Hoy te convertís en (Junior) Data Scientist

Ricardo Colombo Diego Santos Erik Machicado

Departamento de Computación
Universidad de Buenos Aires
C.A.B.A, Argentina

Resumen

El trabajo consiste en aplicar técnicas de Métodos numéricos y Data Science, en particular Regresiones Lineales y Cuadrados Mínimos, a un gran conjunto de datos reales, buscando identificar modelos capaces de predecir comportamientos.

Keywords: Data Science, KPIs, Cuadrados Minimos Lineales (CML)

1. Introducción

En la competitiva industria aeronáutica la eficiencia de los procesos es vital para mantener la calidad del servicio. Cumplir con el servicio no es algo que solo dependa de la empresa que se contrata, ya que año a año la cantidad de vuelos se multipilica y muchas veces existen variables externas que terminen afectando la puntualidad y calidad del servicio. Por eso es importante contar con métricas que analizen los eventos pasados buscando patrones e intentar prevenirlos o desminuir su impacto en el futuro. Estos indicadores son denominados *Key Performance Indicators* (KPIs).

En este trabajo estudiaremos factores que influyen en la organización de las salidas de vuelos para un aeropuerto en particular. Como se mencionó anteriormente, estadísticamente todos los años se registran mas vuelos para

cualquier aeropuerto, esto requiere una sincronización precisa de los tiempos que ocupa un avión dentro del aeropuerto esperando para partir.

Nuestro primer eje de estudio se centra en los factores estacionales que influyen en un vuelo salga en tiempo y forma, es decir estudiaremos la cantidad de retrasos en partidas para un aeropuerto particular, con el objetivo de detectar las estaciones anuales donde el clima pueda afectar el funcionamiento del aeropuerto.

Nuestro segundo eje de estudio es la evolución en la cantidad de tráfico áereo de un aeropuerto, y como la cuota de *market share* se fue concentrando con el correr del tiempo sobre las empresas líderes del segmento. Lo interesante de esta evaluación es esperamos poder proyectar el crecimiento de tráfico sobre el aeropuerto a fin de poder satisfacer la creciente demanda.

Para la realizacin del trabajo analizaremos un set de datos reales correspondientes a vuelos realizados en Estados Unidos durante el período 2004 - 2008 dado a que no habia datos para aos anteriores a 2004 en relacion a nuestras KPI's y la busqueda de encontrar una relacion entre ambos ejes de estudio.

Luego relacionaremos los ejes de estudio para estudiar si es posible predecir temporadas de mayor retraso, esperando que sea de utilidad en la programación de los nuevos vuelos que se incorporn al mercado.

Metodo de Minimos Cuadrados

Minimos cuadrados es una técnica de análisis numérico, en la que, dados un conjunto de pares ordenados intenta encontrar la función que mejor aproxime a los datos de acuerdo con el criterio de Error cuadrático medio. En su forma más simple, intenta minimizar la suma de los cuadrados de las diferencias entre los puntos generados por la función de aproximación y los que corresponden a los datos. CML es el caso en el que se usa a las rectas para la aproximación osea $y = ax + b$, además la generalización del método propone encontrar la función que mejor aproxima de la forma $y = a_1 f_1(x) + \ldots + a_K f_K(x)$ y que no es necesario que las funciones f sean lineales pero si que y sea una combinación lineal de ellas. Para mas informacion consultar [1] (seccion 8.1)

2. Desarrollo

Para la simplificación del problema y dado a que es replicable en otros aeropuertos optamos por centrarnos en el aeropueto *JFK* debido a que hoy día es uno de los aeropuertos que más movimiento presenta por la importancia

que presenta la ciudad donde está ubicado y sabiendo que tiene estaciones invernales marcadas a fines de ano que afectarían los vuelos ayudandonos en nuestro analisis. Como primer eje estudiaremos la cantidad de demoras que se producen por condiciones climáticas. Tomando como definición de demora, un vuelo que sale al menos 15 minutos despues de su horario programado ademas de que este tenga una demora debido al clima. El objetivo de este caso es identificar temporadas cíciclas donde las condiciones climáticas afectan el comportamiento normal del aeropuerto. Luego como segundo eje estudiaremos la evolución de la cantidad de vuelos de ese aeropuerto,entre dos de las empresas mas reconocidas mundialemente *Delta* y *United Airlines* y con mayor presencia de vuelos viendo su evolución a traves de los años.

