

Métodos Numéricos

Catriel Omar D'Elía¹ Gian Franco Lancioni²
Joaquín Perez Centeno³

*Depatamento de Computación
Universidad de Buenos Aires
Buenos Aires, Argentina*

Resumen

En este trabajo práctico se analizará el retraso de vuelos, aplicando técnicas de *Métodos Numéricos* y *Data Science*, en particular *Regresiones Lineales de Cuadrados Mínimos* sobre un conjunto de datos con el objetivo de poder ver información relevante y un modelo que nos permita predecir los retrasos aéreos. Luego se hará un análisis de la calidad de los resultados obtenidos mediante la métrica de *Root Mean Squared Error*, conocido por su sigla como *RMSE*.

Keywords: *RMSE* (*Root Mean Squared Error*), *OTP* (*On-Time Performance*), *KIP* (*Key Performance Indicator*), *Regresiones Lineales de Cuadrados Mínimos*, *Retraso de vuelos: Delays o Cancelación*

¹ Email: catriel.delia@gmail.com

² Email: gianflancioni@gmail.com

³ Email: joaquinpcenteno@gmail.com

1. Introducción

En estos últimos años se ha puesto el foco en hacer un análisis exhaustivo en las actividades periódicas de las empresas aeronáuticas, con el fin de establecer si se han encontrado funcionando correctamente.

El aumento de vuelos creció abruptamente dado la expansión de aeropuertos a lo largo de todo Estados Unidos causando dificultades de organización en la industria y congestión del tráfico aéreo.

[?] Según una investigación de la Universidad Rey Juan Carlos, los problemas de congestión del tráfico aéreo en los aeropuertos de Estados Unidos son cada vez más graves y para mejorar la situación, las técnicas de Gestión de Tráfico Aéreo tratan de anticipar y prevenir la situación de congestión que se puedan producir, asignando retrasos a los vuelos o cancelándolos si fuera preciso.

Debido a esto, en el siguiente trabajo práctico se enfoca en la *predicción de los retrasos de vuelos*. Para este análisis, se aplicarán las técnicas de **Métodos Numéricos** y **Data Science**, en especial **Regresiones Lineales de Cuadrados Mínimos**. También una forma de llevar adelante este enfoque será utilizando **Key Performance Indicator (KPI)** para evaluar la puntualidad del transporte aéreo **On-Time Performance (OTP)**

Se cuenta con un *data set* comprendido con cierta información relacionada a varios vuelos realizados en Estados Unidos entre 1987 y 2008, incluyendo información de la compañía, fecha, horarios planificados de arribo o partida, horarios de entrada o salida, causa del delay, si fueron cancelados o no, su respectiva causa, el tipo de avión utilizado, tiempo de vuelo, y tiempos de *taxi*.

Por último, en este trabajo práctico habrá una experimentación usando la técnica de *cross-validation*. Para ello, vamos a particionar nuestro conjunto de datos y variar la composición de la base de entrenamiento (*training*) y las observaciones consideradas como *test* donde se tendrá en cuenta cómo afecta a variables cuyo valor dependa de la cantidad de observaciones tomadas.

2. Desarrollo

Para estacionalidad en CML fijamos senoides con frecuencias y fases ⁴ de acuerdo a nuestra intuición: buscamos frecuencias relacionadas a las estaciones como puede ser bimensualmente, trimestralmente, semestralmente y anualmente. No granularizamos demasiado tales periodos para evitar *overfitting*: no tiene sentido una estacionalidad cada 10 meses si no existen eventos importantes con esa frecuencia en el mundo real. Para las fases decidimos usar aquellas que permiten que las ondas se inicialicen en puntos intermedios o en picos: $\frac{i\pi}{4}$. Estas frecuencias anulan la necesidad de usar cosenos, que tampoco usamos para prevenir matrices singulares considerando los senos de distintas fases. Viendo que las componentes polinomiales ajustaban mejor al entrenamiento pero a costas de predecir mucho peor decidimos dejar solamente hasta grado 1 para permitir cambios lineales.

Para validación cruzada usamos la librería de scikit-learn que dispone de métodos de partición para tal motivo de series de tiempo. La librería particiona, para n splits, todas las secuencias con los primeros $\frac{i}{n+1}$ ($1 \leq i \leq n$) registros para entrenamiento y el resto para predicciones.

⁴ Al no ser coeficientes CML no los permite resolver.

3. Experimentación

En esa sección, para entender las causas y consecuencias de OTP, se plantean dos ejes de experimentación.

3.1. Primer eje de experimentación

En el primer eje intentaremos analizar a OTP como KPI, es decir qué tan bien evalúa la calidad de los factores involucrados en el transporte aéreo y qué tipo de consideraciones hay que tener sobre su naturaleza para usarla como herramienta predictiva de rendimiento.

