Performance prediction on airlines and flights using Linear Regression

Federico Hosen Julián Palladino Guido Rajngewerc
Damián Silvani

Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Resumen

El presente trabajo consiste en aplicar técnicas de métodos numéricos y ciencia de datos, en particular Regresiones Lineales con Cuadrados Mínimos sobre un (gran) conjunto de datos, buscando proveer información descriptiva y de modelos que puedan ser utilizados para predecir fenómenos que afecten a la puntualidad (OTP), pero no necesariamente limitados a ésta.

Keywords: CML, análisis de datos, tráfico aéreo

1. Introducción

El siguiente trabajo práctico aborda el análisis y discusión del tráfico aéreo doméstico en EEUU y las demoras que sufre. Asimismo, busca encontrar patrones y causalidades para poder generar un modelo que nos permita predecir diversos acontecimientos de interés.

2. Desarrollo

Metodología de trabajo

Al tener que procesar semejante cantidad de datos, utilizamos *sqlite*3. Desde *Python* fueron consultados y se realizaron las estimaciones y análisis necesarios para la experimentación. También se utilizaron las librerías *pandas*, *seaborn*, *matplotlib* y *sklearn*.

En cuanto a la validación de cada modelo de CML, se realizó de la siguiente manera: dada una cantidad de datos T, se fijó K como el cardinal mínimo para el conjunto de training y P como la cantidad de datos usados para predecir. Se evaluó, para distintos valores de $\tau \in$ [K,T], el ECM del algoritmo entrenando con los datos en [K, τ] y prediciendo con el rango [τ, τ+P]. Se eligió cuatro valores diferentes de τ.

Elección de los ejes

Se plantean dos líneas de investigación, con sus respectivos experimentos y conclusiones:

- **Momento de salida y proporción de demorados**
  ¿Es cierto que el momento de salida influye en el retraso de un vuelo? ¿Podemos predecir el comportamiento en cierto periodo de tiempo? ¿Qué granularidad conviene tomar al respecto?

  Comenzamos enfocándonos en la relación entre el momento de salida [1] de los vuelos y la proporción de los demorados. Definimos que un vuelo tiene salida demorada de la misma manera que lo hace la institución [2] que recolectó los datos: se cuenta como *demorado* a aquel vuelo que se retrase 15 minutos o más al salir.

  Se decidió estudiar la **proporción de vuelos demorados** en lugar de la **cantidad total de vuelos demorados**, debido a que esta última se ve afectada por la falta de uniformidad en la cantidad de vuelos totales a lo largo del tiempo: quizás todos los vuelos se ven atrasados en dos períodos disjuntos, pero si uno tiene más caudal de vuelos que el otro, dará que tiene más retrasos, concluyendo que es peor volar en ese momento.

  Además, el momento de salida del vuelo puede ser tomado con diferentes grados de granularidad. Puede ser analizada tanto la hora de salida como el día de la semana, el mes y el año. Desde antes de realizar experimentaciones se puede preveer que cada uno de estos grados se comporta diferente, y es

[1] Notar que se habla del momento de salida programado y no el efectivo.
[2] *Federal Administration of Aviation*

afectado por variables diferentes, por lo que se tendrá en cuenta varios niveles distintos de detalle.

- **Crecimiento de los *carriers* a lo largo del tiempo**
  ¿Los distintos carriers se comportan de igual manera? ¿Es posible predecir la cantidad de vuelos de una aerolínea según el período de tiempo?

  En los datos disponibles hay vuelos relacionados a 1492 empresas distintas. Teniendo eso en cuenta surge la inquietud de saber como se comportan dichas empresas a lo largo del tiempo, especialmente las más grandes. ¿Tienen un comportamiento similar entre ellas? ¿Podemos tomar alguna medida que nos indique el tamaño de la empresa a lo largo del tiempo? ¿Podemos preveer si la empresa va a crecer o no?

  Consideramos de interés la posibilidad de preveer si la cantidad de vuelos de una empresa subirá o bajará, pues creemos que esto es un buen indicio del valor real de dicha compañía. Por lo tanto intentaremos utilizar la técnica de CML teniendo el tiempo y la cantidad de vuelos totales como ejes de análisis.

