Predicting Flight Delays in Commercial Aviation

Mario Alonso López
Aerospace Engineer, Data Science enthusiast
m.alonso.lopez123@gmail.com
https://www.linkedin.com/in/marioalonsolopez/

Published : 2021/05/09

1. Abstract

Airborne transportation is widely accepted as a cornerstone in mobility worldwide. Its meaningful impact on the global community is remarkable from a social and economic viewpoint. Societies all across the globe greatly benefit from this industrial activity. However, in order to do so, flight tickets must be fair. Considering the various costs airlines have to face, margins are undoubtedly tight-fitting; so, in order to make it possible for air travelers to fly at a reasonable price, air carriers require minimizing operational costs. There are several challenging tasks an operator has to tackle to maximize its profit, such as optimizing maintenance activities, or route planning to minimize fuel consumption. Among all these tasks, this project will focus on predicting flights delays; in accordance with the definition convened by the Federal Aviation Administration (FAA), a flight is considered to be delayed when it is 15 minutes behind its schedule time. According to the FAA, the total cost of delays from 2016 to 2019 has steadily increased from $23.7 to $33 billion ($5.6 to $8.3 billion corresponding to air carriers) [1]. By addressing the punctuality issue, operators can better understand what the main reasons are behind delays, and set up corrective measures accordingly in advance. Since meteorological predictions are quite accurate nowadays, the proposed model takes these valuable data into consideration in order to generate actionable insights. A comprehensive analysis is undertaken accounting for various predictors besides weather data. In this paper, predictive modelling approaches are applied based on machine learning techniques using publicly available flight and meteorology datasets from the US for year 2019. A binary classifier is built on such data, and optimized using recall as the primary metrics, achieving a value for delayed flights of 0.65, together with 0.70 for on-time flights. Furthermore, since Artificial Intelligence (AI) interpretability is becoming more and more relevant these days, certain SHAP values have been included to better understand the apparently black-box model's behavior. Finally, an interactive frontend has been developed using Python's Streamlit library in order to ease airlines' daily operation.

2. Introduction

This paper is mainly based on two datasets retrieved from two main sources, both originating on US data for year 2019:

1. On-Time Performance (OTP) [2], provided by the US Bureau of Transportation Statistics (BTS)
2. Local Climatological Data (LCD) [3], provided by the US National Oceanic and Atmospheric Administration (NOAA)

A brief summary of Airline Service Quality Performance is provided below (year 2019) [4].

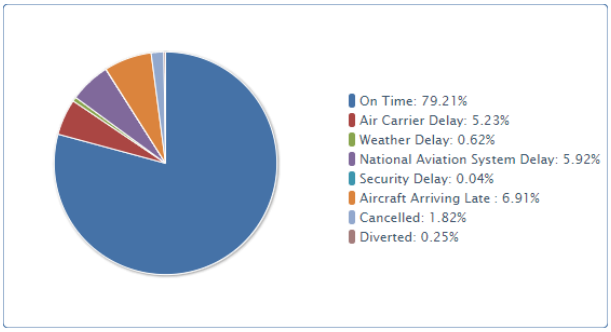

Figure 1. On-Time Arrival Performance National (January - December, 2019)

1.1 Causes of Delays

There are five categories of delays defined by the BTS [5]:

- **Air Carrier**: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- **Extreme Weather**: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- **National Aviation System (NAS)**: Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as **non-extreme weather conditions**, airport operations, heavy traffic volume, and air traffic control.
- **Late-arriving aircraft**: A previous flight with same aircraft arrived late, causing the present flight to depart late.
- **Security**: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

1.2 Weather real share

Adverse weather conditions play a central role in the proposed model. Therefore, an in-depth analysis on how to truly calculate meteorological contributions to total flight delays is necessary. A true picture of total weather-related delays requires several steps. First, the extreme weather delays must be combined with the NAS weather category. Second, a calculation must be made to determine the weather-related delays included in the "late-arriving aircraft" category. Adding the weather-related delays to the extreme weather and NAS weather categories would result in weather's share of all flight delays.

