Escola Tècnica Superior d'Enginyeria ETSE-UV

Máster Universitario en Ciencia de Datos

Trabajo de Fin de Máster

Extracción y predicción de datos de series temporales de reservas de vuelo

AUTOR:
HECTOR MIRETE BLANCO

TUTORES:
JORDI MUÑOZ MARÍ
VALERO LAPARRA PEREZ-MUELAS

SEPTIEMBRE, 2019

VNIVERSITAT DE VALÈNCIA Escola Tècnica Superior d'Enginyeria

Máster Universitario en Ciencia de Datos

Trabajo de Fin de Máster

Extracción y predicción de datos de series temporales de reservas de vuelo

Autor:

Héctor Mirete Blanco

Tutores:

Jordi Muñoz Marí

Valero Laparra Perez-Muelas

Tribunal:

Presidente/Presidenta:

Vocal 1:

Vocal 2:

Fecha de defensa:

Calificación:

Abstract

The following project comes from a request of perform an *origin & destination* analysis of the total traffic of passengers inside one airport in order to analyze the flow of the tourists at the airport with a 3 month forecast of the future movement of the passengers.

The dataset used is build with data of flight bookings with the whole itinerary. The data is presented in a structured format and the information contained is transactional.

To perform the analysis requested, it has been necessary to apply a transformation over the original dataset. Once this transformation has been done, this data has been aggregated, extracting the needed time series. Once this has been done, it has been applied the Prophet algorithm [1] of Facebook to each of the time series, tuning the needed parameters to achieve a good forecast.

In addition, the results of the extracted data have been represented over plots in an interactive dashboard, to visualize the data and extract conclusiones over it. The dashboard was exported as an html file with the data in it, so it was possible to be sent to the client as final result.

As result of the predictive models, the best scores of all the models were between an 8% and a 22% of error, measuring it with the *MAPE* metric [2].

Key words: tourism, flight bookings, time series, machine learning, interactive dashboard.

Resumen

El siguiente proyecto surge de una solicitud de realizar un análisis de tipo *origin – destination* sobre movimientos de pasajeros en un aeropuerto para analizar el flujo de turistas dentro del aeropuerto junto a una estimación de los movimientos de los pasajeros dentro del aeropuerto predicho a 3 meses.

El conjunto de datos a partir del cual se ha realizado el proyecto contiene información de reservas de vuelos junto a todo el itinerario del viaje, es estructurado y de tipo transaccional.

Para poder realizar el análisis requerido ha sido necesario aplicar una transformación sobre el conjunto de datos original. Una vez aplicada la transformación, se han extraído las series temporales necesarias del mismo. Para cada una de las series temporales, se ha optimizado un modelo de estadístico basado en el algoritmo de Prophet [1] de Facebook para lograr una buena predicción.

A su vez, se ha realizado una implementación de un cuadro de mando interactivo para visualizar los datos y poder extraer conclusiones sobre los datos, exportable como un html con los datos incrustados para poder entregarlo al cliente como resultado final.

Como resultados de los modelos predictivos, se han obtenido resultados entre el 8% y 22% de error haciendo uso de la métrica de evaluación *MAPE* [2].

Palabras clave: turismo, reservas de vuelo, series temporales, aprendizaje máquina, cuadro de mando interactivo.

Índice de contenido

Índice de figuras

1

Introducción

El sector del turismo es el área de la economía que opera sobre el movimiento de personas durante un intervalo de tiempo generalmente reducido, hacia un destino diferente al lugar donde viven habitualmente.

Parte de este sector se nutre de los desplazamientos por medios de aeronaves pudiendo impactar genéricamente en los diferentes comercios de la ubicación destino. No obstante, también existen negocios que beben de las estancias de corta duración mientras se realiza algún tipo de transbordo en una localización.

Este es el caso particular de los negocios ubicados en los aeropuertos. Por lo general, estos negocios suelen ser comercios libres de impuestos, tasas locales y nacionales, siendo así un objeto de interés de este sector.

Para la realización de un análisis que ofrezca información de interés sobre turistas que puedan ser potenciales clientes de estos negocios, existe un tipo de análisis en el sector denominado *Origin – Destination.* El propósito de este análisis es aislar y detectar turistas que van a realizar estancias de mayor o menor duración en el aeropuerto, pudiendo así ser objeto y foco de interés para estos comercios.

A su vez, existen organizaciones de ámbito global que se dedican a recoger y almacenar información transaccional de reservas de vuelo. Estas organizaciones son denominadas GDS[1], sistemas de distribución global, donde las agencias de viajes y algunas aerolíneas registran información de sus reservas de vuelo.

[1] Global Distribution Systems: sistema de red informatizado y operado por una compañía que facilita el intercambio transaccional entre proveedores de servicio de la industria del turismo, mayormente aerolíneas, hoteles, compañías de alquiler de vehículos y agencias de viajes.

2

Motivación y objetivos del proyecto

El proyecto desarrollado fue bajo petición de un cliente de la empresa donde se realizó el proyecto. El cliente solicitó conocer el volumen de turistas que llegaban a su aeropuerto en función de las procedencias de los viajeros. La empresa donde se realizó el proyecto contaba con conjuntos de datos de reservas de vuelos facilitados por diferentes GDS´s como fuentes de datos, ingresando una cantidad aproximada 17 millones de reservas de vuelo diarias.

El cliente en cuestión quería un análisis turístico de tipo *Origin and Destination*. Este tipo de análisis consiste en contabilizar los turistas según la estancia en el segmento del viaje, buscando discriminar a los turistas que realizan un transbordo en el aeropuerto de los que realizan un inicio / fin de viaje o una estancia en la ciudad del aeropuerto y permite agregar a aquellos que hacen estancia en el origen y destino indicado indistintamente de si se ha realizado un vuelo directo o uno con *n* escalas. Esta información se solicitó desagregada por un total de 15 países proveedores de turistas y según la direccionalidad del vuelo.

El detalle del análisis solicitado se descomponía en los siguientes filtrados ya sean:

- Vuelos entre el mercado y el aeropuerto con origen en el aeropuerto realizando estancia tanto en el mercado como en el aeropuerto. A estos los denominaremos O&D traffic origin POI[2]

[2] Point of interest: punto o localización de interés

- Vuelos entre el mercado y el aeropuerto con origen en el mercado realizando estancia tanto en el mercado como en el aeropuerto. A estos los denominaremos O&D traffic origin market.

- Vuelos entre el mercado y el aeropuerto realizando estancia tanto en el mercado como en el aeropuerto, pero agnóstico a la direccionalidad, siendo igual al total de los dos previos. A estos los denominaremos O&D traffic.

- Vuelos entre el mercado y el aeropuerto que realizan un transbordo en el aeropuerto donde la estancia previa ha sido el mercado y la próxima estancia no será el mercado. A estos los denominaremos Transfer traffic origin market.

- Vuelos entre el mercado y el aeropuerto que realizan un transbordo en el aeropuerto donde la estancia previa no ha sido el mercado y la próxima estancia será el mercado. A estos los denominaremos Transfer traffic origin others.

- Vuelos entre el mercado y el aeropuerto que realizan un transbordo en el aeropuerto, pero agnóstico de la direccionalidad, siendo igual al total de los dos previos. A estos los denominaremos Transfer traffic.

Junto a estos requisitos, también se realizó una petición para conocer también el comportamiento estimado de los próximos 3 meses para cada una de las diferentes combinaciones solicitadas.

Por parte de la empresa, ésta disponía de información de reservas de vuelo las cuales contienen datos del itinerario del viaje reservado, así como información de la operación de la reserva.

Con todo esto, se plantearon los siguientes objetivos para la realización correcta del proyecto:

1. Desarrollar un proceso de ETL[3] para manipular la información de la empresa de forma que pudiera solucionar las inquietudes del cliente.
2. Estudiar y poner en producción un modelo de predicción de series temporales.
3. Implementar un cuadro de control interactivo que facilite el análisis de los datos y un archivo en formato tabular con los datos para el cliente

[3] Extract, transform and load: proceso de manipulación de datos cuya finalidad es trasladar un conjunto de datos, aplicarles algún tipo de transformación y cargarlos en sistema destino.