Usaremos el promedio de delay por unidad de tiempo, de modo de ajustar por cantidad de vuelos nuestras predicciones, y también contrastaremos la diferencia entre considerar datos con outliers y considerarlos en crudo en términos cuál brinda mejor calidad predictiva.

3.1.1. Estacionalidad

Lo primero que intentamos analizar es la variación estacional de las métricas: la congestión aeroportaria difícilmente sea constante a lo largo del año sino que es de esperarse que en épocas del hemisferio norte como vacaciones de invierno, verano y primavera aumente.

Interesantemente tuvimos en varios casos mejores resultados sin recortar ⁵ outliers para el promedio de delays por día o meses. En general la granularidad diaria funcionó mejor que la mensual para predicciones: se esperaba que la congestión crezca los fines de semana, información que se pierde mensualmente. Obtuvimos los menores errores de predicción en las particiones más cercanas al 50/50, lo que tiene sentido en términos de underfitting y overfitting. Para el caso diario conseguimos el NRMSE de predicción mínimo al rededor de 0.09 con coeficientes de frecuencias mayores principalmente semanales, trimestrales, semestrales y anuales (en orden descendente) y en $\frac{2}{5}$ del dataset para entrenamiento y el resto de predicción, mientras que con $\frac{1}{5}$ para entrenamiento se consiguió el peor a pesar de que fitteaba mejor sobre el set de training.

3.1.2. OTP como KPI para aeropuertos, aerolíneas y rutas aéreas

Luego nos planteamos cómo influye cada factor en el delay, y por consiguiente sobre qué factor es mejor KPI el OTP. Comparamos predicciones

⁵ Se recorta el 10 % de los menores y mayores retrasos

mensuales del top 5 en volumen ⁶ de aeropuertos, aerolíneas y rutas aéreas (pares ciudad destino y ciudad origen). Usamos el top 5 de volumen para asegurarnos de que la presencia de outliers no nos afecte y que la media de delay se comporte bien.

Sobre cada entidad del top 5 de cada categoría se computa el promedio de NMSE en sus particiones sobre los cuales se consideran los boxplots 1 según delay de arribo y 2 según delay de partida para analizar las características la distribución de errores de cada categoría. Inicialmente sospechamos que para delay de arribo la ruta aérea sería lo más determinante por cuestiones climáticas y de distancia relacionadas y que para delay de partida sería la aerolínea por cuestiones de logística claves para organizar abordajes. Los resultados indican errores promedio similares en todas las categorías pero con cuartiles más compactos para aerolíneas y que, por lo tanto, es en función de estas sobre las cuales se predicen mejor (con menor error) los retrasos, seguido por aeropuertos y finalmente rutas aéreas. Que el orden sea el mismo para ambos delays se puede explicar a través de la consecuencia en retrasos de arribo provocada por retrasos de partida.

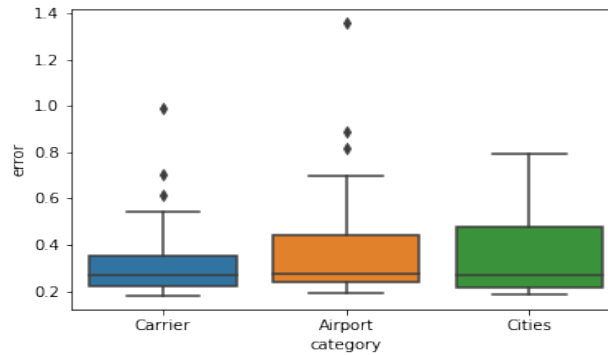


Figura 1. Distribución de errores promedio en predicción de delay de arribos (en minutos) sobre 4 splits de cross-validation para cada entidad del top5 de cada categoría.

3.1.3. El 9/11 y el impacto en predicciones de OTP

Otra pregunta que nos hicimos fue cómo afectó, a través de la reacción en términos de políticas y rigor aeroportuario, el atentado del 9/11 en Estados Unidos a las predicciones sobre delay. Consideramos por un lado que podría

⁶ Volumen medido en los primeros años del dataset, para que no solo sean factores de mayor volumen sino también sostenido a través del tiempo.