3. Experimentación: Primer eje de análisis

Momento de salida y proporción de demorados: por hora diaria

En cuanto a la hora del día, podemos preveer que jugará un papel importante en las chances de que el vuelo se vea demorado: Así como se tienen horas pico en el tráfico de autos en las ciudades, buscamos encontrar un patrón análogo en el tráfico aéreo. Asimismo, se debe poner atención que la cantidad de vuelos que despegan a la madrugada (0:00 - 4:59) serán despreciables en comparación a los del resto del día. En consecuencia, se deberá tener cautela de los *outliers* que esto genere.

En cuanto a cada hora, no resulta interesante un estudio tan minucioso que considere los minutos del momento de salida. Los cambios de proporción de vuelos demorados no serán tan bruscos como para que cambien de un minuto para el otro. Además, notamos que la mayoría de los vuelos son programados para que salgan *en punto*, o en horas terminadas en un múltiplo de 10.

En vista de estos dos puntos, se agruparán los vuelos según su hora de despegue (sin tomar en cuenta los minutos) y se estudiará su relación con la demora que puedan tener.

Teniendo esto en cuenta, se analizó la proporción de demorados en cada franja horaria. En principio, llama la atención el pico de las 3, y los altos valores de 2 y 4. Esto se debe a lo previamente mencionado: dichas franjas

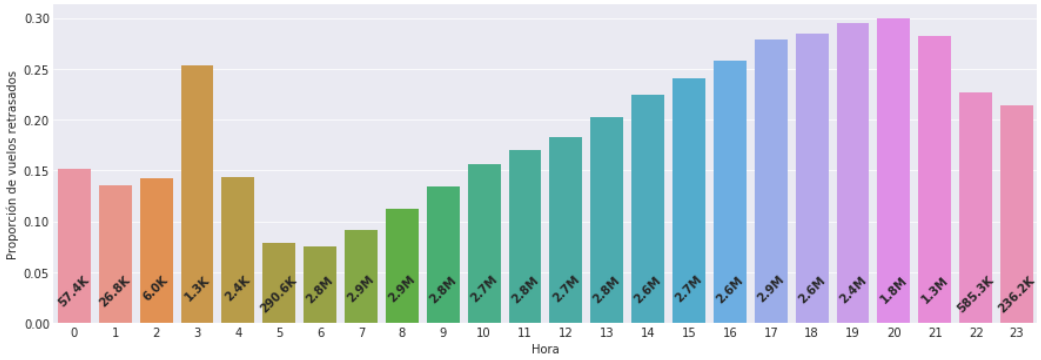

Figura 1. Proporción de vuelos demorados según intervalos de una hora. El eje horizontal indica la franja horaria. El eje vertical indica la proporción de vuelos demorados en cada franja. El número debajo de cada barra representa la cantidad de vuelos totales en dicha franja, notando K como miles y M como millones. Los datos fueron tomados entre los años 2003 y 2008.

horarias manejan un bajo caudal de vuelos, lo cual los hace propensos a generar outliers (a las 3 hay 1.300 vuelos totales, mientras que en la mayor parte del día hay cerca de 3 millones).

Más allá del outlier de la madrugada, se aprecia un comportamiento suave en las horas que acumulan la mayor cantidad de vuelos: desde las 6 hasta las 20 hay entre 2 y 3 millones de vuelos, y se distingue una curva que alcanza el máximo en las 20.

Se optó por no realizar CML en este gráfico, ya que la cantidad de puntos es muy poca (24), y no resulta provechoso aumentar la granularidad.

Momento de salida y proporción de demorados: por hora mensual

Otra manera de lograr tener más puntos para realizar un buen modelado de CML es aumentar el rango abarcado en el análisis. Con esa idea, se extendió el rango desde ver cada hora de un día hasta ver cada hora de cada día durante un mes entero.

Se eligió el mes de Noviembre, y se realizó un gráfico que tiene un punto por cada franja horaria (de una hora) de cada día del mes, es decir, la hora relativa al mes [3]. Dicho punto tiene la proporción de vuelos demorados de esa franja horaria de ese día. Para generalizar el comportamiento se promediaron los datos desde el 2003 hasta el 2008.