3

Estado del arte

3.1. Aprendizaje automático

El *aprendizaje máquina* o *machine learning* [3] [4] es una rama de la inteligencia artificial la cual consiste en una aplicación de la estadística haciendo uso de algoritmos y modelos estadísticos para automatizar la búsqueda de patrones. Su finalidad es crear procesos informáticos que sin estar programados explícitamente para aprender tenga la capacidad de hacerlo a partir de ejemplos dados. Por ello, esta área de conocimiento busca la experimentación y creación de algoritmos que puedan cumplir este propósito a partir de conjuntos de datos que servirán de muestra para que el algoritmo aprenda o se entrene para realizar predicciones futuras sobre nuevas muestras de datos.

Esta rama de conocimientos va de la mano de diferentes campos, en ocasiones superpuestos y entre los que se encuentran la optimización matemática, la analítica de datos, la estadística computacional y la minería de datos.

Para poder aplicar técnicas de machine learning hay que partir del planteamiento de un problema. Este debe de ir ligado a un conjunto de datos que represente una muestra representativa de una población. Sobre este conjunto de datos se debe buscar conocer patrones o propiedades según la pauta de los datos. Estos conjuntos de datos deben estar formados por diferentes muestras y que cada una de estas contenga una o múltiples variables o características.

Dentro del aprendizaje automático existen muchos métodos para diferentes tipos de problemas a resolver. Algunos de los más populares son los problemas de datos de aprendizaje supervisado, no supervisado y aprendizaje por refuerzo.

El aprendizaje supervisado consiste en hacer uso de conjuntos de datos ya etiquetados para intentar predecir nuevas etiquetas de nuevas entradas no procesadas previamente. Dentro de este método, existe una subdivisión en problemas de clasificación y problemas de regresión. En los problemas de clasificación, los valores de las etiquetas son discretos, pudiendo tomar solo *n* valores de posibles clases ya definidos. En los problemas de regresión el espacio de las etiquetas es continuo, por lo que los valores posibles de las etiquetas no están acotados.

El aprendizaje no supervisado consiste en hacer uso de conjuntos de datos que carecen de estas etiquetas. Este tipo de métodos tienen un propósito exploratorio, buscando patrones que existan dentro del conjunto de datos.

El aprendizaje por refuerzo es un método que se basan en ensayo y error, haciendo que el algoritmo mejore su resultado mediante acciones que otorgan recompensas. Generalmente se utiliza para problemas de juegos y robótica.

3.2. Series Temporales

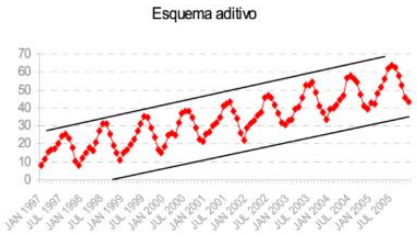
Las series temporales [5] son un listado de diferentes puntos de valores para una variable ordenados cronológicamente. Generalmente, las diferentes entradas de la secuencia de valores de la serie temporal están espaciados equitativamente.

Las series temporales se clasifican en dos tipos, estacionarias y no estacionarias. Las primeras se caracterizan por presentar una media y una variabilidad constantes a lo largo del tiempo. Las segundas no cumplen esta característica, en ellas puede variar la tendencia, la varianza o pueden mostrar estacionalidad.

Estos últimos elementos son los componentes de una serie temporal descompuesta: componente estacional, componente del ciclo de tendencia y componente residual.

Las series estacionarias presentan ventajas frente a las no estacionarias, ya que es más fácil realizar predicciones sobre valores futuros y se pueden calcular intervalos de confianza sobre los valores predichos.

Las series temporales pueden presentar diferentes esquemas según la naturaleza de esta, pudiendo tener un esquema aditivo o un esquema multiplicativo como se puede observar en la ilustración 1.



$$X_t = T_t + S_t + E_t$$

$$X_t = T_t * S_t * E_t \quad o \quad X_t = T_t * S_t + E_t$$

Ilustración 1 Ejemplificación de los diferentes tipos de esquema de una serie temporal, Fuente: Series temporales, Anna Martinez-Gavara, Estadística y optimización, Universidad de Valencia

4

Descripción de algoritmos

4.1. Algoritmos de predicción de series temporales

En este apartado se va a explicar el algoritmo utilizado para predecir la serie temporal. Dentro de los algoritmos existentes para predicciones de series temporales, se ha elegido este el algoritmo de Prophet desarrollado por Facebook.

La elección de este algoritmo de modelado de series temporales se debe a la robustez de este algoritmo frente a la modelación a los eventos móviles (aquellos cuya fecha de celebración varía año a año), ya que las series temporales de movimientos de pasajeros se ven influidas directamente por estos factores.

Por otro lado, el algoritmo ofrece como parámetro de entrada los "puntos de cambio" de la serie temporal para así ajustar mejor los cambios de tendencia, ya sea como cantidad de puntos que debe ajustar el algoritmo o como valor temporal donde la tendencia va a sufrir un cambio. Esta característica es interesante para series temporales de movimientos aéreos, ya que permite indicar al modelo cuando suceden situaciones excepcionales que tenga impacto sobre la tendencia, como la creación y eliminación de rutas aéreas.

4.1.1. Prophet

El algoritmo utilizado para modelar las series temporales y realizar las predicciones ha sido el Prophet [1] desarrollado por Facebook con implementación en R y Python[4].

[4] Implementación del algoritmo: (https://facebook.github.io/prophet/).

El algoritmo trata la serie temporal descomponiéndose en 3 elementos principales: tendencia, estacionalidad y vacaciones. Como resultado, el modelo se explica matemáticamente con la siguiente ecuación:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Ecuación 1 Fórmula del cálculo de la serie temporal del Prophet

Donde *g(t)* es la tendencia de la función que modela cambios no periódicos en la serie temporal, *s(t)* recoge los cambios en la serie temporal con cierta periodicidad, *h(t)* recoge el impacto de las festividades que suceden de forma irregular y donde *e* representa al error.

El cálculo de la tendencia se entiende cómo:

$$g(t) = \frac{C}{1 + \exp(-k(t - m))}$$

Ecuación 2 Fórmula de la tendencia

Dónde C es la capacidad, *k* es un ratio de crecimiento y donde *m* un valor de compensación. No obstante, para el cálculo de la tendencia, el Prophet añade al modelo una definición explícita de ciertos "changepoints" donde el ratio de crecimiento de la tendencia puede variar.

Para el cálculo de la estacionalidad, el algoritmo hace uso de series de Fourier para dar lugar a un modelo flexible a cambios periódicos:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

Ecuación 3 Fórmula transformada de Fourier

Dónde *P* es el periodo de la serie temporal.

Para el cálculo del impacto de las vacaciones, el modelo bebe de una lista dada de festividades, a partir de la cual genera una matriz de regresores a ajustar por el modelo.

Esto se hace incluyendo al modelo información externa sobre los eventos o festividades. Dada una festividad *i,* el conjunto de datos externo lo conforman diferentes entradas de D_i, indicando fechas pasadas y futuras para ese evento. La

representación matemática se realiza con una función temporal *h(t)* para el evento en cuestión con un parámetro k_i para cada evento introducido.

$$h(t) = Z(t)\boldsymbol{\kappa}.$$

$$Z(t) = [\mathbf{1}(t \in D_1), \ldots, \mathbf{1}(t \in D_L)]$$

Ecuación 4 Cálculo del regresor de holidays del algoritmo de Prophet. Fuente: S.J. Taylor, B. Letham, "Forecasting at scale", September, 2017, PeeerJ Preprints

4.2. Técnicas de validación:

En este apartado se va a explicar el método de evaluación del modelo de predicción de series temporales, el cual ha sido la validación cruzada o cross validation.

