Cuadrados Mínimos Lineales

Assenza, Franco [1]

LU: 921/12

Maldonado, Kevin [2]

LU: 018/14

Umfurer, Alfredo [3]

LU: 538/12

Abstract

Este trabajo tiene como objetivo estudiar los retrasos de los vuelos en los aeropuertos de EEUU. Se utilizará los datos de años anteriores e intentará predecir los retrasos usando la técnica de regresión por cuadrados mínimos. Usando esta misma técnica estudiará la posibilidad de estimar el precio de las acciones de una aerolínea en función de la proporción de vuelos que posee.

Keywords: Regresión lineal, Retrasos de vuelos, bolsa

[1] Email: assenza.f@gmail.com
[2] Email: maldonadokevin11@gmail.com
[3] Email: aumfurer@gmail.com

1 Introducción

El análisis de la performance de una aerolínea es importante para distintos agentes: la aerolínea en sí, para analizar aspectos a mejorar; los clientes, para elegir la mejor opción a la hora de realizar un vuelo; la competencia, para tomar mejores decisiones de mercado; los gobiernos, para analizar la aerolínea de bandera y su competitividad, etcétera. El estudio de los retrasos de los vuelos es fundamental: vuelos retrasados pueden causar inconvenientes en la organización de los clientes, en sus combinaciones de vuelos. Retrasos excesivos pueden ser causas de resarcimientos económicos e influir negativamente en la imagen de la empresa.

Los algoritmos de aprendizaje automático nos permiten realizar predicciones de distintas variables en base a registros previos. En este caso, contamos con datos de registros de vuelos en Estados Unidos con información sobre la fecha, el origen y el destino, tiempos estimados y reales de despegue y aterrizaje, entre otros.

El análisis de series temporales tiene amplia difusión en la literatura (ver[3], por ej.), con aplicaciones importantes en muchas áreas del conocimiento. En este trabajo, vamos a enfocarnos en un modelo de regresión lineal para aprender sobre el cuerpo de datos y realizar un pronóstico a futuro.

1.1 Predicción

Para el modelo de predicción utilizamos regresión lineal[1], resuelto utilizando el método de Cuadrados mínimos lineales, con el algoritmo de[2] (pág 621) mediante la descomposición SVD.

1.2 Correlación

Para estudiar la correlación entre distintas variables (como se explica en la sección 3.1), se utilizará el coeficiente de Pearson [4], cuyo valor es 1 para variables directamente correlacionadas, -1 para variables inversamente correlacionadas, y 0 para no-correlación.

2 Desarrollo

2.1 Implementación y ejecución

El método de cuadrados mínimos fue implementado en `C++` utilizando la librería de cálculo numérico `Eigen` para la descomposición SVD, los experi-

mentos fueron realizados en `Python` utilizando las librerías de cálculo numérico `numpy` y `scipy`, y las librerías de visualización `matplotlib` y `seaborn`. Todo se ejecutó en una máquina con procesador `Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz` y `32 GB` de memoria RAM.

2.2 Datos

Los datos pertenecen al *U.S. Departmen of Transportation* [4], utilizados en una competencia de visualización de datos de la *American Statistical Association* [5]. Dado que los datos son susceptibles a problemas en Estados Unidos, en este trabajo estudiamos los datos a partir de 2002 (para evitar el ruido generado por el atentado a las torres gemelas) y antes de septiembre de 2008 (para no considerar los problemas de la recesión de 2008). Además, contamos con datos de los precios de las acciones de algunas aerolíneas, para buscar correlación entre estos y los datos del dataset de vuelos. Estos fueron extraídos de `Yahoo! Finance` [6] y se encuentran en `carrier_stock_prices/`.

2.3 Transformación

Al leer los datos, consideramos el retraso de un vuelo como: 0 si el vuelo está adelantado o no está atrasado, o el máximo entre el retraso de despegue y el retraso de aterrizaje (en minutos) con un valor máximo de 60 minutos. Estas decisiones fueron basadas en la mejor performance del algoritmo.

