Vuelos Domésticos 2021

Rodriguez Gerini, Santiago
Rutsztein, Micaela

Abstract
Los viajes y el turismo componen un tema prioritario tanto para la economía como para los intereses sociales en el país. Es por esto, que se decidió analizar acerca de este tema para que se pueda continuar mejorando el servicio ofrecido y, en consecuencia, la experiencia de quien viaja. Se realizó la predicción del comportamiento de los distintos tipos de vuelos con respecto a la cantidad de pasajeros, a través de un modelo de Machine Learning, más específicamente la Regresión por medio de Aprendizaje Supervisado. Para llegar a un buen resultado, se llevó a cabo previamente un Análisis Exploratorio de los Datos contenidos en el Data Set en cuestión, para poder investigarlo y conocer los comportamientos, la calidad y la consistencia de los datos que lo componen**.**

I. INTRODUCCIÓN

El objetivo de este estudio, fue poder determinar el comportamiento de la demanda de viajes aéreos dentro del país, principalmente de los vuelos regulares, según los distintos factores a analizar. Esto podría ser de gran utilidad para que se pueda planificar con anterioridad la infraestructura y personal requeridos en las distintas situaciones, y reflejar así una disminución en los costos finales.

II. DATA SET

Para realizar el informe se utilizó el siguiente dataset, el cual contiene los detalles correspondientes a los vuelos realizados en Argentina a lo largo de 2021, según fuentes del gobierno.

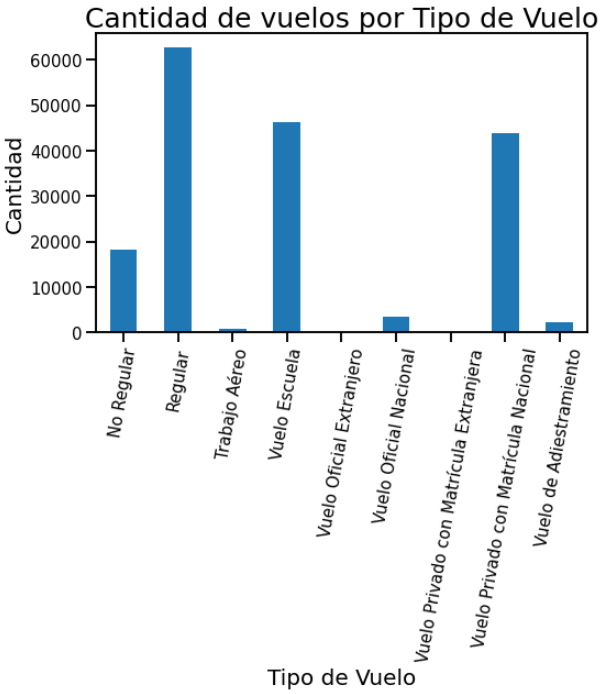
Detalle del link:
https://www.datos.gob.ar/dataset/transporte-aterrizajes-despegues-procesados-por-administracion-nacional-aviacion-civil-anac

Este dataset posee una dimensión inicial de 219.803 filas o muestras.

Previo a cualquier análisis o investigación a realizar sobre estos datos, se debió realizar un preprocesamiento de los mismos, obteniendo un dataframe sin valores nulos y con todos ellos con el formato más claro y útil posible. De esta manera, se obtuvo, finalmente, un total de 177.204 filas y 14 columnas.

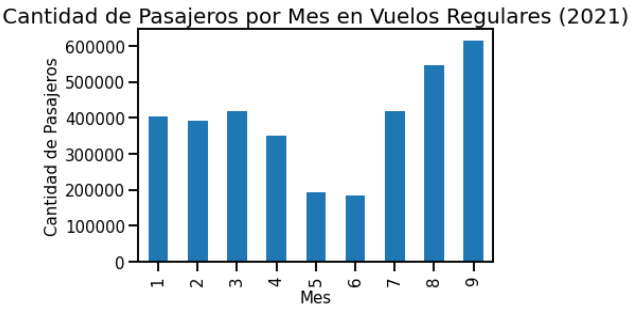
III. ANÁLISIS EXPLORATORIO DE DATOS

Aloritmos utilizados: boxplot, plot, outliers con cuantiles, correlación lineal de pearson.
Comenzando por gráficos simples que otorgan claridad y describen las características principales del dataset en cuestión, se obtuvieron los siguientes resultados:



