

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | Flight Delay Predictor | | | | |  | |
|  |  | | | | | | |  |
|  | | | |  |  | | | |
|  | | | | DATA REPORT |  | | | |
|  | | | | Fecha: 09/09/2024—Autor: Juan Martin Morano—Comisión Coderhouse: 61130 |  | | | |
|  | | |  | | |  | | |



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | |  |  | |
|  | | INTRODUCCIÓN | | | | |  | |
|  | |  |  | | |  |  | |
|  | Descripción de Data Acquisition El dataset que concentra la totalidad de los datos requeridos fue obtenido desde la web de Kaggle, proporcionando el enlace a continuación:  <https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations>  Este dataset (que por cuestiones prácticas de ahora en adelante será llamado “Data Frame” por su composición de dos dimensiones) concentra información de los vuelos de una empresa aérea determinada, llevando en cada registro los vuelos efectuados a cada una de las localidades ubicadas por una longitud y latitud detallada. Tiene 26 columnas y 6.489.062 registros, con los atributos con información detallada sobre los mismos, es decir, cada registro representa un vuelo en particular.  El mismo incluye datos operacionales (como número de asientos y número de vuelos), datos meteorológicos (como precipitación y temperatura), y datos específicos del vuelo (como retrasos y aeropuertos involucrados).  Según se analizó y se estudio acerca del mismo, la demora de la columna 'DEP\_DEL15' indica si el vuelo tuvo un retraso de salida de 15 minutos o más (1 = sí, 0 = no). Se formalizará un mapa para su graficación. | | | | | | |  |
|  |  | | |  |  | | |  |



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | |  |  | |
|  | | CONTEXTO COMERCIAL | | | | |  | |
|  | |  |  | | |  |  | |
|  | *Descripción del Negocio y Punto de Dolor* La importancia de este trabajo radica plenamente en objetivos comerciales.  Conocer si un vuelo saldrá demorado o no permite trabajar y/o modificar y mejorar la planificación operativa, atendiendo a que, el poder atender esos puntos específicos de la operación normal del negocio permitirá disminuir efectos nocivos en la satisfacción de los clientes para con la aerolínea y también las perdidas por devoluciones (para aquellos vuelos que tengan esta opción).  Atento a esto, se hará un mejor uso sobre la empleabilidad de los recursos, en términos de eficiencia operativa, así como también una mejor cobertura sobre el área aeroportuaria con menores concentraciones de gente permitiendo mejorar la planificación operativa.  Esto también podrá devenir en mejoras de las relaciones con empresas comerciales asociadas (tales como Despegar y otras), disminuyendo KPI´s claves de reclamos y bajando niveles de insatisfacción en sus respectivas métricas. *Objetivo* Este proyecto utiliza técnicas de análisis de datos y modelos de machine learning para predecir mediante un análisis de Clasificación para predecir si un vuelo saldrá o no demorado en función de los atributos presentados. El modelo fue entrenado con un conjunto de datos públicos y se emplean diversas técnicas de ingeniería de características y optimización mediante ingeniería de atributos y modificación de hiperparámetros. | | | | | | |  |
|  |  | | |  |  | | |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  | | |  |  | |
|  | | PRE-EXPLORACIÓN DE DATOS | | | | |  | |
|  | |  |  | | |  |  | |
|  | *Conformación de Atributos y su Descripción* Analizando el dataframe, no encontramos nulos y este mismo concentra un porcentaje muy bajo de registros duplicados (0,44% en porcentaje y 28.473 en valores absolutos). Siendo que este porcentaje es demasiado bajo y que su eliminación no afectará la integridad del mismo, se debe verificar con el administrador de las bases porque el mismo está llegando duplicado:  ¿Efectivamente es otra persona que ocupa con exactitud los mismos atributos? ¿O proviene mal de las fuentes de datos por algún inconveniente con el pipeline?  Continuando con esta primera aproximación, este dataset tiene únicamente cuatro variables categóricas y las veintidós restantes (salvo algunas que son numéricas pero que, a efectos prácticos consideraremos discretas por tener muy pocas variaciones en sus valores) son variables continuas.  *Establecimiento de Supuestos*  Siendo el objetivo predecir si un vuelo tendrá o no una demora de 15 minutos o más ('DEP\_DEL15') en base a los demás atributos mediante un modelo de clasificación (árbol de decisión o algun otro algoritmo), podemos asumir en primera instancia que este varie en función de:  Datos Operacionales: Número de vuelos concurrentes, número de vuelos de la aerolínea en el aeropuerto por mes, aeropuerto previo y el bloque de tiempo en el que el vuelo está programado para salir. Estas variables pueden ser consideradas dentro de las controlables por la compañía, siendo fundamental su comprensión y conocimiento.  Datos Climáticos: Precipitación, nieve y velocidad promedio del viento.  Otros Datos Específicos del Vuelo: Edad del avión.  Siguiendo estos supuestos, establecemos la posibilidad de estructurar el modelo de clasificación con una performance entre 85% y 90% de accuracy SIN optimización y entre 90% - incluido - y 95% CON optimización. Esto considerando el Benchmark del 81% dado por el desbalance de clases.  Se deja a continuación la descripción de atributos (también disponible desde el código de la ipynb):   1. MONTH: El mes en el que se realiza el vuelo. 2. DAY\_OF\_WEEK: El día de la semana en que se realiza el vuelo (donde 1 podría representar lunes, 2 martes, etc.). 3. DEP\_DEL15: Indica si el vuelo tuvo un retraso de salida de 15 minutos o más (1 = sí, 0 = no). 4. DEP\_TIME\_BLK: El bloque de tiempo en el que el vuelo está programado para salir (por ejemplo, mañana, tarde, noche). 5. DISTANCE\_GROUP: Un grupo categórico que indica la distancia del vuelo, posiblemente en rangos de millas o kilómetros. 6. SEGMENT\_NUMBER: Número de segmento del vuelo en una ruta específica. 7. CONCURRENT\_FLIGHTS: El número de vuelos que se realizan concurrentemente. 8. NUMBER\_OF\_SEATS: El número de asientos disponibles en el vuelo. 9. CARRIER\_NAME: El nombre de la aerolínea que opera el vuelo. 10. AIRPORT\_FLIGHTS\_MONTH: El número total de vuelos que operan en el aeropuerto durante el mes. 11. AIRLINE\_FLIGHTS\_MONTH: El número total de vuelos operados por la aerolínea durante el mes. 12. AIRLINE\_AIRPORT\_FLIGHTS\_MONTH: El número de vuelos de la aerolínea en un aeropuerto específico durante el mes. 13. AVG\_MONTHLY\_PASS\_AIRPORT: El promedio mensual de pasajeros que pasan por el aeropuerto. 14. AVG\_MONTHLY\_PASS\_AIRLINE: El promedio mensual de pasajeros que vuelan con la aerolínea. 15. FLT\_ATTENDANTS\_PER\_PASS: El número de asistentes de vuelo por pasajero. 16. GROUND\_SERV\_PER\_PASS: El número de personal de servicio en tierra por pasajero. 17. PLANE\_AGE: La edad del avión en años. 18. DEPARTING\_AIRPORT: El código o nombre del aeropuerto de salida. 19. LATITUDE: La latitud del aeropuerto de salida. 20. LONGITUDE: La longitud del aeropuerto de salida. 21. PREVIOUS\_AIRPORT: El código o nombre del aeropuerto anterior al actual en la ruta del vuelo. 22. PRCP: La cantidad de precipitación en el día del vuelo. 23. SNOW: La cantidad de nieve en el día del vuelo. 24. SNWD: La profundidad de la nieve en el día del vuelo. 25. TMAX: La temperatura máxima en el día del vuelo. 26. AWND: La velocidad del viento promedio.   *Funciones Agregadas y su Análisis*  Las variables “FLT\_ATTENDANTS\_PER\_PASS” y “GROUND\_SERV\_PER\_PASS” tienen valores por debajo de cero, pero no son nulos, sino que contienen hasta dos o tres decimales por ser el personal sobre la cantidad de viajantes en el vuelo. En caso de utilizarse a futuro se realizará un cambio de  Se verifican grandes desvíos sobre el promedio de las variables, lo que nos induce a pensar que hay una gran variabilidad en los datos y es algo que hay que tener en cuenta para el modelo. | | | | | | |  |
|  |  | | |  |  | | |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  | |
|  | | ANALISIS EXPLORATORIO DE DATOS | | |  | |
|  | |  |  |  |  | |
