DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Metodos Numericos
Trabajo Practico 3
Cuadrados Minimos Lineales

Alvarez Mon Alicia

aliciaysuerte@gmail.com
LU.: 224/15,
FCEN, UBA,
CABA, Argentina

Ansaldi Nicolas

nansaldi611@gmail.com
LU.: 128/14,
FCEN, UBA,
CABA, Argentina

Castro Luis

castroluis1694@gmail.com,
LU.: 422/14,
FCEN, UBA,
CABA, Argentina

Suarez Romina

romi_de_munro@hotmail.com,
LU.: 182/14,
FCEN, UBA,
CABA, Argentina

Resumen

En este trabajo buscamos extraer información relevante de vuelos en Estados Unidos y usarla para modelar el comportamiento de los mismos. Además, usaremos regresión lineal con cuadrados mínimos para predecir las tendencias de los mismos. Por otro lado, buscamos ver las causas que producen las demoras en los vuelos y analizarlas para determinar los factores influyentes en las mismas.

Keywords: CML, Delay, OTP, Vuelos.

---

1. Introducción

Utilizaremos los datos de los vuelos realizados en Estados Unidos, entre 1987 a 2008, y los analizaremos en los siguientes ejes de estudio:

1. **Cuando conviene viajar:** Vemos el comportamiento del delay(OTP) o demora en las distintas estaciones del año, además vemos cual es el día de la semana con menos retrasos para viajar. La motivación de este eje es ver como se comporta la puntualidad a lo largo del tiempo y encontrar el mejor momento para viajar basándonos en ese comportamiento.
2. **Causas de demoras:** Vemos las demoras causadas por el clima, los aviones, y por último las aerolíneas. Nuestro objetivo es ver como estos factores afectan el delay general, y modelar el comportamiento de los mismos.
3. **Desde donde conviene viajar:** Ver el desempeño(OTP) en distintos aeropuertos ubicados en distintas zonas. El objetivo de este eje es determinar si la zona donde está el aeropuerto afecta el desempeño del mismo.

2. Desarrollo

2.1. Nested Cross-validation

Nosotros implementamos el Cross-validation[1] para verificar la exactitud que tiene nuestro modelo al aproximar con las muestras de entrenamiento, evitando overfittear los datos. El problema es que no toda muestra de entrenamiento se puede usar para predecir ya que las mismas pueden tener problemas con la dependencia temporal. En esta versión tomamos una partición de tiempo anterior al tiempo que queremos predecir y predecimos la primera partición. En la siguiente iteración, hacemos lo mismo con la siguiente partición pero la nueva muestra de entrenamiento pasa a ser la que teníamos antes, más la partición que predijimos anteriormente. Hacemos esto hasta que ya no queden muestras para predecir.

2.2. Cuadrados Mínimos Lineales

Para poder predecir el comportamiento de los datos utilizamos distintas familias de funciones, y resolvimos ecuaciones normales. Utilizamos funciones polinomiales, trigonométricas y la unión de ambas a la cual vamos a llamar fusión. Esta función fusión toma el grado k de una función polinomial y una frecuencia α entre 0 y 2π. Luego la forma de esta es = a*sen(x*α) + c*cos(x*α) + pol(k), donde pol(k) es un polinomio de grado k. La unión de estas dos familias de funciones se ve en la matriz de cuadrados mínimos como una extensión de la otra familia de funciones (si lo vemos desde el punto de vista de un polinomio se agregan 2 columnas, una por el seno y el coseno, y si se ve desde una trigonométrica, se agregan n-1 columnas, n siendo el grado

del polinomio donde la constante no se agrega porque ya tenemos una). En la experimentación utilizamos siempre esta ultima función porque es la que mejor aproximó los datos. Además, usamos a lo largo de toda la experimentación la función predecir, que junto al cross-validation, se encargan de auto-configurar el programa para que este dé como resultado el menor RMSE[2] posible; eligiendo el mejor grado del polinomio y valor α para nuestra función. Analizamos el grado de 0 a 20 para polinomios y la frecuencia de 0 a 2π(con pasos de 0.01 multiplicados por π) para las funciones trigonométricas. Para medir el error de fitteo usamos la métrica RMSE, para medir el error de predicción, usamos la métrica ECM