Luego de ver los datos, se notaron diversos puntos anómalos (*outliers*),

[3] Hora relativa del mes = hora del día + 24 * (número del día del mes - 1)

sobretodo en las horas de la madrugada. Para removerlos, calculamos el valor promedio de cada franja horaria, y si un punto supera cierto umbral relacionado a ese promedio obtenido, concluimos que es un dato anómalo que puede ser obviado.

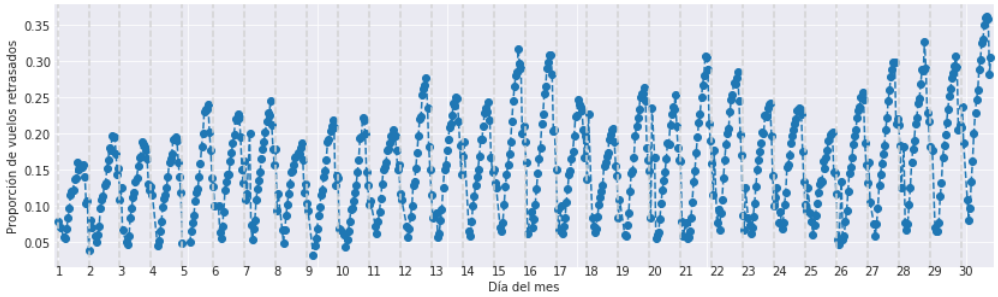

Figura 2. El eje horizontal es la hora del mes. El eje vertical indica la proporción de vuelos demorados en cada franja de una hora. Los datos fueron tomados del mes de Noviembre entre los años 2003 y 2008.

Proponemos la siguiente familia de funciones para realizar CML:

$$\mathcal{F}(x) = a\sin(freq \cdot x) + c$$

Sobre la elección de dicha familia:

- Dado que se observa periodicidad en los datos el factor más importante debe ser una función senoidal.
- Fue necesario ajustar la frecuencia del seno, con lo cual introdujimos una constante $freq$, cuyo valor óptimo fue encontrado en 0,26204.
- Encontramos que utilizar una función lineal ocasionaba que se sobre-entrene según la tendencia de crecimiento (o decrecimiento) de los datos de entrenamiento, lo cual tiende a arruinar predicciones futuras si hay un cambio de tendencia. Por dicho motivo se eliminó el componente lineal.

Usando este modelo, al hacer un análisis usando ideas de *cross-validation* sobre el mes de abril en los años $2003 - 2008$ se obtienen los resultados mostrados en la figura 3.

Probando el mismo modelo con distintos meses arrojan distintos resultados. Los meses más estables como *abril* muestran series más estacionarias, con menos picos. En cambio, meses como *enero* o *diciembre*, que tienen vacaciones, temporadas de invierno y fechas importantes como navidad, muestran comportamientos más inestables, ocasionando que el modelo encontrado funcione peor. Esto se puede observar en la siguiente tabla.

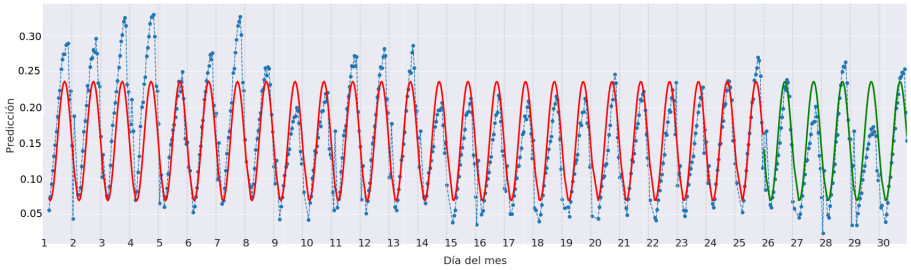

Figura 3. Mes de Abril. En rojo la sección de training, en verde la predicción. ECM promedio: 0.001366

Mes	Enero	Abril	Noviembre	Diciembre
ECM promedio	0.008082	0.001996	0.001327	0.004487

Momento de salida y proporción de demorados: por mes

Como se ha mostrado en el estudio previo, el comportamiento de cada mes es diferente. Con el foco en ello, se cambiará nuevamente la granularidad del análisis, observando el mes de salida de cada vuelo en un lapso de varios años.