|  | A continuación, se expone por cada tipo de gráfico las conclusiones encontradas. Estas valoraciones se realizaron en función de las necesidades de obtención de variables que contribuyan a obtener una primera buena performance denotada en varios modelos distintos, en función de los objetivos de predicción (tanto precisión como generalización).CONCLUSIONES DE LOS GRÁFICOS DE BARRAS:  1. Se puede apreciar por el gráfico del top 10 de aerolíneas que hay una variación significativa en la sumatoria de demoras entre aerolíneas que, como era esperado, algunas cumplen con una mejor calidad de servicio que otras. Esto no implica que no se produzcan incidentes en el aeroparque, sino que debemos añadir de forma nominal esta variable. Es una variable importante y podría ser tomada para el modelo, pero dependemos fuertemente de los recursos que se tenga, ya que contiene 17 valores definidos como únicos. 2. Los horarios con mayor demora cubren una tendencia ascendente durante el día, cubriendo la mayoría de retrasos los horarios de tarde a noche. Esta variable categórica también es sumamente importante. 3. Según el gráfico "Cantidad de Kilómetros/Millas con Mayores Demoras", mientras menores sean las millas recorridas mayor es el número de demoras. Se podría apreciar la posibilidad de que las aerolíneas proveen una mayor relevancia a los vuelos más largos, por un tema de tiempos, esto se debería hablar con el gerente comercial de cada una de las aerolíneas. Sin embargo, estas diferencias (siguiendo el promedio por KM de cantidad de demoras y la cantidad de KM recorridos) por el vuelo, no hay diferencias significativas (tiende a parecer uniforme salvo por pequeñas variaciones). Por esto último, se descartaría del modelo. 4. Estacionalmente notamos, por el gráfico de media de demora mensual, que hay una variación no tan significativa en cuanto a las demoras de vuelos y una tendencia mayor a producirse en los meses de entre el tercer y cuarto trimestre. Este comportamiento de la variable puede verse dado por las vacaciones que se producen en países como Estados Unidos o mismo las vacaciones de verano de Argentina en los primeros meses del año, aumentando para el mes 12 (en fechas festivas como navidad o año nuevo). Sin embargo, estas variaciones no son tan claras por lo que, 5. No se aprecia una diferencia significativa en cuanto al día de la semana. Por lo cual se evaluará eliminar esta variable del entrenamiento o entrenamientos del modelo.  CONCLUSIONES DE LOS GRÁFICOS DE HISTOGRAMAS:  1. Distribución de la variable CONCURRENT\_FLIGHTS:   Sesgo a la derecha (asimetría positiva): La mayor parte de los datos se concentra en los valores más bajos de CONCURRENT\_FLIGHTS, entre 0 y 20, lo que sugiere que los vuelos concurrentes más comunes en tu conjunto de datos son relativamente bajos. A medida que aumentan los vuelos concurrentes, la frecuencia disminuye rápidamente.  Cola larga hacia la derecha: Hay una cola que se extiende a lo largo del eje X, hasta un valor máximo de aproximadamente 100 vuelos concurrentes. Esto indica que hay algunos aeropuertos o períodos con un número más alto de vuelos concurrentes, pero son menos frecuentes. La edad de los aviones repercute en la cantidad de demoras que hubo. Tiene algunas diferencias cuando se realiza el análisis con filtros entre con y sin demoras, lo que le da más relevancia.  Es una variable que se podría someter a una transformación logarítmica para el modelo.   1. Por Precipitaciones podemos visualizar una distribución Log Normal. Parece tener un cierto desbalance en cuanto a demoras por precipitaciones entre 0.1 y 0.5. Es otra candidata a una transformación logarítmica, con el fin de suavizar un poco los datos y que sea aún más funcional al algoritmo, siempre y cuando se deba optimizar el modelo. Se puede visualizar mejor cuando se filtra por la variable target. 2. La variable de velocidad de viento (AWND) tiene más un parecido a una distribución normal. Si bien tiene dos picos, su forma parece aproximarse. Debe tenerse en cuenta su forma Mesocúrtica que, junto con la asimetría claramente positiva (presencia a priori de outliers), tiene mucha varianza sobre sus datos. 