2.4 Extracción de features

La extracción de features se realizó de manera iterativa: se planteó una hipótesis (por ejemplo: "Los datos tienen una tendencia lineal creciente"), se ajustó una función acorde a la hipótesis (por ejemplo: una función lineal), se eliminó esa componente de los datos (por ejemplo: restarle a los datos el resultado de la función lineal) y se procedió a estudiar la siguiente hipótesis. Los features considerados fueron: tendencia lineal al crecimiento, periodicidad con distintas frecuencias, pico de retraso para días cercanos al 31/12, retraso medio de aerolíneas y aeropuertos (ver sección 2.5), día de la semana.

[4] https://www.transtats.bts.gov/Fields.asp?Table_ID=236
[5] http://stat-computing.org/dataexpo/2009/
[6] https://finance.yahoo.com/

2.5 Scores

Le asignamos a cada aerolínea un puntaje por año que es el estimado de retraso de la aerolínea ese año, y a cada aeropuerto dos puntajes por año, que son los estimados de retraso de ese aeropuerto como origen o como destino de vuelo. Los puntajes están entre 0 y 30 minutos como máximo.

2.6 Evaluación

Para la evaluación del algoritmo se utilizó un conjunto de datos distinto al de entrenamiento, se convirtió a los tiempos de retrasos en booleanos: 0 si el retraso es menor que 15 minutos, 1 si el retraso es mayor o igual que 15 minutos, y sobre ese arreglo de booleanos se reportaron las métricas:

- Delay RMSE: RMSE de los tiempos de retraso de los vuelos, en minutos
- Delay NRMSE: RMSE de los tiempos de retraso de los vuelos, normalizado
- RMSE: RMSE del arreglo de booleanos
- Accuracy, precision, recall, balanced accuracy: del arreglo de booleanos

3 Resultados

3.1 Features

Realizamos un *scatter plot* de los datos, exponemos los resultados en la figura 1.

Realizamos un ajuste lineal de los datos, los que nos da un nuevo dataset con media cero (proceso conocido como *detrending*).

Luego, realizamos gráficos que nos muestren cómo se comportan los retrasos dentro de los meses de un año, los días de un mes y los días de una semana. Exponemos los datos en las figuras 2a, 2c y 2b.

Las figuras muestran un comportamiento periódico de los retrasos. Es más claro el de la figura 2a que tiene elevaciones en los meses diciembre/enero y junio/julio, que es esperable pues son meses de vacaciones. También se ve en la figura 2b un aumento en los días miércoles/jueves y domingo/lunes. El comportamiento parece ser menos predecible para la figura 2c.

A modo de ejemplo, se muestran los más altos scores (explicados en la sección 2.5) del año 2008 en las figuras 3a y 3b.

Sobre la base de este análisis se escogieron los features descritos en la sección 2.4. Se ajustó un modelo de cuadrados mínimos (sección 1.1). En la figura 4 se muestra el resultado del modelo con datos de entrenamiento entre

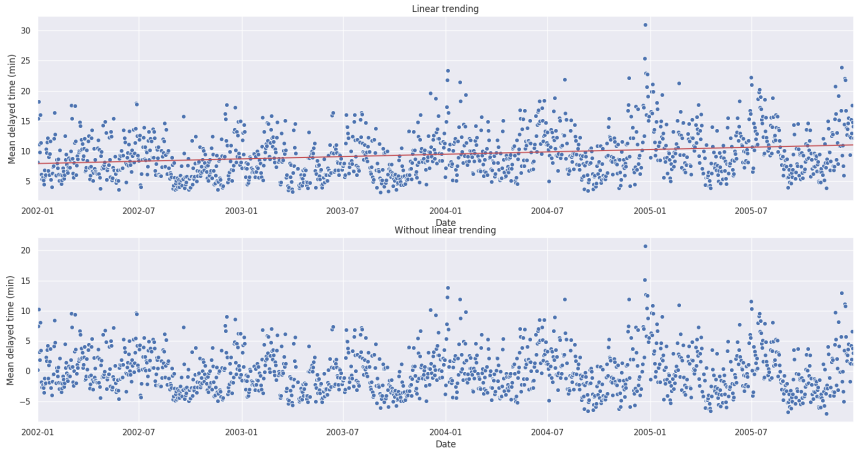
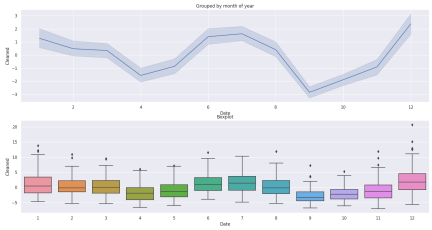
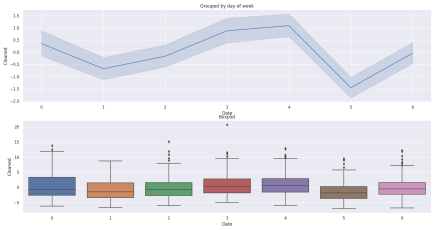