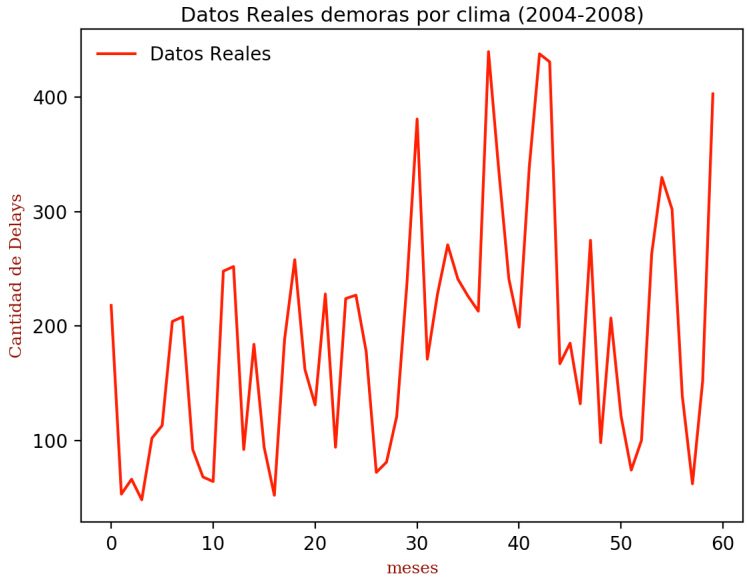
Para finalizar intentaremos probar si hay alguna relación entre el crecimiento de los vuelos y la cantidad de demoras estacionales generadas por el clima, una primera hipótesis que tenemos es que al aumentar el trafico areo la demora por clima debería aumentar pese a que no tienen relación directa creemos que cuando se demora un avión por clima arrastra esta demora a los demás.

Para la realización del estudio implementamos una serie de algoritmos utilizando el lenguaje *Python*. En la primer etapa procesamos el data set inicial filtrando los registros que nos interesa estudiar y los agrupamos por mes. Luego para entrenar nuestra solución propuesta tomamos un subconjunto de meses de los datos conocidos hasta encontrar el que nos de la mejor aproximacion, ejecutamos el proceso para una cantidad n también buscando la mejor combinacion con los meses elegidos y validamos los resultados obtenidos utilizando la técnica de *Coss-Validation*. Quedándonos con los que minímizen el error. Para evitar caer en el conocido overfitting elegiremos distintos subconjuntos del training para hallar los coeficientes de la aproximacion lineal, para ambos entrenamientos usamos el periodo 2004-2007 para entrenamiento y test, luego realizamos una aproximacion sobre el an 2008 y lo comparamos con los resultados reales.

Luego el siguiente paso es ver como con los datos entrenados nuestras funciones que tan bien aproximan al valor real e intentar predecir el comportamiento futuro.

3. Demoras Climáticas

Para comenzar en la busqueda de las funciones a utilizar analizamos los datos reales que nos arrojan la siguiente grafica.

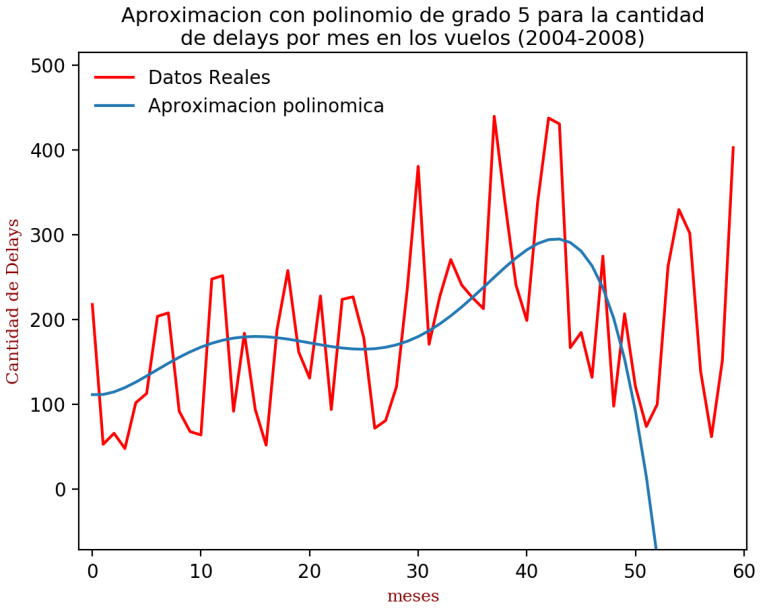


Intentamos buscar funciones que mejor aproximen en base a funciones conocidas como son los polinomios y las funciones trigonometricas.