4.2.1 Validación cruzada:

La *validación cruzada* o *cross-validation* [6] es una técnica utilizada en aprendizaje automático para ajustar los hiper parámetros de los modelos predictores. El método de validación cruzada consiste en particionar el conjunto de datos original en 2 subconjuntos, un subconjunto para entrenamiento del modelo y un subconjunto para validación. Una vez entrenado y validado el modelo con cada subconjunto, se realiza iterativamente nuevas particiones en nuevos subconjuntos para así asegurar que el modelo estadístico generaliza bien sus resultados.

Una de las variantes más populares de este método de validación se denomina *k-fold.* Esta variante consiste en subdividir el conjunto de datos original en *k* subconjuntos aleatorios, y se realizan *k* iteraciones de validación, tomando un subconjunto de los generados como subconjunto de validación y todos los demás como subconjunto de entrenamiento.

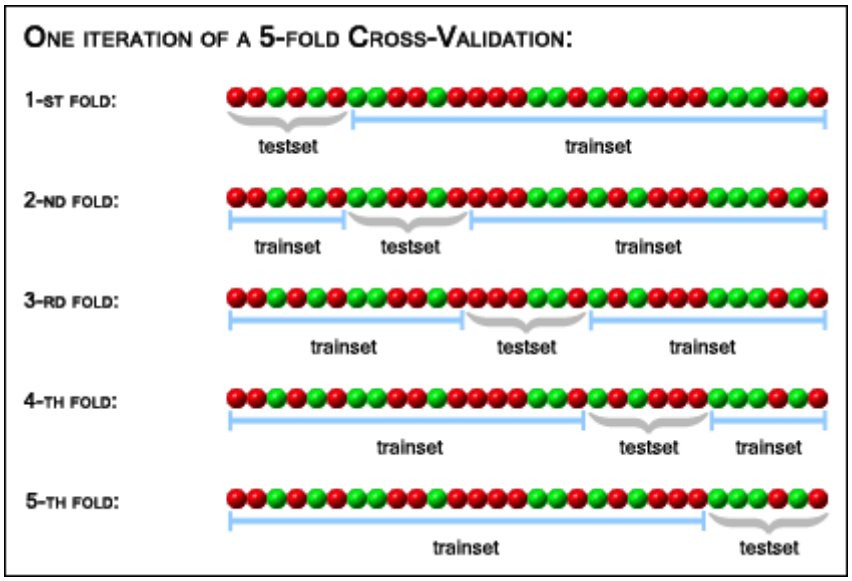


Ilustración 2 Ejemplo de validación cruzada usando k-fold con k=5. Fuente: Cross-Validation Explained, Institute for Genomics and Bioinformatics - FH JOANNEUM - Graz University of Applied Sciences

No obstante, la validación cruzada sobre métodos de predicción de series temporales no funciona exactamente igual debido a la dependencia temporal entre los datos. Por ello, para realizar la validación cruzada con modelos de predicción de series temporales [7] se hace uso de 3 parámetros: un umbral de la ventana inicial *w*, un periodo *p*, y un horizonte *h*. El umbral de la ventana inicial indica en la primera iteración, en qué posición de la serie temporal se corta el conjunto de entrenamiento y el conjunto de validación. El periodo indica cuantos intervalos se debe desplazar el umbral de la ventana inicial para la siguiente iteración. El horizonte indica cuantas unidades temporales debe incluir el conjunto de validación (generalmente en este parámetro se suele fijar la cantidad de intervalos que se desean predecir).

La implementación del algoritmo seria la siguiente, donde *N* es la cantidad de entradas de la serie temporal, *DS* es la serie temporal, y *M* es el modelo estadístico

Mientras que (*w* + *h*) <= *N*:

Entrenamiento = *DS* [hasta *w* - 1]

Validación = *DS* [desde *w* hasta (*w* + *h*)]

Entrenar(M, Entrenamiento)

Validar(M, Validación)

w = *w* + *p*

Habitualmente, cuando se valida el modelo en cada iteración, se aplican el cálculo de diferentes métricas de error, para así poder extraer conclusiones relativas a la calidad de las predicciones del modelo y la capacidad de generalizar del mismo.

El funcionamiento iterativo de la validación cruzada se representa gráficamente en la ilustración 3. En el caso de la ilustración 3, los parámetros de la validación cruzada *h* y *p* tienen el mismo valor, *h=p.*

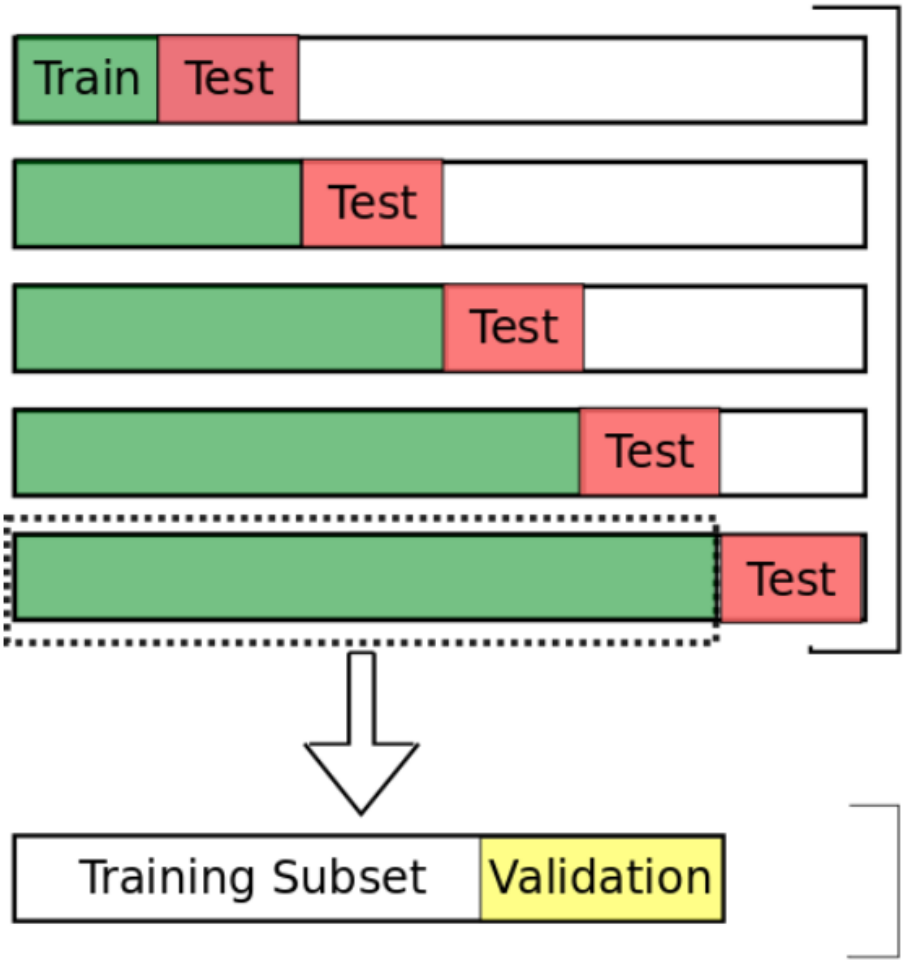


Ilustración 3 Ejemplo de validación cruzada para series temporales. Fuente: Time Series Nested Cross-Validation, Courtney Cochrane, Towards Data Science.

4.3. Métricas de evaluación

Las métricas del cálculo del error basadas únicamente en e_t son dependientes de la escala del error, por ello mismo, no se pueden utilizar como métrica comparativa entre secuencias con diferentes escalas. Los métodos más populares del cálculo del

error dependiente de la escala son el MAE[5] y el RMSE[6]. El MAE es más popular ya que es fácil de calcular y de comprender frente el RMSE que es más difícil de interpretar. Las predicciones que minimicen el MAE provocan un ajuste del modelo hacia la mediana de los datos, mientras que el RMSE provocará un ajuste hacia la media. Por ello, se ha seleccionado el RMSE como métrica dependiente a la escala.

No obstante, ya que las predicciones que se van a realizar en este proyecto son sobre múltiples series temporales con diferentes escalas, se ha decidido realizar también el cálculo de una métrica de error independiente de la escala.

