**CMPE 255 – Data Mining - Spring 2018**

**Group Project**

**Team**

Anirrudh Venkatraman - 011547679

Divyaa Thellore Venkataraman - 011822057

Sadab Qureshi - 011486033

**Project : Predicting flight delays and the reason for the delays**

[**Section 1 Introduction**](#_dg5s0oaynb79) **2**

[1.Motivation](#_oqr5wwv80edv) 2

[2.Objective](#_4zsscgxzwbs8) 2

[For flight delay prediction](#_ejmlf2ardtl2) 2

[For flight delay reason prediction](#_dyww9ldwjofj) 3

[3.In the Market](#_739g7nv57iiz) 3

[**Section 2 System Design & Implementation details**](#_oy2hbxbhkivj) **3**

[1. Algorithm(s) considered](#_ucypbl90y4n1) 3

[2.Technologies & Tools used](#_hrv5bv1nwwgf) 4

[Jupyter Notebook:](#_sryxh2rwwvkh) 4

[Python 2.7:](#_cx24fvgr0k9p) 4

[Dataset:](#_f8cbamojpkv) 4

[**Section 3 Experiments / Proof of concept evaluation**](#_ses42s1hmrmj) **5**

[1.Dataset(s) used](#_9x3ldkiw4633) 5

[2.Methodology followed](#_qbbjv29qobpp) 6

[1. For Delay Prediction](#_xgekmx99am8t) 6

[Steps Followed:](#_g4q5s85zwawn) 6

[Linear Regression](#_phaz9kjjkt9e) 6

[Ridge Regression](#_uydaugckjgwk) 6

[Random Forest](#_tbx2brx7r2br) 7

[Lasso Regression](#_4otduhudu7gt) 7

[2. For delay reason prediction](#_nn3lignolzbz) 7

[Steps Followed:](#_87hc0tvlx709) 7

[Linear Regression](#_o4l1ubgjzaul) 8

[Ridge Regression](#_v2c5ykt1tvu) 8

[3. Graphs](#_mkrn8sp26zq1) 8

[4. Analysis of results](#_z044k6uy20jc) 10

[**Section 4 Discussion & Conclusions (bullet points as applicable)**](#_fpfa0lcxcvu7) **10**

[1.Decisions made](#_d0d5jnntv2v8) 10

[2.Difficulties faced](#_fxztnyy00h5i) 11

[3.Things that worked](#_2e9tp3tjh7xt) 11

[4.Things that didn’t work well](#_rvm6gtpq4wpt) 11

[5.Conclusion](#_1rz1kyo98g9i) 11

[**Section 5 Project Plan / Task Distribution**](#_54j3kviylk1o) **11**

[**References**](#_mppu8d6g5lcf) **13**

# Section 1 Introduction

Name : Predicting Flight delays and the reason for the delays

## 1.Motivation

People who travel a lot know the pain of having to wait at the airport knowing that they are stuck for an indefinite time. Not being able to decide whether or not they can walk around and waiting to hear intently for their boarding call. People can plan their stay in airports properly if they know how long they would have to wait based on their flight delays. They can get food accordingly and make plans to keep them occupied for the time they wait and not get anxious every time they listen to some boarding call. Even a prediction of few minutes beforehand can make a tremendous impact on the waiting time management. People can inform their friends that they will be late and they can either start or park at the right places to avoid having to pay extra when they come to pick them up.

There are several projects apply Data Mining and Machine Learning concepts for various prediction models. However with the increase in the technology people are expecting more precise models for the prediction. We are trying to build a model that will predict the delay. Predicting the delay is just one module of this project. We are also doing an analysis to predict the “reason for the delay” for the flight.