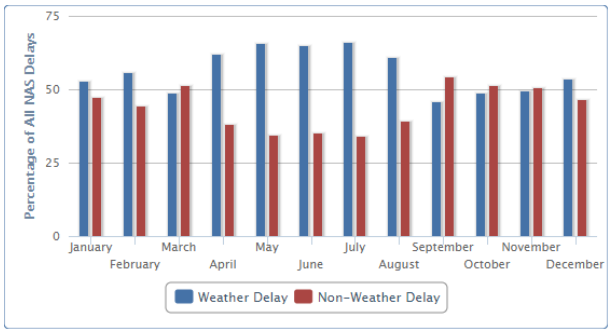

Figure 2. Weather's Share of National Aviation System (NAS) Delays National (January - December, 2019)

Delays or cancellations coded "NAS" are the type of weather delays that could be reduced with corrective action by the airports or the Federal Aviation Administration. During 2019, 56.8% of NAS delays were due to weather. NAS delays were 24.0% of total delays in 2019 [6].

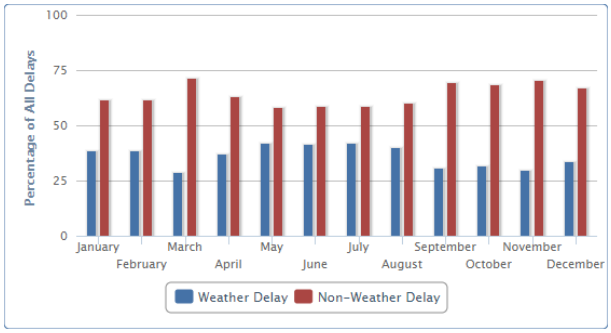

Figure 3. Weather's Share of Delayed Flights National (January - December, 2019)

Weather delay in the figure above is the sum of Extreme Weather delays, NAS delays caused by weather as assigned by the FAA, and the Weather's pro-rata share of late-arriving-aircraft delays based on delay minutes. According to the BTS, the **weather's share of delay as percent of total delay-minutes adds up to 38.7% in 2019** [6].

Once weather contribution to delays has been broken down into manageable pieces and properly explained, it is evident that the model is going to be strongly affected by adverse meteorological conditions.

3. Dataset and Features

3.1. Original datasets

As mentioned above, the paper is built upon two main sources:

1. On-Time : Reporting Carrier On-Time Performance (2019) dataset [2], supplied by the US BTS
2. Local Climatological Data (LCD) [3], provided by the US NOAA

The first dataset, referred to as “OTP” hereafter, collects individual flights performed by air carriers that account for at least one percent of domestic scheduled passenger revenues. It contains on-time data reported directly by airlines based on the flights they operate. Relevant information such as origin, destination, and time-related data is provided.

The second dataset, from this point forward “LCD”, represents all meteorological-related information gathered by each weather station, providing hourly reports.

Additionally, a third database is used to link the above two datasets, which is the Master Station History Report, provided by the Historical Observing Metadata Repository (HOMR), as part of the National Centers for Environmental Information's (NCEI) integrated station history database [7]. This serves as a means to cross-reference each airport and its corresponding weather station.

At this point, some reasonable assumptions had to be made before continuing:

1. Not every airport presented its own weather station. Therefore, after a quick analysis on how many airports would be discarded by this lack of information, it was found that 19 out of a total of 360 airports would have to be rejected. Fortunately, those minor airports only added up to 0.8% of the flights in the entire dataset. As a result, these flights were dismissed.
2. While many weather reports were often provided on an hourly basis, there were a few other which reported more frequently (and not even at the same rate over time). Hence, the observation used in the combined dataset was the first record for each hour. As a result, meteorological conditions at 06:12 and at 06:52 were considered identical. This assumption was required in order to make the analysis feasible.

3.2. Features

3.2.1. OTP

There were originally 109 columns in the dataset and 7.4M records (i.e. flights), so it was initially reduced to 35 to begin with. The data which was held after this initial screening review referred to: time of the flight, airline, origin and destination airports (and their geographic locations), some departure and arrival performance values, and other information such as covered distance and air time.