El MAPE, es una métrica porcentual que normaliza los valores del error a la escala de la serie, y permite hacer una comparativa de las precisiones de las diferentes series. Su cálculo es sencillo y fácil de interpretar como el MAE, con la única diferencia que para el MAPE este valor se normaliza.

A continuación, se van a describir las métricas de evaluación utilizadas en el proyecto, el RMSE y MAPE.

4.3.1 RMSE

Root mean square error o *Raíz del error cuadrático medio* [8] es una métrica de evaluación entre las diferencias de los valores reales de una muestra de datos frente a los valores predichos por un modelo de datos. Es una métrica de evaluación útil para modelos de regresión. La raíz del error cuadrático medio evalúa las desviaciones del modelo frente al valor esperado, calculando la desviación estándar de los residuos permitiendo analizar el error del modelo.

La fórmula que explica el cálculo de la métrica es la siguiente:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Ecuación 5 Fórmula del cálculo de la raíz del error cuadrático medio

Donde *Predicted* es la muestra de los valores predichos, *Actual* es la muestra de los valores originales y *N* es el total de elementos que componen la muestra.

[5] Mean absolute error o error medio absoluto
[6] Root mean square error o Raiz del error cuadrático medio

El valor resultado de la métrica será relativo al error entre los valores reales y los estimados. Este valor puede normalizarse de diferentes maneras:

- Media:

$$\frac{RMSE}{\bar{y}}$$

Ecuación 6 Raíz del error cuadrático medio normalizado con la media

- Diferencia entre máximo y mínimo:

$$\frac{RMSE}{y_{max} - y_{min}}$$

Ecuación 7 Raíz del error cuadrático medio normalizado con la diferencia del máximo y mínimo

- Desviación estándar:

$$\frac{RMSE}{\sigma}$$

Ecuación 8 Raíz del error cuadrático medio normalizado con la desviación estándar

- Rango intercuartílico:

$$\frac{RMSE}{Q1 - Q3}$$

Ecuación 9 Raíz del error cuadrático medio normalizado con el rango intercuartílico

.

4.3.1 MAPE

Mean average percentage error o *error medio absoluto porcentual* [2] es una métrica de evaluación para medir la precisión de métodos de predicción estadística utilizada en problemas de regresión. La métrica se calcula a partir de valores de una muestra de datos originales y sus respectivas predicciones generados por el modelo estadístico. Es una métrica normalizada al ser porcentual, aunque su valor puede salirse del rango 0-1. Valores cercanos al 0 indican un error bajo y valores cercanos al 1 indican un nivel de error importante.

La fórmula para realizar el cálculo de esta métrica es la siguiente:

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Ecuación 10 Fórmula del cálculo del error porcentual medio absoluto

Donde *A* representa los valores actuales, *F* los valores predichos y donde *n* es la cantidad de elementos en la muestra.

5

Descripción del conjunto de datos

El conjunto de datos contiene información de reservas de vuelo, incluyendo datos globales del itinerario del viaje y todas las estancias que se van a realizar. A su vez, dentro del mercado de los datos de movimientos aéreos, el conjunto de datos utilizado representa un sesgo sobre el total de la información. La fracción contenida en el conjunto de datos apunta a reservas de vuelos realizadas a través de agencias de viaje. Esto implica que, dentro del conjunto de datos, aparezcan también itinerarios de viaje más complejos que un simple ida y vuelta a un destino. Los viajes de ida y vuelta suelen realizarse por otros medios de distribución.

Los datos son extraídos de una aplicación web orientada a consultas por punto de interés. Esto quiere decir que se indica un punto geográfico sobre el cual se quiere realizar la consulta de la información del itinerario y el resultado incluye detalles de la reserva para ese segmento junto a información de las estancias previas a el punto dato y las futuras.

Cabe mencionar que, la información dentro del conjunto de datos está organizada de manera que se focaliza la información en las estancias (véase el apartado 5.1.). Es decir, este conjunto de datos incluye información de la fecha de llegada al POI[7] y la duración de la estancia en él junto a información de las estancias previas y posteriores con sus relativas duraciones de la estancia. No obstante, la única fecha que recoge es la fecha de viaje a ese POI y la fecha de reserva

El conjunto de datos utilizado para el desarrollo del proyecto es de tipo estructurado y transaccional. Es estructurado ya que el conjunto de datos contiene un

[7] Point of interest: punto o localización de interés

modelo de datos y esquema definido e idéntico para todas las entradas que contiene. Por otro lado, es transaccional siendo cada entrada contenida dentro del conjunto de datos forma parte de una secuencia de operaciones ordenadas cronológicamente. Esto quiere decir que, dentro este conjunto de datos se incluye el registro de cambios realizados sobre la misma reserva como cancelaciones y nuevas reservas del itinerario.

5.1. Descripción de variables

El conjunto de datos utilizados está formado por estructuras tabulares con un total de 19 variables distintas. En esta sección se va a explicar y describir el significado de todas, e indicar cuáles de ellas han sido necesarias para extraer la información necesaria para el desarrollo del proyecto.

- Pax: Esta variable es de tipo numérica. Incluye información de la cantidad de pasajeros involucrados en la reserva de vuelo. También hace a la par funcionalidad de delta, teniendo valores negativos para las cancelaciones y valores positivos para las nuevas reservas. Es decir, a parte de la información de la cantidad de pasajeros, esta variable incluye información relativa a la transacción, detallando si es una cancelación o una nueva reserva. Esta variable ha sido necesaria para el desarrollo del producto.

- PaxPerBooking: Esta variable es de tipo categórica. Incluye 7 categorías, relativas a la cantidad de pasajeros dentro de la reserva. Esta variable no ha sido necesaria para el desarrollo del proyecto. Las categorías posibles de esta variable son las siguientes:
  - 1 pax: 1 pasajero en la reserva.
  - 2 pax: 2 pasajeros en la reserva.
  - 3 pax: 3 pasajeros en la reserva.
  - 4 pax: 4 pasajeros en la reserva.
  - 5 pax: 5 pasajeros en la reserva.
  - 6-9 pax: de 6 a 9 pasajeros en la reserva.
  - 10 or more pax: 10 o más pasajeros en la reserva.

- BookingDate: Esta variable es de tipo numérica. La información que incluye esta variable es relativa a la fecha de creación de la transacción, siendo esta la fecha de realización o cancelación de la reserva. Esta variable no ha sido necesaria para el desarrollo del proyecto.

- TravelDate: Esta variable es de tipo numérica. La información que incluye esta variable es relativa a la fecha de vuelo de llegada a esa localización. Esta variable ha sido necesaria para el desarrollo del proyecto.

- ArrivalHour: Esta variable es de tipo numérica. La información que incluye esta variable es temporal, siendo relativa a la hora de llegada del vuelo. Esta variable no ha sido necesaria para el desarrollo del proyecto.

- DepartureHour: Esta variable es de tipo numérica. La información que incluye esta variable es temporal, siendo relativa a la hora de salida del vuelo. Esta variable no ha sido necesaria para el desarrollo del proyecto.

- LengthOfTrip: Esta variable es de tipo numérica. La información que incluye esta variable es temporal, siendo la cantidad de días que toma la realización de todo el itinerario del viaje. Esta variable ha sido necesaria para el desarrollo del proyecto.

- TripOrigin: Esta variable es de tipo categórica. La información que incluye esta variable es geográfica, incluyendo el código geográfico correspondiendo al origen del itinerario. Esta variable ha sido necesaria para el desarrollo del proyecto.

- NumInboudGateways: Esta variable es de tipo numérica. La información que incluye esta variable es la cantidad de estancias previas realizadas antes de la llegada al punto de interés. Esta variable no ha sido necesaria para el desarrollo del proyecto.

- InboundGatewaysInfo: Esta variable es una estructura de datos. Incluye información de las estancias previas al punto de interés. La información dentro de esta variable está organizada de tal manera que cada una de las estancias presenta información de la ubicación, la aerolínea con la que se ha viajado y la duración de la estancia en esa ubicación geográfica. A su vez, están ordenadas secuencialmente en el orden que se han visitado las diferentes ubicaciones. La variable presenta la siguiente estructura: Ubicacion1/ Aerolinea1/ DuracionEstancia1: Ubicacion2/ Aerolinea2/ DuracionEstancia2: etc. Esta variable ha sido necesaria para la realización de este proyecto.

