American Airlines fueron las aerolíneas con mayor cantidad de vuelos (de acorde a los datos). En cuanto al delay promedio Jetblue es la aerolínea que tuvo más delays. **Estas son solo algunas de las preguntas o insights que obtuve en mi análisis exploratorio, para algo mas detallados referirse a los notebooks.**

Modelos Utilizados

El proceso de modelamiento para realizar las predicciones incluyo a varios modelos de Machine Learning. Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier. Como se trata de un problema binario de clasificación (1 or 0) se eligieron estos modelos por su habilidad para manejar datos complejos, y relaciones no lineales en los datos.

Se uso principalmente para evaluar los modelos accuracy y roc_auc score, las cuales son métricas comunes al momento de evaluar problemas de clasificación.

	Model	Accuracy	roc_auc_score
0	Logistic Regression	88.50	0.627
1	Decision Tree Classifier	92.94	0.861
2	Random Forest Classifier	94.03	0.823
3	XGBoost Classifier	93.79	0.819

Los datos fueron separados (train, test) antes de realizar algún tipo de preprocesamiento de los datos como los son el scaling (variables numéricas), se usó estándar scale de este modo todos los datos entran en una distribución normal (con un promedio 0 y std 1), hacemos esto para estandarizar los datos a un rango similar, lo que puede incrementar la performance de algunos modelos. Para las variables categóricas se realizo one hot encoding. Esto se realiza después de la separación del training y test set para evitar lo que se conoce como "Data Leakage".

También hice uso de pipelines, estos agilizan el proceso de modelado y se garantiza que el preprocesamiento se aplica de manera consistente a los datos de entrenamiento y los de prueba, reduciendo el riesgo de errores y haciendo que el código sea más eficiente y leíble ya que los pasos del preprocesamiento y modelado se definen en un solo lugar.

A pesar del decente performance de los modelos, hay varias maneras en las cuales se puede mejorar. Por ejemplo, más datos sobre todo de la clase predictora, ya que en nuestra base de datos se observa un desequilibrio, (80% de la clase es 0:no delay). Adicionalmente añadir datos externos como lo pueden ser datos del clima, el clima puede ser de gran influencia al momento de las demoras de un vuelo. Otro tipo de dato que nos ayudaría es el tráfico aéreo, información sobre que tan lleno o que tan ocupado esta el cielo de acorde algunas rutas pueden inclinar a los modelos hacer predicciones más precisas.

De igual manera que en el análisis exploratorio, todo el proceso y el código se encuentra en el notebook models.

Conclusión y Próximos Pasos

Este proyecto ha demostrado el potencial del aprendizaje automático para predecir retrasos en los vuelos en LATAM Airlines. Y demuestra cómo el aprendizaje automático puede ser una herramienta valiosa para apoyar la Estrategia de Sostenibilidad de LATAM Airlines. Al predecir con precisión los retrasos en los vuelos, podemos tomar medidas proactivas para minimizar la quema extra de combustible y así reducir las emisiones de carbono. Los modelos que entrenamos mostraron resultados prometedores. Sin embargo, hay varias áreas donde este trabajo podría mejorarse y expandirse en el futuro:

- Recopilación de Datos: Obtener datos de vuelo propietarios de LATAM podría mejorar la precisión y relevancia de los modelos.
- **Impacto en la Sostenibilidad**: Nuestro enfoque puede ser refinado para cuantificar más directamente el impacto de los retrasos en los vuelos en las emisiones de carbono, lo que permitiría a LATAM tomar decisiones más informadas para alcanzar sus objetivos de sostenibilidad.
- **Incorporación de Más Factores de Sostenibilidad**: Podríamos expandir nuestro análisis para considerar otros factores que afectan la sostenibilidad, como el tipo de combustible utilizado y las condiciones meteorológicas.
- Ajuste de Modelos: Aunque implementamos la afinación de hiperparámetros, un mayor ajuste podría mejorar potencialmente el rendimiento de los modelos.
- Implementación de Modelos: Los modelos podrían implementarse en un entorno de producción para predecir retrasos en los vuelos en tiempo real.

En resumen, aunque los modelos actuales ya proporcionan un buen rendimiento, hay varias direcciones prometedoras para futuras mejoras y expansiones de este trabajo.