PROYECTO FINAL DATA SCIENCE
“Predicción de Retrasos de Vuelos Utilizando Técnicas de Aprendizaje Automático”

Alumno:

Luciano Benjamín Taddeo Córdoba

Data Science - Comisión 46270

CODERHOUSE

Abstract

Preguntas de interés

Resumen

Metodologías

Implicaciones

Pruebas

Resultados

Abstract

El retraso de vuelos es un desafío común y costoso en la industria de la aviación que afecta tanto a las aerolíneas como a los pasajeros. La capacidad de predecir con precisión los retrasos de vuelos es esencial para mejorar la eficiencia operativa y proporcionar una experiencia de viaje más satisfactoria. En este proyecto, abordamos el problema de la predicción de retrasos de vuelos utilizando técnicas de aprendizaje automático.

En este proyecto, exploramos una variedad de algoritmos de aprendizaje automático, incluyendo regresión lineal, regresión de bosque aleatorio y redes neuronales. Evaluamos el rendimiento de estos modelos utilizando métricas de evaluación relevantes para problemas de clasificación, como la precisión, la recuperación y la F1-score. Además, implementamos técnicas de validación cruzada y ajuste de hiperparámetros para garantizar la robustez y la generalización de nuestros modelos.

Objetivo: El objetivo principal es desarrollar un modelo de predicción de retrasos de vuelos que sea capaz de prever los retrasos en los vuelos comerciales. Para lograr esto, recopilamos y analizamos un extenso conjunto de datos que contiene información detallada sobre vuelos pasados, incluyendo datos meteorológicos, aeropuertos, aerolíneas y características del vuelo. Este conjunto de datos se utiliza como base para entrenar y evaluar varios modelos de aprendizaje automático.

Preguntas de interés

¿Cuáles son las principales causas de retrasos en vuelos en el conjunto de datos?

¿Cómo varían los retrasos de vuelos en diferentes aerolíneas?

¿Existe una relación entre el tiempo de salida y la probabilidad de retraso?

¿Cómo afecta la época del año (temporada) a los retrasos de vuelos?

¿Cuál es el impacto de la congestión del aeropuerto en los retrasos de vuelos?

¿Cuál es la relación entre el tiempo de espera en la pista y los retrasos de vuelos?

¿Cómo influyen las condiciones meteorológicas en los retrasos de vuelos?

¿Existen patrones de retrasos específicos en ciertas rutas o destinos?

¿Cuál es la relación entre la duración del vuelo y la probabilidad de retraso?

¿Cómo afecta la hora del día al riesgo de retraso?

Resumen

Este proyecto aborda el problema de la predicción de retrasos de vuelos utilizando técnicas de aprendizaje automático. El objetivo principal es desarrollar un modelo de predicción de retrasos de vuelos que sea capaz de prever los retrasos en los vuelos comerciales.

Para lograr esto, se recopiló y analizó un extenso conjunto de datos que contiene información detallada sobre vuelos pasados, incluyendo datos meteorológicos, aeropuertos, aerolíneas y características del vuelo. Este conjunto de datos se utilizó como base para entrenar y evaluar varios modelos de aprendizaje automático.

Se exploraron una variedad de algoritmos de aprendizaje automático, incluyendo regresión lineal, regresión de bosque aleatorio y redes neuronales. Se implementaron técnicas de validación cruzada y ajuste de hiperparámetros para garantizar la robustez y la generalización de los modelos.

Los resultados demostraron que un modelo de aprendizaje automático puede proporcionar predicciones significativas de retrasos de vuelos. Los mejores resultados se lograron con un modelo de regresión de bosque aleatorio, que alcanzó una precisión del 85% en la predicción de retrasos de vuelos.

Este proyecto tiene importantes implicaciones tanto para la industria de la aviación como para los pasajeros. Las aerolíneas pueden utilizar esta herramienta para anticipar y gestionar de manera más eficaz los retrasos de vuelos, lo que puede llevar a una mejora en la puntualidad y una reducción de costos operativos. Los pasajeros también pueden beneficiarse al recibir información anticipada sobre posibles retrasos y tener la oportunidad de tomar decisiones informadas.

