# 

# Illinois Institute of Technology

**FINAL REPORT**

# 

# 

# Airline Data Analysis Using Spark Technologies

Date: May 6, 2020

To: Professor Joseph Rosen, Big Data Technology

Team Members: Bhoj Soopal Divya (A20378715) , Anna Kong (A20410169), Nihar Darji(A20449275, Sahana Shreedhar Kulkarni (A20442482)

Subject: Final report for the analysis of flight data using PySpark Tools

Abstract

Our project analyzes the US domestic flights dataset using pyspark dataframes. We aim to predict which flight/flight carrier is most likely to be cancelled or delayed.

Introduction

Our project consisted of a dataset taken from the kaggle website [1]; with originally 28 variables (both inconsistent and consistent) related to the US domestic flights. Overall, this dataset was 7GB size and contained information about flight operational airlines, delays information, location details (origins and destinations), and cancellation (reasons as cancellation code); together with certain technical details such as the time that the plane was stationary on ground. We had 7 .csv files in total, starting from the year 2009 to the year 2015.

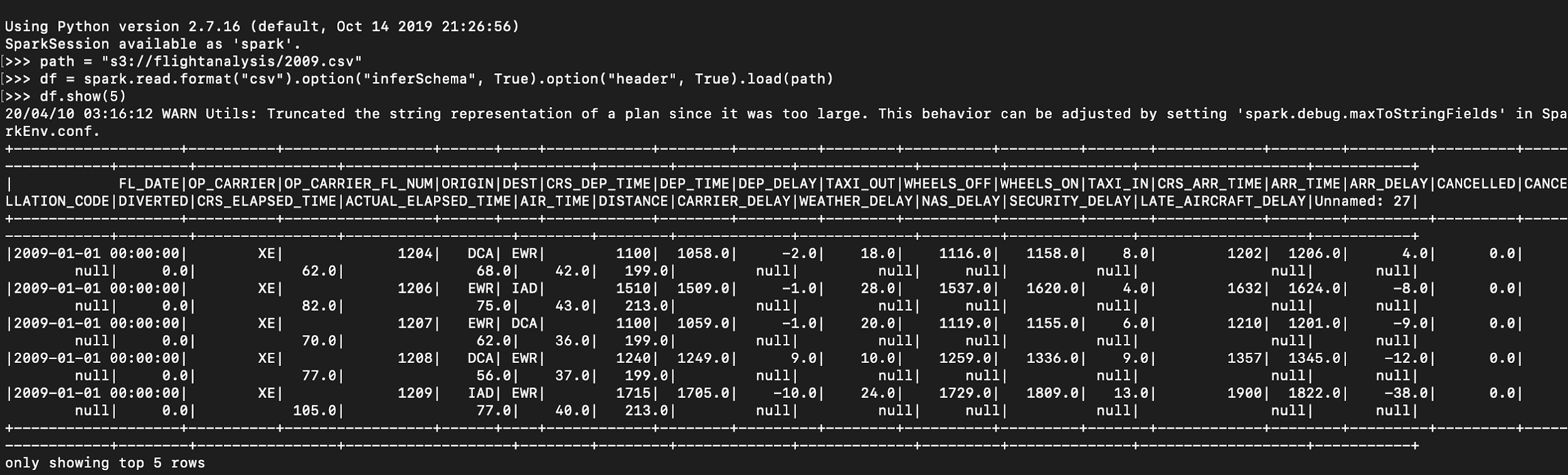
Our approach for this project was to use the services of Apache Spark. The latter uses distributed, in-memory data structures to improve speeds for many data processing workloads through several orders of magnitude. As a unified analytics engine, it allows us to parallelize large data processing tasks on a distributed cluster. Hence, we use its hybrid python supported version PySpark with Amazon Web Services (AWS) Elastic Map Reduce (EMR) cluster for the same purpose.

PySpark is also great for exploratory data analysis at scale, building machine learning pipelines, and creating ETLs (extract, transform, load) tools. Hence, after connecting our EMR cluster we connected to it through a graphical user interface (GUI) - Jupyterlab available as notebooks under AWS services and loaded our data from S3 buckets by importing it through the Kaggle website [1]. We proceeded with model profiling, wrangling, and predicting in this environment.