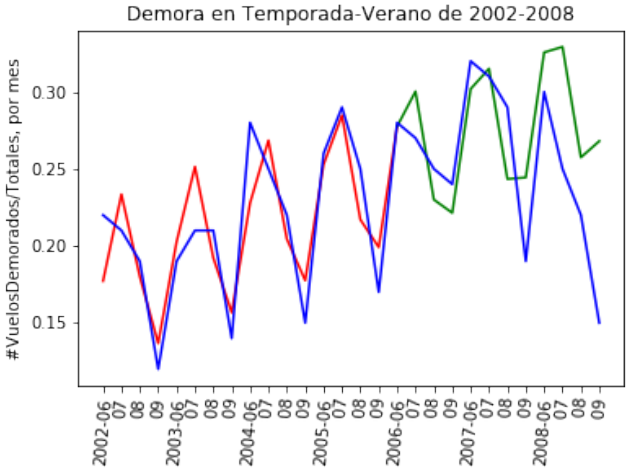
3. Experimentación

3.1. ***Primer Eje:*** *Cuando conviene viajar*

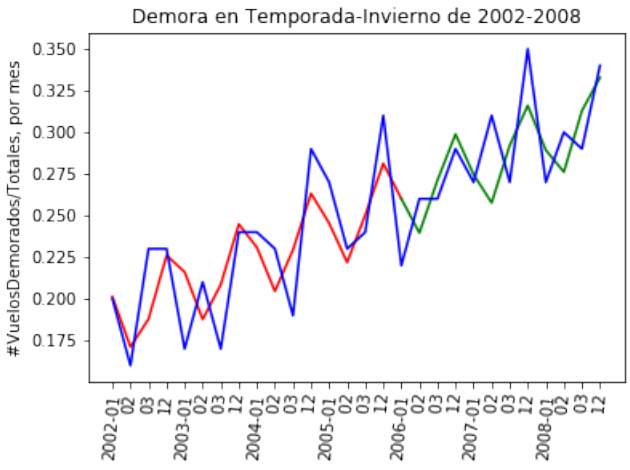
En este eje experimentamos como se comporta el OTP general a lo largo del tiempo, para esto hicimos 2 experimentos; uno basado en estaciones del año y otro centrándonos en los días de la semana.

3.1.1. **Demora según estaciones**

Para este experimento quisimos ver como afectaban las temporadas o épocas del año a la demora de los vuelos en general, bajo la **hipótesis** de que influían de forma consistente. Para esto tomamos 2 temporadas: la de los meses que comprenden Diciembre, Enero, Febrero y Marzo (meses con clima invernal) y la de los meses que comprenden Junio, Julio, Agosto y Septiembre (meses con clima veraniego). Para estas 2 tomamos la cantidad de vuelos demorados, sumados mensualmente, y divididos sobre la cantidad de vuelos totales; esto lo hicimos entre los años 2002 y 2008, y analizamos por separado las temporadas. Entrenamos nuestro modelo por 4 años e intentamos predecir los siguientes 3:



(a) RSME: 0.09740, ECM: 0.00244



(b) RSME: 0.08868, ECM: 0.00067

Figura 1. Predicciones según temporada

Lo primero que vemos es que ambos gráficos comparten la misma escala, con leves diferencias. En el gráfico 1a como en el 1b podemos ver un comportamiento cíclico, por lo tanto, las temporadas mantienen un comportamiento estable. Sin embargo, en el gráfico de la tempora-

da de invierno podemos ver un aumento sostenido de la demora conforme pasan los años, al contrario de la de verano, que parece crecer pero no tan rápido.

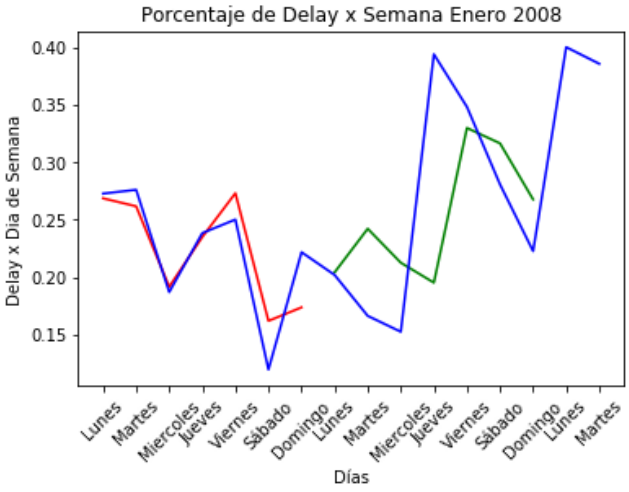
Analicemos ahora este comportamiento: en verano vemos que lo picos comienzan en junio y luego decrecen. Las altas demoras pueden deberse primordialmente a las vacaciones escolares de verano que comienzan en junio, y que terminan en agosto, y la baja demora, a la temporada de huracanes que suele darse en septiembre. En invierno vemos que el aumento de la demora ocurre consistentemente en el mes de diciembre, donde se encuentran las fiestas navideñas, y algunos días de vacaciones que pueden tener los usuarios.