Some of these columns were kept just for exploratory purposes, and dropped later on in the process. Furthermore, some columns were transformed so the model could use them as input features. For example, it is not possible to have the taxi-in and taxi-out times before the flight, since these are known once the flight is performed; however, a statistical approach can be applied to get the median values for each airport-carrier pair. Taxiing procedures are defined by regulators and made available to all carriers by airports; nevertheless, each airline operates their aircraft in very different ways, so times greatly vary among carriers.

3.2.2. LCD

A significant amount of work had to be put into this part of the project. An in-depth analysis was undertaken to assess which features could potentially carry predictive power. At the end, the original number of columns was reduced from 129 to only 14.

The features decided to be kept comprised the station identifier and its geographic location, the reporting date and time, and many various interesting weather variables: altimeter setting (i.e. barometric pressure), dry-bulb temperature, precipitation amount, relative humidity, sky conditions, visibility, and finally wind and gust speeds.

Analogously to what was done on the OTP dataset, various transformations were required to clean this dataset and prepare it for later use in the model. Several reasonable assumptions were made throughout the process, but in order not to excessively elaborate on this topic here, the reader can go into the code for further details.

3.2.3. Master Station History Report

This database served as a means of link between the above two datasets. It contains a list of weather stations spread worldwide and, where applicable, their corresponding airport code. As a result, the OTP origin and destination airports could be linked to their corresponding weather station codes. Finally, the latter were used to merge the LCD weather data to the OTP flights. This last step concluded in the final dataset which would be used later on to feed the model.

3.3. Final dataset

Once the above described datasets were extensively cleaned individually, an inner join was performed in order to merge both of them. Some minor airports had to be dismissed during this process. As a result, a definitive dataset was produced, consisting of 7.2M rows (flights) and 69 columns. However, many of these columns only held exploratory power; therefore, only 27 features were deemed informative and were fed to the model. This will be fully detailed in the "Model and results" chapter.

4. Aim and metrics

Before addressing the model description, the objectives must be identified and clearly explained. According to the FAA, the total cost of delays from 2016 to 2019 has steadily increased from $23.7 to $33 billion ($5.6 to $8.3 billion

corresponding to air carriers) [1]. Therefore, this project aims at addressing the following issues:

1. Help airlines predict potential delays in advance so as to minimize costs incurred from unpunctuality (increased operating expenses comprising extra crew, fuel, or maintenance among others).
2. Provide carriers with causality suggestions to better understand the reason behind these delays. This information is gathered through various methods, such as feature importance or SHAP values.

In order to accomplish these tasks, the model is optimized by recursively maximizing recall. The rationale behind this decision is quite straightforward: the cost-benefit analysis. According to Airlines for America, in 2019 the average cost of aircraft block (taxi plus airborne) time for U.S. passenger airlines was $74.24 per minute [8]. As a result, operators could doubtlessly benefit from putting measures in place to prevent unnecessary delays, when feasible. These measures, which fall out of the scope of this paper, can range from better ground handling methodologies to moving slots, adjusting flight plans or even thinking about other potential routes.

Either way, in most cases cost derived from delays greatly exceeds that allocated to mitigating actions. Hence, **the model focuses on reducing the number of flights that are mistakenly predicted as on-time**. In statistical terms, the model thus aims at minimizing the *false negative* cases (i.e. delayed flights predicted as on-time), since these entail greater costs to the airline. However, a too simple model could just guess that virtually every flight was to be delayed to maximize recall; as a result, a comprehensive evaluation is performed, in which the confusion matrix is painstakingly assessed in order to avoid this misbehavior.

5. Model and results

5.1. Exploration of various estimators

Once the main objectives were defined together with the evaluation metrics, several models were initially explored.

- XGBClassifier
- LogisticRegression
- GaussianNB
- KNeighborsClassifier
- DecisionTreeClassifier
- RandomForestClassifier
- AdaBoostClassifier
- GradientBoostingClassifier

The chosen estimators were evaluated considering only a 100,000-flight fragment of the entire dataset. After examining the outcome from each estimator, the XGBClassifier showed the best results among the preselected models. Consequently, further development was carried out using an XGBoost classifier as the underlying estimator.