Encontramos que la mejor forma de aproximación es una combinación de funciones trigonométricas sobre prolinomios, en la siguiente subseccion pasamos a mostrar el detalle de como llegamos a esta conclusión.

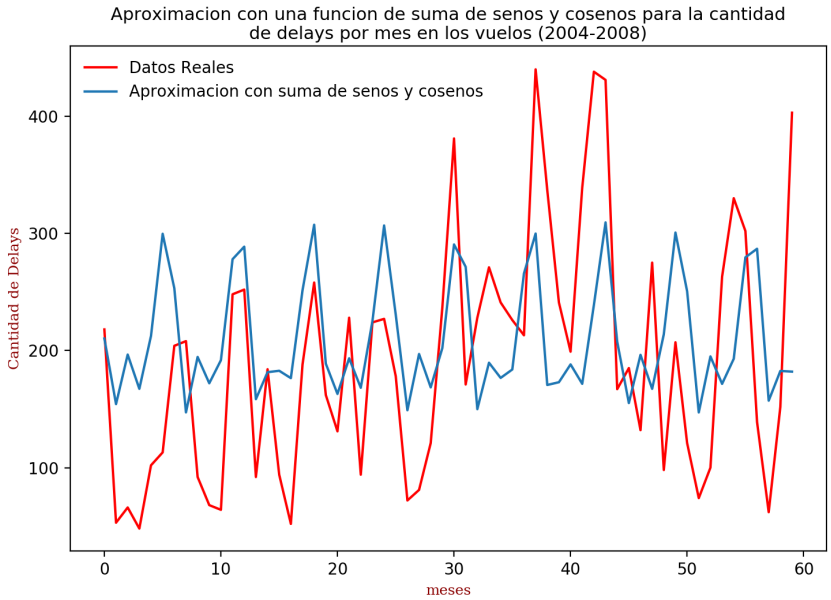
3.1. Resultados

En primer lugar tomamos un polinomio de grado impar ($a * x^5 + b * x^4 + c * x^3 + d * x^2 + e * x + f$, siendo a, b, c, d, e, f valores que nos da cuadrados minimos y x el numero de) , ya que luego de comparar contra uno de grado par notamos que era mas suave y tomaba mejor los puntos, para la aproximación de los meses 0-47, osea los aos 2004-2007, con el fin de ver como resultaba la misma arrojando los primeros resultados.

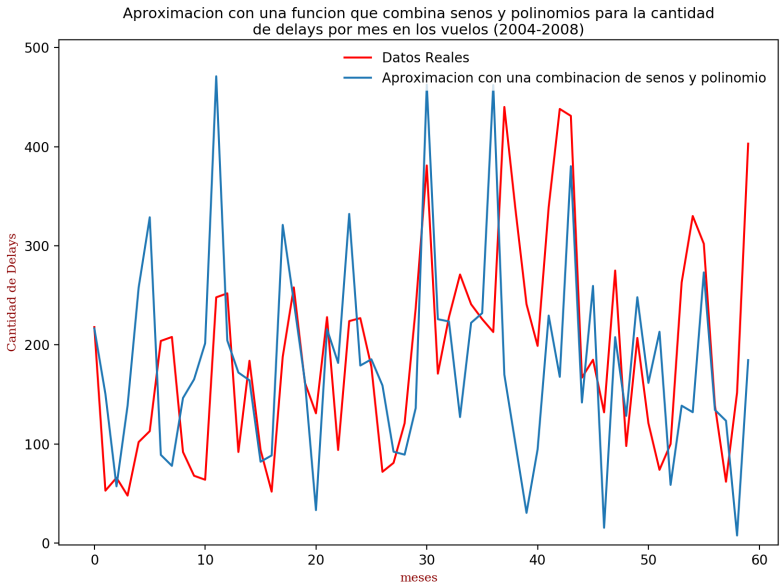


Como podemos ver en el grafico anterior esta aproximación es buena dentro de los meses de entrenamiento pero el problema que surge es que luego en la aproximación no aproxima bien haciendo que la función tienda a menos infinito, siendo este un problema para los polinomios, ya que a lo sumo un polinomio de grado n tiene n derivadas esto nos indica que a partir de un punto va a tender a más o menos infinito, pero por otro lado el problema es que no podemos agarrar los picos que tiene la funcioón real.