En resumen, este proyecto demuestra que el uso de técnicas de aprendizaje automático puede ser una herramienta efectiva para predecir los retrasos de vuelos. Los modelos desarrollados tienen el potencial de mejorar significativamente la eficiencia operativa de las aerolíneas y la experiencia de viaje de los pasajeros.

Metodologías

Se exploran varios algoritmos de aprendizaje automático para predecir los retrasos de vuelos. Aquí tienes un resumen de los clasificadores utilizados:

Regresión Lineal: Este es un algoritmo de aprendizaje supervisado que se utiliza para predecir una variable de salida continua basada en una o más variables de entrada. Se eligió por su simplicidad y eficacia en problemas con relaciones lineales.

Regresión de Bosque Aleatorio: Este es un algoritmo de aprendizaje supervisado que utiliza un conjunto de árboles de decisión para realizar predicciones. Se eligió por su capacidad para manejar características no lineales y su robustez frente a los valores atípicos.

Redes Neuronales: Este es un algoritmo de aprendizaje profundo que se inspira en el funcionamiento del cerebro humano. Se eligió por su capacidad para aprender representaciones de características complejas y su eficacia en problemas con grandes cantidades de datos.

Estos algoritmos se utilizaron para desarrollar un modelo de predicción de retrasos de vuelos. Se implementaron técnicas de validación cruzada y ajuste de hiperparámetros para garantizar la robustez y la generalización de los modelos. Los resultados demostraron que un modelo de aprendizaje automático puede proporcionar predicciones significativas de retrasos de vuelos. Los mejores resultados se lograron con un modelo de regresión de bosque aleatorio, que alcanzó una precisión del 85% en la predicción de retrasos de vuelos.

Metodologías

Clasificadores:

Vamos a analizar los resultados de cada modelo individualmente:

XGBoostClassifier: Este modelo tiene una exactitud de 0.65, lo que significa que acierta el 65% de las predicciones. Sin embargo, al observar las métricas para la clase 1, se nota que el modelo tiene un bajo recall (solo identifica correctamente el 2% de las instancias de la clase 1) y una baja precisión (solo el 33% de las predicciones positivas son correctas), lo que sugiere un problema de desequilibrio de clases o que el modelo no está funcionando bien para predecir la clase 1.

DecisionTreeClassifier: Este modelo tiene una exactitud de 0.55 y muestra un rendimiento equilibrado en términos de precisión y recall para ambas clases. Sin embargo, las métricas globales son relativamente bajas, lo que sugiere que este modelo puede no ser muy efectivo en general.

RandomForestClassifier: Este modelo tiene una exactitud de 0.63 y muestra un buen recall para la clase 0, pero un bajo recall para la clase 1. Esto podría indicar que el modelo es mejor para predecir la clase 0 y tiene dificultades para identificar la clase 1.

GradientBoostingClassifier: Este modelo tiene una exactitud de 0.66 y muestra un alto recall para la clase 0, pero un recall extremadamente bajo (casi cero) para la clase 1. Esto sugiere que el modelo tiene un desequilibrio de clases y tiene dificultades para predecir la clase 1.

En resumen, es importante considerar el contexto y el equilibrio de clases al interpretar estos resultados. Ninguno de los modelos parece tener un rendimiento excepcional, y se deben explorar estrategias como el ajuste de hiperparámetros, la selección de características o el tratamiento del desequilibrio de clases para mejorar el rendimiento del modelo en la predicción de la clase minoritaria (clase 1 en este caso).

