ASSIGNMENT 1 - REPORT

MACHINE LEARNING - FALL 2021

KHUMOYUN AMINADDINOV

M21-DS-01

24th of September, 2021

ABSTRACT

Today, airlines industry is booming as flying from one place to another is considered to be the fastest way of transportation. But it may not be referred to as the most comfortable due to some inconveniences such as inevitable flight delays (due to several factors such as extreme weather conditions or flight durations). Delays can have severe consequences like when you are rushing to get medical care in a different location or trying to be present at an important meeting. All these cases can result in unwanted changes in your plan schedule which is not comfortable at all. Therefore, being able to predict these delays can save lives or help you optimize your schedule which could save you a fortune, thus, there is a huge demand for flight delay prediction models or systems in the industry. This report explains three different ways of predicting flight delays based on duration, departure & destinations airports and time of the day. month.

INTRODUCTION

In this report, I analyze several factors which can cause delays and I use these factors to build 3 machine learning models to predict the new flight delays. Some of the predictors are new features that are calculated from existing columns of the given dataset, some of the unneeded columns are removed and some of them are encoded to numerical values to make them easy for calculations. Then I visualize the outcome with some of the independent variables to get a better idea to choose models. I use fundamental statistics to detect the outliers and remove them in order to fit the models better. Then I use several metrics to evaluate models and compare them accordingly.

MOTIVATION

With this assignment, I try to put all the knowledge that I learned from this course into practice and train myself in solving this challenging task. I hope that this assignment will serve as an interesting practical way of improving my knowledge on outlier detections, regressions and regularization. As this report includes all the practical steps needed to implement data preprocessing, three machine learning models and evaluation - you can expect to gain practical knowledge on how to analyze data and derive new values from it.

PART 1: TASK

1.1. DESCRIPTION

The task begins with reading the given dataset and preprocessing it - cleaning, encoding, adding new calculated features and getting rid of unnecessary ones. Then follows dataset visualization that helps to understand the shape and choose an appropriate model and algorithm. After building models, the task comes to an end by evaluating the models and comparing them.

1.2. DATASET, PREPROCESSING & OUTLIER DETECTION

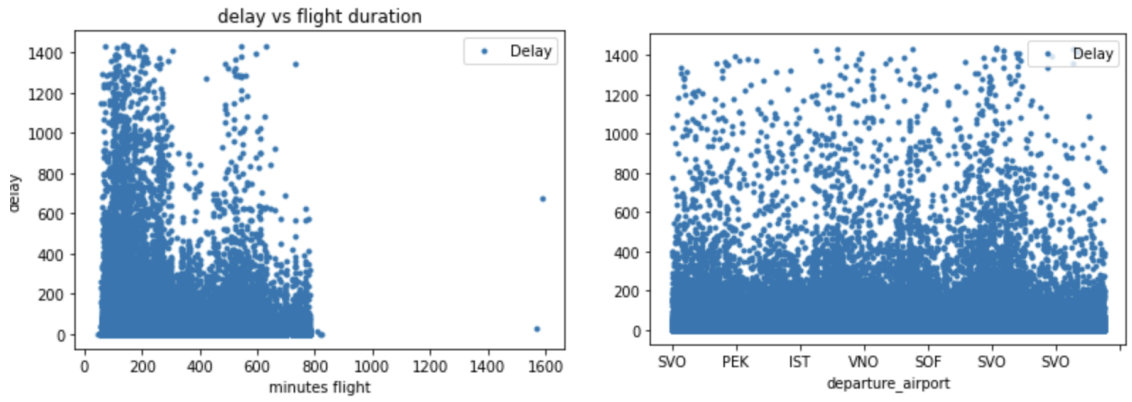
1. We are given with the following flight delay dataset:

	Depature Airport	Scheduled depature time	Destination Airport	Scheduled arrival time	Delay	departure_airport	destination_airport
0	SVO	2015-10-27 07:40:00	HAV	2015-10-27 20:45:00	0.0	SVO	HAV
1	SVO	2015-10-27 09:50:00	JFK	2015-10-27 20:35:00	2.0	SVO	JFK
2	SVO	2015-10-27 10:45:00	MIA	2015-10-27 23:35:00	0.0	SVO	MIA
3	SVO	2015-10-27 12:30:00	LAX	2015-10-28 01:20:00	0.0	SVO	LAX
4	OTP	2015-10-27 14:15:00	SVO	2015-10-27 16:40:00	9.0	OTP	SVO