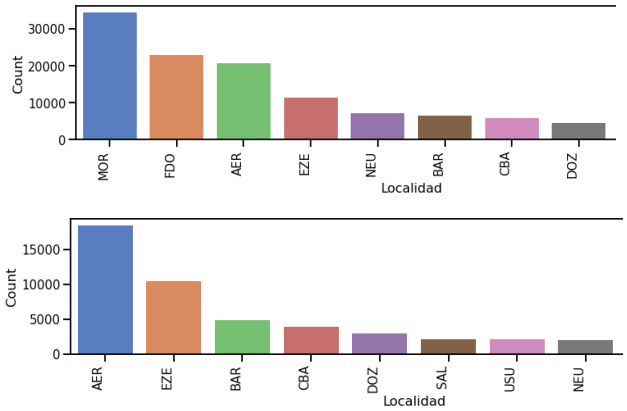
De esta manera, se puede observar que hay una amplia diferencia de cantidad para los 3 tipos de vuelo más frecuentados respecto del resto, y que el tipo con más cantidad es el de “vuelos regulares”.

El dataset contiene datos desde enero hasta septiembre de 2021, es decir, de los 9 primeros meses del año, por lo que se analizó cuáles fueron los meses con mayor cantidad de vuelos de tipo Regular:



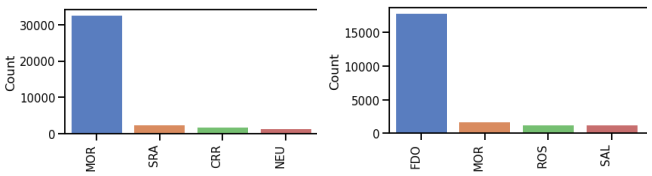
siendo los meses con más vuelos septiembre y agosto respectivamente.

A partir de esto, se quiere conocer cuáles son los aeropuertos en donde más se realizan estos vuelos para el tipo Regular y, además, comparar estos resultados con los obtenidos en total, sin distinguir el tipo de vuelo.



Se obtuvo como resultado que muchos de ellos se repiten en ambos y que Aeroparque (AER) y Ezeiza (EZE), en los cuales obviamente es donde más vuelos se realizan para la clase "Regular", están en tercer y cuarto lugar en la totalidad de vuelos registrados.

Para el caso del aeropuerto de Moreno (MOR) es aquel que tiene la mayor cantidad de "Vuelos Escuela", mientras que el aeropuerto de San Fernando (FDO) es donde se cuenta con la mayor cantidad de "Vuelos Privados con matrícula nacional". Estos dos lugares mencionados son los otros con más cantidad de vuelos para el año 2021.



Análisis de Relacionamiento entre Cantidad de Vuelos y de Pasajeros

Una vez analizado esto, nos centraremos en estudiar, para la clase de vuelo Regular, cómo se relacionan los aeropuertos más importantes (en cuanto a frecuencia de vuelos), con la cantidad de pasajeros que hay en cada uno de ellos. Se creó una tabla Pivot con la cantidad de pasajeros por mes para los 5 aeropuertos más importantes. Los resultados fueron los siguientes:

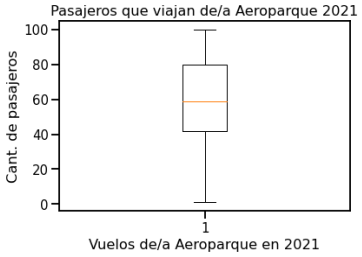
Los primeros 2 meses y medio Aeroparque, el aeropuerto con mayor cantidad de vuelos, permaneció cerrado por reformas. Esto trajo como consecuencia que, indefectiblemente, los vuelos que originalmente se hubiesen realizado con origen/destino Aeroparque, se reprogramaron a Ezeiza. De esta manera, los meses 1, 2 y 3, el aeropuerto que superó ampliamente la cantidad de pasajeros fue Ezeiza. Mientras que, en los restantes 6 meses, la amplia diferencia la marcó Aeroparque, siendo éste el aeropuerto que tuvo tanto la mayor cantidad de pasajeros, como también de vuelos de clase Regular, marcando una gran importancia para el análisis en cuestión.

