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), 8 pages. https://doi.org/0000001.0000001

1 INTRODUCCIÓN

Predecir el retraso de un vuelo es un asunto muy complejo. Pueden entrar en juego numerosas variables que son imposibles de predecir: Un fenómeno meteorológico inesperado, asuntos internos en la compañía de viajes que puedan retrasar un vuelo durante un tiempo, una posible avería, etc. Sin embargo, en este documento se va a intentar hacer un sistema de predicción en retrasos de vuelos con datos conocidos de antemano, como pueden ser el modelo de avión o la hora a la que está programada la salida del avión. Para construir este sistema, hemos seguido la metodología KDD, introducida por Fayyad et al. [1]. Siguiendo esta metodología, el proceso se divide en las siguientes fases: *Selección, Preproceso, Transformación, Data Mining* e *Interpretación*. Este proceso no es unidireccional, si no que se permite, y muchas veces es necesaria, una vuelta hacia atrás en el proceso para revisitar y mejorar las acciones que se tomaron en una etapa anterior.

2 PROCESO KDD

2.1 Selección

Para todo proceso KDD necesitamos una base de datos de suficiente tamaño como para acabar extrayendo conocimiento y patrones de ella. En este caso, la base de datos proviene del **United States Department of Transportation**[1]. Esta base de datos contiene información descriptiva de vuelos entre aeropuertos de Estados Unidos. Esto significa que no tenemos información de vuelos transoceánicos, vuelos que podrían aportar poca información porque se efectúan una vez al día como máximo, y no suelen tener retrasos. Además, los vuelos entre aeropuertos de Estados Unidos podrían ser también una buena aproximación a vuelos entre aeropuertos europeos, por lo que si el sistema acaba prediciendo de forma efectiva retrasos en vuelos estadounidenses, es probable que

[1] https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

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.

XXXX-XXXX/2018/12-ART $15.00
https://doi.org/0000001.0000001

funcione correctamente también en vuelos europeos.

La cantidad de datos proveniente de nuestra fuente es inmensa. Por eso, el primer esfuerzo fue reducir la cantidad de registros que teníamos, quedándonos únicamente con los datos desde Abril del año 2017, hasta mayo de 2018, incluidos. Tras efectuar esta reducción, nos quedamos con un total de 7240901 registros, que siguen siendo una cantidad demasiado grande y difícil de procesar. Por eso, decidimos quedarnos únicamente con los datos del verano de 2017. La idea es intentar construir un modelo para únicamente una etapa del año, y si es suficientemente bueno intentar ampliarlo al resto del año, consiguiendo así una reducción dimensional inicial importante.

Por tanto, la tarjeta de datos inicial consta de 1513787 registros para la primera aproximación (Aunque en modelos posteriores se toma la decisión de incluir todos los datos que discutimos anteriormente). La tarjeta de datos contiene las siguiente características para cada registro:

- Información temporal: el día, mes y año del vuelo, además del día de la semana.
- Información técnica: Compañía, matrícula del avión, número identificador del vuelo, y aeropuertos de salida y llegada.
- Información del vuelo: Horas programadas de salida y llegada, duración y distancia del vuelo. Además también se incluyen las horas de llegada y salida reales del vuelo(teniendo en cuenta si sufrieron retraso o no).
- Información de retrasos: Distinta información respecto al tipo de retraso que sufrió el vuelo: Retraso por controles de seguridad, retraso acumulado por vuelos anteriores, retraso en ser remolcado, etc. Además también contiene información sobre si el vuelo fue cancelado o desviado.

2.2 Preproceso

El preproceso de los datos es una etapa fundamental. En ella nos dimos cuenta de que los vuelos que no sufrían retraso no tenían un valor de cero en el retraso, si no que tenían un valor nulo. Por eso lo que se efectuó fue una limpieza de valores nulos, cambiándolos por un valor por defecto, en este caso de cero.

Al finalizar la etapa de preproceso descubrimos que había algunos valores de retraso con unos valores atípicos. Estos valores estaban muy por encima de los valores normales, hablamos por ejemplo de retrasos de más de 1500 minutos. Estos valores pueden ser debidos a retrasos de varios días debido a condiciones meteorológicas adversas. Estas situaciones no se pueden predecir, y por eso tomamos la decisión de eliminar estos valores atípicos. Para efectuar esta eliminación, usamos una detección estadística: Cómo se puede ver en la Figura 1, la distribución está altamente sesgada a la izquierda (Hay muchos más vuelos que sufren poco o ningún retraso). Además, los valores de la media y la desviación estándar son de 65.9 y 80.0 respectivamente. Decidimos eliminar todos los valores que están más lejos de 5 desviaciones estándar de la media, lo cuál nos deja con retrasos de como máximo 469 minutos.