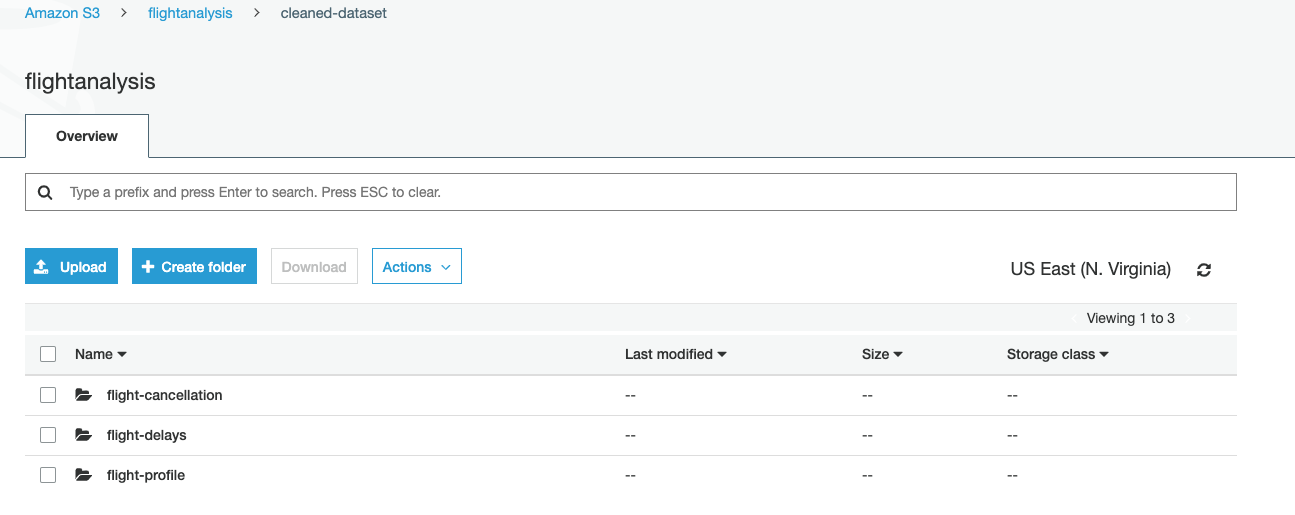
Further, our project predicts if a flight will get cancelled or not. In order to build the predictory system Machine learning models were incorporated, where we modeled a binary classification problem with categorical variables,logistic regression, decision tree classifier, random forest and gradient boosted trees.

Model Profiling

To get a quick glance at our data, we loaded the 2009 csv file into a RDD (Resilient Distributed Dataset) which is the Spark’s representation of a dataset that is distributed across the RAM, in memory of a cluster of many machines. Through this spark session, we discovered that we had about 27 unnamed variables and multiple null values throughout our dataset.



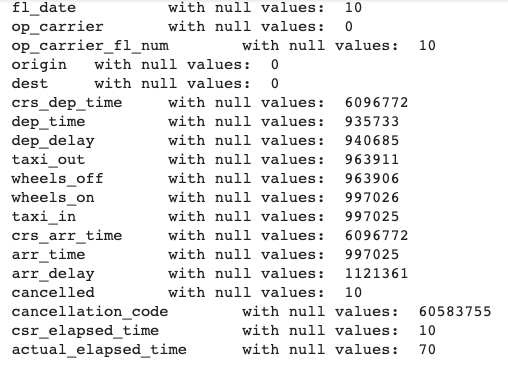
After combining our dataset (2009 to 2015 data), we proceeded into chunking them into 3 major categorical CSV files namely: flight-cancellation, flight-delays, and flight-profile. This was essential such that our team members could work collaboratively and in parallel to keep track of the project milestones.



Exploratory Data Analysis

1. **Data Cleaning and Preparation**

Originally, we had 28 variables to work with. After removing the unnamed columns, we were left with 19 columns of data that we explored and checked for null values (as shown below). We kept columns with only relevant data about the flight profile, delays, and cancellation information.