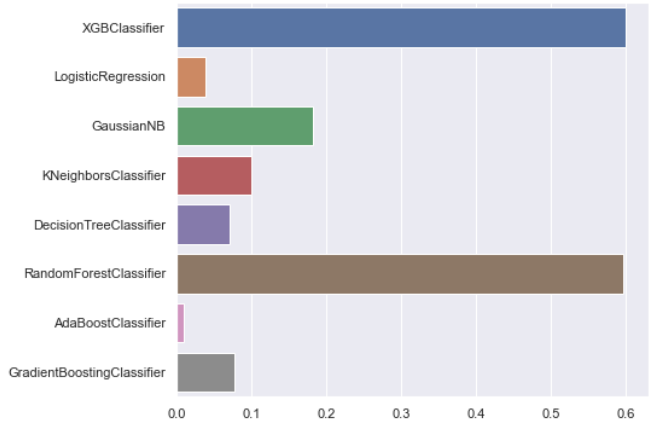

Figure 4. Evaluation of different models. Recall analysis

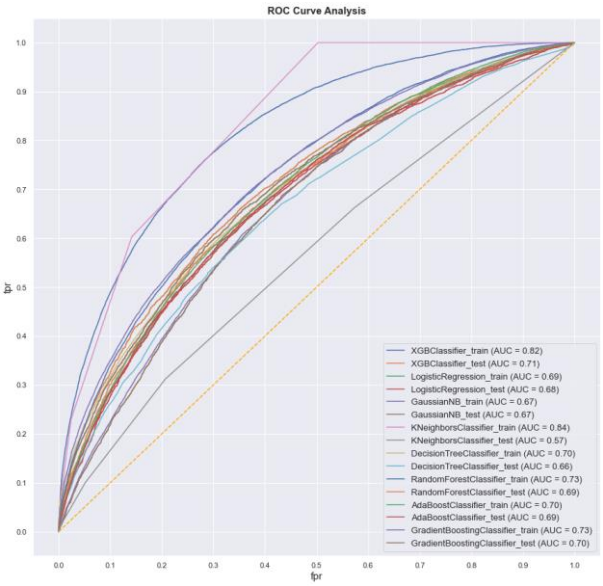

Figure 5. Evaluation of different models. ROC curve analysis

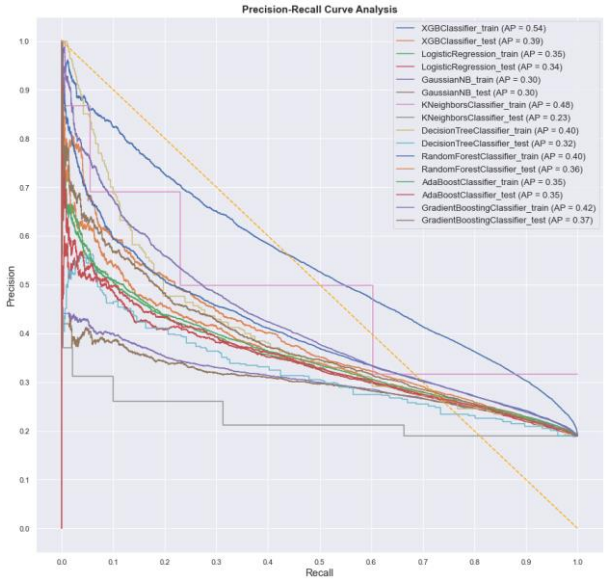

Figure 6. Evaluation of different models. Precision-Recall curve analysis

5.2. Dealing with an imbalanced dataset

The entire dataset, consisting of 7.2M flights and 27 features, was somehow imbalanced. The on-time/delayed flights ratio was around four (i.e. 80% on-time and 20% delayed). Two methods were undertaken in order to overcome this issue:

- Undersampling: considering the dataset was sufficiently big, this technique seemed reasonable
- Weight-scaling: control target imbalance by tuning the *scale_pos_weight* parameter

After running some tests to assess suitability of both approaches, the weight-scaling method offered better results. Furthermore, the philosophy of manipulating the data as little as possible seemed more robust and reliable since it reflects how real data behaves so the model could learn from reproducible patterns.

