Predicting Departure Delays in Flights of USA

ABSTRACT

Using different classifiers and selecting the best one in predicting the departure delay of a flight in USA by using the origin, flight carrier, date and time information.

Sravani Koppala
A25282638

Table of Contents

Introduction:

Project discussion:

Using a dataset from Kaggle which has 100,000 records on flight data has an output column of whether a flight is delayed by 15 minutes (Yes or No). This is chosen for the output having binary classes (Delayed by 15mins? Yes or No). It has various attributes like flight carrier, origin, destination, distance, departure time, month, day of week and day of month. Here, different types of classifiers are used and prediction accuracy is compared for various methods using various training sizes ranging from 10% to 50%. At the end the highest accuracy model is selected for the prediction. Various data visualizations are also done for the analysis of data.

Motivation:

We all have the habit of checking the status of flight which we would be boarding before leaving home. Once, I was similarly checking on Google and found a small description below stating, "this flight is typically on time." I found that very interesting that how depending on the previous statistics Google is predicting whether my flight has any chances of getting delayed or not. So, for the project I picked up the flight delay information dataset from Kaggle to predict whether a flight depending on the origin airport and departure time, gets delayed or not.

Why is it Big Data?

The number of flights flying globally by the airline industry has been steadily increasing since the early 2000's and is expected to reach 39.4 million in 2019. This figure is over one million higher than the prediction for the previous year. Imagine if there are millions of flights, how much would the computational complexity get multiplied with the conditions on which the delay of the flights would get affected. There are number of attributes on which delay can depend like climatic conditions, holiday season, busy airports, taxi-out delay, communication delay etc.

Thus, as it deals with huge amounts of data which is always changing(**velocity**), high **volume** (millions), **variety** (different types of flight delays), this indeed is Big Data!

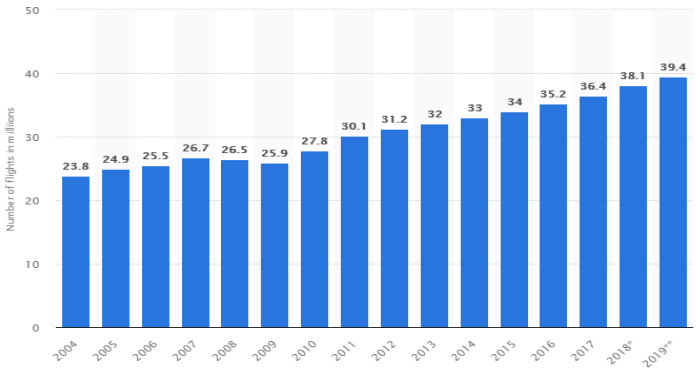

Fig 1: Bar plot of No. of flights per year globally.

Hypothesis outline:

The maximum number of flight delays should occur during the holiday season and at the busiest airports.

Discussion of Methods:

This dataset contains huge amount of data with 8 attributes. Having irrelevant features in data can decrease the accuracy of many models. It reduces Overfitting, Improves Accuracy, Reduces Training Time. Hence, I have used ANOVA for selecting best features. Also, I have used dimensionality reduction technique using PCA. But the data should be numerical to use these methods or classifiers. Hence, label encoder method was used to give number codes to columns like Origin, Destination cities and Carrier name.

The aim of classification methods is to assign true label to a new observation. Even though classification is one of the oldest statistical methods, finding the mechanism by which new observations are classified with the lowest error is still challenging. For predicting the delay in flights, I have compared 4 different classifiers like SVM, Naïve Bayes Gaussian Model, K Nearest Neighbor and Logistic regression. Whichever method gives the best accuracy, that can be used for predicting delay in flight departure. Also, I have tried using the dimensionality reduced data from PCA as input to the four different classifiers listed above and done the same with ANOVA's feature scaled data as well. Comparing the accuracy scores of different classifiers with and without reduced data will help in choosing a better classifier for this dataset.

Label Encoder: Input to the classifiers should be continuous. This dataset contains few categorical data which would not be accepted by the following models. Hence, this method is used to encode Origin and Destination airports as well as the Flight carrier name columns. It is a very efficient tool for encoding the levels of categorical features into numeric values. This method encodes labels with value between 0 and n_classes-1 where n is the number of distinct class labels.