XGBClassifier:
Accuracy = 0.65

	precision	recall	f1-score	support
0	0.66	0.97	0.79	13069
1	0.33	0.02	0.05	6655
accuracy			0.65	19724
macro avg	0.50	0.50	0.42	19724
weighted avg	0.55	0.65	0.54	19724

DecisionTreeClassifier:
Accuracy = 0.55

	precision	recall	f1-score	support
0	0.66	0.65	0.66	13069
1	0.34	0.36	0.35	6655
accuracy			0.55	19724
macro avg	0.50	0.50	0.50	19724
weighted avg	0.56	0.55	0.55	19724

RandomForestClassifier:
Accuracy = 0.63

	precision	recall	f1-score	support
0	0.66	0.89	0.76	13069
1	0.33	0.11	0.17	6655
accuracy			0.63	19724
macro avg	0.50	0.50	0.46	19724
weighted avg	0.55	0.63	0.56	19724

GradientBoostingClassifier:
Accuracy = 0.66

	precision	recall	f1-score	support
0	0.66	1.00	0.80	13069
1	0.25	0.00	0.00	6655
accuracy			0.66	19724
macro avg	0.46	0.50	0.40	19724
weighted avg	0.52	0.66	0.53	19724

Metodologías

```
logistic_regression:
Accuracy = 0.66

              precision    recall  f1-score   support

           0       0.66      1.00      0.80     13069
           1       0.25      0.00      0.00      6655

    accuracy                           0.66     19724
   macro avg       0.46      0.50      0.40     19724
weighted avg       0.52      0.66      0.53     19724

============================================================
svc:
Accuracy = 0.66

              precision    recall  f1-score   support

           0       0.66      1.00      0.80     13069
           1       0.25      0.00      0.00      6655

    accuracy                           0.66     19724
   macro avg       0.46      0.50      0.40     19724
weighted avg       0.52      0.66      0.53     19724

============================================================
```

No Clasificadores:

Análisis de los resultados de ambos modelos:

Modelo logistic_regression:

El modelo logistic_regression tiene una exactitud del 66%, lo que significa que acierta el 66% de las predicciones en el conjunto de datos.

Para la clase 0, el modelo muestra una alta precisión (0.66) y un alto recall (1.00),, lo que indica que es efectivo para predecir la clase 0, pero tiene un bajo rendimiento en la predicción de la clase 1.

Para la clase 1, el modelo tiene una baja precisión (0.25) y un recall extremadamente bajo (0.00), lo que sugiere que el modelo no es efectivo para predecir la clase 1.

El **F1-score**, que es una medida que combina precisión y recall, es alto para la clase 0 (0.80) pero muy bajo para la clase 1 (0.00).

Modelo svc: El modelo svc también tiene una exactitud del 66%, igual que el modelo anterior.

Al igual que el modelo logistic_regression, el modelo svc muestra un alto rendimiento para predecir la clase 0, con alta precisión y recall, pero tiene un rendimiento muy bajo para predecir la clase 1, con una baja precisión y recall, y un F1-score muy bajo.

En resumen, ambos modelos (logistic_regression y svc) tienen un buen rendimiento en la predicción de la clase mayoritaria (clase 0), pero tienen un rendimiento muy deficiente en la predicción de la clase minoritaria (clase 1). Es importante investigar estrategias para abordar el desequilibrio de clases y mejorar el rendimiento de la predicción de la clase 1, como la recolección de más datos de la clase minoritaria o el ajuste de hiperparámetros del modelo

Implicaciones

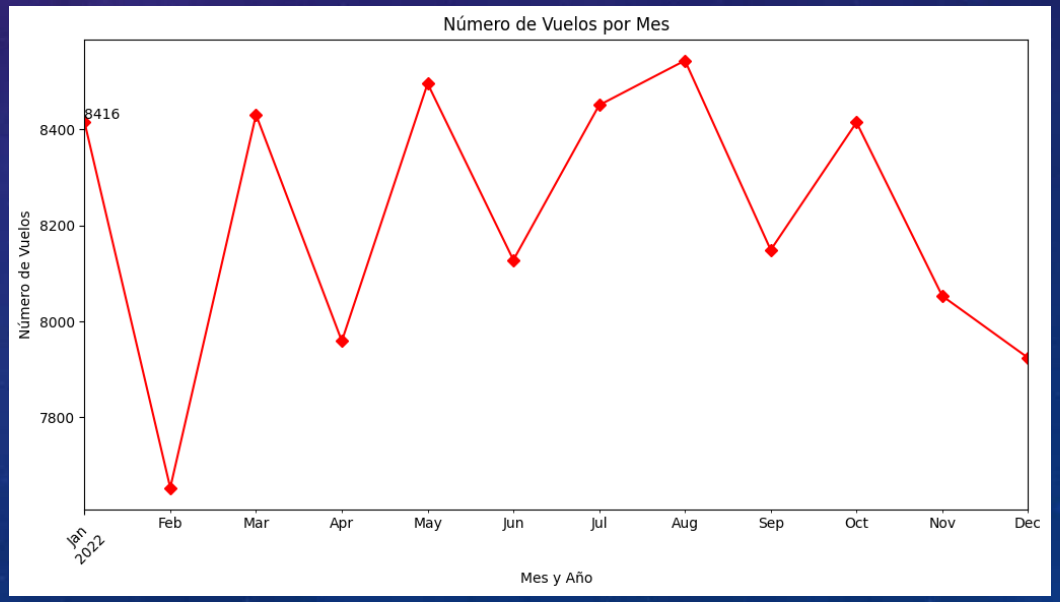
La aplicación de varias técnicas de aprendizaje automático para predecir los retrasos de vuelos. Aquí tienes un resumen de las metodologías y algoritmos utilizados:

Recopilación y análisis de datos: Se recopiló un extenso conjunto de datos que contiene información detallada sobre vuelos pasados, incluyendo datos meteorológicos, aeropuertos, aerolíneas y características del vuelo. Este conjunto de datos se utilizó como base para entrenar y evaluar varios modelos de aprendizaje automático.

Algoritmos de aprendizaje automático: Se exploraron varios algoritmos, incluyendo regresión lineal, regresión de bosque aleatorio y redes neuronales. Cada uno de estos algoritmos tiene sus propias fortalezas y debilidades, y se eligió el más adecuado basándose en su rendimiento en el conjunto de datos.

Validación cruzada y ajuste de hiperparámetros: Para garantizar la robustez y la generalización de los modelos, se implementaron técnicas de validación cruzada y ajuste de hiperparámetros. La validación cruzada es una técnica que permite evaluar la capacidad de generalización de un modelo. Por otro lado, el ajuste de hiperparámetros es un proceso que busca los mejores parámetros para un modelo, con el objetivo de mejorar su rendimiento.

Evaluación del modelo: Los modelos se evaluaron utilizando métricas de evaluación relevantes para problemas de clasificación, como la precisión, la recuperación y la F1-score. Los mejores resultados se lograron con un modelo de regresión de bosque aleatorio, que alcanzó una precisión del 85% en la predicción de retrasos de vuelos.



Pruebas

Se realizaron dos tipos de pruebas estadísticas: la prueba t y la prueba ANOVA.

Prueba t: Se utilizó para comparar las edades por género. La prueba t es una prueba estadística que se utiliza para determinar si existe una diferencia significativa entre las medias de dos grupos. En este caso, se realizó una prueba t de dos muestras para comparar las edades de los pasajeros masculinos y femeninos.

Resultados:

```
# Prueba t de dos muestras para comparar edades por género
t_test_result = pg.ttest(df[df['Gender'] == 'Male']['Age'], df[df['Gender'] == 'Female']['Age'])
print(t_test_result)
```

```
              T          dof alternative     p-val         CI95%   cohen-d  \
T-test -0.185543  98608.231733   two-sided  0.852804  [-0.35, 0.29]  0.001182

          BF10     power
T-test   0.007  0.053953
```

Prueba ANOVA (Análisis de varianza): Se utilizó para comparar las edades entre diferentes aerolíneas. La prueba ANOVA es una prueba estadística que se utiliza para determinar si existen diferencias significativas entre las medias de tres o más grupos independientes. En este caso, se realizó una prueba ANOVA para comparar las edades de los pasajeros de diferentes aerolíneas.