3. La edad del avión tiene un pico que podría dar una muy poca aproximación a una distribución normal, pero sus valores siendo una variable discreta, tiene mucha variancia y asimetría tanto positiva como negativa. Se debe estudiar a fondo su incorporación al modelo.  CONCLUSIONES DEL ANALISIS DE BOXPLOT:Encontramos algunos valores outliers en algunas variables de interés. Estos se concentran mayormente en las variables climáticas como la de precipitaciones, cantidad y profundidad de la nieve en el día del vuelo.Lo mismo sucede sobre el número de vuelos de la aerolínea en un aeropuerto específico durante el mes, dentro de las variables operativas. *Se debe analizar que tratamiento a fin de cuentas se le dará, por el momento se tomará la media de esas variables.* CONCLUSIONES DEL ANALISIS DE DISPERSIÓN:Para empezar, se analizaron por arriba mediante código PairPlot todos los gráficos en su conjunto, con el objetivo de obtener una visualización rápida de las relaciones entre dos variables, con puntos que representen a la variable target de “demora de vuelo” (DEP\_DEL15).Con respecto a las variables continuas Operativas encontramos las siguientes conclusiones: Hay variables con una correlación que tienen una relación bastante importante que van desde las operativas, como:  ("CONCURRENT\_FLIGHTS","AIRPORT\_FLIGHTS\_MONTH"), ("CONCURRENT\_FLIGHTS","AIRLINE\_AIRPORT\_FLIGHTS\_MONTH"),  ("CONCURRENT\_FLIGHTS","AVG\_MONTHLY\_PASS\_AIRPORT"), ("AIRPORT\_FLIGHTS\_MONTH","AIRLINE\_AIRPORT\_FLIGHTS\_MONTH"),  ("AIRPORT\_FLIGHTS\_MONTH","AVG\_MONTHLY\_PASS\_AIRPORT"), ("AIRLINE\_AIRPORT\_FLIGHTS\_MONTH","AVG\_MONTHLY\_PASS\_AIRPORT").  ("FLT\_ATTENDANTS\_PER\_PASS","GROUND\_SERV\_PER\_PASS"), ("NUMBER\_OF\_SEATS","AIRLINE\_FLIGHTS\_MONTH"),  ("NUMBER\_OF\_SEATS","GROUND\_SERV\_PER\_PASS")  Si bien no tienen una relación de puntos tan clara en relación a que los puntos de demora a modo general se encuentran mayormente en el centro de los ejes, frente a una superposición de los mismos, ahí pueden haber insights que el modelo adopte en su predicción.  Con respecto a las variables climáticas sucede lo contrario, pero podríamos considerarlo normal, ya que frente a altas variaciones en los datos climáticos que no se dan en condiciones “normales”, suceden una serie de eventos que terminan por infringir o alterar las condiciones y terminar en una demora. CONCLUSIONES DEL ANALISIS DE CORRELACIONES: Análisis de las *correlaciones positivas*:  CONCURRENT\_FLIGHTS - AIRPORT\_FLIGHTS\_MONTH (0.85): Esta correlación alta sugiere que el número de vuelos concurrentes (aquellos que ocurren al mismo tiempo) está fuertemente relacionado con el total de vuelos mensuales en un aeropuerto. Esto es intuitivo, ya que en aeropuertos con mayor tráfico mensual, es lógico que haya más vuelos ocurriendo simultáneamente.  CONCURRENT\_FLIGHTS - AIRLINE\_AIRPORT\_FLIGHTS\_MONTH (0.58): Aunque más débil que la correlación anterior, sigue siendo significativa. A medida que aumenta el número de vuelos de una aerolínea específica en un aeropuerto, también aumenta el número de vuelos concurrentes, lo que sugiere que las aerolíneas con mayor presencia en un aeropuerto contribuyen a la congestión en la cantidad de vuelos simultáneos.  CONCURRENT\_FLIGHTS - AVG\_MONTHLY\_PASS\_AIRPORT (0.85): Los aeropuertos con un mayor promedio de pasajeros mensuales tienden a tener más vuelos concurrentes. Esto tiene sentido, ya que más pasajeros probablemente impliquen más vuelos operando al mismo tiempo.  AIRPORT\_FLIGHTS\_MONTH - AIRLINE\_AIRPORT\_FLIGHTS\_MONTH (0.65): A medida que aumentan los vuelos mensuales de un aeropuerto, también aumentan los vuelos de una aerolínea específica en ese aeropuerto. Esto indica que las aerolíneas con una mayor presencia tienden a contribuir significativamente al tráfico global del aeropuerto.  AIRPORT\_FLIGHTS\_MONTH - AVG\_MONTHLY\_PASS\_AIRPORT (0.97): Esta es la correlación más alta de todas, lo que sugiere una relación extremadamente fuerte entre el número de vuelos mensuales de un aeropuerto y el número promedio de pasajeros. Esto es esperable, ya que más vuelos generalmente implican un mayor número de pasajeros.  