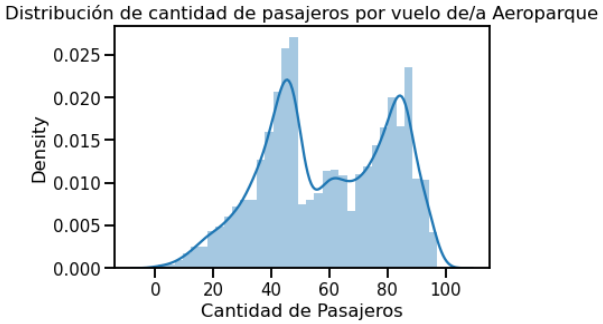
Durante todo el período de análisis (los 9 meses en cuestión), Bariloche fue el aeropuerto que sumó la segunda mayor cantidad de pasajeros. Esto significa que podría existir una relación entre la cantidad de vuelos y la cantidad de pasajeros por vuelo, ya que coinciden los 3 aeropuertos con más vuelos con los que tienen más pasajeros.

Por la gran importancia de Aeroparque mencionada anteriormente, se procedió a analizar este aeropuerto en particular.

Análisis para Aeroparque

Se llevó a cabo un estudio para analizar el comportamiento de las cantidades de pasajeros durante el año 2021 para este aeropuerto en particular. Para esto, se realizó, por un lado, un histograma con la distribución de los mismos y, por otro, un boxplot, a través de los cuales se puede concluir que la mayor cantidad de vuelos realizados a este aeropuerto tuvieron entre 40 y 80 pasajeros.



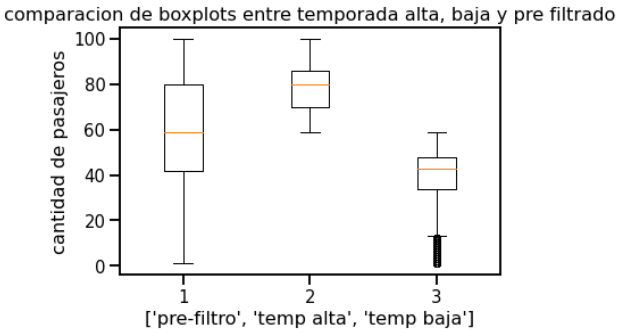


Análisis a través de Outliers

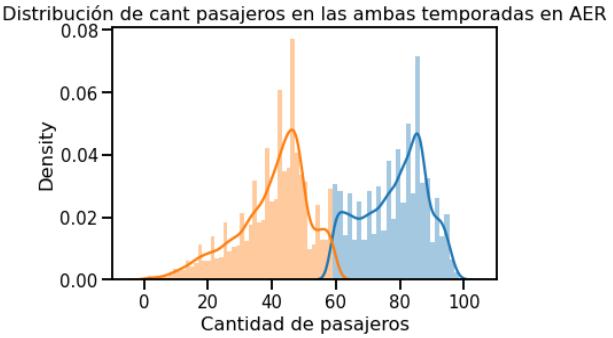
Los Outliers son una herramienta importante para detectar valores atípicos, es decir que es numéricamente distante del resto de los datos. Se debe decidir criteriosamente si conviene o no eliminarlos, según el posterior análisis a realizar.

En este caso, denominamos "Temporada Alta" a aquellos vuelos con una cantidad de pasajeros por sobre el cuantil 0.5, y "Temporada Baja" a aquellos vuelos con una cantidad por debajo.

Se obtuvo la siguiente distribución:

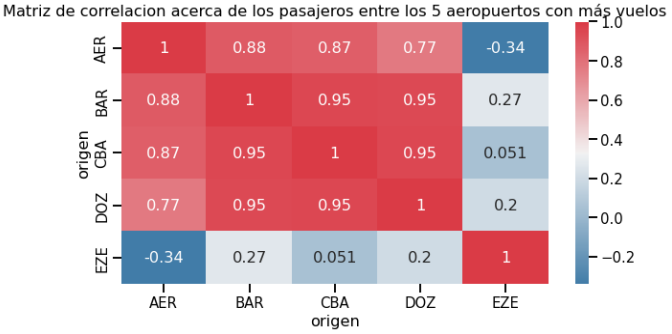


Si comparamos los boxplots de "temporada alta" y "temporada baja" con aquel del pre-filtrado, podemos observar que ambos presentan una gran disminución de la dispersión de los datos, con el primero centrado en la media de 78 pasajeros y el segundo centrado en una media de 40. Ambos coinciden con los picos de pasajeros que presenta la distribución de Aeroparque con los datos pre-filtrados.



Relación entre aeropuertos con mayor cantidad de vuelos

Finalizando el análisis exploratorio de datos, se procedió a realizar un análisis de correlación acerca de la cantidad de vuelos a lo largo del año 2021 para los aeropuertos de Bariloche, Córdoba, Mendoza, Ezeiza y Aeroparque.