En un principio, para decidir la cantidad de años que se toman parecería ser que, cuanto más años, mejor. Esta idea se debe a que, en teoría, cuanto más datos tenga de entrenamiento CML, mejor estimación podrá dar. Por tanto, se analizaron los años 1988-2008, salteando 2001 y 2002, ya que sus comportamientos anómalos provocados por el atentado del 9/11 producen un outliers.

No obstante, lo que nos ocurrió en la práctica es que **tomar un rango tan amplio como 1988-2008 generó ruido en la búsqueda de un modelo estimativo** [4] .

De esta manera, decidimos acotar nuestro rango de años a 2003-2008, el cual ofrece una buena cantidad de datos sin involucrar épocas tan diferentes.

En primer lugar, se debe prestar atención a la irregularidad de los datos. Ya no se tiene una periodicidad tan bien definida como al analizar franjas horarias. Esto causará mayor dificultad a la hora de encontrar el modelo que lo represente.

En cuanto al modelado, la familia de funciones utilizada fue:

$$\mathcal{F}(x) = a\sin(x) + b\cos(0{,}1 \cdot x^2) + c\sin(x^2) + d\ln(x) + e$$

[4] Esto ocurre porque el tráfico aéreo entre épocas tan distintas no se rige por la misma función.

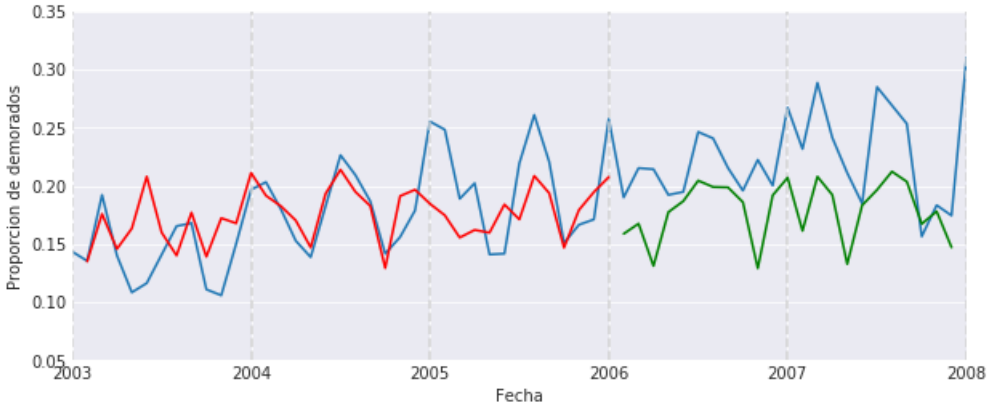

Figura 4. El eje horizontal representa los meses a lo largo de los años, cada mes es un punto y cada barra vertical denota el fin de un año. El eje vertical indica la proporción de vuelos demorados en cada mes (siendo 1.0 el 100 %). Los datos fueron tomados del período 2003-2008. ECM promedio: 0.002349

La misma fue encontrada por medio de prueba y error, a partir de las familias utilizadas en los análisis anteriores, ya que sigue la similitud de ser una función periódica sobre un eje horizontal (no se inclina).

Como se había previsto, es más difícil encontrar un buen modelo para esta granularidad. A ojo, vemos que tanto el entrenamiento como la predicción ajustan bien a ciertos picos, mientras que otros no. Asimismo, el ECM de 0.002349 resulta cerca del triple que al analizar horas mensuales.

4. Experimentación: Segundo eje de análisis

Como segundo eje de discusión, se optó por lo sugerido por la cátedra: variar la perspectiva con respecto a los datos, y analizar un aspecto diferente de ellos. De esta manera, se calcularon los 5 carriers con más cantidad de vuelos en el período 2003-2008, y se estudiaron sus cantidades de vuelos en dicho período.

Análisis de los datos

Como podemos observar en la figura 5 parecería que los carriers tienen comportamientos similares a lo largo del tiempo, a excepción de una aerolínea en particular, Southwest, a la cuál se le puede atribuir un comportamiento a la alza más marcado.

