Reporte de Análisis y Modelo de IA para Predecir Tarifas Aéreas en los EE.UU.

A01368818 Joel Sánchez Olvera

Tecnológico de Monterrey Campus Querétaro

Abstracto .-

Este informe presenta el desarrollo y análisis de un modelo predictivo usado para determinar si las tarifas aéreas en los Estados Unidos superarán un umbral específico en los años futuros. Utilizando un conjunto de datos que incluye variables como el año, trimestre, ciudades, aeropuertos, número de millas recorridas, y tarifas, se realizó un exhaustivo proceso de preparación de datos, que incluyó la transformación de variables categóricas y la normalización de las características.

1. Introducción

El proyecto aborda dos objetivos fundamentales dentro del análisis predictivo de tarifas aéreas en los Estados Unidos. Ambos objetivos están diseñados para ayudar a los consumidores de la aerolíneas a anticipar comportamientos futuros en los precios de los boletos aéreos.

Predicción de Tarifas: El primer objetivo es construir un modelo de regresión lineal que permita predecir el valor exacto de las tarifas aéreas futuras. Utilizando un modelo de regresión para predecir los valores exactos de las tarifas aéreas basándose en datos históricos.

Clasificación de Tarifas: El segundo objetivo es implementar un modelo de regresión logística que clasifique si una tarifa aérea superará un umbral específico en el futuro, como por ejemplo: $1000, en los próximos años.

2. Preparación del Dataset y Limpieza de Datos

El conjunto de datos seleccionado contiene información detallada sobre tarifas aéreas recientes en vuelos dentro de los Estados Unidos, incluyendo rutas específicas, lo que proporciona una base sólida para analizar tendencias en tarifas y los factores que pueden influir en las variaciones de precios con el tiempo. Este conjunto de datos fue utilizado tanto para desarrollar un

modelo de regresión logística que predice si las tarifas aéreas superarán un umbral específico en los próximos años, como para crear modelos predictivos que anticipen tarifas futuras, lo cual es de gran utilidad para aerolíneas y consumidores.

Columnas clave del dataset:

- Year: Año en que se registraron las tarifas aéreas.
- Quarter: Trimestre del año en que se registraron las tarifas.
- City1: Ciudad de origen del vuelo.
- City2: Ciudad de destino del vuelo.
- Airport_1: Código del aeropuerto de origen.
- Airport_2: Código del aeropuerto de destino.
- Nsmiles: Distancia en millas entre los aeropuertos de origen y destino.
- Passengers: Número de pasajeros en la ruta durante el período de registro.
- Fare: Tarifa aérea promedio registrada para la ruta.
- Carrier_lg: Código de la aerolínea estándar.
- Fare_lg: Tarifa promedio registrada para la aerolínea estándar.
- Carrier_low: Código de la aerolínea de bajo costo.
- Fare_low: Tarifa promedio registrada para la aerolínea de bajo costo.

2.1.- Limpieza de Datos

El proceso de limpieza de datos es un paso crucial en nuestro proyecto para garantizar que los modelos predictivos operen con la máxima precisión y sin sesgos en su entrenamiento. Para la limpieza de datos, las acciones implementadas fueron:

Manejo de Valores Faltantes: Se eliminó cualquier fila del conjunto de datos que contuviera valores faltantes. Los valores faltantes pueden distorsionar los resultados de los modelos. Ésta desición fue tomada basándonos en que tenemos un gran número de datos para manejar en nuestro dataset, y el conteo de datos faltantes no representaba un porcentaje importante comparado con los datos totales, al eliminar estas filas, aseguramos que el modelo se entrene

únicamente con datos completos y precisos.

Codificación de Variables Categóricas: Las variables categóricas, como nombres de ciudades, códigos de aeropuertos y aerolíneas, fueron convertidas a valores numéricos utilizando diccionarios predefinidos. Esta transformación es esencial porque la codificación numérica permite al modelo interpretar y utilizar estas variables para realizar predicciones.

