

# Sistema de predicción de retrasos en vuelos

FERNANDO VALLEJO, Universidad de Castilla-La Mancha, Spain

JUAN JOSÉ CORROTO, Universidad de Castilla-La Mancha, Spain

ALVARO GUERRERO, Universidad de Castilla-La Mancha, Spain

JAVIER CORDOBA, Universidad de Castilla-La Mancha, Spain

## ACM Reference Format:

Fernando Vallejo, Juan José Corroto, Alvaro Guerrero, and Javier Cordoba. 2018. Sistema de predicción de retrasos en vuelos. 1, 1 (December 2018), 6 pages. <https://doi.org/0000001.0000001>

## 1 INTRODUCCIÓN

## 2 PROCESO KDD

### 2.1 Selección

### 2.2 Preproceso

### 2.3 Transformation

### 2.4 Minería de Datos (Primera Aproximación)

En esta primera aproximación realizaremos una clasificación binaria de forma que intentaremos predecir cuando un vuelo se va a retrasar (1) o no (0), en base a los datos que sabemos a priori en un vuelo. En este caso, usaremos los siguientes datos:

- El nombre de la aerolínea encargada del vuelo.
- El día de verano, es decir, dentro de la temporada del verano (Junio, Julio y Agosto), en qué día nos situamos.
- Hora programada de salida.
- Hora programada de llegada.
- Día de la semana, del 1 (Lunes) al 7 (Domingo).

Para crear este modelo de entrenamiento, hemos usado un árbol de decisión con el 70% de los datos respectivos al verano [1] que son más de 1 millón de registros.

### 2.5 Interpretación (Primera Aproximación)

Al realizar el *testing* con el resto de los datos (el 30% restante) obtenemos una matriz de confusión [2], como podemos observar no son datos muy buenos, si bien acertamos el 60% de los retrasos hay un 40% que debemos minimizar lo máximo posible ya que el objetivo de nuestro sistema es predecir todos los retrasos, es decir, minimizar los Falsos Negativos.

---

Authors' addresses: Fernando Vallejo, Universidad de Castilla-La Mancha, Ciudad Real, Ciudad Real, Spain, Fernando.Vallejo@alu.uclm.es; Juan José Corroto, Universidad de Castilla-La Mancha, Ciudad Real, Ciudad Real, Spain, JuanJose.Corroto@alu.uclm.es; Alvaro Guerrero, Universidad de Castilla-La Mancha, Ciudad Real, Ciudad Real, Spain, Alvaro.Guerrero1@alu.uclm.es; Javier Cordoba, Universidad de Castilla-La Mancha, Ciudad Real, Ciudad Real, Spain, Javier.Cordoba1@alu.uclm.es.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2018 Association for Computing Machinery.