5.3. Final model: building the pipeline

Among the dataset features, nine of them were string typed, thus some kind of encoding was required prior to feed the data to the XGBoost classifier. Due to the high cardinality of some of those features, a target encoder was selected to transform string to numeric columns. The airport features (namely "ORIGIN" and "DEST") presented 341 different values, hence a one-hot-encoding approach was dismissed; on the one hand, there was a high risk of suffering from the so-called *Curse of Dimensionality*; on the other hand, computation times were unfeasible.

Once target encoding was deemed as a reasonable approach to address the issue of having string features, some measures had to be put in place in order to prevent data leakage from appearing. For the purpose of dealing with this issue, a pipeline was constructed.

First of all, no transformations were set for numerical columns given that XGBoost is not sensitive to monotonic transformations of its features since it is a tree-based model; neither is it impacted by outliers, since scores used to split the data are calculated based on the homogeneity of the resulting data points.

Concerning the categorical variables, a target encoding is included as part of the pre-processing steps of the pipeline. By doing so, the test fold for each cross validation stage is transformed using only known targets from the training set; in this way, data leakage is prevented.

Afterwards, the full pipeline is built drawing upon the pre-processing pipeline together with the inclusion of the base model (XGBoost classifier in this case).

Finally, the resulting pipeline is fed into a *RandomizedSearchCV* so that the stratified cross-validation is performed trying out different hyperparameters combinations in order to optimize recall. Computation times follow a geometric progression: the bigger the search space, the longer it takes to train the model. Hence, the model was trained in a little smarter way:

1. Choose a relatively high learning rate so that fitting times do not take too long.
2. Optimize the *max_depth* hyperparameter first, since this is one of the most relevant settings to tune.
3. Toggle *min_child_weight* : too low values lead to overfitting.
4. Set the *gamma* parameter to an appropriate value so that a minimum loss reduction is guaranteed before splitting the leaf node of a tree.
5. Tune regularization parameters *lambda* and *alpha*, also to prevent overfitting.
6. Finally, refit the model with lower learning rates to see if better results are obtained.

5.4. Results

As indicated above, this project aimed at maximizing recall, while keeping reasonably high values of precision. In

other words, greater importance has been given to minimize delayed flights wrongly predicted as on-time.

Having this in mind, below are presented the results of the model.

	precision	recall	f1-score
on-time	0.89	0.70	0.78
delayed	0.34	0.65	0.44

Table 1. Final model. Classification report

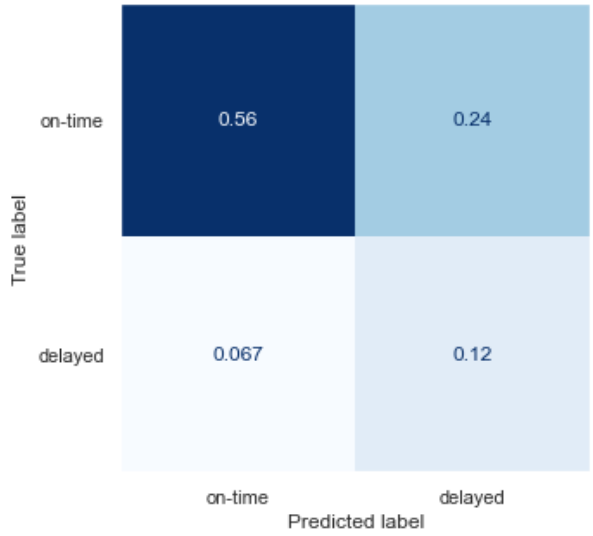

Figure 7. Final model. Confusion matrix

The confusion matrix shows that only 6.7% of flights are delays which are "missed" by the model. Considering that:

1. Only 38.7% of total delay minutes in 2019 were related to weather conditions, while the remainder were heavily dependent on airline and airport operations, and that
2. Around 19% of the total number of flights were delayed in 2019,

It is safe to state that correctly predicting 12.3 out of a total of 19 percent of delayed flights implies that the model successfully captures the inherent complexity of the dataset. Satisfactory results are obtained considering the type of data publicly available. Further work could be carried out in order to improve the model; several research proposals will be outlined in the following section.