PCA: This is done for dimensionality reduction and to turn the related variables into some unrelated variables and make the classes separable. It linearly projects a high dimensional data to its lower dimensions such that all the important details are preserved. It chooses a projection in a way such that the mean squared error of data and projection is minimized, and variance of projected data is maximized. PCA uses linear algebra to transform the dataset into a compressed form. Generally, this is called a data reduction technique. Thus, this reduced data can be used as input to different classifiers to increase the accuracy and performance, and reduce the training time.

ANOVA: The data features that are used to train the machine learning models have a huge influence on the performance. Irrelevant or partially relevant features can negatively impact model performance. ANOVA (Analysis of Variance) helps us to complete our job of selecting the best features. It is a statistical method, used to check the means of two or more groups that are significantly different from each other. Number of features wanted can be instructed to the model and it gives that many best features from our input data. Then the output of ANOVA with n-best features can be inputted to our models listed below to check performance and accuracy.

Logistic Regression: Applying logistic regression over 100,000 records to obtain a "binary classifier", using data about each flight to predict whether it was delayed, takes a fraction of a second. It is the appropriate regression analysis to conduct when the dependent variable is

dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Depending on the accuracy of the training model, we will get to know if this classifier can be used for this dataset.

Naïve Bayes classifier: Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. It is the fast, accurate and reliable algorithm. It has high accuracy and speed on large datasets. Naïve Bayes classification used here is in Gaussian model and deals with continuous values and it's also assumed that all features are following Gaussian distribution, i.e., normal distribution. This is more of a probabilistic approach and works well for binary classes. As this dataset also has binary classes where we must predict whether the flight is delayed or not, depending on the accuracy of the training model, we will get to know if this classifier can be used for this dataset.

SVM: Support Vector Machines (SVMs) are an example of a maximum margin classifier, which finds the linear classifier that maximizes the margin. It is a supervised machine learning algorithm which can be used for both classification or regression challenges. I have chosen this method to compare as this dataset has exactly two classes and SVM helps separate it. Depending on the accuracy of the training model, we will get to know if this classifier can be used for this dataset.

KNN: K nearest neighbors (KNN) are known as one of the simplest nonparametric classifiers. KNN makes predictions using the training dataset directly. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. In classification this might be the mode (or most common) class value. Depending on the accuracy of the training model, we will get to know if this classifier can be used for this dataset.

Random Forest Classifier: Random forests is a supervised learning algorithm. It is also the most flexible and easy to use algorithm. It creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance. Depending on the performance of the training model, we will get to know if this classifier can be used for this dataset.

By using the above classifiers, the model with highest performance that is more accuracy and less training time will be selected as the best classifier for this dataset. Also, the accuracy scores of original data in these classifiers are compared with input of PCA reduced data and feature selected data.

Results and Analysis:

Firstly, this dataset had categorical values like the airport names, carrier names etc., which were changed or encoded to numerical values using label encoder. Then in order to reduce the dimensions, PCA was used. The variance ratio obtained for the first two components was observed to be 97.9% and that of three components as 97.91%. When the data was visualized into 2D and 3D axes, as per the scatter plots, the classes are overlapping and cannot be distinguished clearly. Hence, I have used different classifiers to predict the delays and analyze performance by calculating the accuracy of each classifier.

Here, different training sizes were used ranging from 10% to 50%. Starting with the Random Forest Classifier, the accuracy ranged in between 80-81% but when PCA reduced data was used to fit and transform the model, the accuracy dropped a bit to 78-80%. For Naïve Bayes Classifier using a Gaussian model, the accuracy always dropped from 81-80% when it was iterated many times. Though it is just a one percent change and gave a pretty good accuracy of 80%, the model is not able to get trained properly with increasing training size. Normally, the scenario is like, as the training size is increased, accuracy for a model increases until it is overfitted.

Generally, it is believed that Logistic Regression is best and easy to be used for binary classification. Hence, upon trying this dataset, it does give an accuracy ranging from 80.9-81.1% for 10% of training size, but the accuracy scores act weird even though with many number of iterations. As the training size increases, the accuracy drops then shoots for original as well as PCA reduced data. From this it can be inferred that, Logistic Regression may not work well with huge amounts of data as this dataset has 100,000 records.

The next classifier used is K Nearest Neighbor classifier. The performance of the model is increasing with the increase in training size and it has the same case with PCA reduced data. But, the accuracy score is in the range of 76-78% for this model compared to the other classifiers described above. Hence this model can be discarded. May there are not enough counts for similar records in the dataset to model to KNN. But as the training size increases, accuracy is increasing. Thus, it can be inferred that it needs more data which has similar values to get trained.