**The main intent of this project is to make use of the raw data and derive useful flight delay information from it and also predict the reason for the delay.**

## 2.Objective

### For flight delay prediction

* Using the available Data mining algorithms to find out the right model that would predict the flight delay.
* Applying the right preprocessing before applying the algorithm to increase the precision.
* Selecting the right attributes and fitting them to the model
* Choosing the right regression algorithms to predict the delays
* Dividing the dataset into train and test set
* Train based on the training dataset and fitting to the test dataset.
* Find the accuracy of the model based on metrics

### For flight delay reason prediction

* Grouping the required attributes together and also applying the proper preprocessing on the data to improve the accuracy
* Ways to deal with missing values in the fields
* Finding the correlation between the individual reasons and also between all the reasons to see how each affects the other
* Splitting the dataset into training and test dataset and applying the regression techniques to find the reason for the delay.
* Applying metrics to evaluate the precision of the model

## 3.In the Market

* Google has updated its Flights app with a pair of new features that should help weary (and wary) travelers get to grips with the next trip to the airport. The first uses machine learning to predict upcoming flight delays, and the second breaks down exactly what different airlines mean by “basic economy” — explaining what amenities are and are not included in so-called last class. [2]

Click on the link for the article :

<https://www.theverge.com/2018/1/31/16955580/google-flights-app-delays-machine-learning-economy>

* More reads for similar projects.

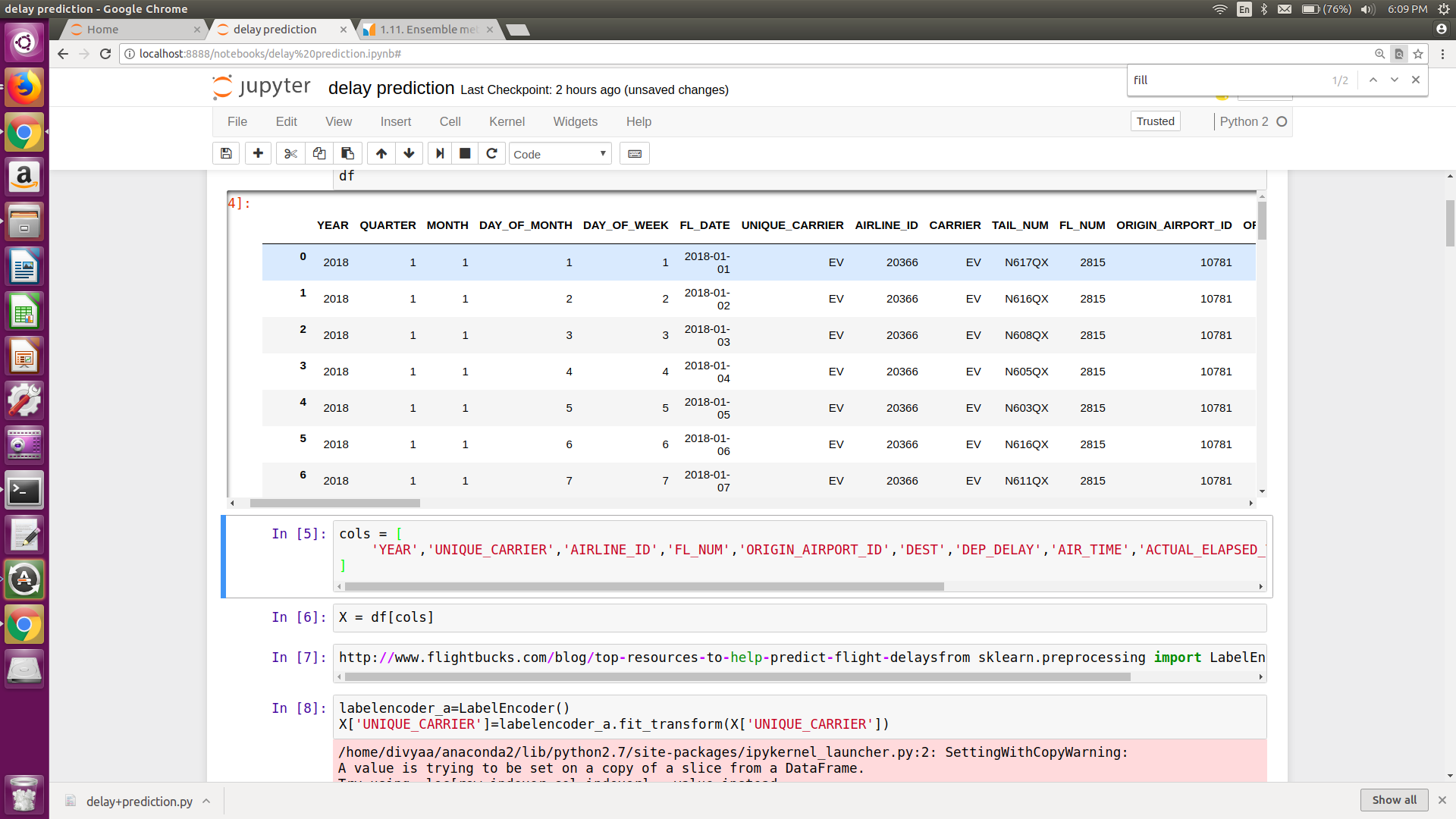
<http://www.flightbucks.com/blog/top-resources-to-help-predict-flight-delays>