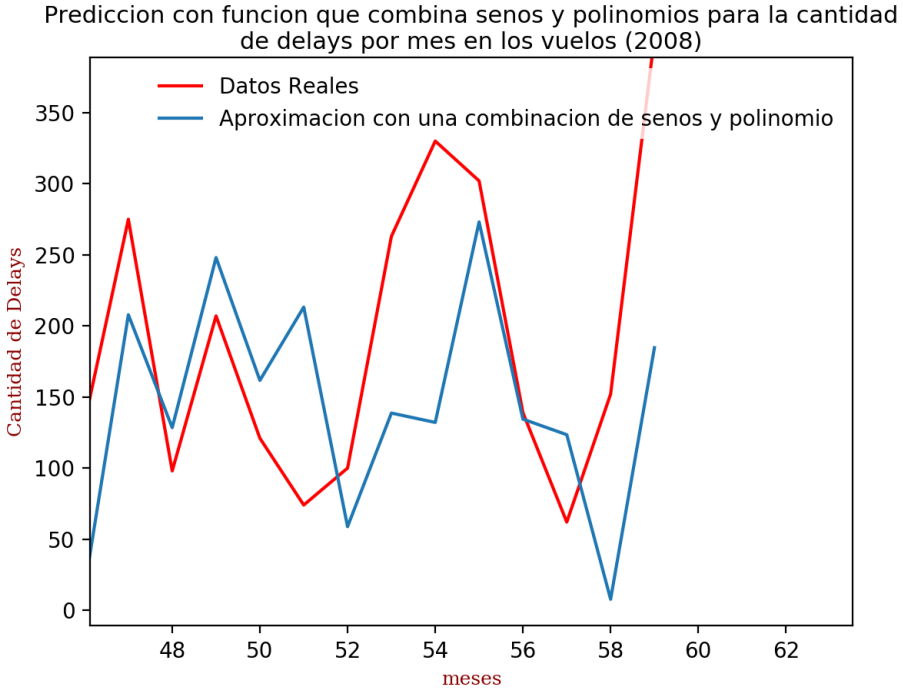
Para evitar este problema intentamos aproximar mediante funciones trigonométricas, mas puntualmente con senos y cosenos ya que estas estn acotadas tanto superior como inferiormente, la funcion utilizada fue $a * sin(x) + b * cos(x) + c * sin(2 * x) + d * sin(3 * x) + e * cos(2 * x) + f$.



Como se puede ver en el siguiente gráfico, el problema de utilizar estas funciones es que se mantienen dentro de esas cotas en línea recta, con lo cual si nuestra función a aproximar crece levemente durante toda su grafica no vamos a poder aproximarla bien. Por lo tanto decidimos hacer una combinación de ambas funciones para poder aprovechar ambas necesidades que teníamos, que la misma no tienda a infinito luego de los meses de entrenamiento y que en cierta forma podamos captar los picos de la función a aproximar, con lo que tuvimos una mejor aproximación.



Si vemos más en detalle lo que es la aproximación para 2008 notamos que no tenemos tanto margen de error resultando una buena aproximación.



Resultando dicha función $a * sin(x)^5 + b * sin(x^2)^4 + c * sin(x^2)^3 + d * sin(x^2)^2 + e * sin(x^2) + f$ y la combinacion de 20 ejecuciones para el cross validation y 24 meses de training.

4. Comparación entre vuelos de Delta - American Airlines

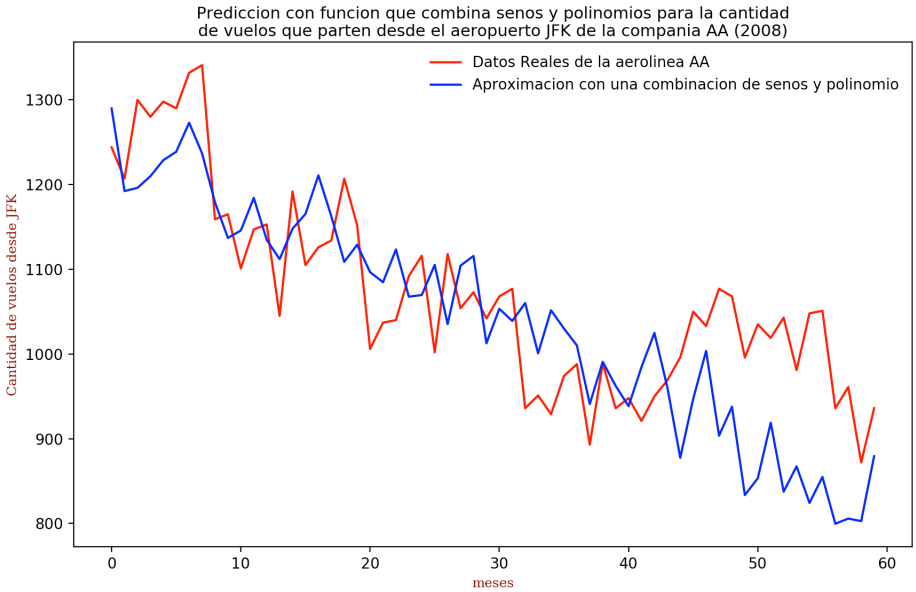
Luego de analizar los datos de vuelo reales de las empresas, comenzamos con la búsqueda de funciones que nos permitan predecir de forma relativamente certera los valores reales.