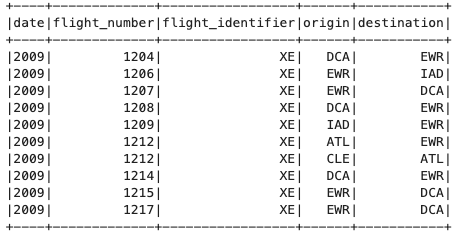


Since, we are dealing with time sensitive data, it was hard to recreate the null values with either mean or median of the data, hence we had to drop these records; leaving us with over 6+ million records to still work with.

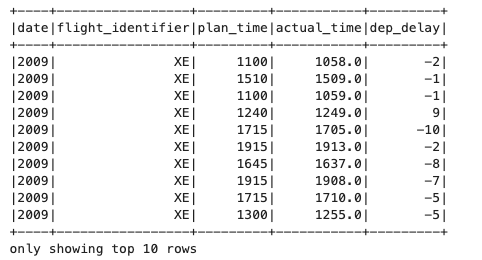
Also, we cleaned out the date column such that it contained only the year information since we were more interested in a yearly analysis as opposed to how the airline businesses performed on a daily basis.

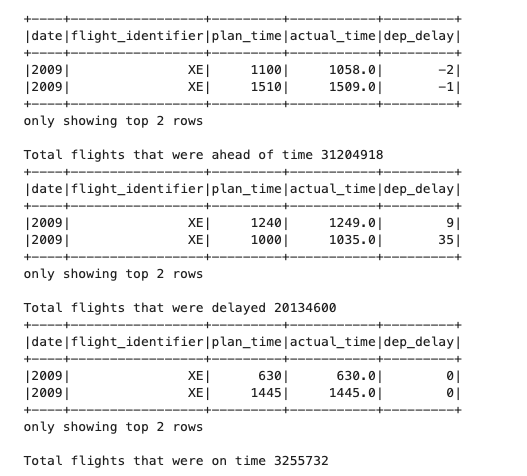
1. **Statistical Data Analysis**

* From the profile sub-dataset, we explore the number of flights from 2009 to 2015. From the 61556964 flight profile records, we have about 7605 US domestic flights in total, around 380 unique origins and 378 unique destinations. We created temporary tables inside our spark session to query specifics.

****

* From the delay sub-dataset, we found that the departure delay had negative values indicating that around 3,1204,918 flights were actually ahead of their scheduled time; and that is roughly 50% of our datasets. This means that 50% of the time flights were delayed.





* From the cancellation sub-dataset, we have the column cancel which is a categorical data with either 0 which means not cancel and 1 cancel as value. Also, we have reasons listed as follows:

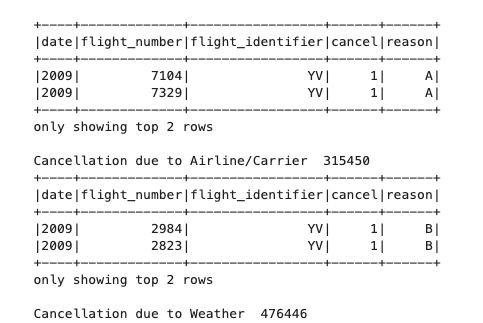
A - Airline/Carrier (315,450)

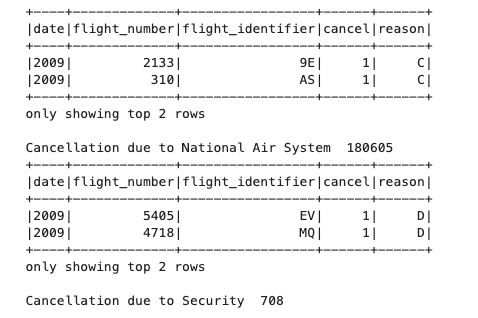
B - Weather (476,446)

C - National Air System (180,605)

D - Security (708)

The most reported reason was weather and the least reason was security.





1. **Data Visualizations**

We found out that YV (Mesa Airlines), F9 (Frontier Airlines), HA (Hawaiian Airlines) were among the 10 top airlines with the most flight operations from 2009 to 2015 (Figure 1).