XXXX-XXXX/2018/12-ART \$15.00

<https://doi.org/0000001.0000001>

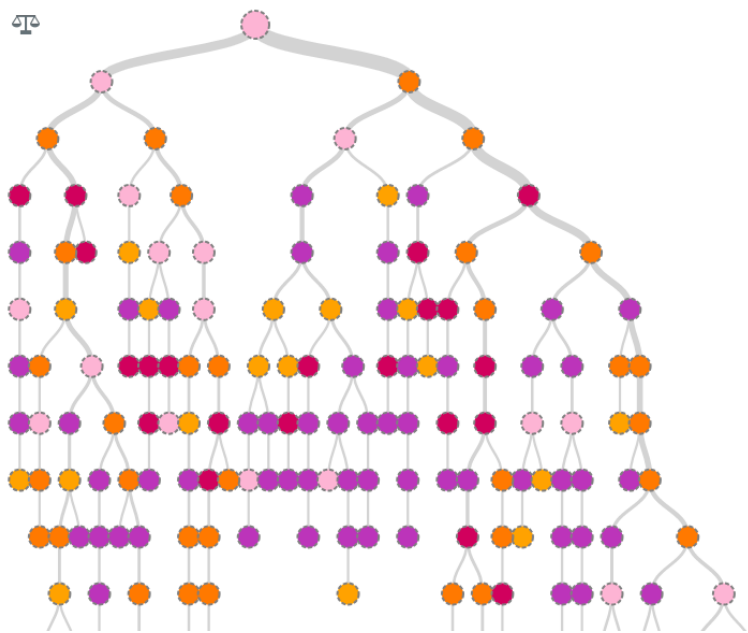


Fig. 1. Fragmento de árbol de decisión de la primera aproximación

ACTUAL VS. PREDICTED		0	1	ACTUAL	RECALL
0		152,075	80,348	232,423	65.43%
1		22,820	40,558	63,378	63.99%

Fig. 2. Matriz de confusión de la primera aproximación

Pero el objetivo de crear este árbol no era realmente extraer un buen modelo de predicción sino aumentar nuestro conocimiento sobre la importancia de las variables de nuestra tarjeta de datos [3]. Podemos ver como la variables más importante es la aerolínea que realiza el vuelo, pero tampoco es un dato excesivamente relevante ya que es un hecho que hay aerolíneas con más retraso que otras. Es decir, nuestro árbol de decisión ha realizado la mayoría de sus predicciones basándose en si ese vuelo lo realizaba una aerolínea con muchos retrasos (los marcaba como 1), o con pocos retrasos (los marcaba como 0).

Otro dato que cobra mucha importancia es el día del verano y las horas programadas de llegada y salida, esto ya nos da unas primeras pistas de que puede existir cierta estacionalidad, por lo que ya empezamos a intuir que una Serie Temporal podría darnos unos mejores resultados.

2.6 Minería de Datos (Segunda Aproximación)

Como una segunda aproximación realizamos otro árbol de decisión [4] para hacer una clasificación binaria de forma análoga a la anterior aproximación. En este caso nos quedamos solo con las variables relativas a la hora programada (llegada y salida), con una particularidad, en este caso es una variable categórica que representa el intervalo de tiempo en el que el vuelo se realiza (esta división fue hecha mediante uso de cuantiles para crear una distribución uniforme). Además incluimos el modelo de avión que del vuelo en concreto, por falta de datos referentes al modelo

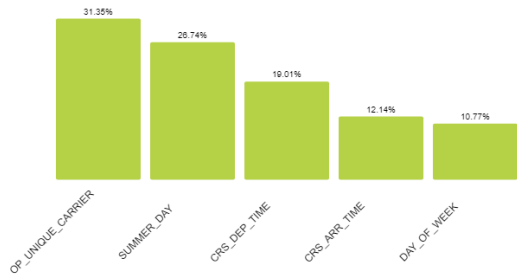


Fig. 3. Importancia de las variables de la primera aproximación

la cantidad de datos se redujo a la mitad, por lo que vimos conveniente empezar a usar todos los datos referentes al año completo, no solo el verano como en la aproximación anterior.