Sobre nuestro modelo de predicción podemos ver que falla más al aproximar el gráfico 1a pues el entrenamiento es más cíclico que el período de predicción y en estos últimos años la función decrece rompiendo el comportamiento lineal que se observaba en el entrenamiento. Mientras que en el gráfico 1b, el error de aproximación es menor, lo que muestra que en estos años hubo un comportamiento más regular con una tendencia de aumento, que se mantiene desde el período de entrenamiento al de predicción, por lo que la función puede predecirlo de mejor manera.

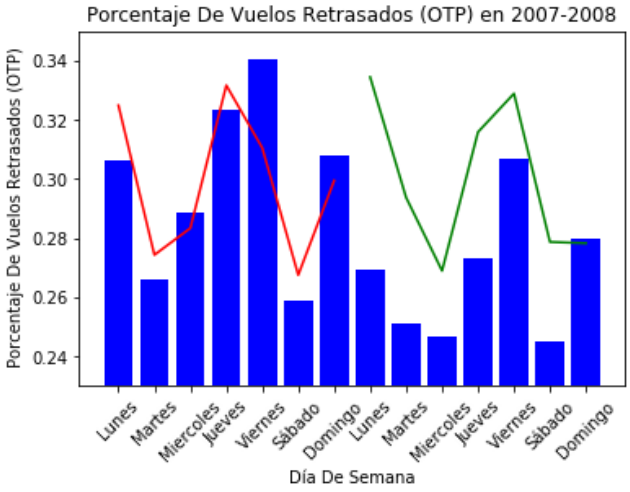
En conclusión a este experimento podemos ver como la época del año, y en especial las fiestas y vacaciones, alteran la demora de los vuelos. Además vemos que este comportamiento se mantiene a lo largo de los años, especialmente en invierno, por lo que las aerolíneas y los usuarios deberían estar preparados para estos retrasos usuales.

3.1.2. **Demora por día de semana**

En este experimento nos propusimos ver si existe un comportamiento predecible del porcentaje de demora (OTP) en los días de la semana, y si es así, determinar el día más propicio para viajar. Nuestra **hipótesis** es que los miércoles van a tener menos demora por ser un día de mitad de semana y donde no suele haber feriados. En la figura 2a, tomamos una semana de Enero de 2008 y tratamos de predecir la siguiente. Para ilustrar y ver si era general este punto, en la figura 2b graficamos los porcentajes de vuelos demorados de todo del año 2007 por semana, y tratamos de predecir los del año 2008.



(a) RSME: 0.191, ECM: 0.00750



(b) RSME: 0.058, ECM: 0.00142

Figura 2. Predicciones según día de la semana

En la figura 2a, nuestra función que mejor determina nuestro error tiene un componente

polinómico de grado 1, que muestra que el comportamiento semanal puede ir aumentando semana a semana(o disminuyendo, pero nuestra predicción muestra que en ésta, en particular, se supuso un aumento), que es lo que efectivamente ocurre en la siguiente, donde ambas semanas muestran comportamiento similar. Los picos superiores se dan los viernes y los inferiores se dan los miércoles y sábados.

Realizamos otra experimentación tomando tres semanas e intentando predecir la siguiente, esto nos da un peor resultado, donde el error de entrenamiento(RMSE) es igual a 0.223 y el de predicción(ECM) es igual a 0.01069. La causa de esto puede deberse a que el modelo aproxima más a una tendencia mensual, mientras que la regularidad cíclica que observamos en 2a es semanal. En cuanto al gráfico 2b, del mejor día de la semana observado anualmente, podemos notar que la predicción tiene menor porcentaje de retrasos en 2008, sin embargo, nuestro modelo puede predecir su comportamiento (picos en los mismos días).

Finalmente, podemos concluir que los días de mayores retrasos son los viernes y los días de menores retrasos son los miércoles y sábados, siendo este último el de menor retraso, reafirmando lo visto en las semanas de enero, del gráfico 2a.

3.2. ***Segundo Eje:*** *Causas de demora*

En el segundo eje, nos enfocamos en las causas de demora por separado, focalizándonos en las demoras generadas por clima, la generada por las aerolíneas, y la generada por los aviones. Nuestro objetivo es ver: como las causas climáticas afectan el comportamiento de los vuelos, como se comporta comparativamente las demoras en las distintas aerolíneas y ver si la elección de la aerolínea es un factor determinante de demora, y por último, ver como la antigüedad de los aviones influye en las demoras de los arribos tardíos (Late Aircraft Delay).