AIRLINE\_AIRPORT\_FLIGHTS\_MONTH - AVG\_MONTHLY\_PASS\_AIRPORT (0.64): A medida que aumentan los vuelos mensuales de una aerolínea específica, también tiende a aumentar el número promedio de pasajeros en el aeropuerto. Sin embargo, no es una correlación tan fuerte como la del tráfico total del aeropuerto, lo que indica que los pasajeros no necesariamente están distribuidos de manera uniforme entre todas las aerolíneas.  FLT\_ATTENDANTS\_PER\_PASS - GROUND\_SERV\_PER\_PASS (0.48): La correlación moderada sugiere que cuando hay más asistentes de vuelo por pasajero, también hay un mayor número de servicios en tierra por pasajero. Esto puede reflejar aerolíneas que ofrecen un mejor servicio o mayor personal para vuelos de alto tráfico.  NUMBER OF SEATS - AIRLINE\_FLIGHTS\_MONTH (0.31): Aunque la correlación es baja, sugiere que a medida que el número de asientos de una aerolínea aumenta, también lo hace la cantidad de vuelos mensuales de esa aerolínea. Esto podría indicar que aerolíneas con flotas más grandes tienden a volar con más frecuencia.  NUMBER OF SEATS - GROUND\_SERV\_PER\_PASS (0.34): Una correlación baja que indica que un mayor número de asientos está débilmente asociado con más servicios de tierra por pasajero. Las aeronaves más grandes pueden requerir más apoyo logístico, aunque no es una relación muy fuerte.  PLANE AGE - FLT\_ATTENDANTS\_PER\_PASS (0.20): Una correlación muy débil, lo que sugiere que el aumento de la edad de un avión tiene poca relación con la cantidad de asistentes de vuelo por pasajero.  Análisis de las *correlaciones negativas*:  LATITUDE - TMAX (-0.36): Hay una correlación negativa moderada entre la latitud y la temperatura máxima, lo que indica que a medida que aumenta la latitud (se avanza hacia el norte en el hemisferio norte), la temperatura máxima tiende a disminuir, lo cual es lógico desde una perspectiva geográfica.  SNWD (snow depth - profundidad de nieve) - TMAX (-0.28): Esta correlación negativa sugiere que cuando la temperatura máxima es más alta, la profundidad de la nieve tiende a ser menor. Esto es razonable, ya que temperaturas más altas tienden a derretir la nieve.  SNOW - SNWD (0.28): Aunque es una correlación positiva moderada, sugiere que a medida que cae más nieve, la profundidad de la nieve tiende a aumentar, aunque no es una relación perfecta (lo que puede depender de otros factores, como el derretimiento o la compactación).  TMAX - AWND (-0.17): Esta correlación negativa muy baja sugiere que cuando la temperatura máxima es más alta, la velocidad del viento promedio tiende a disminuir ligeramente. Sin embargo, es una relación muy débil y probablemente no sea muy significativa.  Conclusión general de las correlaciones:  Las correlaciones positivas fuertes sugieren relaciones predecibles y esperadas entre variables de tráfico aéreo, como el número de vuelos y pasajeros, que están fuertemente conectadas. Estas variables pueden ser útiles para modelar fenómenos relacionados con el tráfico en aeropuertos.  Las correlaciones negativas también son lógicas, especialmente en variables climáticas como la temperatura y la nieve, aunque no son tan fuertes como las positivas.  Las correlaciones débiles indican relaciones que pueden no ser muy útiles por sí solas para el análisis o modelado. CONCLUSIONES DE ESTA PRIMERA PARTE A partir de todos estos estudios de las variables, reducimos la exploración analítica de datos a dos puntos importantes:   1. De las variables categóricas, las más importantes son: la aerolínea en particular que tiene la demora en particular, con sus eficiencias y deficiencias correspondientes, y el horario determinado del vuelo, recordemos con una tendencia ascendente hacia la tarde noche. La variable de estacionalidad (por mes) tiene alguna variación entre distintos trimestres del año, pero es muy poca, por el momento, como para asumirla variable para ingresar al modelo. 2. De las variables operacionales y climáticas hay varias correlacionadas (más en las operacionales que en las climáticas). La mayoría tiene una distribución log-normal a normal, salvo la de Temperatura Máxima que tiene más una distribución con asimetría por izquierda. Las demoras se suelen concentrar más en el centro de los datos para las operativas que para las climáticas, que están más hacia afuera. Este diferencial de comportamiento, junto con la variable de “Edad del Avión”, nos permitirá concluir el feature selection al día de hoy. | | | | |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  | |