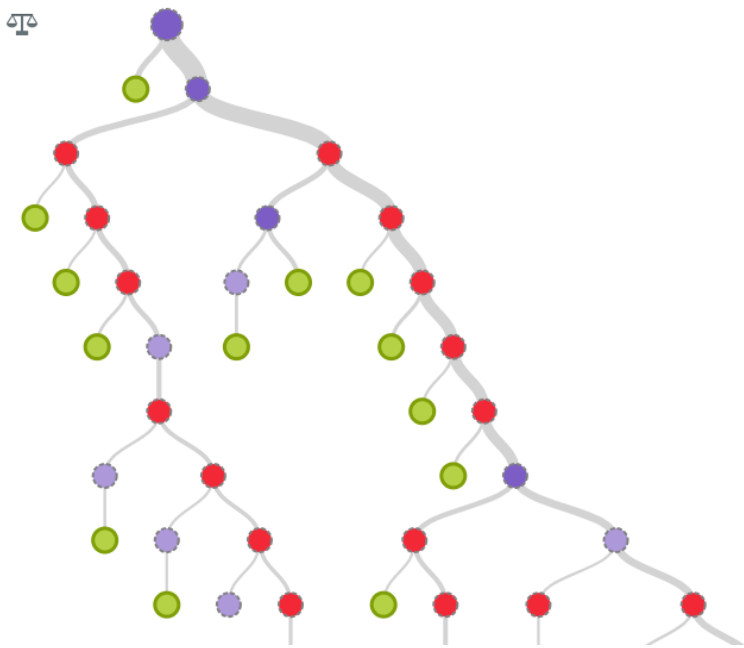


Fig. 4. Fragmento de árbol de decisión de la segunda aproximación

2.7 Interpretación (Segunda Aproximación)

Como se puede observar en el árbol [4] en los nodos de color rojo (modelo de avión), esta es una característica que parece cobrar mucha importancia (esto también se puede ver en [6]) en este caso es porque hay ciertos modelos de avión que realizan viajes más cortos y que otros, por lo que en un día pueden realizar muchos viajes y por tanto, acumular más retraso en el caso que se produzca. De nuevo el tiempo vuelve a cobrar cierta importancia por lo antes mencionado.

Los resultados mostrados en la matriz de confusión [5] vuelven a no ser muy buenos, si bien la proporción de falsos negativos se reduce, los falsos positivos aumentan. Aunque dar un falso

positivo no es algo crítico, sí que se puede observar un gran exceso de estos y habría que reducirlos, pero no es necesario una minimización de los mismos.


ACTUAL VS. PREDICTED		0	1	ACTUAL	RECALL
0		476,199	368,941		
1		65,312	117,937	183,249	64.36%

Fig. 5. Matriz de confusión de la primera aproximación

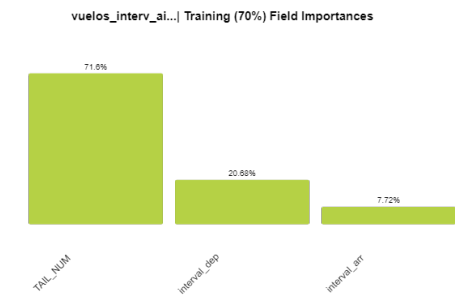


Fig. 6. Importancia de las variables de la primera aproximación

2.8 Regresión

Una vez terminadas las anteriores aproximaciones destinadas a conseguir clasificar los vuelos entre aquellos que se retrasaron y aquellos que no, con mejor o peor resultado, procedemos a intentar predecir cuanto se retrasará un determinado vuelo.

Para ello, utilizaremos series temporales, en concreto, series temporales regulares. Por este motivo, necesitamos obtener registros que contengan información sobre intervalos de tiempo espaciados regularmente, que en nuestro caso, serán días. Así, obtendremos un nuevo *Dataset*, que contiene 426 registros, dado que estamos utilizando los datos correspondientes a un año completo y dos meses. El retraso correspondiente a cada uno de estos registros será el retraso medio de ese día.

Adicionalmente, utilizaremos los datos correspondientes a un solo modelo de avión. Por ello, ya tenemos nuestros datos para la regresión, los retrasos medios de un modelo de avión en cada día desde Abril de 2017 a Mayo de 2018. Utilizaremos un 90% de los datos para *training* y un 10% en *testing*, lo que se corresponde con utilizar los meses el primer año completo, de Abril a Marzo, en *training* y el resto en *testing*. En [7] vemos un ejemplo utilizando los datos del modelo Boeing 737-7H4.

Como podemos ver en [7], en la parte derecha (correspondiente a los datos de *testing*), el modelo (en verde) predice correctamente los retrasos, siguiendo los máximos y mínimos reales (en blanco).

Ahora bien, para obtener estos resultados, es decir, predecir correctamente el retraso que tendrá un vuelo, necesitamos tan solo dos datos, el día de dicho vuelo y el modelo de avión que se utilizará. Dado que el modelo de avión podría ser, o no, conocido de antemano, cabe preguntarse si los datos obtenidos utilizando un determinado modelo de avión pueden ser utilizados para otros modelos, considerados similares.