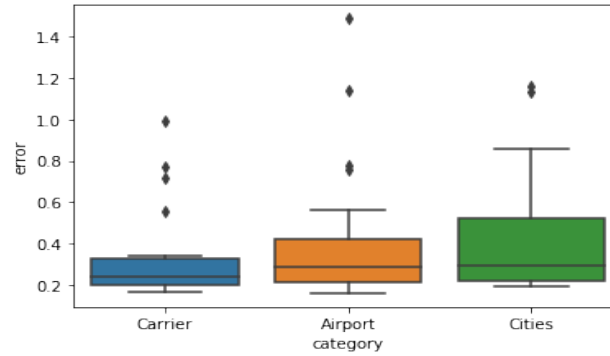


Figura 2. Distribución de errores promedio en predicción de delay de partidas (en minutos) sobre 4 splits de cross-validation para cada entidad del top 5 de cada categoría.

ser posible ver aumentos en delay por cuestiones de mayores exigencias de seguridad antes de salir a pista, o sino ver un decrecimiento en delays provocado por mayor planificación en respuesta, en ambos casos generando errores de predicción.

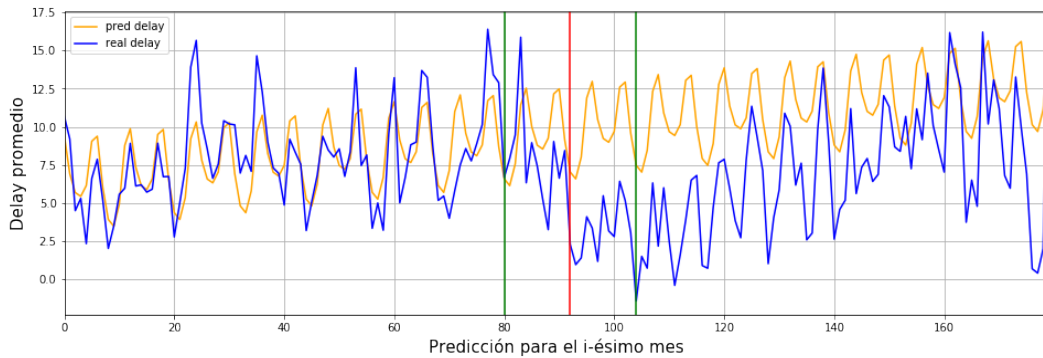


Figura 3. Predicción de delay en minutos contra delay histórico. Las líneas verdes marcan el intervalo de predicción sobre el cual se considera el error antes y después del atentado, representado por la línea roja.

Para experimentar entrenamos al predictor de CML con datos anteriores a un año antes del atentado y comparamos el NMSE de las predicciones antes del atentado contra un año después del mencionado. En predicciones diarias notamos una diferencia entre aproximadamente 0.11 y 0.17 de RMSE antes y después del atentado respectivamente mientras que la diferencia en predicciones mensuales resultó de 0.28 y 1.26, con lo que resultaría que el impacto a nivel predicciones es importante (como se puede apreciar en la figura 3). Inda-

gando más profundo descubrimos que el promedio de delay pasó de al rededor de 8 minutos antes del atentado a 3 luego de este, validando la hipótesis de una mayor planificación.

3.2. Segundo eje de experimentación

En el segundo eje intentamos contemplar otros factores de impacto que no hayamos considerado originalmente para el primero.

3.2.1. Demoras en el arribo causados por demoras en los despegues

Es razonable pensar que un vuelo que se demora al despegar, va a demorarse también al aterrizar. Se estudió la correlación entre la demora al despegue, `DepDelay` y la demora en el momento del aterrizaje `ArrDelay`. Se obtuvo una correlación alta, de $\rho = 0,92$.

Para limpiar los datos, se utilizó un filtro de *outliers* basado en una regla escrita a mano que descarta los vuelos que se adelantan varias horas a su hora de despegue programada. Creemos que se tratan de errores de medición.

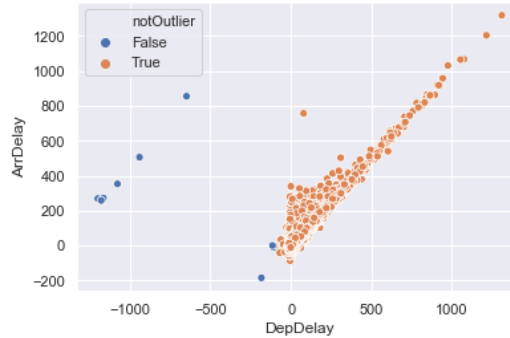


Figura 4. Relación entre la demora al despegar y la demora al aterrizar.

Finalmente, se entrenó un modelo de regresión lineal entre ambas variables obteniendo un NRMSE de 0,00646. La mayor densidad de puntos se encuentra sobre la recta del modelo. Por otro lado, en el rango de $-200 \leq \text{DepDelay} \leq 800$ pueden observarse vuelos con un mayor *delay* al predicho por el modelo. Se propone para un futuro experimento, añadir regresores adicionales para mejorar la predicción en dicha zona.