3.2.1. **Demora por Clima**

En este experimento queríamos ver como afectaba el clima, en particular, a la demora de los vuelos. Para ello, tomamos los años 2004 y 2005 y analizamos de enero a diciembre, entrenando nuestro modelo con un año, e intentando predecir el siguiente, bajo la **hipótesis** de que verano e invierno tendrían la mayor demora, en comparación con otoño y primavera, donde el clima es menos extremo:

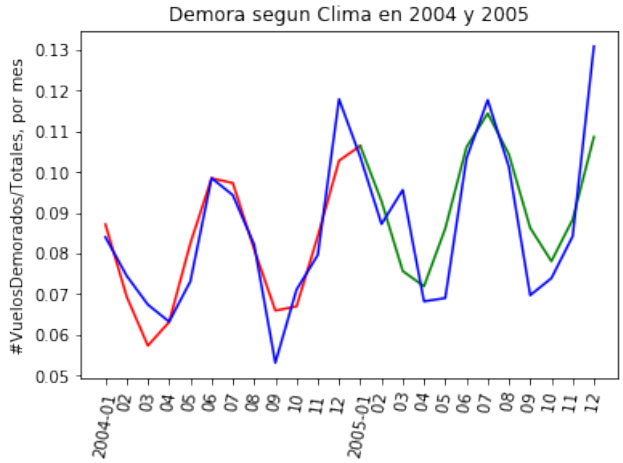


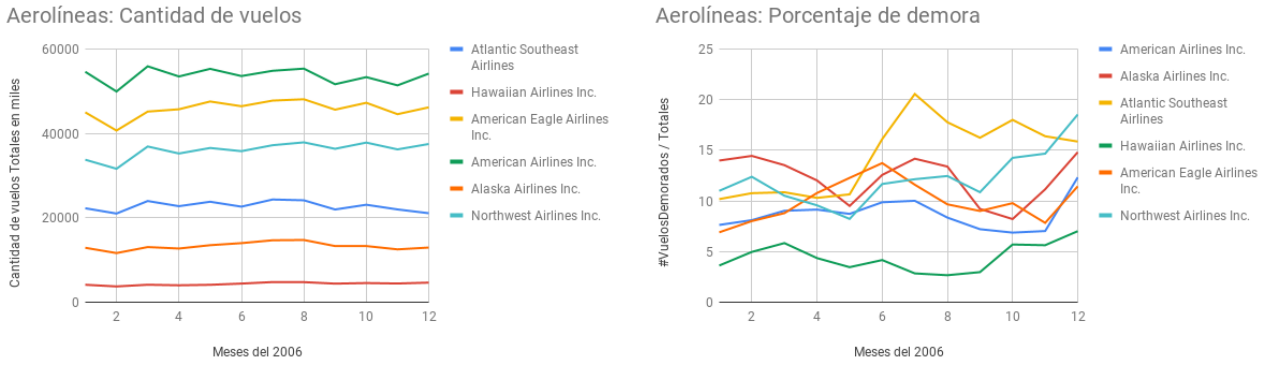
Figura 3. RMSE: 0.04094, ECM: 0.00013

Podemos ver en la figura 3 el comportamiento cíclico y creciente de la demora por clima. Podemos notar a su vez, que los picos de demora se encuentran en junio/julio y diciembre. Las demoras de verano pueden deberse a la necesidad de viajar a lugares más frescos para evitar las altas temperaturas, además estas fechas coinciden con la temporada de vacaciones, permitiendo a la gente, tener tiempo libre para viajar. Y bajo la premisa de evitar el clima extremo, podemos decir lo mismo de diciembre, el clima fuerte y las tormentas de nieve logran influir en las demoras de los aviones. Notar que las demoras por clima y por temporada encuentran sus picos en los mismos meses, pero recordemos que este experimento solo toma la demora por clima de los vuelos, mientras que el de temporadas toma las demoras generales (podrían haber dado resultados distintos).

Para concluir este experimento, podemos decir que efectivamente, el clima afecta la eficiencia de los vuelos, y tantos los aeropuertos, las aerolíneas y los usuarios deberían actuar en consecuencia. Sobre el comportamiento de nuestro modelo de predicción, podemos ver que este comportamiento es bastante suave y cíclico, por lo tanto, aproxima bien y da un error bajo.