|  | | FEATURE SELECTING AND ENGINEERING | | |  | |
|  | |  |  |  |  | |
|  | *Variables seleccionadas para el modelo* Las siguientes variables fueron las seleccionadas:  VARIABLES OPERACIONALES:  - CONCURRENT\_FLIGHTS  - AIRLINE\_AIRPORT\_FLIGHTS\_MONTH  - NUMBER\_OF\_SEATS  VARIABLES CLIMÁTICAS:  - AWND  - PRCP  - TMAX  - SNOW  - SNWD  OTRAS VARIABLES:  - CARRIER\_NAME  - DEP\_TIME\_BLK  - MONTH  - PLANE\_AGE *Variables Transformadas mediante Normalización* Las siguientes variables fueron transformadas:  - TMAX: Normalización mediante aplicación de la Raíz Cuadrada.  - AIRLINE\_AIRPORT\_FLIGHTS\_MONTH: Normalización mediante aplicación de Logaritmo.  - CONCURRENT\_FLIGHTS: Normalización mediante aplicación de Logaritmo. | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  | |
|  | | ENTRENAMIENTO Y EV. DE MODELOS | | |  | |
|  | |  |  |  |  | |
|  | *Entrenamiento de Modelos de Machine Learning Aplicados* *Objetivo basado en Trade-OFF sobre el Accuracy:* 86% a 90%.  MODELO DT01  ***Breve Descripción***  En esta instancia se pasan a entrenar varios modelos entre los que se incluyen los modelos simples, tales como son los “Árboles de Decisión”, que toman decisiones en función de “nodos de decisión” (y sus correspondientes ramas o bifurcaciones). Respecto del método de selección de primeras y subsiguientes variables, el método aplicado será el que toma el argumento por default que es el del coeficiente de “Gini”.  ***Optimización***  Respecto a su optimización se utilizó la siguiente grilla:   * criterion=[‘entropy’]: Limitar a una opción reducirá el número de combinaciones. * max\_depth=[20, 30]: Limitar la profundidad del árbol es clave para evitar árboles muy grandes, lo que reduce el tiempo de entrenamiento. Estas profundidades menores también tenderán a generalizar mejor en este caso. * min\_samples\_split=[20]: Un número más alto para la división mínima en los nodos nos ayudará a prevenir que el árbol se vuelva muy detallado, lo que acelerará el proceso. * min\_samples\_leaf=[10]: Fijamos un valor relativamente alto para el número mínimo de muestras en las hojas también limita el crecimiento del árbol y acelera el entrenamiento.   Los resultados no fueron los óptimos, manteniendo el mismo accuracy que el baseline.  ***Evaluación del Modelo***   1. Modelo SIN Optimización: En un primer modelo sin optimización de hiperparámetros, este modelo presento una performance bastante similar al baseline (81% de desbalance aproximado de ahora en adelante). El área bajo la curva nos dejo un valor de 53,22 lo que nos indica que no es muy bueno discriminando entre las clases positivas y negativas. 2. Modelo CON Optimización: No hubo cambios relevantes tanto en el Clasification Report como en el Área Bajo la Curva.   MODELO RF01  ***Breve Descripción***  Similar a los Árboles de Decisión, este modelo trabaja con multiples árboles de decisión que, por medio de una votación de manera independiente entre sí, se llega a una decisión sobre la clase a predecir. Este modelo, al igual que el anterior, tiene una performance similar al baseline.  ***Optimización***  Por motivos de falta de recursos este modelo no tendrá un intento de optimización desde sus hiperparámetros.  ***Evaluación del Modelo***   1. Modelo SIN Optimización: En un primer modelo sin optimización de hiperparámetros, este modelo presento una performance bastante similar al baseline (81% de desbalance aproximado de ahora en adelante). El área bajo la curva nos dejo un valor de 53,22 lo que nos indica que no es muy bueno discriminando entre las clases positivas y negativas. 2. Modelo CON Optimización: No hubo cambios relevantes tanto en el Clasification Report como en el Área Bajo la Curva.   