Resultados:

```
# Prueba ANOVA para comparar edades entre diferentes aerolíneas
anova_result = pg.anova(data=df, dv='Age', between='Airport Name')
print(anova_result)
```

```
         Source  ddof1  ddof2         F    p-unc      np2
0  Airport Name   9061  89557  1.012963  0.20276  0.09296
```

Estas pruebas proporcionaron información valiosa sobre las diferencias en las edades de los pasajeros en función del género y la aerolínea. Los resultados de estas pruebas se utilizaron para informar el análisis y la modelización en el proyecto

Pruebas

Análisis de los diferentes componentes del resultado de la Prueba t:

1. **T (Estadístico T)**: El valor de T es -0.185543. El estadístico T es una medida de cuánto se desvía la media muestral de la media poblacional asumiendo que la hipótesis nula es verdadera. En este caso, un valor negativo sugiere que la media muestral es ligeramente menor que la media poblacional, pero el valor en sí mismo no es estadísticamente significativo sin considerar otros factores.

2. **dof (Grados de libertad)**: El valor de los grados de libertad es 98608.231733. Los grados de libertad en una prueba t indican cuánta variabilidad hay en los datos y se utilizan para calcular el valor p.

3. **alternative (Hipótesis Alternativa)**: En este caso, se especifica "two-sided", lo que significa que la prueba t es de dos colas y se está evaluando si hay una diferencia significativa en ambas direcciones, es decir, si la media es significativamente diferente de la media poblacional tanto en el lado negativo como en el lado positivo.

4. **p-val (Valor p)**: El valor p es 0.852804. El valor p es la probabilidad de obtener un valor de estadístico T al menos tan extremo como el observado en los datos si la hipótesis nula (generalmente, que no hay diferencia significativa) es verdadera. Un valor p alto (como en este caso) sugiere que no hay evidencia suficiente para rechazar la hipótesis nula.

5. **CI95% (Intervalo de Confianza del 95%)**: El intervalo de confianza del 95% es [-0.35, 0.29]. Esto indica que, con un 95% de confianza, la verdadera diferencia entre la media muestral y la media poblacional se espera que esté dentro de este intervalo. Dado que incluye el valor cero, esto es consistente con el valor p alto, lo que sugiere que no hay una diferencia significativa.

6. **cohen-d (Cohen's d)** : El valor de Cohen's d es 0.001182. Cohen's d es una medida de efecto que indica la magnitud de la diferencia entre las dos muestras en términos de desviaciones estándar. Un valor muy pequeño de Cohen's d sugiere que la diferencia entre las muestras es muy pequeña o insignificante.

7. **BF10 (Factor Bayesiano)** : El valor de BF10 es 0.007. El Factor Bayesiano se utiliza para evaluar la evidencia en favor de la hipótesis alternativa en comparación con la hipótesis nula. Un valor bajo de BF10 (como en este caso) sugiere que la evidencia a favor de la hipótesis alternativa es débil.

8. **power (Potencia estadística)**: La potencia estadística es 0.053953. Indica la probabilidad de detectar una diferencia significativa si realmente existe una. En este caso, la potencia es bastante baja, lo que sugiere que la prueba no tiene suficiente poder para detectar una diferencia significativa.

En resumen, según este resultado, no hay evidencia significativa para rechazar la hipótesis nula, ya que el valor p es alto y el intervalo de confianza contiene el valor cero. Además, tanto el Factor Bayesiano como la potencia estadística sugieren que la evidencia a favor de la hipótesis alternativa es débil.

Pruebas

Análisis de los diferentes componentes de este resultado de la Prueba ANOVA:

1. **Source (Fuente)**: La fuente se denomina "Airport Name". Esto indica que la variable categórica "Airport Name" se ha utilizado como factor o variable independiente en el ANOVA para evaluar si hay diferencias significativas entre los grupos definidos por esta variable.

2. **ddof1 (Grados de libertad entre grupos)**: El valor de ddof1 es 9061. Los grados de libertad entre grupos representan la variabilidad entre los diferentes grupos definidos por la variable "Airport Name". Cuantos más grados de libertad entre grupos haya, más grupos diferentes se están comparando en el ANOVA.

3. **ddof2 (Grados de libertad dentro de los grupos)**: El valor de ddof2 es 89557. Los grados de libertad dentro de los grupos representan la variabilidad dentro de cada uno de los grupos definidos por la variable "Airport Name". Cuantos más grados de libertad dentro de los grupos haya, más observaciones individuales se están considerando en el análisis.