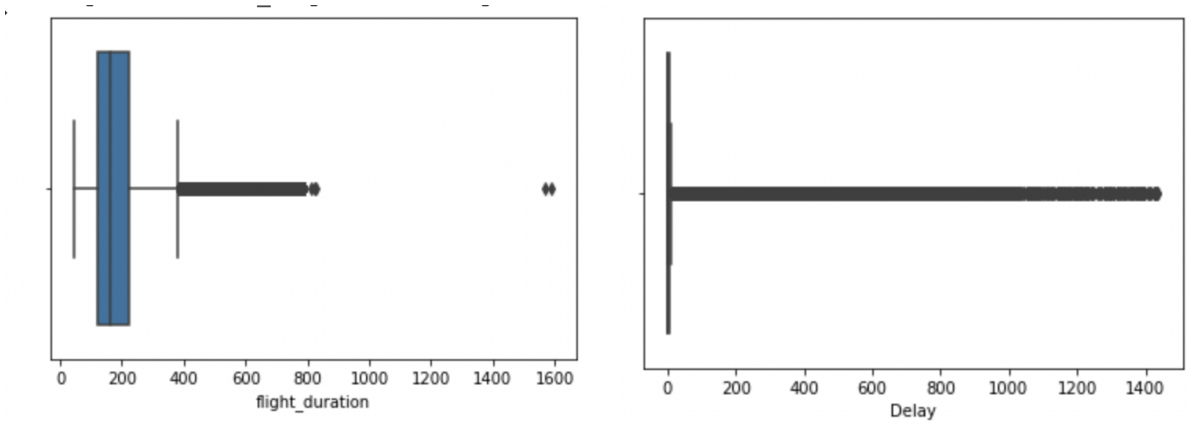
For delay prediction I used linear regression, multiple linear regression and polynomial regression with 1, 5 and 5 features correspondingly. So first I split date and time and remove date and time merged column, then I encode departure and destination airports using **sklearn.processing.LabelEncoder** and at last I create new columns - Weekday(from date) and duration time, by subtracting departure time and scheduled arrival and so it looks like this:

	0	1	2
Scheduled depature time	2015-10-27 07:40:00	2015-10-27 09:50:00	2015-10-27 10:45:00
Scheduled arrival time	2015-10-27 20:45:00	2015-10-27 20:35:00	2015-10-27 23:35:00
Delay	0	2	0
departure_airport	SVO	SVO	SVO
destination_airport	HAV	JFK	MIA
departure_date	2015-10-27 00:00:00	2015-10-27 00:00:00	2015-10-27 00:00:00
departure_time	07:40:00	09:50:00	10:45:00
departure_dayofweek	1	1	1
arrival_date	2015-10-27 00:00:00	2015-10-27 00:00:00	2015-10-27 00:00:00
arrival_time	20:45:00	20:35:00	23:35:00
arrival_dayofweek	1	1	1
timedate_departure	2015-10-27 07:40:00	2015-10-27 09:50:00	2015-10-27 10:45:00
timedate_arrival	2015-10-27 20:45:00	2015-10-27 20:35:00	2015-10-27 23:35:00
flight_duration	785	645	770

2. Then I plot **flight_duration vs. delay** and **departure_airport vs. delay** using matplotlib.



I realize that I need to get rid of some points too far from majority so I use seaborn library and use boxplot to plot the outliers:



Using different threshold values I picked 3 and removed outliers from the dataset using z-score.