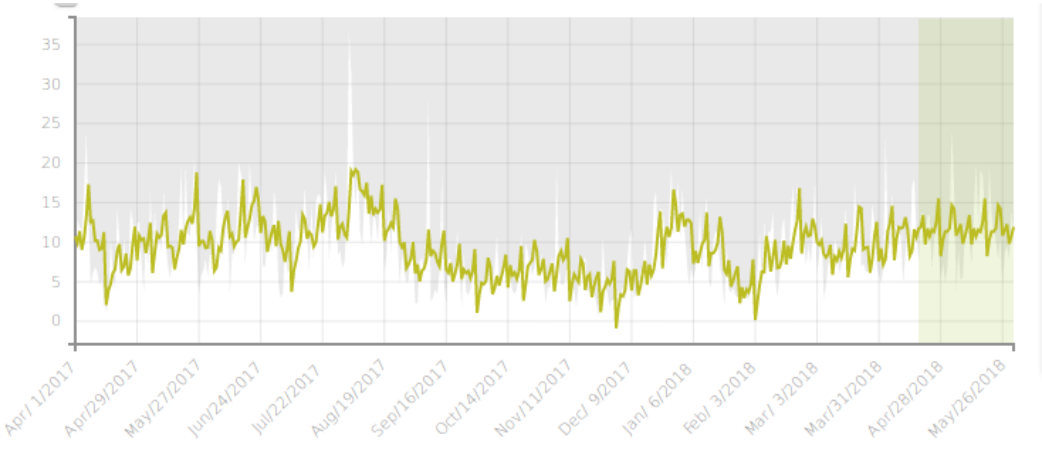


Fig. 7. Regresión obtenida utilizando los datos del modelo 737-7H4

Para ello, utilizaremos los modelos 737-7H4 y 717-200, que bajo nuestro conocimiento del problema cumplen funciones similares. Ambos son aviones comerciales, que transportan pasajeros con su equipaje, y que tienen capacidad para alrededor de 150 pasajeros. Los resultados obtenidos pueden verse en [8a] y [8b]. Los resultados no son prometedores, y queda patente que no sería posible utilizar los datos de un modelo para predecir el retraso de otros modelos distintos.

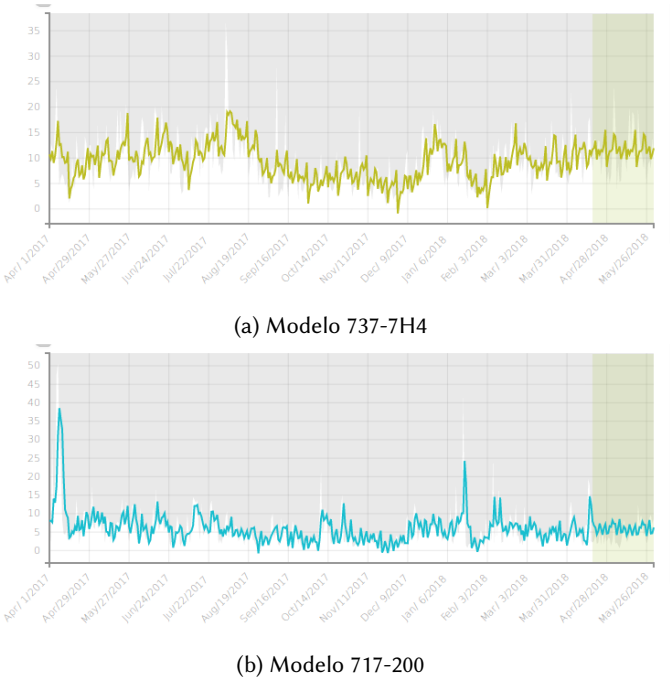


Fig. 8. Regresión utilizando dos modelos diferentes

### 3 CONCLUSIONES

A lo largo de este documento hemos obtenido diversos resultados, en nuestro intento de predecir retrasos en vuelos de avión, en Estados Unidos. Las primeras aproximaciones a clasificar los vuelos entre aquellos con retraso y aquellos sin él, dieron buenos resultados, si bien mejorables. En el caso de la regresión, no se obtuvo ningún resultado satisfactorio.

De cara al futuro, y a mejorar los resultados, consideramos posibles dos mejoras. La primera de ellas es obtener información adicional relativa a los vuelos, como por ejemplo, la climatología. La segunda mejora consiste en aumentar nuestro conocimiento del dominio, que nos permitiría un mejor uso de los datos que ya disponemos.