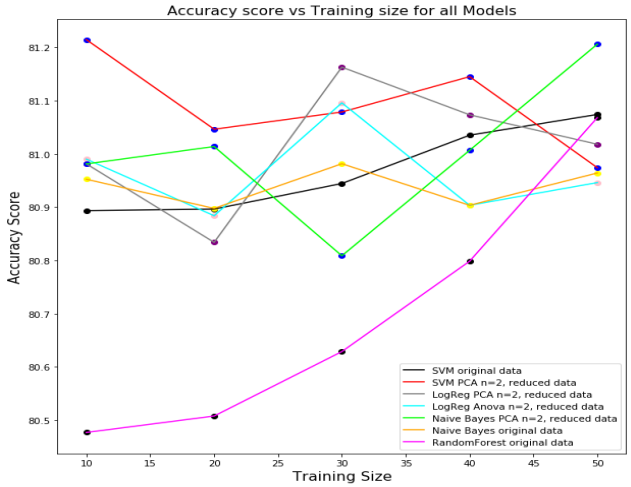

Fig 2: Performance comparison of different classifiers.

SVM takes a lot of time to run as compared to any other classifier described above. This is the biggest drawback if you have to work with large amounts of data. However, SVM gave the highest accuracy of 81.3% with PCA reduced data to 3 components. I have tried using different regularization parameters from c= 1, 2, 2.5, 3 etc. All gave almost similar results and there was no much difference. Hence, I have used c=1 here. SVM however gave a steady increase in the accuracy score for original data input ranging from 80.9-81.1%. SVM being a supervised learning, performs better!

Anova feature selection method was also used to select 2 or 3 best features. However, no classifier performed better than the original data input or the PCA reduced input and hence discarded this method.

It can be seen from the graph plot in Fig 2 that even with PCA reduced data the accuracy is similar to original data in all the classifiers. Hence, this reduction technique will be very helpful when the dataset has larger dimensions and this will become a huge plus as it will reduce the training time.

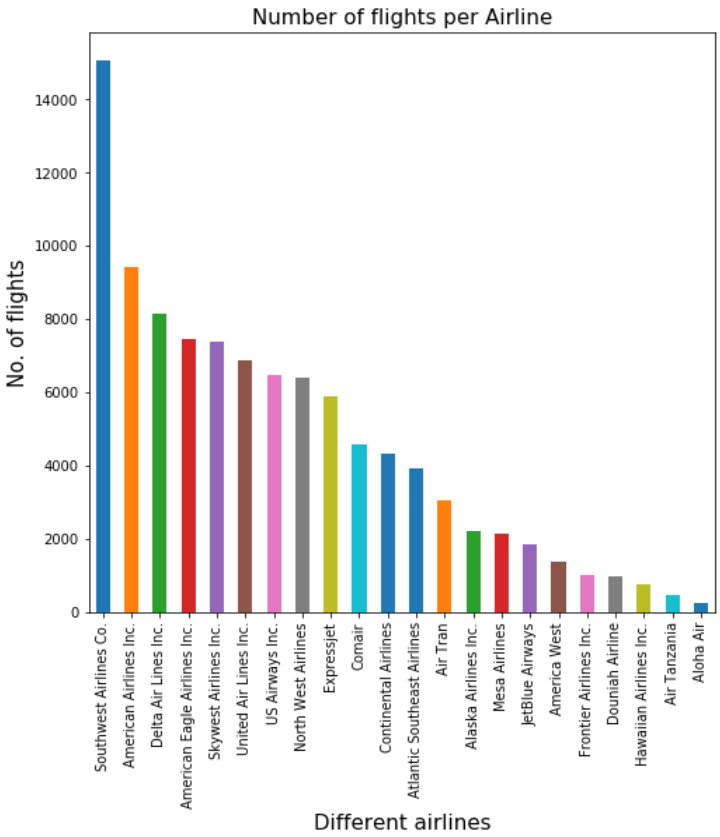

Figure 3 on the left-hand side shows a bar plot of total number of flights operated by a particular airline during an year according to the dataset. It can be inferred easily from the graph that the Southwest Airlines have highest number of flights and the Aloha Air the least.

Fig 3: The total number of flights flying per year for an airline.