Selección de Variables: No todas las variables del dataset fueron utilizadas para los modelos predictivos. Se realizó una selección de las más relevantes:

- **Year**: Año en el que se registraron las tarifas aéreas. Tipo de dato: Entero.
- **quarter**: Trimestre del año en el que se registraron las tarifas. Tipo de dato: Entero.
- **city1**: Ciudad de origen del vuelo. Tipo de dato: Categórico (mapeado a valores numéricos).
- **city2**: Ciudad de destino del vuelo. Tipo de dato: Categórico (mapeado a valores numéricos).
- **airport_1:** Código del aeropuerto de origen. Tipo de dato: Categórico (mapeado a valores numéricos).
- **airport_2:** Código del aeropuerto de destino. Tipo de dato: Categórico (mapeado a valores numéricos).
- **nsmiles:** Distancia en millas entre los aeropuertos de origen y destino. Tipo de dato: Entero.
- **fare**: Tarifa aérea promedio registrada para la ruta. Tipo de **dato**: Decimal.
- **fare_lg:** Tarifa promedio registrada para la aerolínea de tarifa estándar. Tipo de dato: Decimal.
- **fare_low:** Tarifa promedio registrada para la aerolínea de bajo costo. Tipo de dato: Decimal.

La selección de estas variables fue guiada tanto por el conocimiento del dominio como por análisis exploratorios iniciales que mostraron su relevancia en la predicción de tarifas.

3. Modelado

3.1- Modelo de Regresión Lineal

Propósito:
El objetivo del modelo de regresión lineal es predecir el valor exacto de las tarifas aéreas futuras. La regresión lineal es una técnica de modelado predictivo ampliamente utilizada que intenta modelar la relación entre una variable dependiente (la tarifa) y una o más variables independientes (como el año, trimestre, distancia, etc.).

Enfoque:
El modelo de regresión lineal busca encontrar la "mejor línea recta" que pase a través de los puntos de datos en un espacio multidimensional. Esta línea se

define por un conjunto de coeficientes que representan el peso de la modificación de cada variable independiente al valor predicho de la tarifa. El modelo ajusta estos coeficientes de manera que se minimice la suma de los errores cuadrados (diferencias al cuadrado entre los valores predichos y los valores reales).

- **Función Objetivo:** El modelo minimiza el Error Cuadrático Medio (MSE), que se calcula como la media de los errores cuadrados entre las predicciones del modelo y los valores reales observados.

Entrenamiento y Validación: Para entrenar y validar el modelo, el conjunto de datos fue dividido en dos partes: el 80% se utilizó para entrenar el modelo y el 20% restante para probarlo. Esta división asegura que el modelo no solo se ajuste a los datos de entrenamiento, sino que también generalice bien a datos no vistos (de prueba).

3.2- Modelo de Regresión Logística

Propósito:
El modelo de regresión logística se diseñó para resolver un problema de clasificación binaria: determinar si una tarifa aérea será superior o inferior a un umbral específico, en este caso, $1000. Este tipo de modelo es útil en situaciones donde se requiere una decisión clara (sí/no), como en sistemas de alerta de precios o en la evaluación de riesgos financieros.

Enfoque:
La regresión logística es un tipo de análisis de regresión que se utiliza cuando la variable dependiente es categórica, en este caso, binaria (es decir, tarifa superior o inferior a $1000). A diferencia de la regresión lineal, que predice un valor continuo, la regresión logística predice la probabilidad de que un evento ocurra. Este modelo utiliza la función sigmoide para convertir una combinación lineal de las variables independientes en una probabilidad que se mapea entre 0 y 1.

- **Función Sigmoide:** La función sigmoide es una curva en forma de "S" que transforma cualquier valor real en un rango de [0, 1]. Esta función es clave en la regresión logística, ya que permite interpretar la salida del modelo como una probabilidad. Si la probabilidad calculada es mayor que un umbral (comúnmente 0.5), el modelo predice que la tarifa estará por encima del umbral de $1000; de lo contrario, predice que estará por debajo.
- **Función de Costo:** En lugar del MSE utilizado en la regresión lineal, la regresión logística utiliza la pérdida de entropía cruzada (cross-entropy loss) como su función objetivo. Esta función penaliza fuertemente las

predicciones incorrectas con alta confianza, lo que mejora la precisión del modelo al aprender.