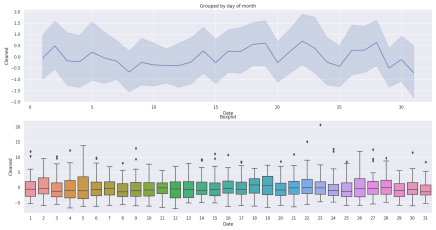
Fig. 1. *Detrending* de los datos



(a) Retraso medio por mes



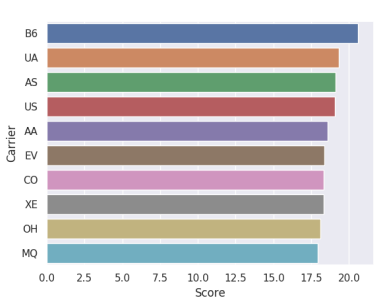
(b) Retraso medio por día de la semana



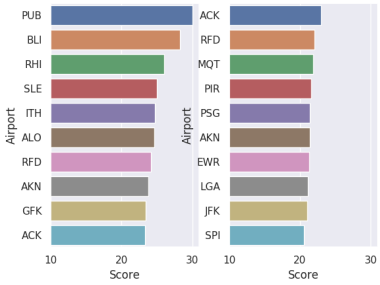
(c) Retraso medio por día del mes

2002 y 2006 y datos de test a partir de 2006.

Nos planteamos la pregunta: ¿Aprende el algoritmo efectivamente de los datos? ¿Mejora la predicción si entrenamos con más datos? Para responder a esta pregunta, ejecutamos el modelo entrenando con una cantidad creciente de años, desde uno a todos menos uno, y testeamos contra los demás años. Exponemos los resultados de las métricas en la figura 5



(a) Primeros puntajes de aerolíneas



(b) Primeros puntajes de aeropuertos (origen y destino)

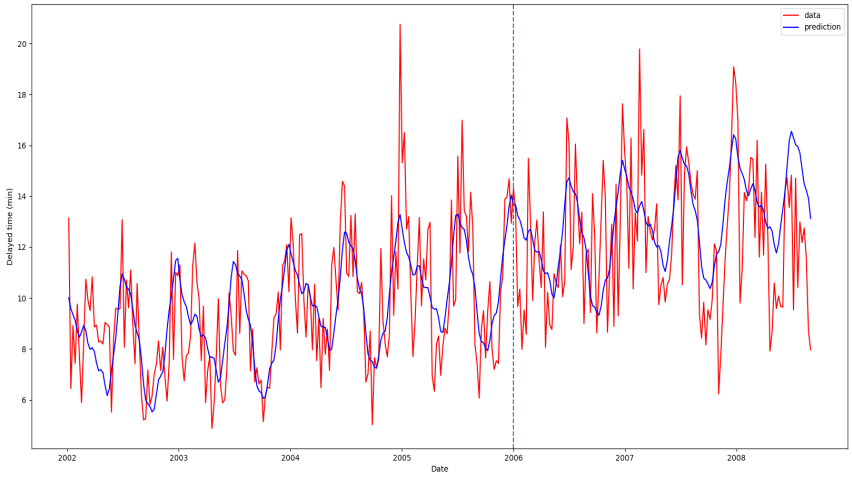


Fig. 4. Ejemplo de predicción. La línea punteada delimita el fin de los datos de entrenamiento y el comienzo de los datos de test

Training years	Delay RMSE	Delay NRMSE	RMSE	Accuracy	Precision	Recall	Balanced Accuracy
1	14.19	0.24	0.53	0.72	0.32	0	0.5
2	13.94	0.23	0.54	0.71	0.43	0	0.5
3	14.01	0.23	0.57	0.68	0.42	0.21	0.54
4	14.42	0.24	0.58	0.66	0.42	0.26	0.55
5	14.62	0.24	0.59	0.65	0.42	0.27	0.55
6	14.81	0.27	0.6	0.64	0.40	0.34	0.55