Figure 4 on the right-hand side shows which month has the highest amount of delay in terms of percentage of delayed flights for that particular month. It can be easily visualized that December has the highest amount of delay followed by the month of July. It can be inferred that Christmas being a holiday season and having large number of passengers commuting with more number of flights, this month may experience the maximum number of delays. Also, December comes into Winter season and may have the highest snowfall disrupting the flight take off. It can be confirmed that it is holiday season with the next highest being July which has the Independence Day of USA being 4th of July.

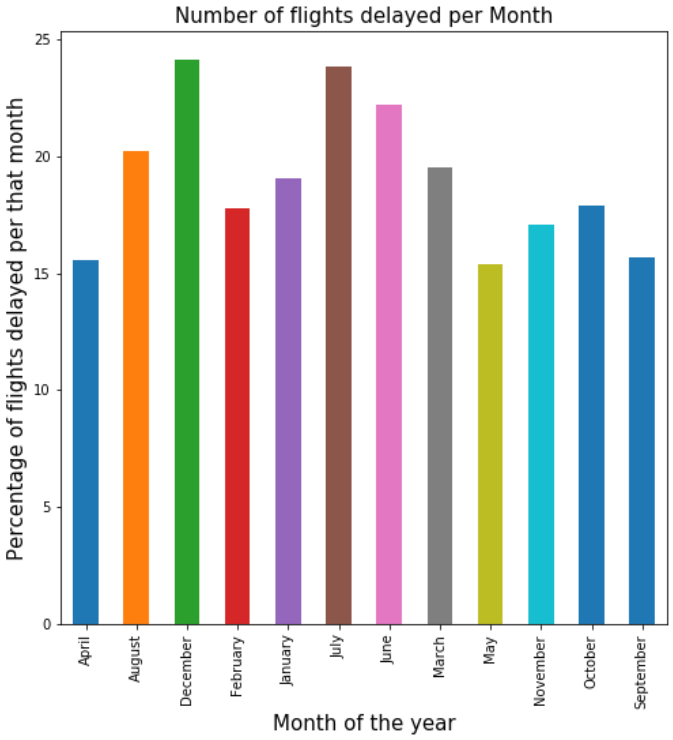

Fig 4: Bar plot depicting the percentage of number of fligths delayed in each month.

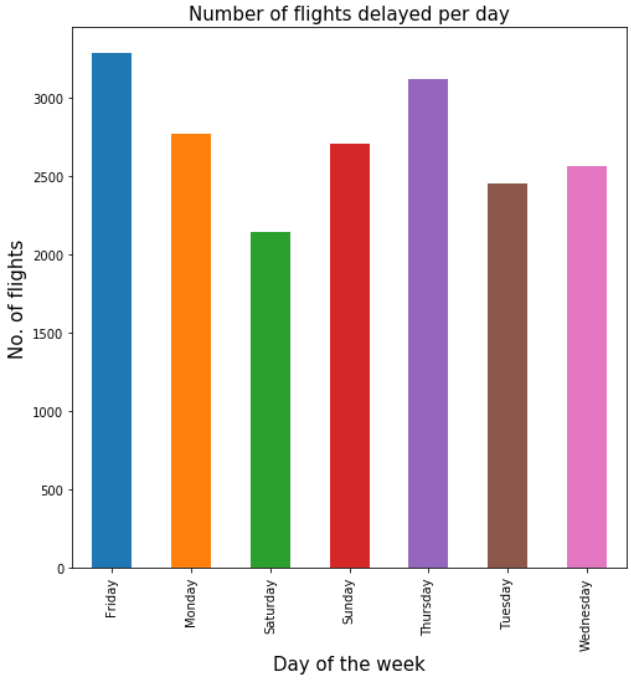

Fig 5: Bar plot depicting the number of flights flying on each day of the week.

Figure 5 on the left-hand side shows which day has the highest amount of delay in terms of number of delayed flights for that particular day. It can be easily visualized that Friday has the highest amount of delayed flights followed by Thursday. It can be inferred that Friday is when weekend starts and working people or students want to visit their families and would like to go home. Or it is weekend time and people would like to go on a short trip. Again, being a holiday time and having large number of passengers commuting with more number of flights, this day may experience the maximum number of delays. Second highest is Thursday as many take off on Friday and have a long weekend. And the third highest is on Monday where everybody who flew over the weekend would come back to their destinations and hence have large passengers and in turn exhibiting delays.