These show that there is a need for flight delay prediction softwares in the market.

# Section 2 System Design & Implementation details

## 1. Algorithm(s) considered

The data is loaded from the TIME! File. The loaded data is presented in the following screen-shot.



Algorithms are chosen predominantly to suit the data. The dataset for this project suggest regression techniques. Regression techniques were applied to predict the delay and for each the metrics are calculated to see which algorithm performs best for the chosen dataset.

## 2.Technologies & Tools used

Data Mining algorithms are used to predict the fight delays and the reason for the delays.

### Jupyter Notebook:

The **Jupyter Notebook** is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.[1]

### Python 2.7:

Python is the preferred language for Data Science as it has many in-built libraries for the implementation. For this project we chose to use python 2.7 and used some of the libraries from sklearn like OneHotEncoder, Ridge, Lasso etc.

### Dataset:

- <https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236>

# Section 3 Experiments / Proof of concept evaluation

## 1.Dataset(s) used

* Source of data:

<https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236>

There are 72 attributes in the dataset. Listed below are some of those.

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Field Name** | **Description** |
| 1 | Year | Year |
| 2 | Month | Month |
| 3 | DayofMonth | Day of Month |
| 4 | DayOfWeek | Day of Week |
| 5 | OriginAirportID | Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused |
| 6 | DestAirportID | Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused. |
| 7 | DepTime | Actual Departure Time (local time: hhmm) |
| 8 | DepDelay | Difference in minutes between scheduled and actual departure time. Early departures show negative numbers. |
| 9 | ArrTime | Actual Arrival Time (local time: hhmm) |
| 10 | ArrDelay | Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers. |
| 11 | CarrierDelay | Carrier Delay, in Minutes |
| 12 | WeatherDelay | Weather Delay, in Minutes |
| 13 | NASDelay | National Air System Delay, in Minutes |
| 14 | SecurityDelay | Security Delay, in Minutes |
| 15 | LateAircraftDelay | Late Aircraft Delay, in Minutes |

## 2.Methodology followed

### 1. For Delay Prediction

#### Steps Followed:

* The dataset is loaded
* Encode the labels of those columns that have string values
* Find the correlation to see how ‘ARR\_DELAY’ is correlated to all the other attributes.
* Drop all unwanted or un correlated attributes
* Using “OneHotEncoder” to give values to label encoder.
* Creating new list of fields with all the required fields and the encoded labels

#### Linear Regression

* Splitting the dataset to train and test dataset
* Applying linear regression to train the model
* Fit the test data to the trained model
* Evaluate metrics:

- Mean Absolute Error - 8.00

- root mean square - 11.3120800394

- Mean squared error - 127.96

#### Ridge Regression

* Splitting the dataset to train and test dataset
* Applying linear regression to train the model
* Fit the test data to the trained model
* Evaluate metrics:

- Mean Absolute Error - 8.00

- root mean square - 11.3119565739

- Mean squared error - 127.96

#### Random Forest

* Splitting the dataset to train and test dataset
* Applying linear regression to train the model
* Fit the test data to the trained model
* Evaluate metrics:

- Mean Absolute Error - 21.68

- root mean square - 46.5137545878

- Mean squared error - 2163.53

#### Lasso Regression

* Splitting the dataset to train and test dataset
* Applying linear regression to train the model
* Fit the test data to the trained model
* Evaluate metrics:

- Mean Absolute Error - 0.23

- root mean square - 0.285507606833

- Mean squared error - 0.08

### 2. For delay reason prediction

#### Steps Followed:

* The dataset is loaded
* Remove those entries without delays
* There are 5 reasons that can lead to a flight delay
* The delay in flight can be due to a single reason or due to combination of reasons. The delay reason can also be due to some other reason.
* Correlation between the delays each reason and all the other attributes are found and are stored in a separate data frame.
* For those attributes that have good correlation but do not have values in a specific row those are filled with mean values.
* Repeat the same for all the reason data frames

#### 

#### Linear Regression

* Split the data into train dataset with those rows from the attributes that have non empty values.
* Those with empty values will form the test dataset
* Using linear regression fit the model to the train dataset
* Apply the trained model to the test set to predict the minutes delay caused by the reason
* Repeat the same for all the reason
* Sum up all the minutes of delay from the predicted minutes to create a new attribute by name ‘New\_Delay’
* Calculate metrics :
  + - * - Accuracy - 42.719 %
      * - rms - 6.844

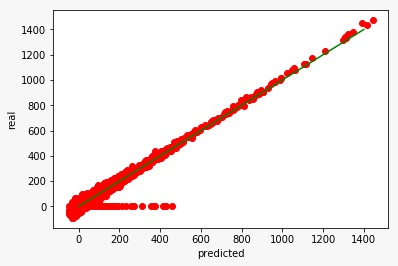
#### Ridge Regression

* Split the data into train dataset with those rows from the attributes that have non empty values.
* Those with empty values will form the test dataset
* Using ridge regression fit the model to the train dataset
* Apply the trained model to the test set to predict the minutes delay caused by the reason
* Repeat the same for all the reason
* Sum up all the minutes of delay from the predicted minutes to create a new attribute by name ‘New\_Delay’
* Calculate metrics :

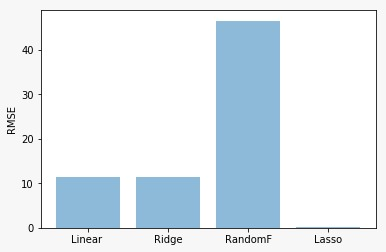
- Accuracy - 42.717 %

- rms - 6.8

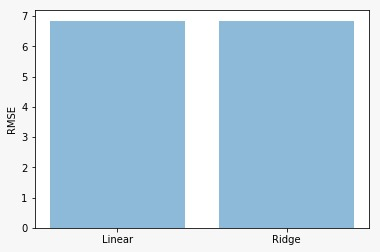
## 3. Graphs



Linear Regression for Delay Prediction



Comparison of regression models in delay prediction



Comparison of the regression models for rason Prediction

## 4. Analysis of results

|  |  |  |
| --- | --- | --- |
| **Module** | **Applied Technique** | **Observation** |
| **For flight delay prediction** | Linear Regression | Mean Absolute Error - 8.00  root mean square - 11.3120800394  Mean squared error - 127.96 |
| **For flight delay prediction** | Ridge Regression | Mean Absolute Error - 8.00  root mean square - 11.3119565739  Mean squared error - 127.96 |
| **For flight delay prediction** | Random Forest | Mean Absolute Error - 21.68  root mean square - 46.5137545878  Mean squared error - 2163.53 |
| **For flight delay prediction** | Lasso Regression | Mean Absolute Error - 0.23  root mean square - 0.285507606  Mean squared error - 0.08 |
| **Reason for delay** | Linear Regression | Accuracy: 42.719 %  root mean square: 6.844 |
| **Reason for delay** | Ridge regression | Accuracy: 42.717 %  root mean square: 6.8 |
| **Reason for delay** | Polynomial regression | Accuracy: 42.717 %  root mean square: 6.8 |