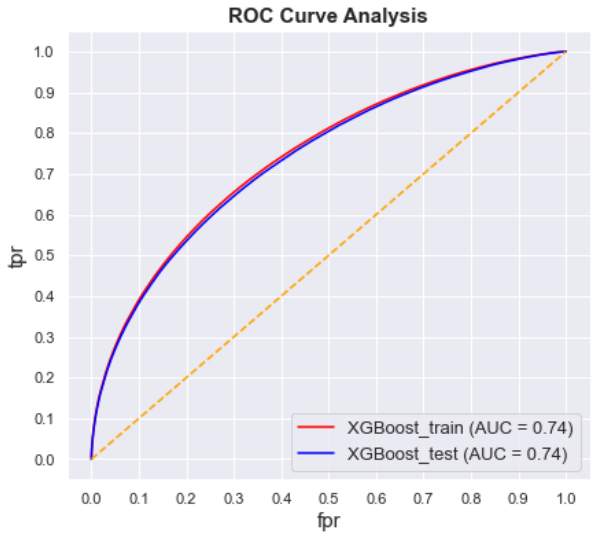

Figure 8. Final model. ROC curve analysis

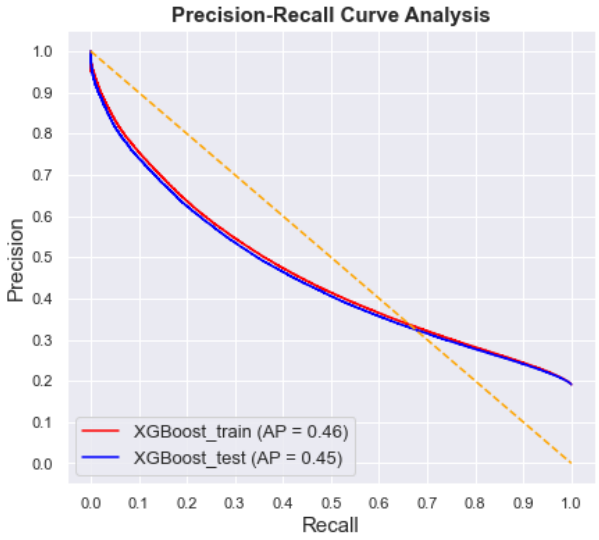

Figure 9. Final model. Precision-Recall curve analysis

Furthermore, the reader can observe that there is no sign of noticeable overfitting, since both numeric and graphical metrics for training and test datasets closely match.

5.5. Model interpretability. SHAP values

A significant effort was made in order to explore the model interpretability. Both global and local (individual flights) interpretability were examined so as to better understand the underlying complexity and model's internal decisions. This work is the cornerstone for suggesting business data-driven actions to airlines. Global interpretation methods are depicted mainly through summary plots, which

serve as a means to convey general behavior of the model. On the other hand, local explanations provide a detailed view of each feature's contribution to an individual flight's prediction.