3.2.2. Demoras de arribos y Cancelaciones

Otro eje interesante es el de usar la cantidad de cancelaciones como KPI. Se hipotetizó una relación dinámica entre las demoras y las cancelaciones, ya

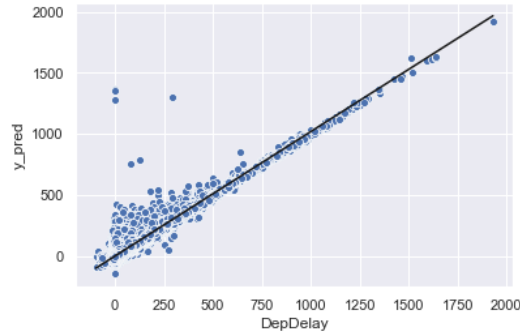


Figura 5. Regresión lineal entre la demora al despegar y la demora al aterrizar.

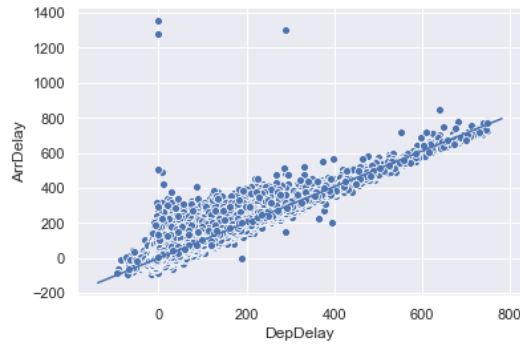


Figura 6. Detalle de la regresión lineal para la demora al despegar y la demora al aterrizar en el rango $-200 \leq \text{DepDelay} \leq 800$.

que ambas pueden influirse mutuamente.

Para el análisis se agregaron los datos de todo el dataset por día obviándose los datos del 2001, que presentan características propias por la situación política y de seguridad.

Se eliminaron del dataset todas las cancelaciones debidas a factores climáticos y de razones de seguridad, para disminuir el peso de factores ajenos a la calidad del servicio.

En la figura 8, vemos una buena correlación entre ambas variables con un *nrmse* que si bien no es despreciable, nos permite establecer una correlación lineal.

También nos preguntamos si las demoras o las cancelaciones, podían afectar periodos posteriores. Para esto generamos una matriz de correlación, para los mismos datos anteriores, tomando las cancelaciones del día actual vs la demora y también las cancelaciones de dos días en el futuro y en el pasado. En la figura 8. Se ve que la mayor correlación se encuentran entre días con-

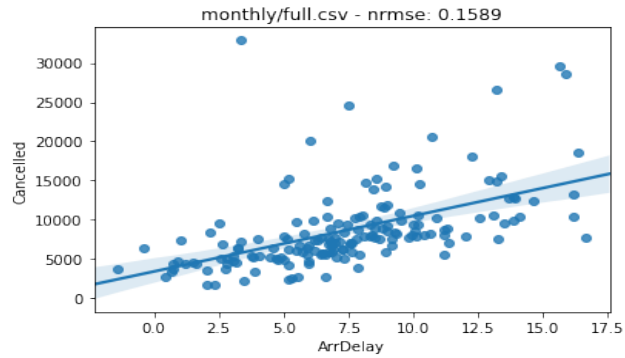


Figura 7. Relación entre las demoras de arribo y las cancelaciones, para todo el dataset desde 1994 hasta 2008 sin incluir 2001. Los datos están agregados por día. Las cancelaciones se sumaron, las demoras se promediaron.

cecutivos para las cancelaciones, incluso superando la alta correlación entre demora y cancelaciones del día actual mostradas en la fig. 8 ⁷

	ArrDelay	Cancelled	Cancelled-1	Cancelled-2	Cancelled+1	Cancelled+2
ArrDelay	1	0.49	0.32	0.21	0.32	0.17
Cancelled	0.49	1	0.62	0.39	0.62	0.39
Cancelled-1	0.32	0.62	1	0.62	0.39	0.31
Cancelled-2	0.21	0.39	0.62	1	0.31	0.29
Cancelled+1	0.32	0.62	0.39	0.31	1	0.62
Cancelled+2	0.17	0.39	0.31	0.29	0.62	1

Figura 8. Correlacion entre las demoras de arribo y las cancelaciones para días cercanos. Las cancelaciones están shiteadas hacia adelante y hacia atrás. Los datos se tomaron para todo el dataset desde 1994 hasta 2008 sin incluir 2001. Los datos están agregados por día. Las cancelaciones se sumaron, las demoras se promediaron.

⁷ Las métricas no son iguales en ambas figuras.

4. Referencias

Referencias

- [1] A Alonso-Ayuso. Mathematical optimization in air traffic management. *Boletín de Estadística e Investigación Operativa. BEIO*, 26(1):25–41, 2010.