A través del gráfico Heat Map, se puede observar que existe una relación estrecha en la gran mayoría de las combinaciones, destacándose la correlación entre CBA-BAR, CBA-DOZ y BAR-DOZ con un R = 0,95. Esto significa que los aeropuertos varían de forma similar respecto a la cantidad de vuelos mensuales a lo largo del año, respetando las variaciones de la temporada.

El mayor desentono se marca en el aeropuerto de Ezeiza, que no tiene correlación con ningún otro aeropuerto. Esto sucede debido a que EZE tuvo que tomar los vuelos "regulares" provenientes de AER para los primeros meses del año. Cuando AER retomó su actividad normal, EZE disminuyó considerablemente la cantidad de vuelos "regulares".

IV. APLICACIÓN DE MACHINE LEARNING

El tipo de aprendizaje que nos pareció más adecuado fue el Aprendizaje Supervisado. Esto se debe a que, una vez conocido el dataset y determinado el objetivo del trabajo, definimos que se quiere utilizar estos datos como datos de entrada para que el algoritmo aprenda, comparando su resultado real con los "correctos" introducidos.

Una vez definido esto, se decidió realizar Regresión, ya que no se trata de variables categóricas. Sin embargo, tampoco son totalmente variables continuas, ya que son números enteros (cantidad de pasajeros), pero al resultado de la predicción se puede tomar el número entero más próximo, por lo que lo consideramos el proceso más adecuado.

Se realizaron, entonces, 3 modelos de regresión distintos, para poder comparar sus resultados y elegir, así, el más preciso: Linear Regression, Ridge Regression y Support Vector Regression.

Linear Regression

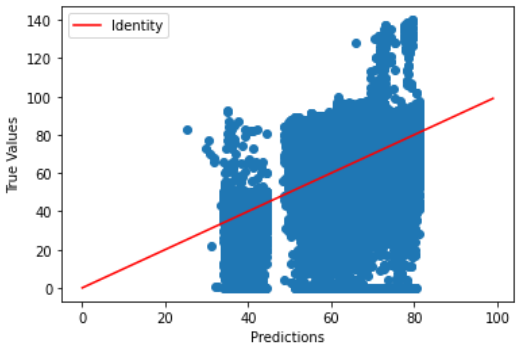
Para realizar este método, se dividieron los datos entre Features y Target. La variable target será la que se quiere predecir, en este caso, "Pasajeros"; y las variables Features serán las que condicionan los pasajeros: mes, aeronave, origen y destino.

Cómo estas últimas 3 variables son strings, se procedió a realizar un Label Encoder para cada una de ellas, obteniendo así un número que representa cada tipo de aeronave, y para cada aeropuerto correspondiente al origen y el destino.

Luego, se procede a entrenar los datos con un 60% de test y, finalmente, se obtiene como salida una función lineal, que intenta asemejarse lo máximo posible a los valores reales y, de esta manera, disminuir el error.

$$\hat{y} = f(x, w)$$

El resultado fue el siguiente:



Support Vector Regression

Este modelo consiste en buscar el hiperplano que maximiza el margen, obteniendo de output un número real. Para esto, se ingresa una cantidad significativa de opciones para cada hiper parámetro, para que encuentre la combinación que tendrá el mejor resultado.

En este caso, el mejor resultado fue:

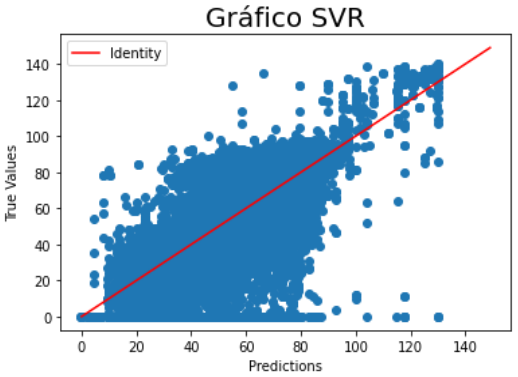
```
SVR(C=100, epsilon=10, gamma=0.1, max_iter=25000)

{'C': 100, 'epsilon': 10, 'gamma': 0.1}

-176.63057192763154
```

(la imagen muestra el mejor estimador, con los valores de los hiperparámetros 'C', 'epsilon' y 'gamma' seleccionados).