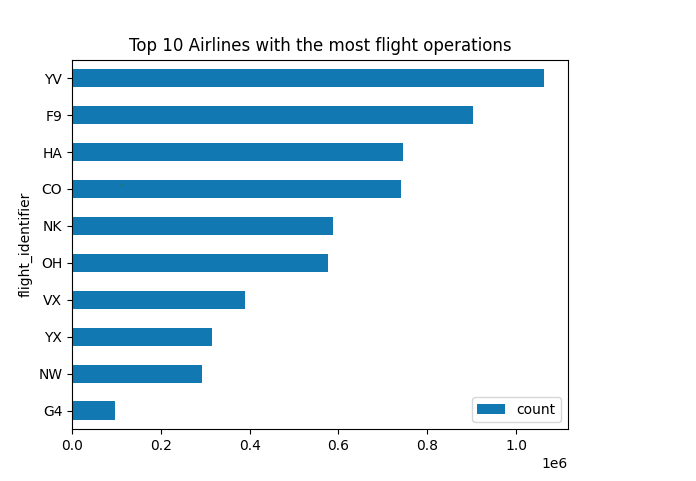


Fig 1: Top 10 Airlines with the most flight operations from 2009 to 2015

Also, we wanted to understand the variations of the flights across the years. We noticed that 2013 had the most flights, 2015 showed the lowest one since probably we had less data for that year compared to the others. Overall, there were slight variations with the total number of flights per year; that is the number stayed in between 6 to 6.5 millions (Figure 2).

We found out the proportion for the total flight cancellation reasons across 2009 to 2015 (Figure 3).



Fig 2: Number of flights across the years 2009 to 2015

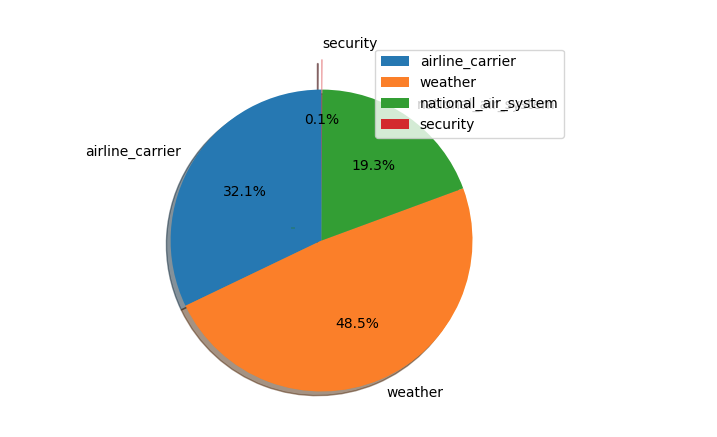


Fig 3: Proportion of flight cancellation reasons across the years 2009 to 2015

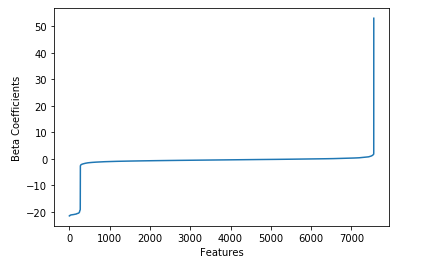
Model Prediction

We want to predict whether the flight will be cancelled or not. To do this, we model it as a binary classification problem with categorical variables.

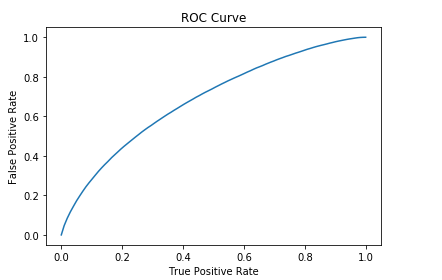
First, in order to prepare the data for machine learning, we use StringIndexer, OneHotEncoder and VectorAssembler to transform our features. Our categorical features - date, flight\_number and flight\_identifier are converted into one-hot encoded variables and then the binary vectors are appended to the end of the row. Our labels, in this case 0 for not cancelled and 1 for cancelled, are also converted. Then, all of the feature columns are combined in VectorAssembler into one vector. A pipeline allows us to chain these multiple transformations into our workflow.

The data was split into a 70/30 test and train. We test three models: logistic regression, decision tree classifier, random forest and gradient boosted trees to compare the best accuracy for predicting cancellations.