# Section 4 Discussion & Conclusions (bullet points as applicable)

## 1.Decisions made

The following decision were made as a team to proceed with this project

* Choosing which project to work on
* What modules will be included
* What preprocessing techniques to be applied
* What models to be tested
* What metrics to be chosen to evaluate the results
* How to deal with the fields with empty values

## 2.Difficulties faced

The dataset that we chose had values for delay. Though there is an attribute for the reason for the delay and 5 fields for reasons most of the values are empty.

It was challenging to deal with the empty fields. We had two options:

1. To ignore the empty values (which was nearly 80% of the database)- which will lead to high approximation
2. To fill the values in the empty fields using regression techniques from those values that were present in the database.

## 3.Things that worked

1. Finding the flight delay after data preprocessing predicting the delay of the flights worked
2. Finding the correlation between the fields helped a lot for finding the dependency between the fields.
3. Regression techniques applied were very effective
4. Filling the empty values with values improved the accuracy
5. Eliminating the outliers improved the model

## 4.Things that didn’t work well

1. It took a long time for us to decide how to deal with the missing values

## 5.Conclusion

1. Preprocessing applied depends on the dataset
2. It is based on the dataset that we choose what algorithms to apply
3. Using metrics to choose the best model for the dataset is very effective
4. There is dependency between the various reasons for flight delay

# Section 5 Project Plan / Task Distribution

**Modules**:

FD- Flight Delay

RP - Reason Prediction

|  |  |  |
| --- | --- | --- |
| **Name** | **Module** | **Assigned Task** |
| TEAM | FD | Research for the dataset |
| TEAM | FD | Analysis on the dataset |
| TEAM | FD | Choosing techniques |
| TEAM | RP | Analysis on the Dataset |
| TEAM | RP | Choosing preprocessing to apply |
| TEAM | RP | Agreeing on how to fill the empty values |
| Anirrudh | FD | Loading the dataset |
| Anirrudh | FD | Using label encoder for String labels |
| Anirrudh | FD | Loading the required fields |
| Anirrudh | FD | Dropping the unwanted attributes |
| Sadab | FD | OneHotEncoder to encode the labels |
| Divyaa | FD | Creating the df |
| Divyaa | FD | Filling the empty values |
| Divyaa | FD | Splitting the train and the test dataset |
| Anirrudh | FD | Applying the Linear Regression model to find the delay |
| Anirrudh | FD | Calculating the various metrics to evaluate the accuracy of the model |
| Divyaa | FD | Applying Ridge regression model |
| Divyaa | FD | Calculating the various metrics to evaluate the accuracy of the model |
| Divyaa | FD | Applying Random forest model |
| Divyaa | FD | Calculating the various metrics to evaluate the accuracy of the model |
| Sadab | FD | Applying lasso regression model |
| Sadab | FD | Calculating the various metrics to evaluate the accuracy of the model |
| Anirrudh | RP | Dropping rows without delay |
| Divyaa | RP | Found correlation between each reason and rest of the dimensions |
| Divyaa | RP | Created dataframe for each reason with the correlated dimensions |
| Sadab | RP | Split the data into train and test set |
| Sadab | RP | Filled the empty values with mean value |
| Anirrudh | RP | Applied linear regression to train the model |
| Anirrudh | RP | Fit the model to the test set |
| Divyaa | RP | Applied linear regression technique for all the other reasons |
| Sadab | RP | Summed up all the results that get generated for total delay calculation |
| Divyaa | RP | Computed accuracy |
| Anirrudh | RP | Computed rms |

# References

[1] <http://jupyter.org/>

[2]<https://www.theverge.com/2018/1/31/16955580/google-flights-app-delays-machine-learning-economy>