- PointOfInterest: Esta variable es de tipo categórica. Incluye información relativa al punto geográfico sobre el cual se ha realizado la consulta, siendo el código geográfico de la ubicación sobre la que se han

consultado datos. Esta variable ha sido necesaria para la realización del proyecto.

- LengthOfStayAtPOI: Esta variable es de tipo categórica y numérica. Incluye información relativa a la duración de la estancia, y puede tomar valores categóricos o numéricos. Esta variable ha sido utilizada para el desarrollo del proyecto. Esta variable toma valores de tipo numéricos si se ha realizado una estancia en la ubicación, detallando la cantidad de días alojados. Esta variable toma valores categóricos en caso de no haberse realizado una estancia en la ubicación y sus posibles valores son los siguientes:

    - Return home: indica que se ha terminado el itinerario y la ubicación final es la misma que la inicial.
    - End of trip: indica que se ha terminado el itinerario y la ubicación final es distinta a la inicial.
    - Short Transfer: indica que se está realizado un transbordo de duración corta en el aeropuerto.
    - Mid Transfer: indica que se está realizando un transbordo de duración media en el aeropuerto.
    - Long Transfer: indica que se está realizando un transbordo de duración larga en el aeropuerto
    - Day Trip: indica que se está realizando un viaje de un día en la ubicación.
    - StopOver: indica que se está realizando una parada técnica en el aeropuerto (generalmente para repostar combustible), pero los pasajeros no cambian de avión.

- TripHighestCabin: Esta variable indica es de tipo categórica. Indica el tipo de asiento de mayor precio dentro del itinerario. Esta variable no ha sido necesaria para el desarrollo del proyecto. Los valores posibles para esta variable son:
    - Economy: clase económica
    - Economy premium: clase económica con beneficios premium.
    - Business: clase empresarial.
    - First: primera clase.

- PaxProfile: Esta variable es de tipo categórica. Indica el tipo de perfil de los pasajeros dentro de la reserva. Esta variable no ha sido necesaria para el desarrollo del proyecto. Los valores posibles para esta variable son:
    - Business: el perfil es de tipo empresarial
    - Leisure: el perfil es de tipo ocio y vacacional.

    - Group: el perfil es de tipo grupo.
    - VFR: el perfil es de tipo visita a familiares.

- DistChannel: Esta variable es de tipo categórica. Indica el tipo de medio de distribución por el cual se ha realizado la reserva. Esta variable no ha sido necesaria para el desarrollo del proyecto. Los valores posibles para esta variable son *Retail, Online, Corporate, Other.*

- AgencyIATA: Esta variable es de tipo categórica. Indica el código IATA[8] de la agencia proveedora del dato. Esta variable no ha sido necesaria para el desarrollo del proyecto.

- AirlinePOI: Esta variable es de tipo categórica. Indica el código de la aerolínea utilizado para transportarse al punto geográfico. Esta variable no ha sido necesaria para el desarrollo del proyecto.

- NumFurtherDestiations: Esta variable es de tipo numérica. Indica la cantidad de estancias posteriores a la ubicación consultada. Esta variable no ha sido necesaria para el desarrollo del proyecto.

- FurtherDestiationsInfo: Esta variable es una estructura de datos. Incluye información de las estancias posteriores al punto de interés. La información dentro de esta variable está organizada de tal manera que cada una de las estancias presenta información de la ubicación, la aerolínea con la que se ha viajado y la duración de la estancia en esa ubicación geográfica. A su vez, están ordenadas secuencialmente en el orden que se han visitado las diferentes ubicaciones. La variable presenta la siguiente estructura: Ubicacion1/ Aerolinea1/ DuracionEstancia1: Ubicacion2/ Aerolinea2/ DuracionEstancia2: etc. Esta variable ha sido necesaria para la realización de este proyecto.

Debido al tipo de análisis que se ha solicitado y la organización de la información en el conjunto de datos, las variables necesarias han sido todas aquellas relacionadas con información de las estancias previas y posteriores con su duración de la estancia, junto a fecha de llegada al POI permitiendo así reconstruir las fechas de los vuelos dentro del itinerario de viaje. Además de las mencionadas, se ha hecho uso de la variable pax, ya que recoge información sobre la cantidad de pasajeros en el vuelo y es, por tanto, la variable sobre la que se realizarán los agregados.

[8] International Air Transport Association: asociación comercial de las compañías aéreas del mundo.

6

Desarrollo

El análisis solicitado por el cliente era un análisis de los datos de tipo *Origin – Destination.* Es decir, el tipo de análisis solicitado busca aislar dentro de los segmentos del itinerario, cuáles son un origen o un destino 'real', realizando una estancia en la localización y cuáles son simplemente puntos intermedios en los que se realiza un transbordo. El propósito del análisis es agregar conjuntamente a los pasajeros que se desplacen a un mercado independientemente de la cantidad de escalas que incluya el itinerario de vuelo. Esto permite contabilizar conjuntamente a los pasajeros que vuelen directos al mercado y a los que hagan *n* escalas entre el origen y el destino.

Partiendo de la descripción del conjunto de datos y sabiendo al análisis que se desea realizar, el formato inicial de los datos no es adecuado para realizar este tipo de análisis ya que la información de las ubicaciones previas y posteriores están encapsuladas dentro de dos variables.

A su vez, la información de la fecha de viaje se tiene solo para la llegada al POI, pero este análisis es independiente de llegadas y salidas sobre el punto de interés, por lo que es necesario también la información de la fecha de vuelo del siguiente segmento del viaje (en caso de existir). Las fechas de vuelo se pueden reconstruir haciendo uso de la duración de las estancias, ya que están ordenadas secuencialmente, y la fecha de llegada al POI.

Además, este conjunto de datos no incluye directamente la información de los orígenes y destinos 'reales'. Es decir, entre cada origen y destino sobre el que se realiza estancia, los pasajeros pueden realizar *n* escalas (habitualmente no más de 3). Por ello, es necesario calcular para cada segmento de vuelo del itinerario los orígenes y destinos 'reales' de los vuelos, y poder así aplicar los filtrados requeridos para el análisis (mencionados en el apartado 2.). La información de los orígenes y

destinos reales se pueden calcular haciendo uso de las estancias visitadas y la duración de la estancia.

Finalmente, es necesaria realizar una transformación por segmento de vuelo entre dos puntos, ya que algunos filtrados (indicados en el apartado 2.) requieren añadir información sobre la direccionalidad del vuelo.

En conclusión, por los argumentos mencionados en esta sección, es necesario aplicar una transformación sobre el conjunto de datos que facilite su manipulación y la extracción de información requerida.

6.1. Transformación de los datos

En este proceso se aplica una transformación sobre los datos para darles un formato más manipulable de cara a las fases posteriores. En el formato inicial, cada transacción incluye información del itinerario íntegramente. Por tanto, se va a desglosar el itinerario en los segmentos de los vuelos que componen el itinerario, dándole un nuevo esquema a los datos.

El nuevo esquema de datos incluirá como resultado las siguientes variables: PaxPerBooking, TripOrigin, CountryOrigin, AirportOrigin, Airline, CountryDestination, AirportDestination, TrueOrigin, TrueDestination, LengthOfStay, Flag, TravelDate.