Por otro lado se nota también a simple vista un cambio brusco en el comportamiento del carrier Delta en el franja de los meses finales de 2005. Supo-

niendo que dicha tendencia no se corresponde con el funcionamiento normal de una aerolínea, investigamos por fuera del contexto de los datos proporcionados por la cátedra y pudimos relacionar esta anomalía con el pedido de quiebra de la compañía, golpeada por la suba de precios del petróleo luego del huracán *Katrina* en 2005[1].

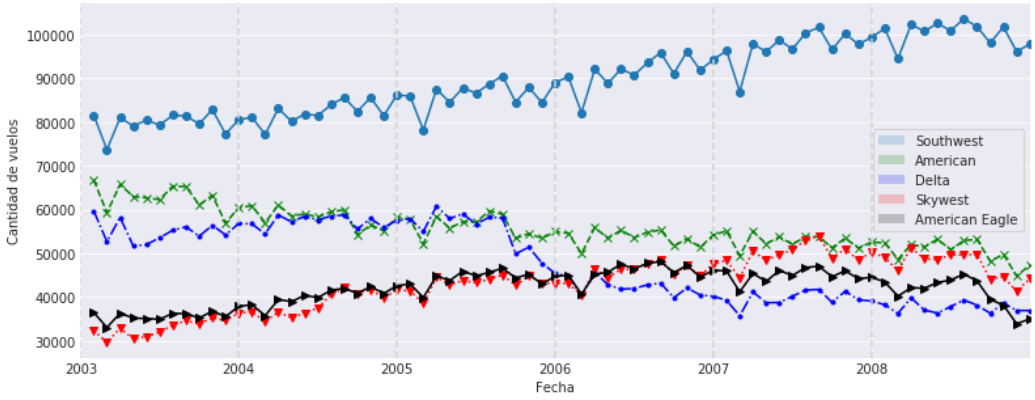

Figura 5. Cantidad de vuelos realizados por distintos carriers a lo largo de los años (desde 2003 a 2008)

Modelado del comportamiento de una aerolínea

Ahora se busca realizar un modelo fidedigno del comportamiento de alguna de las 5 aerolíneas. Luego de probar distintas combinaciones, se encontraron buenos resultados para modelar las aerolíneas más regulares, a partir del año 2004. En la figura 6 vemos la instancia de cross-validation sobre *American Airlines* que arrojó mejor ECM: estimando con 2004 y 2005, y prediciendo 2006. Esto se debe a que 2004, 2005 y 2006 tienen comportamiento más parecido que el resto de los años. El segundo modelo fidedigno que se logró realizar fue el de *Southwest Airlines*: en la figura 6 se puede ver la instancia de CV que dio mejor ECM, la cual se comporta de manera parecida que el modelo antes realizado. [5] Vemos que algunos picos son tan agudos que le es imposible al modelo imitar esos saltos [6].

Además en concordancia con lo comentado anteriormente notamos que el modelo de CML no aproxima como esperamos en el caso de *Delta Airlines*

[5] Notar que seguimos usando cross-validation, pero no mostramos todos los gráficos por falta de espacio.

[6] Esto, quizás, se podría evitar utilizando Fourier para encontrar una mejor familia de funciones. Sin embargo, esto excede el trabajo presentado.

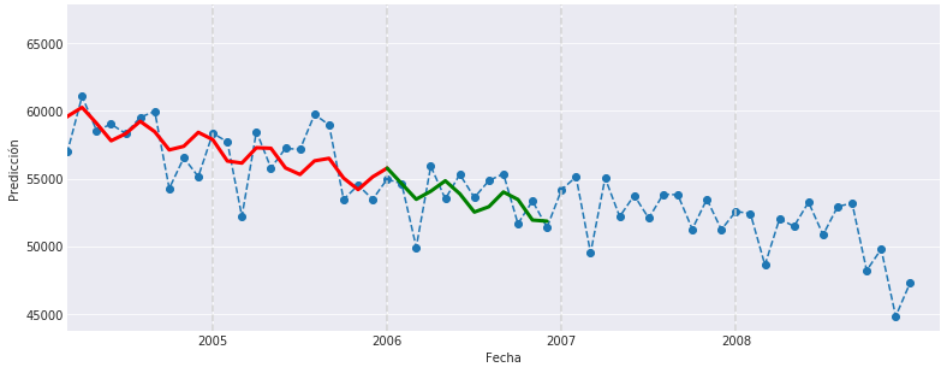

Figura 6. Cantidad de vuelos de American Airlines (2004 a 2008). En rojo la predicción, en verde la estimación. ECM promedio: 5758333.8679

