Machine Learning Engineer Nanodegree

Miguel Ángel Huitrado Alemán

Capstone Proposal

June 26th, 2018

Domain Background

In a globalized world travel is getting as easy as possible. It is possible to go from one country to another in less time and with less effort than 20 years ago. This makes even more evident the problems that exist in the travel process, in which one common problem is the delay of the flights. There are a few papers trying to understand the impact of this problem, which seems to affect travelers who are forced to wait several hours or even miss a connection flight, which is the worst scenario[1]. This also affects companies and airports with increased block times on routes and higher carrier costs and airfares[2], and finally affecting other segments of economy such as industries that rely on air transportation to conduct business, and more[3].

This creates the opportunity for machine learning algorithms to make an effort and explain why does this happen and also, when this is more probable to happen and by this helping the customers to choose best flights and companies planning and avoiding the reasons why a delay can happen.

In this project the objective will be to analyse the data from flights in the united states in 2015 and its departure times, both planned and real, to help predicting flight delay based on location, hour, flight length and other features.

Problem Statement

A flight delay can be caused by a lot of features, which can be mechanical, human, meteorological, etc. Understanding the causes of it can lead to improvements on the flight

[1] https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1017&context=ttra

[2] https://www.researchgate.net/publication/267398020_The_Impact_of_Flight_Delays_on_Passenger_Demand_and_Consumer_Welfare

[3] https://www.isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf

planning for customers and the workflow for airlines, and even the companies that offer flights by making a better pricing strategy on flights with high-probability of delay.

Datasets and Inputs

The dataset used comes from The U.S. Department of Transportation, it is available from Kaggle and contains the records for a year of flights. The features are going to be cleared, and probably reduced to have the most information gain while not falling in the curse of dimensionality.

The dataset have the following features:

1. YEAR - Year of the Flight Trip
2. MONTH - Month of the Flight Trip
3. DAY - Day of the Flight Trip
4. DAY_OF_WEEK - Day of week of the Flight Trip
5. AIRLINE - Airline Identifier
6. FLIGHT_NUMBER - Flight Identifier
7. TAIL_NUMBER - Aircraft Identifier
8. ORIGIN_AIRPORT - Starting Airport
9. DESTINATION_AIRPORT - Destination Airport
10. SCHEDULED_DEPARTURE - Planned Departure Time
11. DEPARTURE_TIME - WHEEL_OFF - TAXI_OUT
12. DEPARTURE_DELAY - Total Delay on Departure
13. TAXI_OUT - The time duration elapsed between departure from the origin airport gate and wheels off
14. WHEELS_OFF - The time point that the aircraft's wheels leave the ground
15. SCHEDULED_TIME - Planned time amount needed for the flight trip
16. ELAPSED_TIME - AIR_TIME + TAXI_IN + TAXI_OUT
17. AIR_TIME - The time duration between wheels_off and wheels_on time
18. DISTANCE - Distance between two airports
19. WHEELS_ON - The time point that the aircraft's wheels touch on the ground
20. TAXI_IN - The time duration elapsed between wheels-on and gate arrival at the destination airport
21. SCHEDULED_ARRIVAL - Planned arrival time
22. ARRIVAL_TIME - WHEELS_ON + TAXI_IN
23. ARRIVAL_DELAY - ARRIVAL_TIME - SCHEDULED_ARRIVAL
24. DIVERTED - Aircraft landed on airport that out of schedule
25. CANCELLED - Flight Cancelled (1 = cancelled)

26. CANCELLATION_REASON - Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
27. AIR_SYSTEM_DELAY - Delay caused by air system
28. SECURITY_DELAY - Delay caused by security
29. AIRLINE_DELAY - Delay caused by the airline
30. LATE_AIRCRAFT_DELAY - Delay caused by aircraft
31. WEATHER_DELAY - Delay caused by weather

The dataset contains 5819079 data points with 31 features each. During 2015 there were 2125618 flights delayed and 3607308 flights on time on the dataset.
The distribution of classes is Delayed: 36% VS Not Delayed 61%

Solution Statement

Using the dataset some models can be trained to predict whether a flight is going to be delayed or not. The binary classifiers can include Decision Trees, Random Forests and XGboost classifiers, and each of this can be tested on sets of unseen data to get the accuracy for each class to see if the information provided by the model can be used or not.

Benchmark Model

A simple benchmark model that can be used is to think that every single flight is not going to be delayed, such naive way of thinking is the way services are offered on the market so if we can have even few information about the probability of a flight being delayed is going to represent a big change for the customers, airlines and sellers.

This naive classificator uses the distribution of the classes in the dataset to “classify” everything as not delayed.

Evaluation Metrics

Since the distribution of the classes in the dataset is modestly imbalanced it will be necessary to look not only at its accuracy but use a confusion matrix to get a better look at how the model is behaving within each class. Other measures can be used like Kappa and F1 score, which also can be implemented.

Since it will be impacting the demand of a service it is important to have the performance of the model against false positives and false negatives; this can be measured using True Positive Rate and False Positive Rate[4], and later the model can be tuned to meet the desired specifications.

[4] https://www.kdnuggets.com/2017/04/must-know-evaluate-binary-classifier.html

True Positive Rate (TPR) = TP / (TP + FN)

False Positive Rate(FPR) = 1 - Specificity = 1 - (TN / (TN + FP))

Accuracy = (TP + TN) / (TP + TN + FP + FN)

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Project Design

he project will follow the workflow stated below:

1) Data Analysis

This step is for data understanding and preparation for usage. Understanding the data will have a great impact on the future models since it is important to check if features are related, which types of features the dataset contains, maximum and minimum values for the numerical values, and more.

2) Data Processing

With the data analysis will come some questions about how the features can be used and improved, cleaning NaN values, outliers, selecting ranges for use when the data is big enough, and also create new features using feature engineering.

3) Model Building

Selecting the best model is only possible by knowing the data enough to detect which behaviours can be generalized and how. In this step a Decision Tree, a Random Forest and a XGBoost Classifier will be trained and tested using Cross Validation, and tuning each of its parameters.

4) Model Evaluation

Each model will be evaluated by getting its Accuracy score and verifying its Confussion Matrix to know how the model is performing on classifying True Positives and True Negatives against False Positives and False Negatives.

5) Results Explanation

Based on model evaluation a single model will be selected and its caracteristics explanied, why it was selected and if the model can be used in a real life situation. If possible, the behaviour on flight delay will be detailed based on the models classifications.