Entrenamiento y Validación: El conjunto de datos también fue dividido en partes de entrenamiento y prueba (80% para entrenamiento y 20% para prueba). Durante el entrenamiento, el modelo ajusta los coeficientes de las variables independientes para minimizar la pérdida de entropía cruzada, utilizando métodos de optimización como el descenso de gradiente.

4. Resultados

Comparación de Modelos:

- **Regresión Lineal:** Este modelo demostró ser efectivo en la predicción de valores exactos de las tarifas aéreas. Aunque es capaz de capturar tendencias generales en los precios de los boletos, tiene algunas limitaciones cuando se trata de predecir variaciones extremas o valores atípicos. En general, el modelo es útil para análisis de tendencias y para proporcionar estimaciones de precios precisos que pueden ser utilizados en la planificación de ingresos de las aerolíneas.

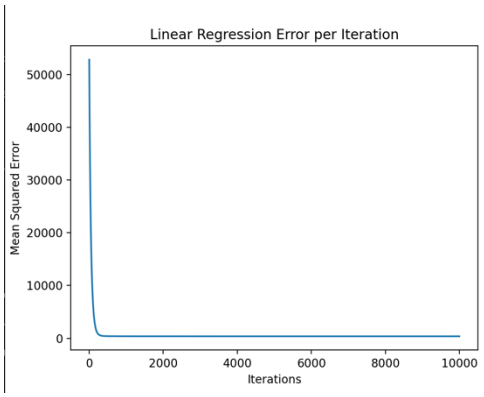

Figura 1. Gráfica de Error por Iteración

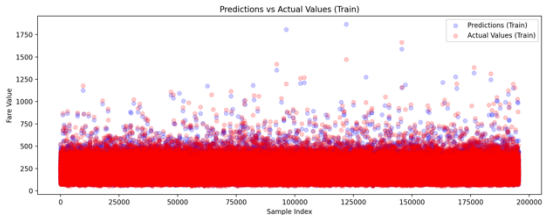

Figura 2. Resultados de Regresión Lineal (Entrenamiento)

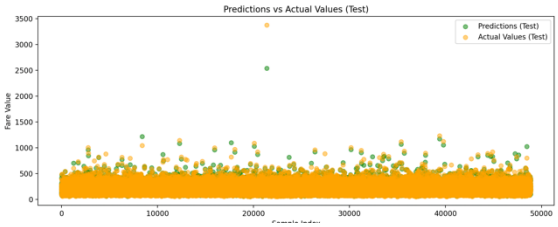

Figura 3. Resultados de regresión Lineal (Prueba)

- **Regresión Logística:** El modelo de regresión logística mostró un buen rendimiento en la clasificación de tarifas por encima de un umbral específico. Esto es especialmente útil para escenarios donde es más importante conocer si una tarifa cruzará un cierto límite. El modelo se comporta bien en identificar casos donde las tarifas serán inusualmente altas, lo que es valioso para los consumidores que buscan evitar vuelos costosos.