De forma esquemática, la transformación aplicada se explica con la siguiente figura, desglosando una entrada en *n* entradas según el itinerario

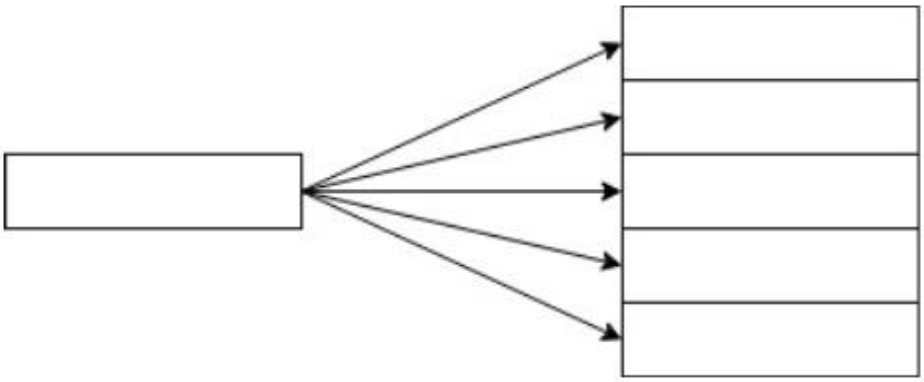


Ilustración 4 Ejemplo desglose de un itinerario de 5 segmentos en sus diferentes vuelos, Fuente: propia

El resultado de la transformación contiene variables calculadas en el proceso a partir del itinerario entero, indicando para el segmento actual, cual es el origen de

su origen y destino reales sobre los que hace una estancia, aunque el segmento actual sea un transbordo. Estas variables son TrueOrigin y True Destination.

A su vez, se añade una variable *Flag* donde se indica si el segmento es un segmento intermedio o es un segmento *Origin-Destination.*

Este proceso de transformación de los datos se ha realizado en *Python* haciendo uso del paquete *multiprocessing* para paralelizar la ejecución del proceso.

6.2. Extracción de las series temporales

En este proceso se extraen las series temporales a partir del nuevo esquema de datos. No obstante, previamente hay que entender el concepto de mercado y punto de interés o POI. El POI representa la localización geográfica sobre la cual se va a realizar el análisis, en este caso particular se trata de un aeropuerto. Los mercados son las diferentes localizaciones que se quieren analizar y su relación del flujo de pasajeros con el POI seleccionado. En este caso particular, los mercados son 15 países distintos.

Para este proceso, se realizan diferentes filtrados en función de los mercados de seleccionados para el análisis. Iterativamente se va filtrando el mercado de interés y se hacen diferentes agregados sobre los datos para extraer las series temporales, ya que los datos que incluye este nuevo conjunto de datos siguen siendo transaccional.

Los diferentes agregados que se realizan son los siguientes:

- Tráfico O&D con origen ubicación de interés: vuelos cuyo True Origin sea la localización de interés y cuyo True Destination sea el mercado.

- Tráfico O&D con origen el mercado: vuelos cuyo True Origin sea el mercado y cuyo True Destination sea la localización de interés.

- Tráfico transfer con origen en el mercado: vuelos cuyo True Origin sea el mercado, su True Destination no sea la localización de interés pero el destino del vuelo sea la localización de interés

- Tráfico transfer con origen otros mercados: vuelos cuyo True Origin no sea el mercado ni la localización de interés, su destino sea la localización de interés, pero esta no sea True Destination del vuelo.

Todas las agregaciones diferentes se realizan también por fecha de vuelo para así generar la serie temporal sobre el total de pasajeros para cada una de las combinaciones de los vuelos.

Como resultado del problema, son necesarias 2 series temporales. No obstante, pueden ser calculadas a partir de estas 4 extraídas (mencionado en el apartado 2.). Por lo que las series temporales restantes serán calculadas a partir de estas una vez obtenido el resultado de las predicciones de la serie temporal, para evitar que estas den resultados no consistentes.

Así pues, la salida de esta fase son 4 series temporales por cada mercado, con un total de 60 series temporales con comportamientos diferentes, sin compartir siempre tendencia, estacionalidad o incluso la escala/volumen de pasajeros.

6.3. Optimización de los parámetros del modelo

El modelo del prophet implementado en R y Python requiere como entrada un dataframe[9] con dos columnas, una con información de las fechas que debe denominarse *ds* (date sequence) y otra que recoja el valor de la variable de la serie temporal denominada *y.*

A su vez, el modelo acepta diversos parámetros de entrada para optimizar el ajuste a la serie temporal y obtener unas predicciones con mayor precisión. Estos pueden ser:

- Growth: el tipo de crecimiento de la tendencia, pudiendo ser *lineal* o *logística.*

- Changepoints: un vector de fechas con las fechas donde se produzca un cambio de tendencia. Sino se da ningún valor el modelo lo ajusta automáticamente.

- Number of changepoints: Si el modelo ajusta automáticamente los puntos de cambio, este parámetro indica la cantidad de puntos de cambio de la serie temporal.

[9] Objeto tabular utilizado en lenguajes de programación para la manipulación de datos estructurados

- Holidays: un dataframe con las festividades, fechas en las que suceden y una ventana o intervalo de posible impacto alrededor a la fecha de la festividad.

- Seasonality mode: modo de la estacionalidad, pudiendo ser aditiva o multiplicativa

- Seasonality prior scale: parámetro que modula la fuerza de la estacionalidad del modelo. Valores grandes ajustan mejor la estacionalidad del modelo y valores bajos tienen el efecto contrario.

- Holiday prior scale: parámetro que modula el impacto de las vacaciones sobre el modelo salvo que se sobreescriba por hacer uso del parámetro Holidays.

- Changepoints prior scale: parámetro que modula la flexibilidad de la selección automática de los puntos de cambio de tendencia. Valores elevados permitirán muchos puntos de cambio y valores bajos sólo unos pocos.

- Interval width: ancho del intervalo de confianza proveído como resultado del modelo.

Para la optimización del modelo, se ha añadido un dataframe externo con información de las vacaciones, eventos o festividades que pudieran influir sobre la serie temporal. Al ser para este caso concreto, series temporales de vuelos que pasaban por Singapur, se han seleccionado como eventos: *new year, chinesse new year, good friday, labour day, vesak day, hari raya puasa, national day, hari raya haji, deepavali, christmas day*.

Esta información es importante añadirla, porque no todos los eventos ocurren con la misma periodicidad sobre la serie temporal, como es el caso del *año nuevo chino*.

Con el dataframe de vacaciones fijo para el modelo, los parámetros que se han optimizado para cada serie temporal han sido: *changepoint prior scale, seasonality prior scale, seasonality mode, number of changepoints.* Todos ellos con un valor de *Growth = "linear"* y el resto de valores por defecto.

Para el *Growth* se ha seleccionado este valor debido a la naturaleza de las series temporales. El parámetro *Changepoints* no ha sido utilizado, ya que es complementario del *number of changepoints* y no se tenía información externa de posibles puntos de cambio de la tendencia. El parámetro *holidays prior scale* se

sobrescribe por la información externa de los eventos, por lo que tampoco tenía sentido ajustarlo. El parámetro del *interval width* no se ha modificado ya que no se iba a hacer uso del intervalo de confianza ofrecido por el modelo.

El grid que se ha explorado, ha sido sobre las siguientes combinaciones de parámetros:

- Change prior scale: 0.001, 0.01, 0.2, 0.5, 0.9
- Seasonality prior scale: 1, 4, 8, 10, 20, 50
- Number of change points: 25, 30, 50, 80, 100
- Growth: lineal
- Seasonality mode: additive, multiplicative

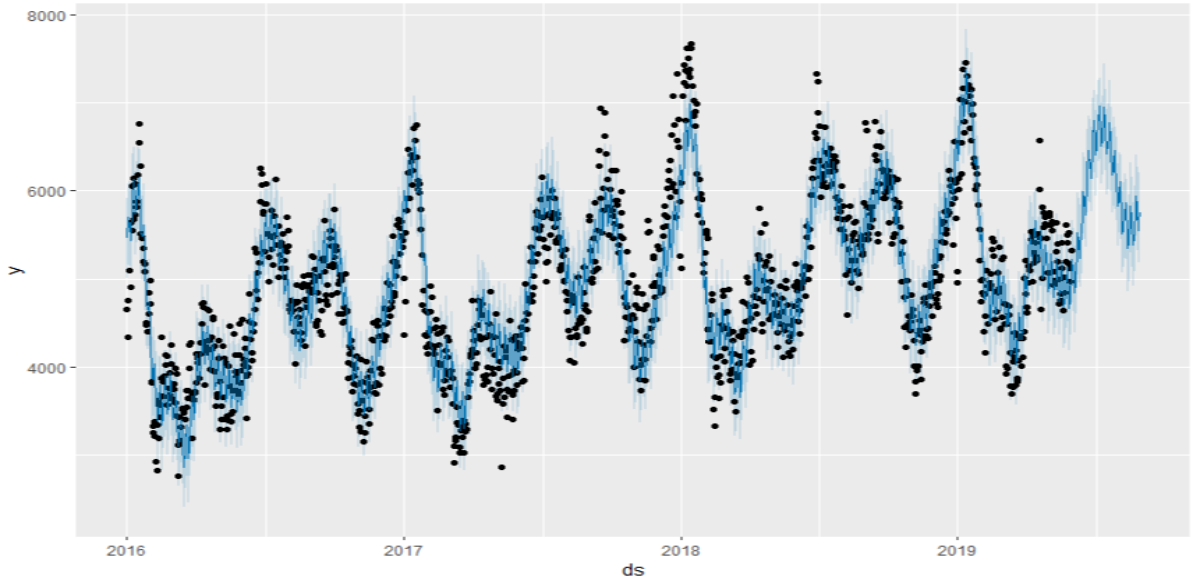


Ilustración 5 Gráfico ejemplo del ajuste del modelo a la serie temporal, Fuente: propia.