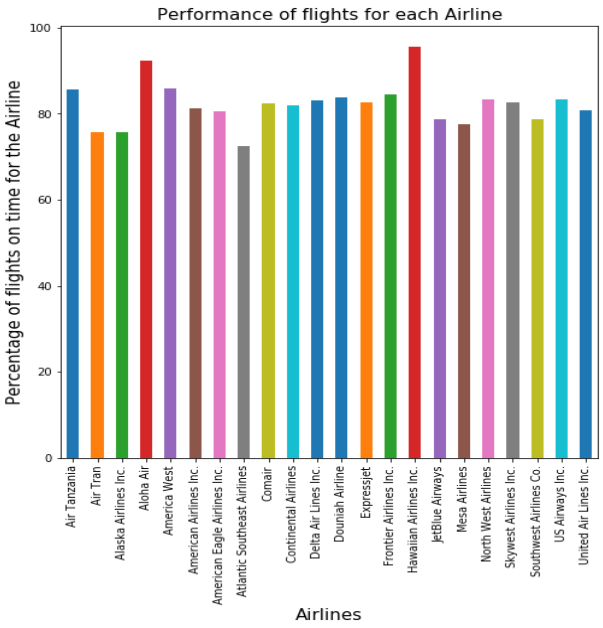

Fig 6: Percentage of flights on time per airline

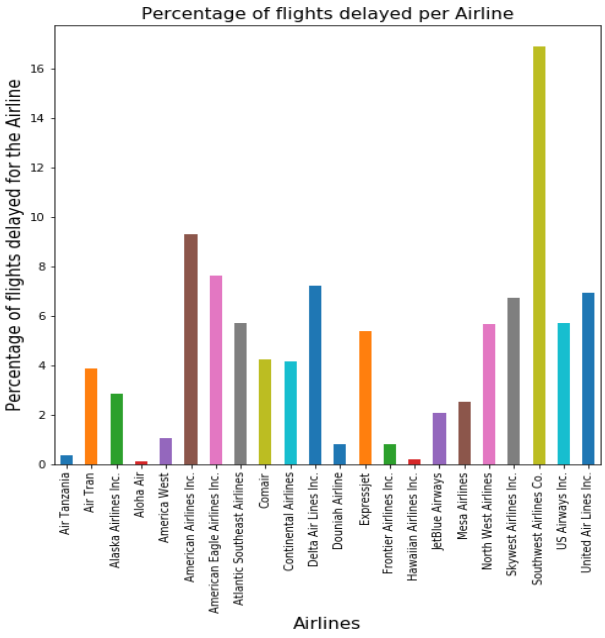

Fig 7: Bar plot depicting the pecentage of delays per airline

The above two bar plots tell us the performance of each airline. From Fig 6, it can be seen that the Hawaiian and Aloha airlines have the best on time performance and from fig 7 it can be inferred that Southwest airlines have the least performance which mean they have the highest number of flight delays. However, on comparing with Fig 3, we can say that Southwest having the highest number of flights and being the busiest to handle, it is natural to have high delays. And, vice versa, Fig 3 shows Hawaiian airlines and Aloha airlines have the least number of operations.

Also, using the basemap package, the airports were plotted on USA map with altitudes and longitudes dataset. This lets us know the range of number of flights flying from each airport. It can be inferred from the plot that more airports are cluttered in the east coast and have high number of flights flying. And the big cities/busiest airports have the highest number of flights like Atlanta.

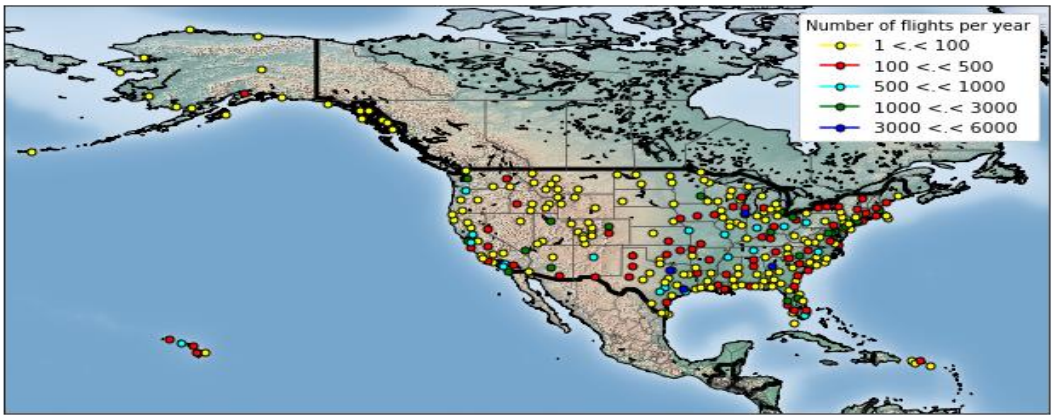

Fig 8: Visualization of airports and number of flights flying from it per year using Heatmap

Conclusion:

What was done?