2.3 Transformación

En la etapa de transformación se prepara la tarjeta de datos para ser procesada en la siguiente etapa. En esta etapa nos hemos centrado en eliminar aquellas características que no resultan apropiadas desde un punto de vista cognitivo para un modelo predictivo, y añadir nuevas características que creamos que pueden ser importantes a la hora de predecir.

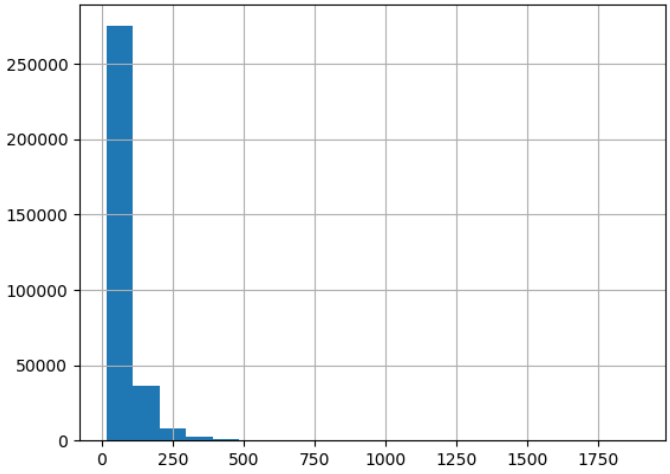


Fig. 1. Distribución del retraso de los vuelos

En cuanto a columnas que hemos eliminado, para la primera aproximación se eliminaron los datos que no se pueden conocer antes del vuelo. Estos son los datos de salida y llegada reales. Sólo deberíamos usar la hora programada porque es lo único que se sabe *a priori*. En aproximaciones posteriores también se eliminó la columna de compañía. También eliminamos columnas que no aportan información, como la del mes o año. También se combinaron todas las columnas de retrasos para crear una única característica, que después usaremos como variable a predecir.

En cuanto a añadir nuevas características, en el caso de los modelos donde nos centrábamos en los datos del verano, decidimos crear una nueva columna que representaba el día del verano. Esto es debido a que el 1 de junio no es el mismo día que el 1 de agosto. Esta columna nos deja diferenciar estos dos días mientras mantenemos información acerca del día en que se efectúa el vuelo. Esta característica es fácilmente expandible a los datos de un único año, donde el 1 de enero sería el día 1 y el 31 de diciembre sería el día 366. También decidimos cruzar nuestra base de datos con una base de datos de aviones, con el objetivo de conseguir datos acerca del modelo de cada avión, mientras que antes teníamos sólo su matrícula.

En posteriores aproximaciones, se decidió cambiar las características de hora de llegada y salida de un valor numérico a un valor categórico, transformando las horas en intervalos horarios. Los intervalos se hicieron manteniendo la distribución, por lo que unos intervalos podían representar franjas horarias más grandes que otras. En la última aproximación, era necesario realizar una transformación total de los datos, para poder efectuar un estudio de series temporales. Para ello, primero se aislaron los datos según su modelo, obteniendo tantas tablas como modelos teníamos. Más tarde, se hizo una media de los retrasos por día, de tal manera que cada tabla contenga los datos del retraso medio para cada día. Con los datos de esta forma, se puede efectuar un estudio de series temporales por modelo de avión.

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 [2] que son más de 1 millón de registros.