Para el ajuste de estos parámetros, se ha optado por aplicar un método de *grid search*, dando una serie de posibles valores para todos los parámetros y calculando todas las combinaciones posibles. Para cada iteración de diferentes combinaciones de los parámetros, se han calculado las métricas *RMSE* y *MAPE*, sobre el cual se buscaba minimizar el *RMSE* y se calculaba el *MAPE.* En caso de no haber mejora en el *RMSE*, se comprobaba si había mejora en el *MAPE*, aunque este principalmente se ha calculado para tener una métrica independiente de la escala del error.

Esto se ha realizado aplicando un método de validación cruzada sobre series temporales. Para ello se ha establecido una ventana inicial, donde todos los datos previos fueran de entrenamiento. Seguido de un horizonte con valor de 90 días, los 3 meses que se desean predecir, y un periodo de separación de días del nuevo conjunto de validación con el valor previo de la ventana inicial.

Para obtener el mejor ajuste, se ha inicializado los resultados del RMSE y MAPE a valores elevados, e iterativamente se enfrentaba el resultado del nuevo ajuste contra le almacenado como mejor resultado. En caso de haber mejora sobre el valor del RMSE, se almacena el nuevo resultado de ambas métricas y los valores de las predicciones obtenidas. En caso de dar un mismo valor de RMSE, se comprueba si hay mejora en el MAPE, e igualmente se almacenan los mismos resultados.

6.4. Visualización de resultados

Para la visualización de los datos y facilitar una herramienta interactiva que facilite el análisis de las diferentes series temporales se realizó un cuadro de control interactivo.

Los datos embebidos en el cuadro de control son los referentes a los últimos meses de la serie temporal con 6 meses de histórico y 3 meses de datos predichos, y se agregan mensualmente

El cuadro de control interactivo se ha realizado en R haciendo uso de Rmarkdown y flexdashboard[10]. Esto permitía que los datos vayan embebidos en el dashboard, pudiendo entregar como resultado un archivo formato html, sin necesidad de un servidor web detrás ejecutando una aplicación para cargar los datos.

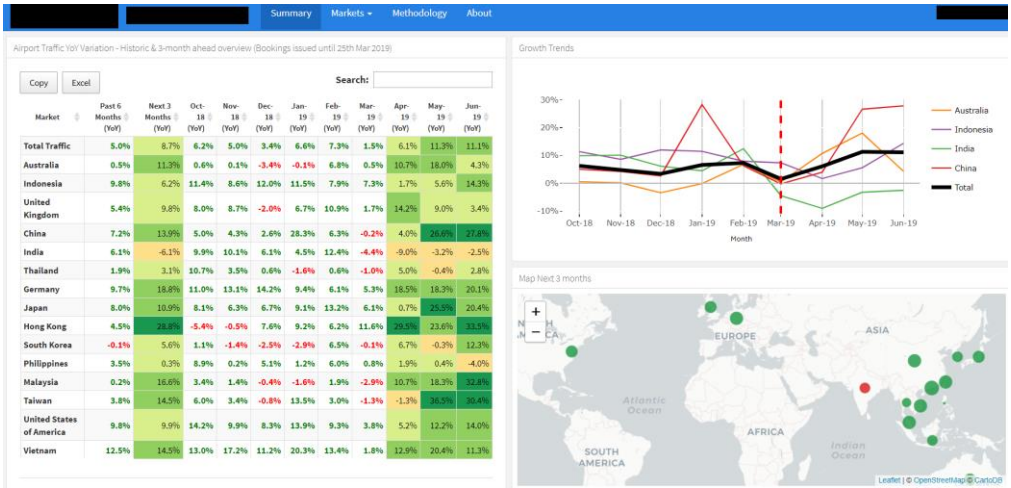


Ilustración 6 Ejemplo del dashboard resultado, Fuente: propia

[10] Implementación de la herramienta: https://github.com/rstudio/flexdashboard

Para la representación de las tablas se han hecho uso de *datatables* con los resultados. Para la representación de las series temporales se ha hecho uso de gráficos con ggplot2. Para la representación del mapa se ha hecho uso del paquete *leaflet.*

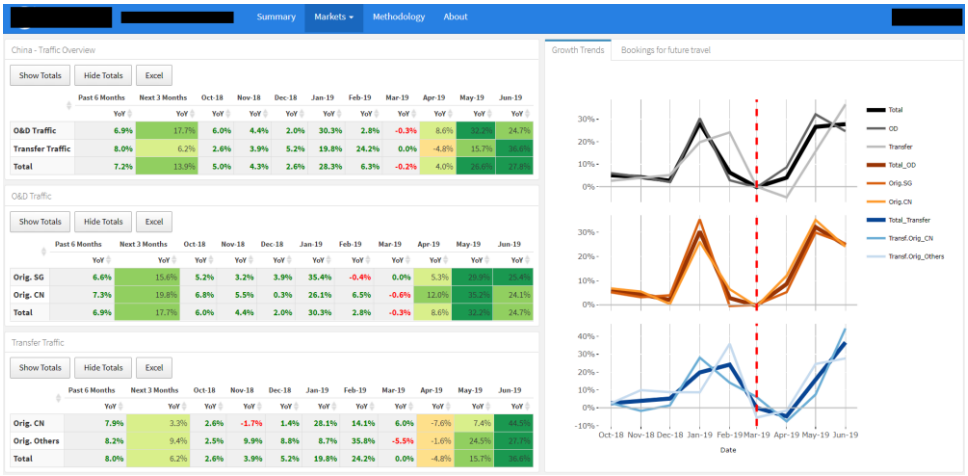


Ilustración 7 Ejemplo del dashboard resultado, Fuente: propia

7

Evaluación y resultados

Como resultado de todo el proceso de ajuste del modelo, se han realizado un total de 300 ajustes del modelo para cada serie temporal, habiendo 4 series temporales por cada mercado de interés. Con una cifra de 15 mercados calculados, como resultado se ha realizado un total 18.000 ajustes distintos del modelo, pero solo con 60 ajustes finales válidos.

La siguiente tabla recoge los mejores resultados obtenidos. Tanto el *MAPE* como el *RMSE* son métricas indicativas del error del modelo. No obstante, la única métrica normaliza que permite comparar el error de los modelos entre las diferentes series temporales es el *MAPE* por lo que se han ordenado los resultados por esta métrica.