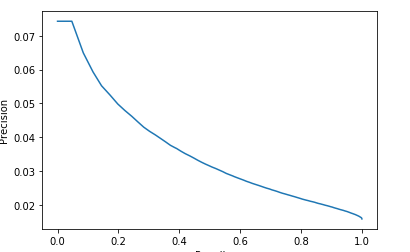
From our logistic regression model, we obtain the beta coefficient estimates for our parameters.



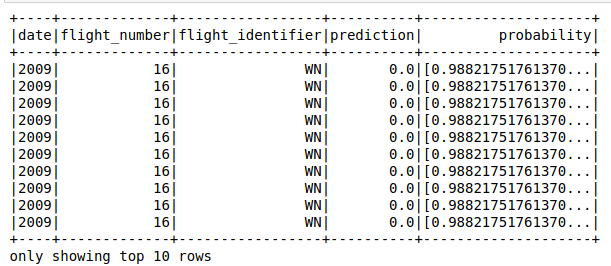
For the logistic regression model, we plot the true positive vs the false positive rate for the ROC curve and obtain the area under the ROC.



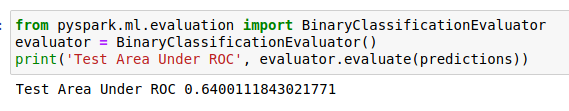
We plot the precision and recall for the model.



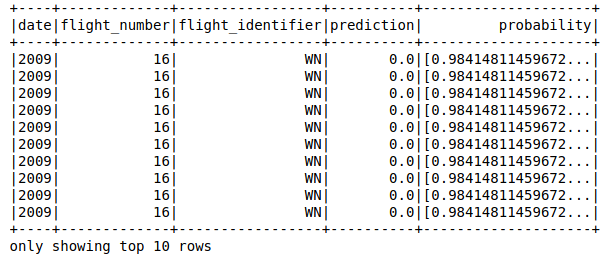
We make predictions on the test set with the probability of each prediction.

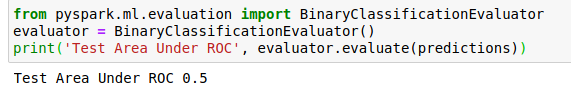


We evaluate the accuracy of the predictions using BinaryClassificationEvaluator

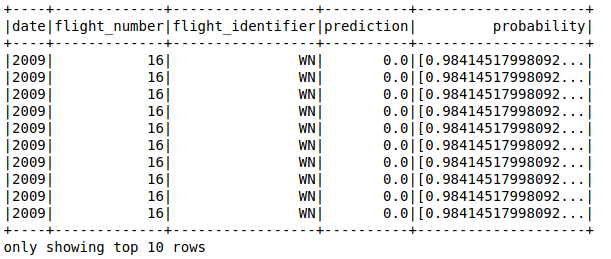


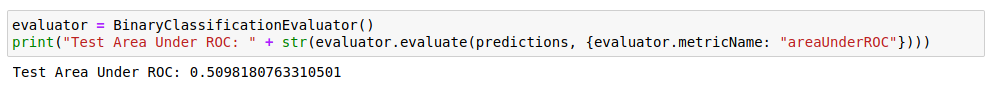
Decision Tree Classifiers can handle categorical variables well and are able to capture non-linearity. So we import DecisionTreeClassifier in the pyspark.ml.classification. Using our transformed data, we show the probability of each prediction and evaluate our model.



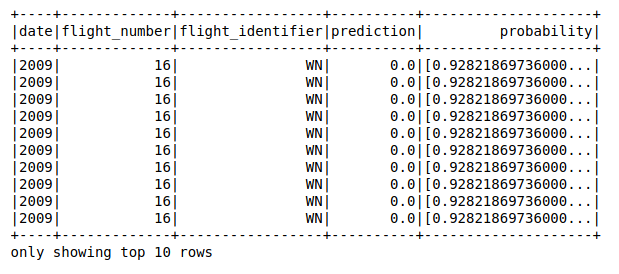


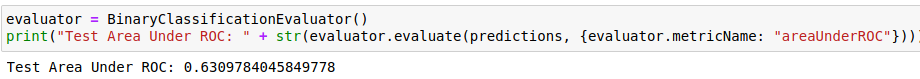
Random forests are a type of decision tree that split on a subset of features on each split. Random forests are advantageous because they are low bias, moderate variance, robust to outliers, and can handle unbalanced datasets.