Fig. 5. Métricas para distintos datos de entrenamiento

Los resultados de las ejecuciones para distintos conjuntos de entrenamiento son paradójicos, ya que algunas métricas mejoran (como Recall y Balanced Accuracy), pero otras empeoran (como Delay NRMSE). Nuestra hipótesis es que se puede deber a que el modelo con más datos de entrenamiento es más susceptible a *overfitting*, o que al entrenar con conjunto de datos más grande se toma en consideración datos viejos que no son tan influyentes en los datos más nuevos.

En la figura 6 se grafican la cantidad de vuelos y el precio de las acciones de algunas aerolíneas. En ella se puede ver algunas cosas notables, como por ejemplo el gran desplome de los precios de las acciones con la crisis de 2008, o la aparición importante de la aerolínea `AA`, y su caída en el 2008 aún más grande que la de las otras aerolíneas.

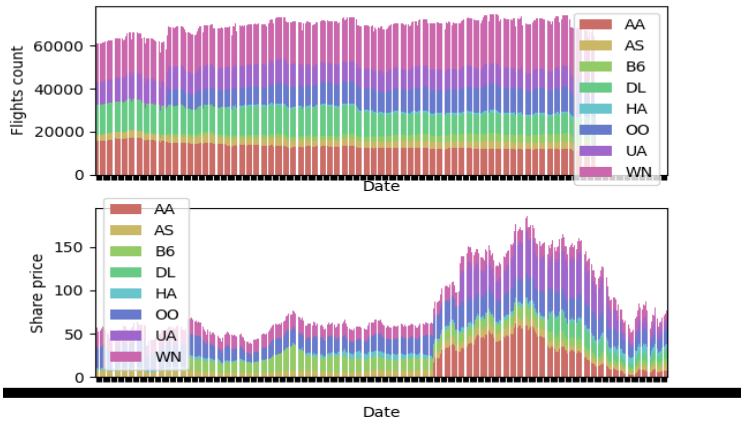


Fig. 6. Cantidad de vuelos y precio de las acciones por aerolínea

Una pregunta que nos planteamos es: ¿Están correlacionadas ambas variables? Para estudiarlo, consideramos, para cada aerolínea, qué porcentaje de vuelos le corresponda a esa aerolínea, y qué porcentaje de precio de la acción le corresponde (sobre el conjunto de aerolíneas para las que tenemos datos de precios de acciones), y estudiamos la correlación (vía coeficiente de Pearson) de ambas variables a lo largo del tiempo. Exponemos los resultados del coeficiente de Pearson en la figura 7. De la figura se desprende que no hay una correlación evidente entre estas variables, ya que los valores del coeficiente de Pearson son variables entre 0.74 y -0.94.

Carrier	Pearson
AA	0.74
AS	0.33
B6	-0.94
DL	0.53
HA	0.05
OO	-0.6
UA	0.29
WN	-0.62

Fig. 7. Coeficiente de Pearson entre el porcentaje de vuelos y el porcentaje de precio de las acciones para cada aerolínea

4 Conclusión

En este trabajo pudimos apreciar las dificultades del problema de intentar predecir los retrasos en los vuelos. Observamos que hay factores que incrementan la posibilidad de que un vuelo se retrase, como por ejemplo el dia de la semana o eventos estacionales, pero aún con estos es difícil poder tener certeza en las prediciones obtenidas, lo cual se vió reflejado en los valores que obtuvimos en metricas que usamos, en particular la recall: De la mayoría de las veces que se predijo un retraso, el mismo no existió.

References

[1] Hastie, T., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd Ed., Springer, 2019.

[2] Burden, R., and Faires, J., "Numerical Analysis," 9th Ed., Brooks/Cole, 2011.

[3] Shumway, R., and Stoffer, D., "Time Series Analysis and Its Applications," 3rd Ed., Springer, 2010.

[4] Witte, R., and Witte, J., "Statistics," 9th Ed., John Wiley & sons, Inc., 2010.