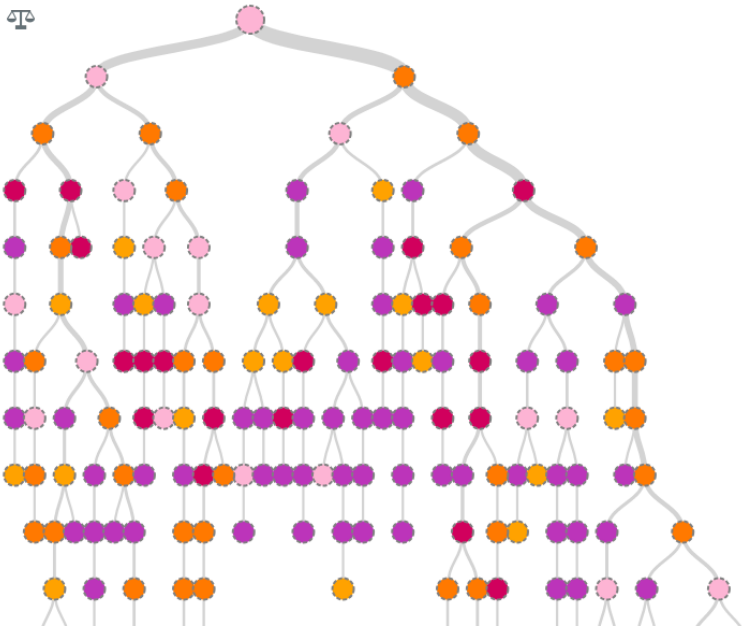


Fig. 2. Fragmento de árbol de decisión de la primera aproximación

2.5 Interpretación (Primera Aproximación)

Al realizar el *testing* con el resto de los datos (el 30% restante) obtenemos una matriz de confusión [3], como podemos observar no son resultados 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.

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 [4]. 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

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. 3. Matriz de confusión de la primera aproximación

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.

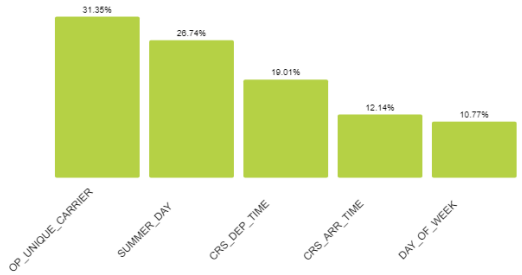


Fig. 4. Importancia de las variables de la primera aproximación

2.6 Minería de Datos (Segunda Aproximación)

Como una segunda aproximación realizamos otro árbol de decisión [5] 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 del vuelo en concreto, por falta de datos referentes al modelo 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.

2.7 Interpretación (Segunda Aproximación)

Como se puede observar en el árbol [5] en los nodos de color rojo (modelo de avión), esta es una característica que parece cobrar mucha importancia (esto también se puede ver en [7]) en este caso es porque hay ciertos modelos de avión que realizan viajes más cortos 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 [6] 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.

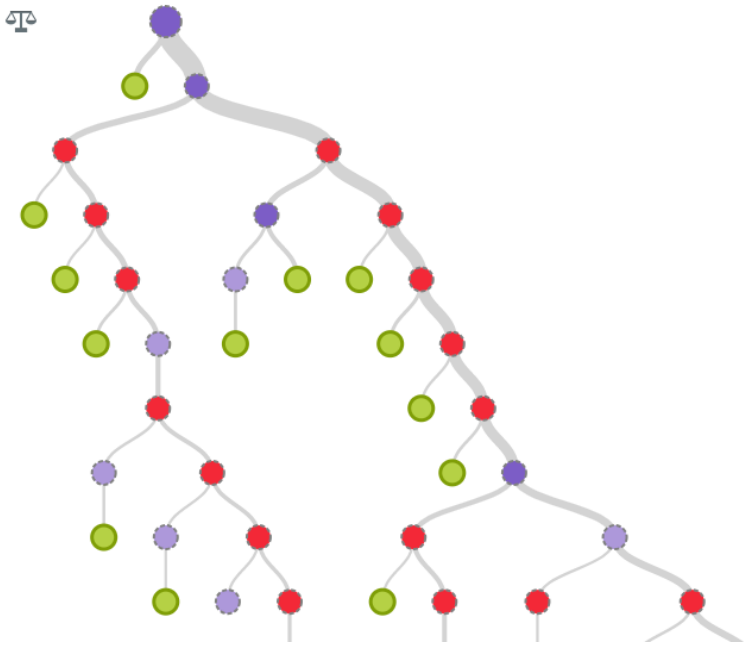


Fig. 5. Fragmento de árbol de decisión de la segunda aproximación

ACTUAL VS. PREDICTED	0	1	ACTUAL	RECALL
0	476,199	368,941	845,140	56.35%
1	65,312	117,937	183,249	64.36%

Fig. 6. Matriz de confusión de la primera aproximación

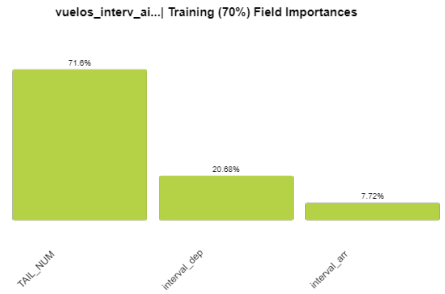


Fig. 7. Importancia de las variables de la primera aproximación

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 [8] vemos un ejemplo utilizando los datos del modelo Boeing 737-7H4.