MODELO KN01  ***Breve Descripción***  En este modelo se trabaja con una matriz de distancias con centroides que se mueven por el espacio en función de la cercanía con los datos. Para poder tomar un mismo nivel de peso sobre las variables elegidas (y que sus valores nominales no apliquen un desbalance en el peso del algoritmo) se decide escalar los datos mediante el “MinMaxScaler” (con un rango de dominio entre 0 y 1).  ***Optimización***  Por motivos de falta de recursos este modelo no tendrá un intento de optimización desde sus hiperparámetros.  ***Evaluación del Modelo***  El modelo no alcanza los estándares propuestos quedando en un 81%.  MODELO BN01  ***Breve Descripción***  Adaptive Boosting es un algoritmo de ensamble que combina varios clasificadores débiles, como árboles de decisión simples, para mejorar el rendimiento. Funciona ajustando iterativamente los pesos de las muestras mal clasificadas, enfocándose más en los errores de iteraciones anteriores.  ***Optimización***  No se prefirió optar por optimizar este modelo.  ***Evaluación del Modelo***  El modelo no alcanza los estándares propuestos quedando en un 81%.  MODELO BG01  ***Breve Descripción***  Extreme Gradient Boosting es una implementación optimizada de gradient boosting que usa regularización y técnicas avanzadas para mejorar la precisión y la eficiencia. Es altamente escalable y es conocido por su rendimiento superior en competiciones de machine learning.  ***Optimización***  Respecto a su optimización se utilizó la siguiente grilla:   * 'n\_estimators'=[150, 300]: Número de árboles en el modelo. Más árboles pueden mejorar la precisión, pero también aumentan el tiempo de entrenamiento y el riesgo de sobreajuste. En este caso, se prueban 150 y 300 árboles. * 'max\_depth'=[3, 5]: Limitar la profundidad del árbol es clave para evitar árboles muy grandes, lo que reduce el tiempo de entrenamiento. Estas profundidades menores también tenderán a generalizar mejor en este caso. Aquí se prueban profundidades de 3 y 5. * 'learning\_rate'=[0.01, 0.05]: Tasa de aprendizaje o "step size" en cada iteración. Valores más pequeños hacen el entrenamiento más lento pero pueden mejorar la precisión. Se prueban 0.01 y 0.05. * 'subsample'=[1]: Porcentaje de muestras utilizadas para entrenar cada árbol. Evita el sobreajuste reduciendo la correlación entre árboles. Se usa el 100% de los datos. * 'colsample\_bytree'=[0.4, 0.5]: Proporción de características seleccionadas al construir cada árbol. Limita el número de columnas (features) para evitar el sobreajuste. Aquí se prueban 0.4 y 0.5. * 'min\_child\_weight'=[1, 3]: Peso mínimo de una hoja. Controla la cantidad mínima de muestras que debe tener una hoja para dividirse. Valores más grandes tienden a hacer los árboles más conservadores. Se prueban 1 y 3.   ***Evaluación del Modelo***   1. Modelo SIN Optimización: El resultado nos deja un 5% más arriba en el accuracy, lo que de forma inicial es bastante importante. Representa cerca de 250 mil vuelos de los cuales ya se puede obtener un beneficio inicial, y obtuvo muy buenas métricas de precisión para la clasificación de los vuelos sin demoras, lo que permitirá a la empresa continuar con sus operaciones (manteniendo un plan de acción de contingencia acorde a la probabilidad que nos otorga el algoritmo). 2. Modelo CON Optimización: Dado que no se alcanzan los estandares esperados en el Best Model (dado que tampoco hay representatividad), se toma el modelo con hiperparámetros calculados con el fin de realizar una mejor estimación de los datos. Los pocos valores para entrenar el Best Model con el Paramgrid durante la optimización tiene que ver con la pequeña proporción de datos para training. Lamentablemente, no se cuentan con los recursos necesarios para tomar más contundencia y poder tomar el 80% que se tomó originalmente, con vistas a realizar más pruebas y tratar de llevar al 90% a 95% por lo menos la performance sobre el accuracy.   CONCLUSIONES FINALES: Dados todos estos modelos entrenados, se decide tomar como funcional el modelo nombrado como “XG01” dada su performance actual de Accuracy del 86%. | | | | |