3.2.2. Demora según Aerolíneas

Para este experimento, quisimos ver como afectaba al usuario tomar una u otra aerolínea(carrier), basándonos en la demora general propias que estas suelen tener, independientemente de en que zonas trabajan, el clima, los aeropuertos visitados, etc. Siendo nuestra **hipótesis** de que tendrán distintas eficiencias y conviene tomar la de mejor comportamiento. Para ello, primero, tomamos 6 aerolíneas y analizamos su comportamiento en el año 2006. Para eso, averiguamos la cantidad de vuelos y la cantidad de vuelos demorados de cada una, por mes. El gráfico 4a nos muestra la cantidad de vuelos por mes de las 6, y el gráfico 4b nos muestra el porcentaje de vuelos demorados sobre totales, por cada mes:

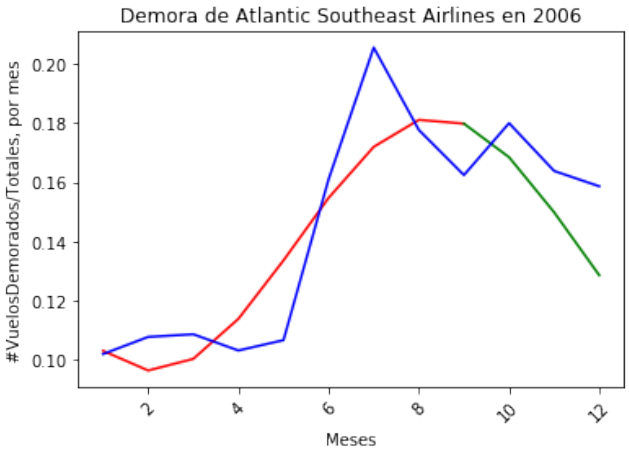


(a) Cantidad general

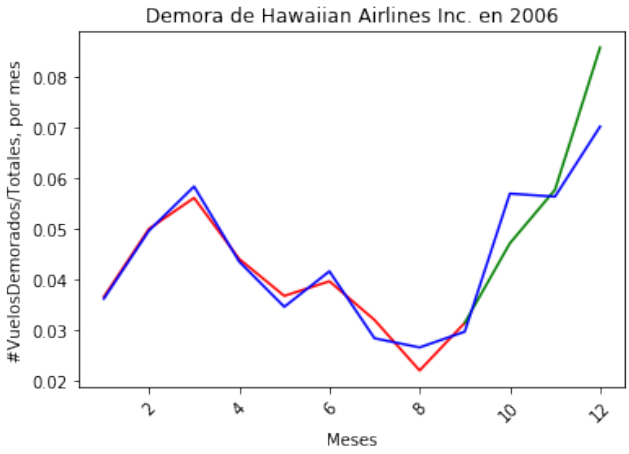
(b) Porcentaje

Figura 4. Análisis de Aerolíneas

Luego de analizar estos datos, quisimos predecir los siguientes 2 comportamientos: el de peor comportamiento, que es Atlantic Southeast Airlines, como vimos en 4b, y el de la aerolínea con mejor comportamiento, el cual es Hawaiian Airlines. Para esto tomamos los vuelos de ambas aerolíneas durante el 2006, sumados por mes y luego, entrenamos nuestro algoritmo por 9 meses e intentamos predecir los restantes 3. Con esto obtuvimos los siguientes 2 gráficos:



(a) RMSE: 0.04781, ECM: 0.00038



(b) RMSE: 0.01375, ECM: 0.00008

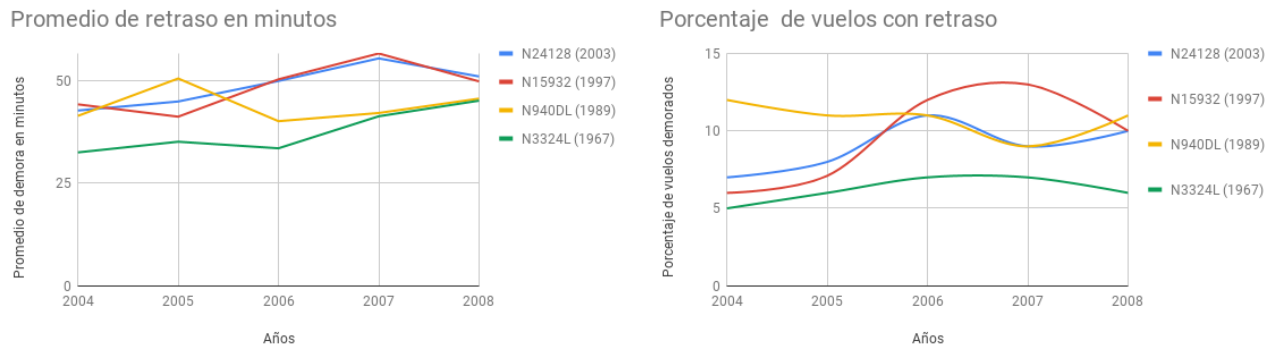
Figura 5. Predicciones por Aerolínea

Primero, observemos que la escala del eje y de ambos gráficos difieren, ya que el porcentaje comparado de Hawaiian airlines es menor que el de Atlantic Southeast, como se ve en el gráfico 4b. Luego, analicemos que nos muestran: el gráfico 5a da un error de aproximación mayor que el gráfico 5b, y se debe el gran pico en el mes de julio, que nuestro algoritmo aproxima de forma suave y por lo tanto, no predice correctamente en el resto de los meses, ya que, recordemos que nuestra función es una mezcla entre trigonométrica y polinomial. Este pico de demora puede deberse a mala administración de recursos de la aerolínea, a la temporada de vuelos, al clima, a los aeropuertos visitados, etc. Pero lo importante de ver es que en este mismo mes, la otra aerolínea no presenta este mal comportamiento, por lo tanto, podríamos descartar los factores externos. Para por decir esto último con más confianza, hicimos un pequeño análisis comparando que aeropuertos visitan principalmente, para ver si el clima podría afectarles de distinta manera, produciendo esta diferencia que vimos. Pero, en general, ambas aerolíneas se manejan en el oeste del país. Atlantic Southeast tiene sus principales hangares en Georgia y Michigan, y Hawaiian airlines, en Hawaii.

Por esto, concluimos que esta mayor demora puede deberse al manejo de la aerolínea, y por ello, es conveniente que el usuario conozca el porcentaje de eficiencia de las mismas para tomar una buena decisión. De parte de nuestro modelo de predicción, podemos ver que aproxima mejor si el conjunto de datos es suave, como se ve en el gráfico 5b.

3.2.3. Demora por Antigüedad de aviones

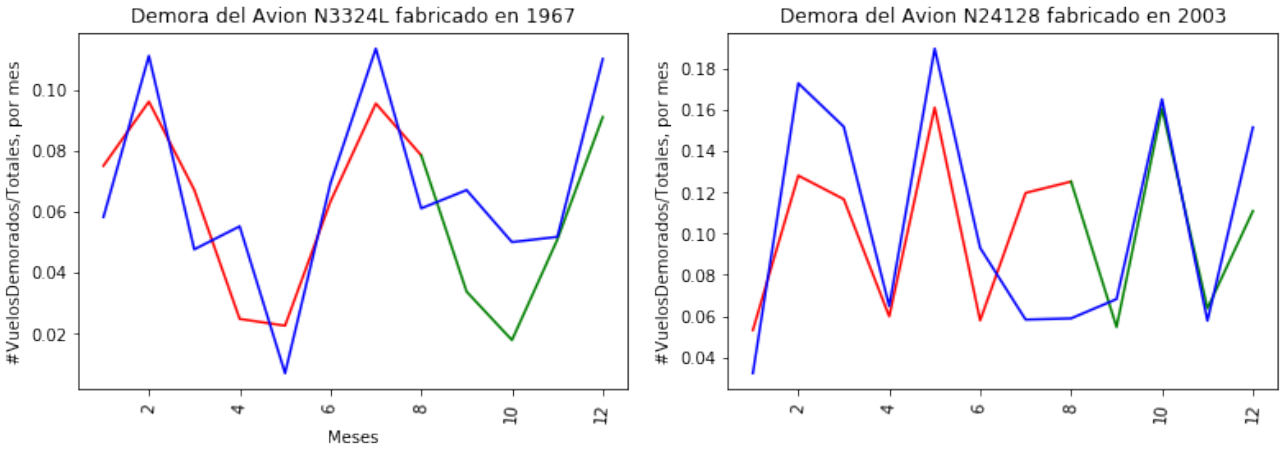
La motivación de este experimento es que la antigüedad de los aviones puede ser un factor influyente en el retraso de los vuelos. Luego, la **hipótesis** es que los aviones más antiguos tienen mayor demora que los aviones más recientes. Para este experimento elegimos cuatro aviones con diferentes años de fabricación que tuvieron vuelos en el periodo 2004-2008. Tomamos los vuelos que tuvieran demoras por arribo tardío o late aircraft delay (notar que esta demora se debe a que el avión llegó tarde por causas del avión en si mismo) y los dividimos por los vuelos totales:



(a) Promedio (b) Porcentaje

Figura 6. Análisis en Antigüedad de Aviones

Por lo que podemos ver en 6b el avión que menos sufre este tipo de demora es el más antiguo mientras que a medida que vamos viendo aviones más modernos estos son los que más lo sufren. Además de esto quisimos ver en 6a cuánto era el promedio de minutos de demora para ver si había alguna diferencia que apoyara nuestra hipótesis, una vez más resultó que el avión más viejo era el que menos promedio tenía mientras que el más nuevo era el que más minutos en promedio tenía. Como conclusión sacamos que la evolución de la tecnología afecta a la hora de preparar un avión para un vuelo, es decir, que esta tarea es cada vez más compleja por lo que lleva a un aumento de demora para los aviones más nuevos.