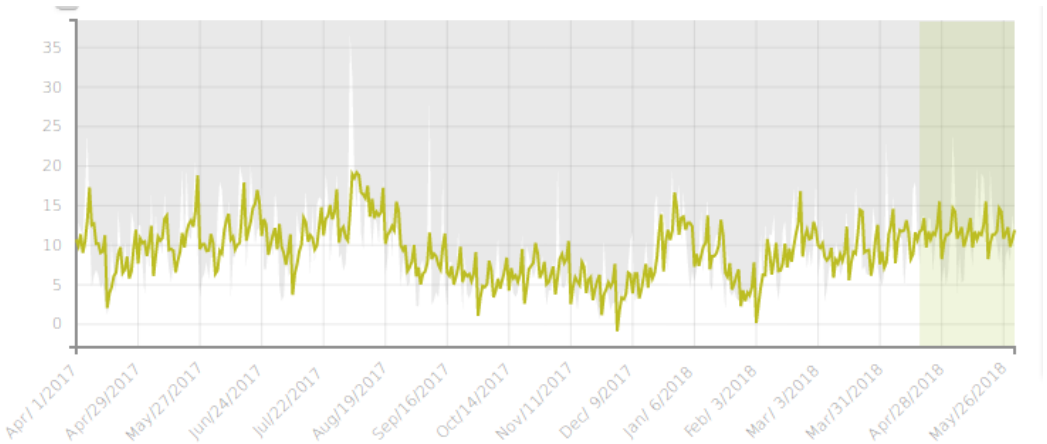


Fig. 8. Regresión obtenida utilizando los datos del modelo 737-7H4

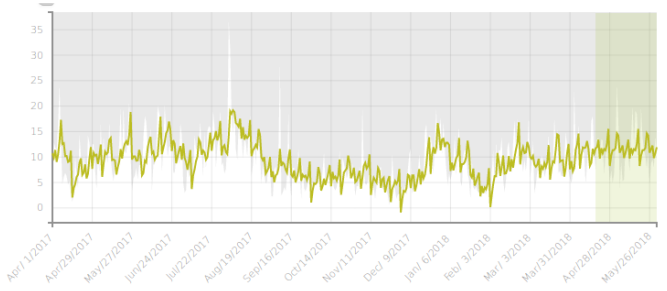
Como podemos ver en [8], en la parte derecha (correspondiente a los datos de *testing*), el modelo (en verde) predice correctamente los retrasos, siguiendo los máximos y mínimos reales (en blanco).

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.

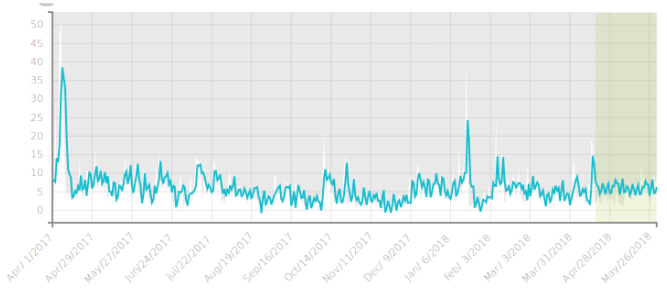
Para ello, utilizaremos los modelos 737-7H4 y 717-200, que bajo nuestro conocimiento del problema cumplen funciones similares. Ambos son aviones comerciales, que transportan pasajeros con su equipaje, y que tienen capacidad para alrededor de 150 pasajeros. Los resultados obtenidos pueden verse en [9a] y [9b]. 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.

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 resultados mejorables. En el caso de la regresión, no se obtuvo ningún resultado satisfactorio.



(a) Modelo 737-7H4



(b) Modelo 717-200

Fig. 9. Regresión utilizando dos modelos diferentes

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.

REFERENCES

[1] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996, November). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM 39*(11), 27–34.