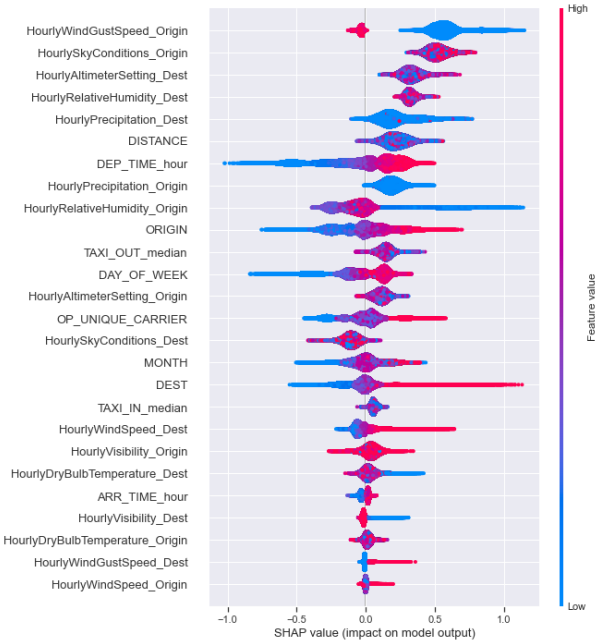

Figure 10. SHAP values. Global interpretability

In addition, the origin's sky conditions and the destination's pressure and relative humidity always push predictions to delays. This illogical result probably stems from a faulty model concerning those features.

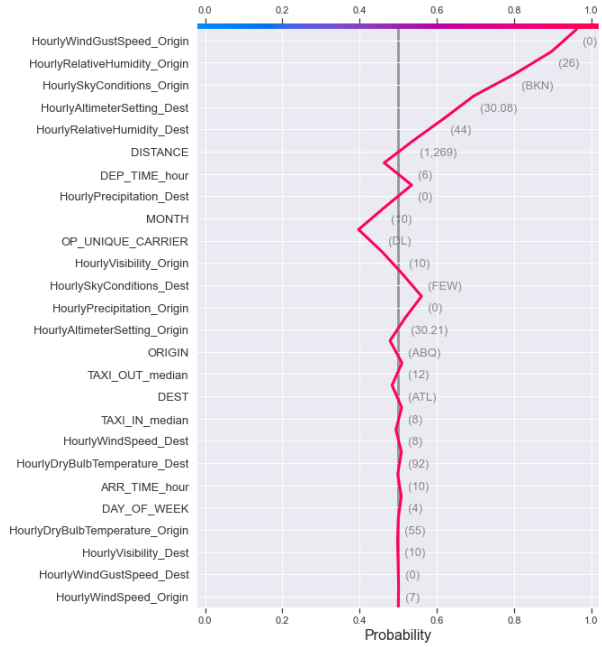

Figure 11. SHAP values. Local interpretability

The figure above shows the global SHAP values in a summary plot. Some features closely match common intuition, e.g. early morning departing flights tend to be on-time much more often than afternoon ones, since delays accumulate throughout the day. However, there are some other features which cannot be easily explained at first glance just by observing this plot. For this reason, local interpretability plays a key role in explaining more complex behaviors based on individual combinations of features.

On the other hand, there are some limitations in the model which can be appreciated through this figure too; this is the case of the wind gust speed at the origin airport. As can be observed, high values of this feature slightly lead to on-time predictions, whereas null gust values (present in the vast majority of the flight records) drive predictions to delays. This counterintuitive interpretation of the feature's contribution to the model implies a deficiency in its internal decision-making. However, an attempt was made to rerun the model after eliminating these features (i.e. origin and destination gusts), and metrics showed worse results. Since the figure above manifests an intuitive explanation for its origin counterpart, it was decided to keep them both despite their contradictory interpretation.

Figure above exhibits the local interpretability power of the model. A SHAP decision plot has been used to comprehensively display each feature's share to the final prediction of a particular flight. The plot successfully breaks down the prediction into intuitive additions to the final probability. For example, it can be observed that a value of "DL" for the carrier feature pushes the prediction to on-time, whilst a remarkably low value of relative humidity (44%) at destination has the opposite effect, driving the prediction to delay. This can be explained by stating that Delta Airlines (DL) presents significantly good on-time performance records; on the contrary, low relative humidity values (44%) are normally associated to poor weather-related conditions in terms of punctuality.

However, previously mentioned model limitations are clearly manifested here. Even for a null gust value at origin airport, the model is wrongfully leading the prediction towards delay. This undesired effect underscores the importance of putting in place solid procedures to test and review the model, and of making an in-depth research of its interpretability in order to have a clearer view of what is going on inside the estimator.

6. Frontend

A web-based application under the name of *CLEAR SKY* was both developed and deployed using Streamlit. The main purpose of this service is to ease airlines workload concerning flight planning operations.