(a) RMSE: 0.047320, ECM: 0.00056 (b) RMSE: 0.09019, ECM: 0.00125

Figura 7. Predicciones por Antigüedad de Aviones

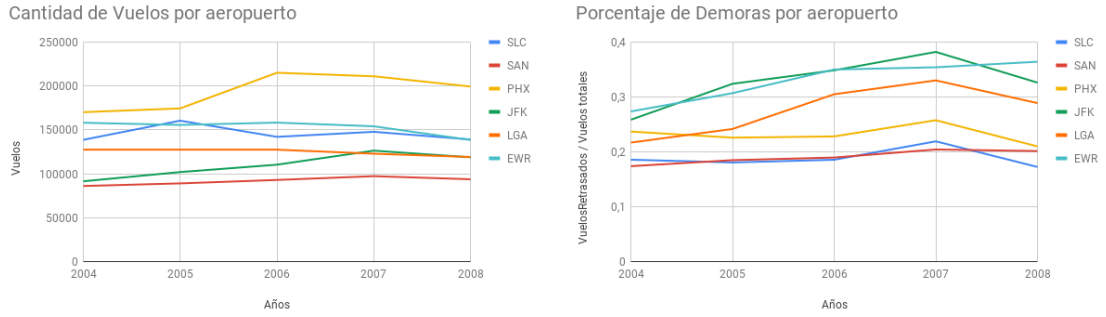
Luego tomamos el avión más nuevo y el más viejo y, usando nuestro modelo, quisimos ver si era posible predecir el comportamiento de los mismos en un año. Para eso, tomamos 8 meses de entrenamiento para el avión más moderno y 7 para el más viejo, e intentamos predecir los siguientes 5 meses para el avión más viejo y 4 para el más nuevo. En ambos casos vemos una función parecida, en 7a tenemos un ECM menor, producto de que la función tiene menos picos, pero en general la función es similar. Lo que podemos concluir es que como podemos predecir con mucha certeza el desempeño de ambos aviones entonces podemos esperar que la diferencia de demoras se mantenga en un futuro cercano, con lo que en general los aviones más modernos sufren más este tipo de retraso que los aviones más viejos.

3.3. ***Tercer Eje:*** *Desde donde conviene viajar*

En el tercer eje, comparamos el comportamiento de los retrasos en aeropuertos de distintas zonas, con el objetivo de determinar si el factor geográfico afecta a la demora, y cuál es la zona con la menor demora para viajar.

3.3.1. **Demora por Aeropuerto**

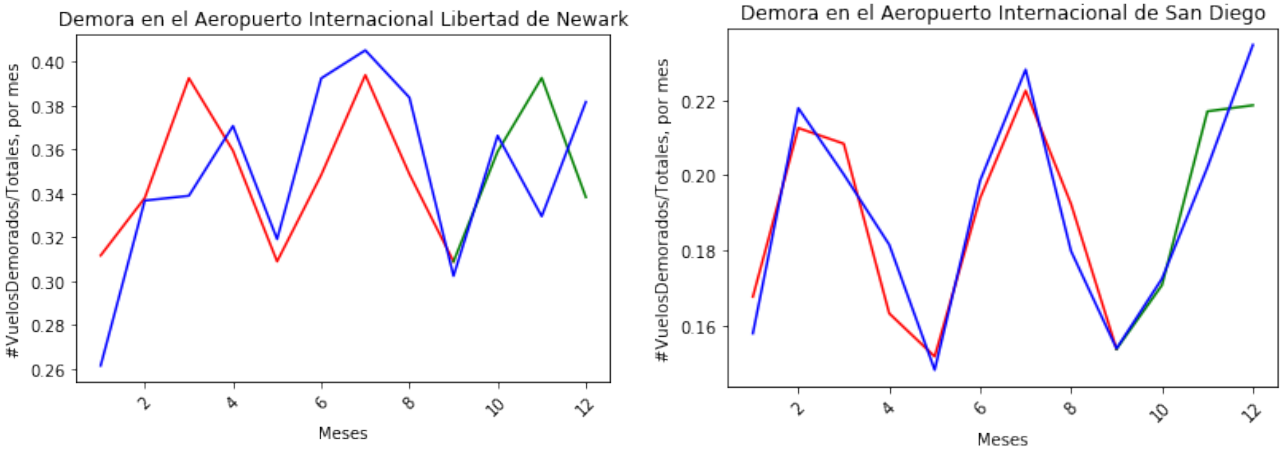
Para este experimento quisimos ver si la zona geográfica afectaba el comportamiento general de un aeropuerto, ya sea por la densidad de población como por la cantidad de vuelos. Para esto tomamos 6 aeropuertos, 3 al Oeste de Estados Unidos y 3 al Este, los cuales son: SLC (Salt Lake City Intl), SAN (San Diego International-Lindbergh), PHX (Phoenix Sky Harbor International), JFK (John F Kennedy Intl), LGA (LaGuardia, New York) y EWR (Newark Intl). Luego, tomamos los vuelos demorados por más de 15 minutos en un año y los dividimos por los vuelos totales de ese mismo. Esto lo hicimos en el período comprendido entre 2004 a 2008. Finalmente, tomamos la cantidad total de vuelos en ese período de cada aeropuerto y la comparamos:



(a) Cantidad General (b) Porcentaje

Figura 8. Análisis de Aeropuertos

Como podemos ver en 8b los aeropuertos ubicados en el Oeste tienen menos porcentaje de demoras que los ubicados al Este y en 8a vemos, que los aeropuertos del Este son los que menos vuelos tienen (exceptuando a SAN) pero, a su vez, los que **peor** desempeño tienen.



(a) RMSE: 0.06640, ECM: 0.00149 (b) RMSE: 0.03397, ECM: 0.00012

Figura 9. Predicciones en Aeropuertos

Finalmente, tomamos el mejor aeropuerto (SAN) y el peor (EWR) y quisimos ver si tomando 9 meses de entrenamiento de ambos podíamos predecir los 3 meses siguientes del año, para este experimento tomamos el año 2006. En 9b vemos que el error de predicción es muy bajo, esto se debe a que la función se comporta de manera estable en todo el año. Por otro lado a la hora de aproximar 9a no llegamos a predecir con exactitud el comportamiento ya que es una función menos estable que la anterior; aun así podemos concluir que en general, el porcentaje de demoras del aeropuerto del Oeste va a ser menor que el ubicado en el Este.

4. Trabajo a futuro

Pudimos conjeturar que el experimento de temporadas (1a y 1b), sería, quizás, más relevante hacerlo por días, en vez de meses, para concluir de forma más contundente sobre las demoras provocadas, en particular, por las fiestas y las vacaciones. Además, nos hubiese gustado ver la demora según la franja horaria del día, porque intuimos que este factor también influye considerablemente en las elecciones de los usuarios y en el manejo de las aerolíneas y los aeropuertos.

5. Conclusiones

Luego de este estudio, podemos concluir que:

1. En cuanto a los mejores momentos para viajar, determinamos que hay menos proporción de retraso en la temporada de invierno, y que los días de menor retraso en la semana son los miércoles y sábados.
2. En nuestro análisis de la demora producida por clima, determinamos que las mayores demoras de clima se producen en junio(verano), y en diciembre(invierno).
3. En la demora de distintas aerolíneas, vimos que la Atlantic Southest Airlines tiene peor comportamiento que el de Hawaiian Airlines, a pesar de trabajar en la misma zona geográfica. Por lo tanto, es conveniente, para el usuario, elegir bien la aerolínea que utiliza.
4. Sobre la antigüedad de los aviones, concluimos que los aviones nuevos pueden tener un poco más de demora en minutos que los viejos, pero la diferencia en el comportamiento de demora no era muy significativa.
5. Por último, sobre la zona donde conviene viajar, el análisis de distintos aeropuertos determinó que el promedio de demoras suele ser menor en el Oeste del país que en el Este.

Sobre nuestro modelo podemos decir que es útil para predecir comportamiento a futuro, si es que los datos de entrenamiento no presentan picos exagerados y sirve para aportar datos a las aerolíneas y a los aeropuertos, de modo que organicen sus recursos de forma eficiente. Además, permite a los usuarios realizar mejores decisiones a la hora de tomar un vuelo.

Referencias

[1] $Cross - validation$, Refaeilzadeh P., Tang L., Liu H. (2009) Cross-Validation. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA

[2] $RMSE$ Hyndman, Rob J.; Koehler, Anne B. (2006). " Another look at measures of forecast accuracy ". International Journal of Forecasting