By dimensionality reduction of input data from PCA, found that classes were overlapping. Different classifiers were used to predict the data. The PCA reduced data with 2 components and 3 components was inputted to different types of classifiers. KNN, SVM, Naïve Bayes, Random Forest and Logistic Regression methods were used amongst many different classifiers. The accuracy of the predicted outcomes were compared with all the methods and also with PCA reduced data. Inferences were drawn based on this to select which classifier and if reduced data is of any use or not. Also, Anova was used to select 2 and 3 best features out of the sample data. But Anova output data when used as input to different classifiers, they didn't perform well as compared to PCA reduced and original data. Hence, this method was dropped.

And, different kinds of comparisons were done using histograms for data visualization. Namely No. of flights vs the carrier, percentage of delayed flights for each month and each day of week. Performance of flights per each airline etc. Also, a basemap package was used to visualize the airports and number of flights flying from each airport on USA's map using latitudes and longitudes of the airports.

What were the outcomes?

It was observed that SVM gave the highest accuracy for PCA n=2 reduced data. But, SVM takes a lot of time like atleast 20 mins to run my dataset. Most of the classifiers except KNN performed well and gave around 80-81% accuracy. KNN may have not performed well in the initial stages because of less training sizes as it needs similar data to get trained, thus by increasing the training size, it performed better. The random forest classifier also had an increasing curve for the accuracy where as other classifiers didn't give a smooth curve when the training sizes increased. It can be inferred that as the training size is increasing the data can get overfitted in these models.

The important thing noticed here is that PCA reduced data has a high variance of about 98% and when this data is fed into the classifiers, it performed almost equally well compared to the original data. It can be inferred that for high number of dimensions, if I reduce the dataset's components to 2 or 3 using PCA, it still performs well. This is very helpful in reducing the training time and avoiding overfitting with irrelevant features. This is a boon to save time!

I can conclude that if you have time and need more accuracy, SVM can be used to get a stable classification, but it gets overfitted for high training sizes which must be kept in mind. If small amounts of data, Logistic regression can be used as this dataset has binary classes as it is simple, easy and quick. Naïve Bayes classifier can be used for PCA reduced data. Depending on the features and the size of the dataset, classifier has to be chosen.

And, from the data visualizations, I can infer that the holiday season in December and July have the highest number of delays as stated in the hypothesis. Hypothesis has been proven right. Also,

as Friday is the time when more passengers fly, even that day has highest number of delays. Another factor observed was the more the number of flights per carrier, more they have delays and vice versa. The delays are directly proportional to the number of flights per carrier. Finally, using heatmaps it can be seen that East Coast has the highest number of airports with highest number of flights flying from each airport.

This project has been very interesting and engaging to me. I have learnt mainly how to do the data analysis. I had lots of fun trying to visualize data assuming different circumstances and playing around with data. Starting from scratch, without knowing before the class started, now I understand a lot about data science and data analysis.

References:

1. Dataset: "https://www.kaggle.com/aenik97/flight-delays"
2. Label encoder: "https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features"
   "https://towardsdatascience.com/encoding-categorical-features-21a2651a065c"
3. PCA : "https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_iris.html"
   "https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60"
4. ANOVA: "https://chrisalbon.com/machine_learning/feature_selection/anova_f-value_for_feature_selection/"
5. KNeighboursClassifier: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
6. Logistic regression: "https://en.wikipedia.org/wiki/Logistic_regression"
   "https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html"
7. SVM: "https://scikit-learn.org/stable/modules/svm.html"
8. Random Forest Classifier" https://www.datacamp.com/community/tutorials/random-forests-classifier-python"
   https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html"
9. Train test split: "https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html"
10. CS588 Fall 2019 lecture slides

Appendix:

The code is implemented in Python using Jupyter notebook. The files are zipped and attached below:

Implementation in Python.zip

The datasets used are zipped in the file below:

dataset.zip