Como en el caso anterior empezamos probando con funciones polinómicas, y con el mismo resultado observamos que para obtener predicciones mejores requeriamos una combinación de funciones trigonométricas y polinómincas.

Para la construcción de nuestra predicción tomamos como set de entrenamiento los años 2004-2007 y realizamos una proyección sobre el año 2008 comparando los datos reales con la proyección obtenida por nuestra función.

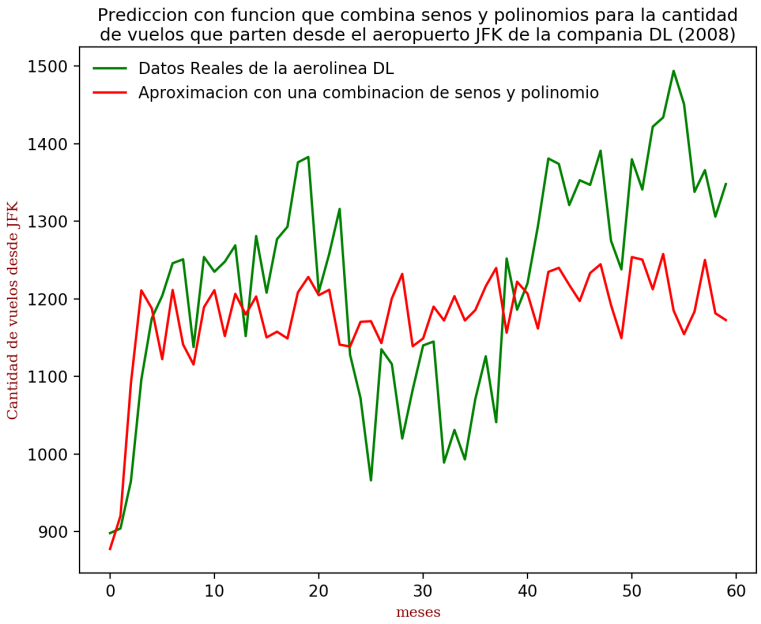
En el caso de la empresa *American Airlines* la función que hacia mejor fit sobre el set de datos fue la dada por: $a * sen(mes)^5 + b * sen(mes^2)^4 + c * sen(mes^2)^3 + d * sen(mes^2)^2 + e * mes + f$

Presentamos los resultados obtenidos para la proyección del 2008 en el siguiente gráfico



En el gráfico se puede observar que la función elegida se comporto de forma similar a la funcion dada por los valores reales, pero de forma mas suave. Es decir, la función propuesta no es sensible a los valores outliers que se pueden dar por picos de baja o alta demanda, por lo que en meses promedios acompañara la pendiente de los datos reales, pero en mese excepcionales quedara desfazada para ir reacomodandose en el correr del tiempo.

Y para el caso de la empresa *Delta* la función que hacia mejor fit sobre el set de datos fue la dada por: $a * tan(mes^6)^4 + b * sen(mes^8) + c * sen(mes^2) + d * sen(mes) + e$



En este gráfico se puede observar mas el efecto de haber tomado el la función seno en varios de los coeficientes. Los saltos en la venta de pasajes son suavizados pero siguiendo la misma tendencia. Tomamos la tangente para el coeficiente asociado al x_0 porque nos permitía poder representar mejor el crecimiento de de los primeros meses, sin usar una función exponencial, que en el largo plazo sería contraproducente ya que el creciemiento de ventas exponencial nunca podría ser sostenido.

5. Conclusiones

Referencias

[1] Burden, R. L., and J. D. Faires, "Análisis numérico, " Math. Surveys **7**, Amer. Math. Soc., Providence, R.I., 1961.

[2] ASA Section on Statistical Computing. 2009 data expor competition, URL: http://stat-computing.org/dataexpo/2009/the-data.html.