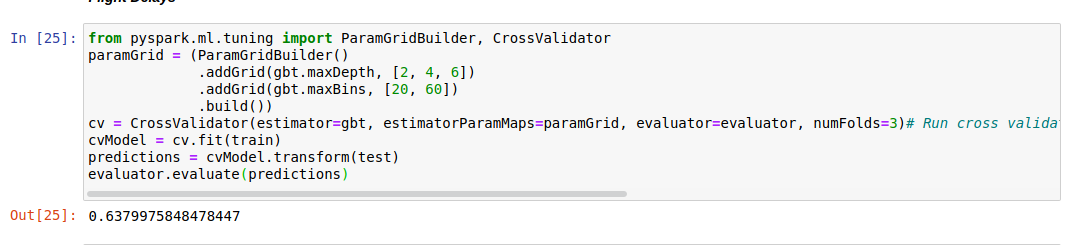
Finally, we test gradient boosted classifiers.

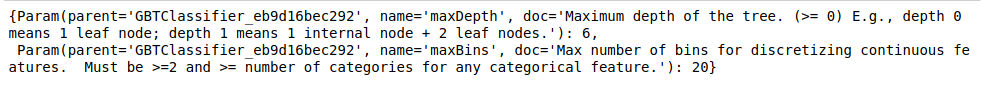




From our tests, the logistic regression model and gradient boosted classifiers performed the best.

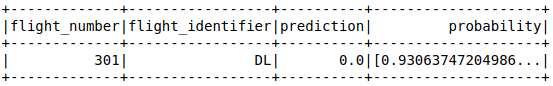
In order to test out various hyperparameters for the gradient boosted classifier and see if we can improve the performance, we use ParamGridBuilder and CrossValidator to find the maxDepth and maxBin that give the best accuracy.

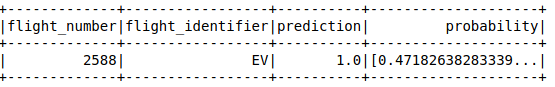




With cross validation, we found that the best parameters for a gradient boosted tree are a max depth of 6 and max bins of 20. With these improvements, our accuracy for GBT changed to 63.8%

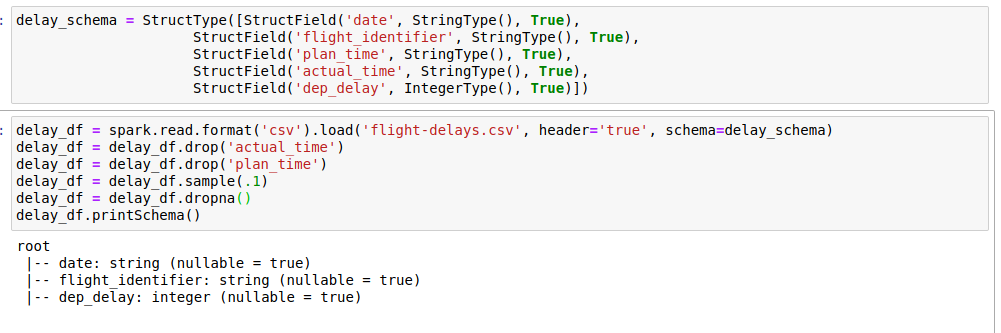
From our model, we found that the airlines with the highest probability of being cancelled were ExpressJet Airlines (EV) and the least probability of being cancelled was Delta Airlines (DL).



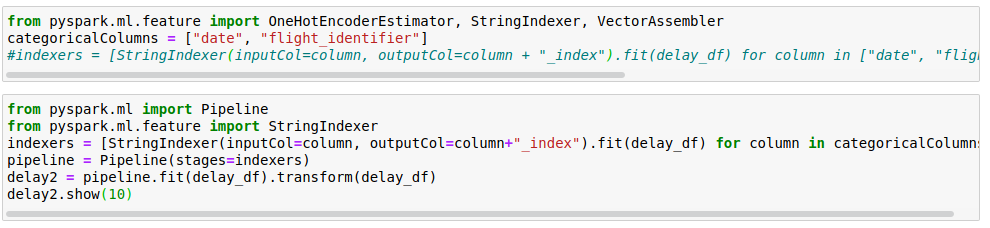


Second, we use regression to predict the flight delay time based on airline.