MAPE	RMSE	Mercado	Serie temporal
0.0784	57.34	AU	Transfer origin others
0.0822	55.11	CN	O&D origin market
0.0830	85.53	MY	Transfer origin market
0.0864	71.86	ID	Transfer origin others
0.0892	65.27	DE	Transfer origin others
0.0904	148.89	CN	Transfer origin market
0.0904	64.33	AU	O&D origin POI
0.0935	174.99	MY	Transfer origin others
0.0935	153.57	CN	O&D origin POI
0.0938	90.43	IN	Transfer origin others
0.0943	206.94	TH	O&D origin POI
0.0949	230.08	UK	Transfer origin market
0.0975	176.90	UK	Transfer origin others
0.1005	77.23	AU	O&D origin market
0.1008	132.09	DE	O&D origin POI
0.1016	139.57	JP	Transfer origin market

0.1020	233.56	CN	Transfer origin others
0.1021	255.30	MY	O&D origin POI
0.1022	196.99	TH	Transfer origin others
0.1030	281.08	AU	Transfer origin market
0.1036	158.78	JP	Transfer origin others
0.1051	133.54	IN	O&D origin market
0.1056	213.70	US	Transfer origin others
0.1069	254.75	UK	O&D origin market
0.1076	287.84	TH	Transfer origin market
0.1114	346.59	HK	O&D origin POI
0.1126	255.93	JP	O&D origin POI
0.1133	206.64	IN	Transfer origin market
0.1143	366.03	TH	O&D origin market
0.1143	181.49	VN	Transfer origin market
0.1161	199.59	US	O&D origin market
0.1169	274.95	DE	O&D origin market
0.1184	169.36	HK	Transfer origin others
0.1184	357.69	JP	O&D origin market
0.1185	384.60	ID	O&D origin POI
0.1188	179.45	IN	O&D origin POI
0.1207	592.06	DE	Transfer origin market
0.1215	549.78	UK	O&D origin POI
0.1219	139.60	TW	Transfer origin market
0.1259	146.57	VN	O&D origin market
0.1264	188.46	HK	Transfer origin market
0.1274	492.11	PH	Transfer origin others
0.1286	220.52	MY	O&D origin market
0.1302	398.26	TW	O&D origin market
0.1320	127.40	ID	Transfer origin market
0.1332	224.69	HK	O&D origin market
0.1338	626.07	VN	O&D origin POI
0.1350	403.14	TW	Transfer origin others
0.1356	417.95	ID	O&D origin market
0.1359	362.90	PH	Transfer origin market
0.1359	627.14	US	O&D origin POI
0.1397	225.83	US	Transfer origin market
0.1400	144.71	TW	O&D origin POI
0.1439	711.04	VN	Transfer origin others
0.1765	490.84	PH	O&D origin market
0.1886	285.08	KR	O&D origin market
0.1889	617.56	KR	Transfer origin others

0.2149	442.36	PH	O&D origin POI
0.2201	411.44	KR	O&D origin POI
0.2222	494.49	KR	Transfer origin market

Ilustración 8 Tabla con los mejores resultados de las métricas para cada modelo generado, Fuente: propia

Por lo general, casi todas las series temporales se predicen con un nivel de precisión elevado, rondando casi todos los valores del *MAPE* con un error entre el 7-14%. No obstante, algunas de las últimas series temporales no se ajustan tan bien al modelo, llegando a errores del 17-22%.

Las series temporales que peor predice el modelo son aquellas cuyo mercado es Korea del Sur. Observando los valores de sus series temporales (ilustración 9) se puede observar un cambio de tendencia positivo a partir del segundo tercio de la serie temporal, provocando un aumento del error al realizar la predicción.

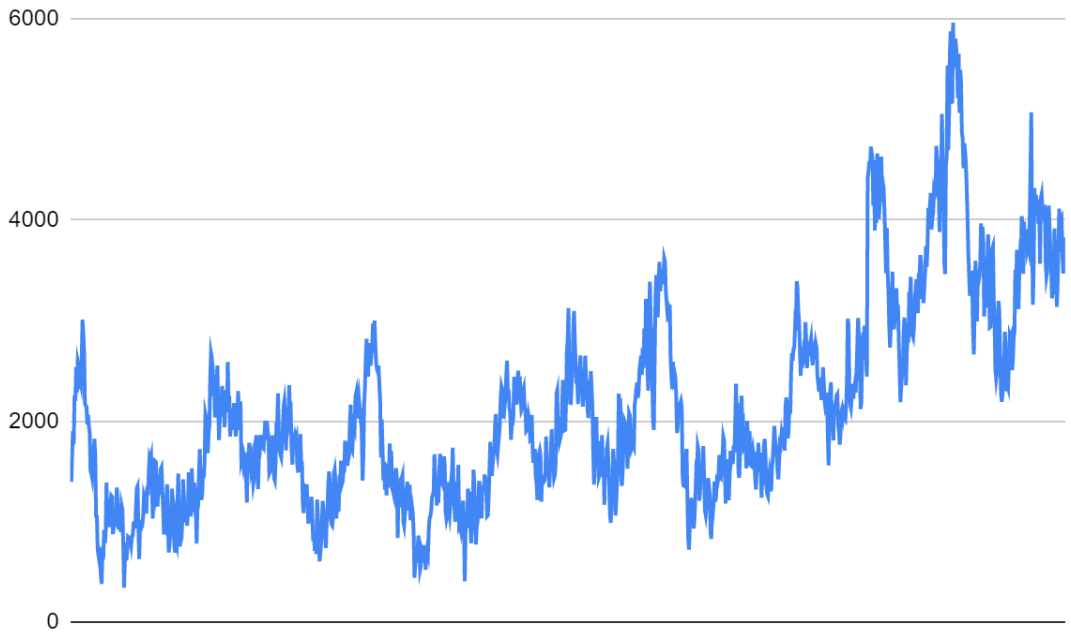


Ilustración 9 Serie temporal: KR-Transfer origin market, Fuente: propia

8

Conclusiones

La planificación de este proyecto partía de la finalidad de realizar un análisis relacionado con el flujo de pasajeros de aviones y el tiempo de estancia en un aeropuerto. A su vez, se pretendía predecir mediante información histórica, el comportamiento y el flujo de estos pasajeros con una ventana de 3 meses hacia adelante.

Como conclusiones del proyecto, se puede decir afirmativamente que es posible extraer la información requerida para este tipo de análisis a partir de información transaccional de reservas de vuelo. Esta información es posible filtrarla según los diferentes tipos de estancias dentro del aeropuerto y según los diferentes orígenes o destinos que puedan tener.

Por otro lado, se puede afirmar positivamente que, sobre estas series temporales, el algoritmo implementado por Facebook, Prophet, realiza predicciones sobre la serie temporal con unos bajos niveles de error en función de información histórica.

A su vez, es posible corroborar que la librería de flexdashboard junto a Rmarkdown, facilitan la creación de un cuadro de mando interactivo exportable como archivo html. Además, es una herramienta que no requiere ni necesita un servidor de aplicación para dar sustento al cuadro de mando. Por otro lado, como desventaja a incrustar los datos en el cuadro de mando, si este es exportado como archivo html, este pasa a tener un tamaño considerable. Por ello, sería recomendable no hacer uso de esta herramienta si el tamaño del resultado es considerablemente grande.

Como conclusión final, el cuadro de mando generado permite analizar el comportamiento de los diferentes mercados de origen sobre el aeropuerto a analizar, facilita comparar la conducta de estos y permite estudiar cada uno de ellos individualmente.

9

Referencias bibliográficas

[1] S.J. Taylor, B. Letham, “Forecasting at scale”, September, 2017, PeeerJ Preprints.

[2] Kim, Sungil and Heeyoung Kim, 2016, "A new metric of absolute percentage error for intermittent demand forecasts." International Journal of Forecasting.

[3] Bishop, Christopher, 2008, “Pattern Recognition and Machine Learning”, Springer Verlag. ISBN=978-0-3873-1073-2.

[4] Flach, Peter, 2012, “Machine Learning: The Art and Science of Algorithms that Make Sense of Data”, Cambridge University Press. ISBN 978-1-107-42222-3.

[5] A. Martinez-Gavara, “Estadística y optimización: Series temporales”, Universidad de Valencia.

[6] FH JOANNEUM , 2005-2006, “Cross-Validation Explained”, Institute for Genomics and Bioinformatics, Graz University of Applied Sciences

[7] Courtney Cochrane, 2018, “Time Series Nested Cross-Validation", Towards Data Science.

[8] Hyndman, Rob J.; Koehler, Anne B., 2006, “Another look at measures of forecast accuracy”, International Journal of Forecasting.