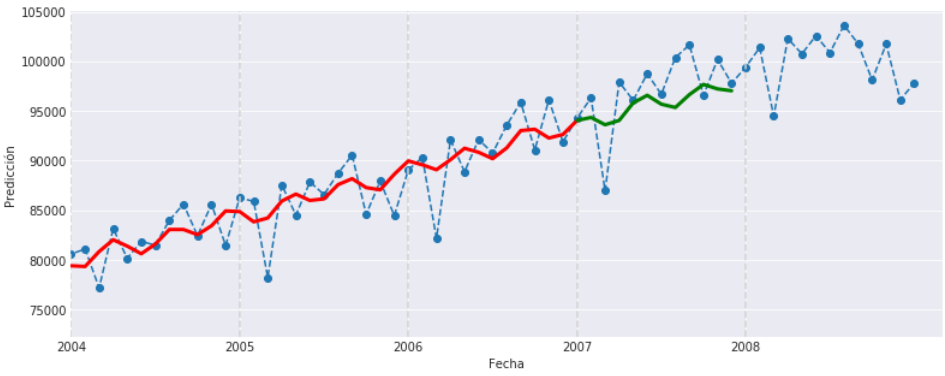

Figura 7. Cantidad de vuelos de Southwest Airlines (2003 a 2008). En rojo la predicción, en verde la estimación. ECM promedio: 10880616.5496

pues su bajada de vuelos debido a la quiebra es muy pronunciada y no se corresponde a ninguna instancia de entrenamiento que hayamos encontrado en el dataset.

5. Conclusión

En cuanto al análisis granular del momento de salida y la proporción de demoras, hemos visto que:

- El análisis más granular de todos (por minuto de salida) no era fructífero, y nos vimos forzados a agrupar por franjas de una hora para obtener datos más coherentes(3). Luego, a pesar de encontrarnos con claros outliers [7] debido

[7] El pico de las 3AM

la poca cantidad de datos, se obtuvo una periodicidad bastante uniforme, al extender el rango y analizar por horas mensuales (3).

- A reducir la granularidad se trabaja con mayor cantidad de datos, lo que involucra más factores que los afectan [8] . Esto dificulta el análisis y el armado de un buen modelo, como vimos al analizar los meses en varios años(3).

En conclusión a estos dos ítems, si tuviéramos que elegir un nivel de granularidad, nos quedaríamos con el de **horas en un mes**(3), que posee una suficiente cantidad de datos, periodicidad uniforme, y una cantidad de factores acotada. Se podría aplicar el modelo encontrado a cada mes, con el objetivo de predecir dentro del mismo mes, pues ya hemos visto que cada mes tiene un comportamiento distinto.

En cuanto al estudio de los carriers, nuevamente nos encontramos con un estudio que involucraba muchos y diversos factores. En el período de un año, se encontraban los picos usuales debido a las fechas festivas. Al mismo tiempo, a lo largo del tiempo se encontraban crecimientos y decrecimientos dependientes del precio del petróleo, los desastres naturales y de las decisiones tomadas por cada compañía. Por tanto, armar un modelo tan ajustado como en el análisis de hora mensual se hizo difícil. Además, nos encontramos con una función periódica particular, por lo que fue más inefectivo realizar prueba y error para encontrar un modelo correcto para CML (un posible trabajo futuro sería realizar Fourier para encontrarla).

A pesar de esto, fue posible realizar una estimación sobre el crecimiento de cada aerolínea, y se pudo apreciar diversos comportamientos, como la bancarrota de una, y el crecimiento de otra.

Referencias

[1] Delta Air Lines files for bankruptcy, *CNN Fortune 500*, http://money.cnn.com/2005/09/14/news/fortune500/delta/

[8] Ya no se trabaja sólo con el horario, sino también con fechas festivas, eventos climáticos, económicos, políticos y sociales.