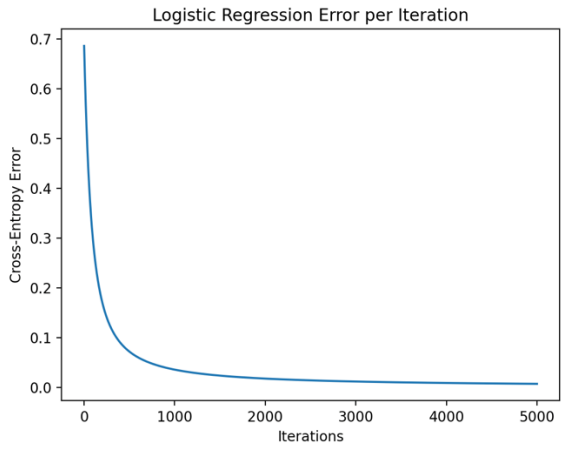

Figura 1. Gráfica de Error por Iteración

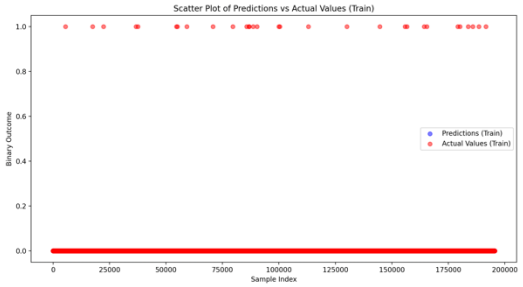

Figura 2. Resultados de Regresión Logística (Entrenamiento)

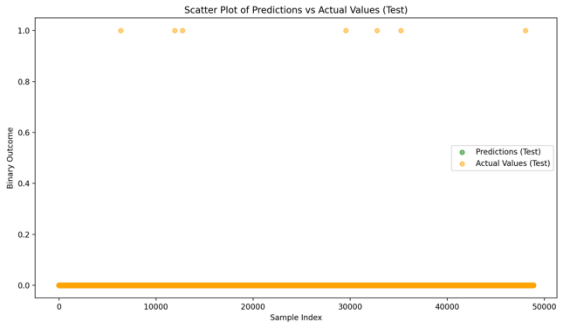

Figura 3. Resultados de Regresión Logística (Prueba)

Visualizaciones:

- **Análisis de Errores:** El error en el modelo de regresión lineal disminuyó de manera constante a lo largo de las iteraciones, como se muestra en la gráfica de error (MSE por iteración). Esto sugiere que el modelo está aprendiendo de manera efectiva a ajustar sus predicciones a los datos de entrenamiento. Los valores finales del MSE en los datos de prueba también indican que el modelo tiene una buena capacidad de generalización.
- **Gráficas de Dispersión:** Las comparaciones visuales entre los valores predichos y los valores reales tanto en el conjunto de entrenamiento como en el de prueba fueron clave para entender las fortalezas y debilidades de los modelos. En el caso del modelo de regresión lineal, se observó una buena alineación entre las predicciones y los valores reales, aunque con algunas desviaciones en casos extremos. Para el modelo de regresión logística, las gráficas de dispersión mostraron cómo el modelo separa las tarifas por encima y por debajo del umbral de manera efectiva, reflejando su precisión en la clasificación binaria.

5. Conclusiones

Resumen de Hallazgos:

El proyecto logró desarrollar y evaluar con éxito dos modelos de predicción de tarifas aéreas:

1. **Modelo de Regresión Lineal:** Este modelo fue efectivo para predecir los valores exactos de las tarifas aéreas. Aunque demostró ser útil para capturar las tendencias generales y proporcionar estimaciones precisas, tuvo dificultades en manejar variaciones extremas en los datos. No obstante, este modelo es valioso para la predicción continua de tarifas y puede ser utilizado por las aerolíneas para planificar precios de manera estratégica.
2. **Modelo de Regresión Logística:** Este modelo fue exitoso en clasificar tarifas como superiores o inferiores a un umbral predefinido. La regresión logística es especialmente útil en contextos donde se requiere tomar decisiones binarias, como la activación de alertas de precios altos para los consumidores. El modelo mostró un buen rendimiento en identificar correctamente los casos de tarifas elevadas, lo que puede ser crítico tanto para las aerolíneas como para los viajeros.

Implicaciones:

Los resultados de este proyecto sugieren que ambos modelos pueden ser herramientas valiosas:

- **Para las Aerolíneas:** Los modelos pueden ayudar a optimizar la gestión de ingresos mediante la predicción de tarifas futuras y la identificación de posibles picos de precios, permitiendo ajustar estrategias de precios en consecuencia.
- **Para los Consumidores:** Los modelos pueden proporcionar información valiosa sobre cuándo es probable que las tarifas suban o bajen, lo que puede ayudar a los viajeros a tomar decisiones informadas sobre la compra de boletos.