Since the application goes beyond the main scope of the data science project, it was decided to leave it out of this paper. The interested reader can further explore this tool by following the next link:

https://github.com/malonsol/TFM_KSchool/tree/main/frontend

Detailed examination is kindly encouraged. Should any doubt may arise during tests, contact details can be found in the front page of this paper.

7. Conclusion and future work

7.1. Conclusion

A noteworthy industry problem has been fully laid out. Relevant metrics have been outlined so as to highlight the predicting capabilities of the model. Predictions successfully meet the main project objective, i.e. consider the impact of weather conditions together with basic airport and airline idiosyncrasies.

Moreover, SHAP values have been calculated on both a global and a single-flight basis.

Furthermore, an additional line of work has been carried out to develop a Streamlit light app as an effort to ease airlines daily operations planning. Some interesting features have been incorporated into the app, such as a model explanation section. As a result, this inclusion ultimately serves as a means for carriers to assess key factors and turn these predictions into actionable insights.

Overall, obtained results fulfil the author's expectations considering the publicly available data sources on which this paper has been built.

7.2. Future work

Many potentially interesting ideas have been disregarded throughout the course of this project in order to meet the delivery deadline. Some of them will be briefly outlined below so as to inspire the reader to further extend the scope of this project and motivate a global effort to improve the model.

7.2.1. Exhaustive model selection and hyperparameter fine-tuning

As succinctly stated before, several models have been lightly explored and XGBoost classifier turned out to be the best estimator. As a result, subsequent work was performed with it. An attempt was made using Google Cloud Platform (GCP) services. An AI Platform JupyterLab instance was explored in order to run an XGBoost exhaustive parameter grid space. However, the allowed quotas for computation engines were quite low. Lots of errors were triggered due to lack of memory, and fitting times were even slower than with a common laptop. After several trials and communications with GCP Support Team, increasing user quotas was not possible given the short history of the project in the platform. Therefore, this approach had to be dismissed.

7.2.2. Airport and carrier related features

The main focus of this project was to explore the meteorological impact on flight delays, which accounts for a significant share of the whole figure (38.7%). However, most flight delays are related to airline and airport operating procedures. Therefore, an in-depth analysis of airline and airport data could substantially improve the model's metrics:

- Airline: fleet age, route plans, scheduled/unscheduled maintenance tasks, crew resource management…
- Airport: ground handling (cabin service, catering, ramp or passenger service), platform distribution, taxiing procedures…

Some of these characteristics have been inherently considered by the model, such as the taxi-out/taxi-in times, which depend on both airline and airport. Nevertheless, a comprehensive low-level investigation should be performed in order to thoroughly understand the problem's complexity.

8. Acknowledgements

Many people have contributed to the satisfactory development of this project. From an academic point of view, I would like to thank KSchool teachers for their support, and to the whole data science online community to help me dispel doubts.

On the other hand, an even greater support has been required from family and friends. Not only they have occasionally provided me with their feedback, but have also been during hard times. This is my first data science project, and emotional support has been essential when I got stuck at any point (which by the way happened quite a few times).

To all of you, thank you very much.

9. References

[1] Michael Lukacs, FAA - APO-100 Deputy Division Manager, [2019], *Cost of Delay Estimates*
[2] On-Time Performance (OTP) [2019], *US Bureau of Transportation Statistics*
[3] Local Climatological Data (LCD) [2019], *US National Oceanic and Atmospheric Administration*
[4] Airline Service Quality Performance 234, [2019], *US Bureau of Transportation Statistics*
[5] Airline On-Time Performance and Causes of Flight Delays, [March 5, 2020], *US Bureau of Transportation Statistics*
[6] Understanding the Reporting of Causes of Flight Delays and Cancellations, [March 5, 2020], *US Bureau of Transportation Statistics*
[7] MASTER STATION HISTORY REPORT, Historical Observing Metadata Repository (HOMR), National Centers for Environmental Information's (NCEI) integrated station history database
[8] U.S. Passenger Carrier Delay Costs, [September 22, 2020], *Airlines for America (A4A)*