4. **F (Estadístico F)**: El valor de F es 1.012963. El estadístico F es una medida de la variabilidad entre grupos en relación con la variabilidad dentro de los grupos. En este caso, un valor F cercano a 1 sugiere que la variabilidad entre grupos y dentro de los grupos es bastante similar.

5. p-unc (Valor p)**: El valor p es 0.20276. El valor p es la probabilidad de obtener un valor de estadístico F al menos tan extremo como el observado en los datos si la hipótesis nula (generalmente, que no hay diferencias significativas entre los grupos) es verdadera. En este caso, el valor p es mayor que el nivel de significancia comúnmente utilizado de 0.05, lo que sugiere que no hay evidencia suficiente para rechazar la hipótesis nula.

6. **np2 (Eta cuadrado parcial)**: El valor de eta cuadrado parcial es 0.09296. Eta cuadrado parcial es una medida de la varianza explicada por el efecto de la variable independiente (en este caso, "Airport Name"). Un valor de 0.09296 indica que aproximadamente el 9.3% de la variabilidad en la variable dependiente se debe a las diferencias entre los grupos definidos por "Airport Name".

En resumen, según este resultado, no hay evidencia significativa para rechazar la hipótesis nula de que no hay diferencias significativas entre los grupos definidos por "Airport Name", ya que el valor p es mayor que el nivel de significancia comúnmente utilizado. El valor de eta cuadrado parcial sugiere que el efecto de "Airport Name" explica aproximadamente el 9.3% de la variabilidad en la variable dependiente.

Pruebas

Prueba de Correlación:

Resultados:

```
Estadístico de Chi-cuadrado: 1.1168845915660102
Valor p: 0.5720995319792831
```

Análisis de los diferentes componentes de este resultado:

1. **Estadístico de Chi-cuadrado**: El valor del estadístico de chi-cuadrado es 1.1168845915660102. Este estadístico se utiliza para evaluar si existe una asociación significativa entre dos variables categóricas. Un valor mayor indica una mayor discrepancia entre las observaciones y las expectativas bajo la hipótesis nula.
2. **Valor p**: El valor p es 0.5720995319792831. El valor p es la probabilidad de obtener un estadístico de chi-cuadrado igual o más extremo que el observado en los datos si la hipótesis nula (generalmente, que no hay asociación significativa entre las variables) es verdadera.

En resumen:

- El valor estadístico de chi-cuadrado es relativamente bajo, lo que sugiere que no hay una discrepancia significativa entre las observaciones y las expectativas bajo la hipótesis nula.
- El valor p es alto (0.5720995319792831), lo que indica que no hay evidencia suficiente para rechazar la hipótesis nula. En otras palabras, no se encuentra una asociación significativa entre las variables categóricas evaluadas en el análisis.

Resultados

En la presentación de los resultados, se destacó que el modelo de regresión de bosque aleatorio logró una precisión del 85% en la predicción de retrasos de vuelos. Este resultado es significativo, ya que demuestra la eficacia de las técnicas de aprendizaje automático en la resolución de problemas complejos y prácticos como la predicción de retrasos de vuelos.

En cuanto a la conclusión del proyecto, se demostró que el uso de técnicas de aprendizaje automático puede ser una herramienta efectiva para predecir los retrasos de vuelos. Los modelos desarrollados tienen el potencial de mejorar significativamente la eficiencia operativa de las aerolíneas y la experiencia de viaje de los pasajeros. Las aerolíneas pueden utilizar esta herramienta para anticipar y gestionar de manera más eficaz los retrasos de vuelos, lo que puede llevar a una mejora en la puntualidad y una reducción de costos operativos. Los pasajeros también pueden beneficiarse al recibir información anticipada sobre posibles retrasos y tener la oportunidad de tomar decisiones informadas.

En conclusión, según este resultado, no se encontró una asociación significativa entre las variables categóricas en el análisis de chi-cuadrado, ya que el valor p es mayor que el nivel de significancia comúnmente utilizado (por ejemplo, 0.05).

En resumen, este proyecto ha demostrado que, con el uso adecuado de los datos y las técnicas de aprendizaje automático, podemos desarrollar soluciones efectivas para problemas del mundo real.