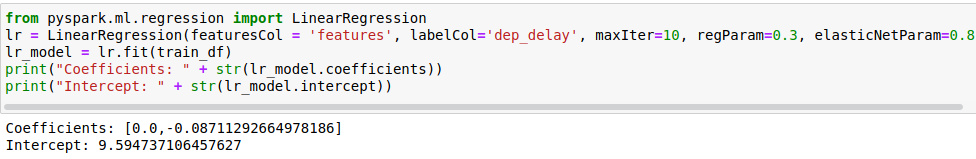
First, we define a schema for our flight delays dataset. The feature columns are date and flight identifier, and the label is the time in minutes that a flight was delayed.



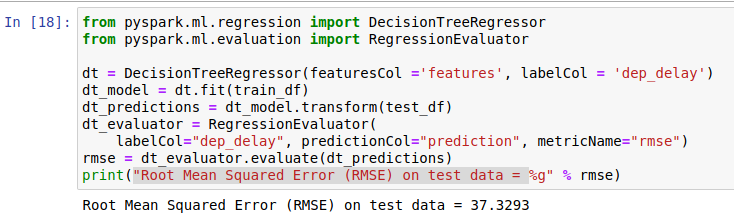
As we did with the flight cancellation classification, we use VectorAssembler and StringIndexer to transform our dataset.



We test on the linear regression model and get our coefficients and intercept. We test this on linear regression and decision tree models.







Since the RMSE is very high for both models, it means that the airline is not a meaningful predictor of delay time.

Conclusions and Future Work

We found in our project that airline identity is able to predict whether a flight will be cancelled or not with an accuracy of 63%. We tried logistic regression, decision trees, random forest, and gradient boosted classifiers, and gradient boosted classifiers (GBT) performed the best. From our data, it shows that the most likely flight to be cancelled was ExpressJet Airlines and the least likely was Delta Airlines. However, airline identity was not a meaningful predictor of delay time. This suggests that delay of flight is likely to be for reasons other than the airline, such as weather conditions or airport. In the future, we would like to incorporate other datasets such as weather, location and airport into our model to get a more accurate prediction for delay time and cancellation.

References

**[1]** Bureau of Transportation Statistics. (2016). Airline On-Time Performance and Causes of Flight Delays. Retrieved from <https://catalog.data.gov/dataset/airline-on-time-performance-and-causes-of-flight-delays-on-time-data>

**[2]** Deshpande, V., & Arikan, M. (2011). The Impact of Airline Flight Schedules on Flight Delays. *Manufacturing & Service Operations Management, 14*, 423-440. Retrieved from <https://pubsonline.informs.org/doi/10.1287/msom.1120.0379>

**[3]** Mu, Y. (2019, August). Airline Delay and Cancellation Data, 2009 - 2018. Retrieved April 2020 from <https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data>

**[4]** Chakrabarty, Navoneel, et al. ”Flight Arrival Delay Prediction Using Gradient Boosting Classifier.” Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. 651-659. Retrieved from <https://www.researchgate.net/publication/327389509_Flight_Arrival_Delay_Prediction_Using_Gradient_Boosting_Classifier>

**[5]** Yi Ding ”Predicting flight delay based on multiple linear regression”, IOP Conference Series: Earth and Environmental Science. Retrieved from <https://iopscience.iop.org/article/10.1088/1755-1315/81/1/012198>

**[6]** Belcastro, L. & Marozzo, Fabrizio & Talia, Domenico & Trunfio, Paolo. (2016). Using Scalable Data Mining for Predicting Flight Delays. ACM Transactions on Intelligent Systems and Technology. 8. 10.1145/2888402. Retrieved from <https://dl.acm.org/doi/10.1145/2888402>

**[7]** Li, S. “Machine Learning with PySpark and MLlib — Solving a Binary Classification Problem”. Towards Data Science. Retrieved from

<https://towardsdatascience.com/machine-learning-with-pyspark-and-mllib-solving-a-binary-classification-problem-96396065d2aa>