El resultado fue:



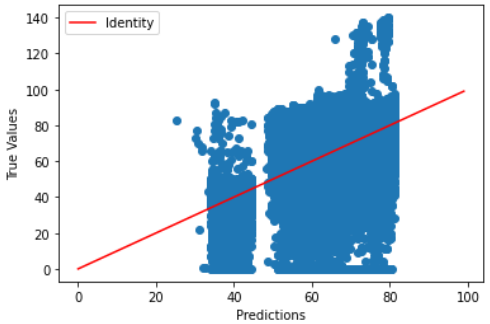
Como se puede ver, las predicciones en este caso se asemejan más a los valores reales que en el caso de Linear Regression.

Ridge Regression

Este modelo consiste en utilizar un hiperparámetro (λ) llamado L2, el cual es elegido para penalizar los parámetros "w".

Cuanto mayor sea la penalización, más parámetros del vector 'w' se aproximarán a cero.

El resultado, en este caso, fue similar al de Linear Regression:



V. RESULTADOS

Una vez que aplicamos los tres modelos lineales, se observaron los resultados finales para cada uno y se

determinó el modelo más conveniente para realizar las predicciones.

Las variables a analizar para la comparación fueron:

R2 (con valores entre 0 y 1): explica la proporción de la varianza de "y" que explica el modelo de regresión.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

MSE: Mean Squared Error (error cuadrático medio)

$$\boldsymbol{MSE} = \frac{\sum (\hat{y}_t - y_t)^2}{n}$$

MAE: Mean Average Error (Media del error)

$$\boldsymbol{MAE} = \frac{|\sum (\hat{y}_t - y_t)|}{n}$$

ambos, para calcular el error (el MSE, al ser cuadrático, permite sumar en términos absolutos, siendo sensible a predicciones muy malas).

El resultado de la comparación fue el siguiente:

	Model	Features	R2	MSE	MAE
0	LR	Lineal	0.349	364.472	14.668
1	SVR	Linear	0.688	174.577	9.407
2	Ridge	Lineal	0.349	364.470	14.668

Realizando el análisis de los datos, podemos indicar que el modelo que mejor los explica es el SVR, debido a que si comparamos los resultados finales de los modelos analizados, es aquel que tiene un R2 de ampliamente mayor magnitud, además que el quien tiene un menor valor tanto de MSE como de MAE.

VI. PREDICCIONES

Realizando las predicciones, se tomaron de imput dos combinaciones distintas de mes-aeronave-origen-destino.

Para los dos ejemplos siguientes:

Mes	Aeronave	Origen	Destino
1	BO-B737-800	DOZ	EZE
9	BO-B737-8LP	BAR	AER

Se comparan Datos Reales vs Predicciones:

Mes	Aeronave	Orig	Dest	Datos Reales	Predicciones
1	BO-B737-800	DOZ	EZE	75	71
9	BO-B737-8LP	BAR	AER	87	73

VII. CONCLUSIONES

Para este informe se utilizó el método de Machine Learning para poder predecir la cantidad de personas que realizarán un viaje "regular" en la Argentina. Se pudo observar que la gran mayoría de los aeropuertos del país conservaron una fuerte relación en cuanto a la variación de vuelos mensuales, salvo el caso de Ezeiza debido al cierre de Aeroparque en los meses de enero y febrero.

En cuanto a las predicciones, se utilizó el método de SVR debido a que es aquel que explica mejor la variabilidad de los datos, al compararlo con los métodos de Linear Regression y Ridge Regression.

Se quiere resaltar que el año en el que se redacta el informe se trata de uno atípico debido a la pandemia por Covid-19 y que algunos de los datos provenientes del dataset pudieron haber sido afectados (por ejemplo, disminución de vuelos o pasajeros por el virus). Es probable que para futuros informes relacionados a esta temática, se puedan obtener datos y predicciones con mayor precisión. Aún así, creemos que los resultados obtenidos de las predicciones son acordes a lo pretendido.

Cabe destacar que no se tomaron como dataset informes de años anteriores ya que éstos no tenían el dato de los pasajeros por vuelo, haciendo imposible el análisis en los casos anteriores

VIII. REFERENCIAS

[1] Amat, C., Michalski, T., & Stoltz, G. (2018). Fundamentals and exchange rate forecastability with simple machine learning methods. Journal of International Money and Finance, 88, 1 - 24.

[2] Drucker, H., Burges, C. J., Kaufman L., Smola, A. J., & Vapnik, V. (1997). *Support vector regression machines*. *In Advances in neural information processing systems* (pp. 155-161).

[3] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.