PART 2: SOLUTION

2.1. SPLIT DATA

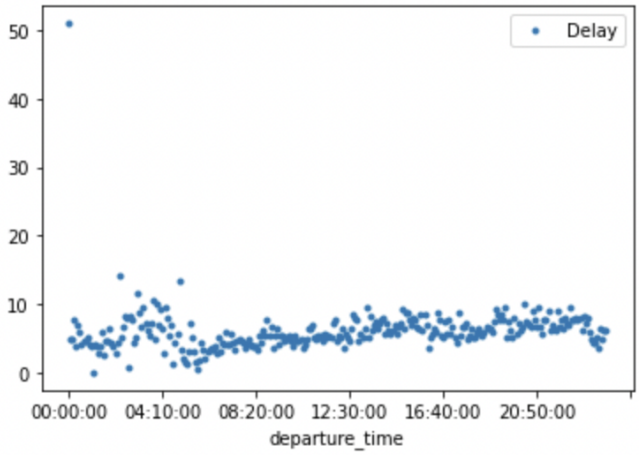
I divide the dataframe on X_train, X_test and y_train, y_test based on *departure_year*.

```
X = flight_delay_df.drop(['Delay', 'Scheduled depature time',
                          'Scheduled arrival time', 'departure_date',
                          'departure_time', 'arrival_date', 'arrival_time',
                          'timedate_departure', 'timedate_arrival'], axis=1)
y = flight_delay_df.loc[:, ['Delay']]
X_train = X.loc[flight_delay_df['departure_date'].dt.year<2018]
y_train = y.loc[flight_delay_df['departure_date'].dt.year<2018]
X_test = X.loc[flight_delay_df['departure_date'].dt.year==2018]
y_test = y.loc[flight_delay_df['departure_date'].dt.year==2018]
X_test
```

2.2. MACHINE LEARNING MODELS

Linear Regression

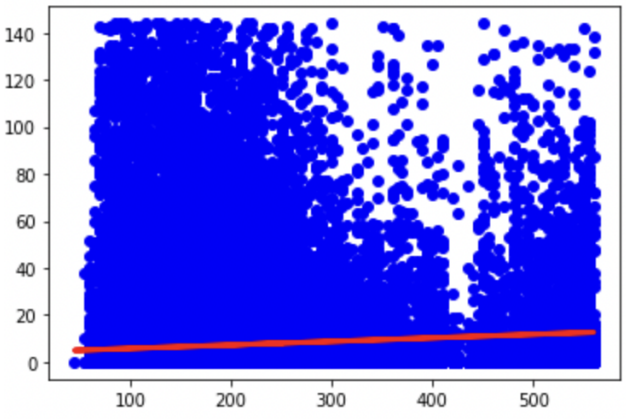
Looking at the dataset I thought it is a mess and there is no way to predict delay from these features(airport departure, destination, date, time) until I found some correlation shape on **departure_time vs. delay:**



I can clearly see a polynomial shape on this plot. The delay values are calculated as average over all flights at a particular time.

As it was proposed in task description, for one feature model I should use **flight_delay** so my linear regression looks like this:

Blue points are delay time and red line is a model.

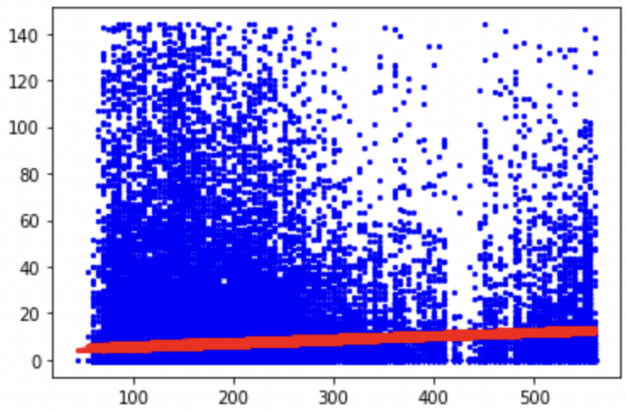



```
Model intercept : [4.28632133]

Model coefficient : [[0.01499899]]

Mean Absolute Error: 8.3478

Mean Squared Error: 178.1137

Root Mean Squared Error: 13.3459

R score -0.0711
```

It seems like the left bottom corner is very high density and that is why the line goes with a positive slope.

Multiple Linear Regression

Second model I used was multiple polynomial regression. Because we have several features I thought it will have some effect on prediction.

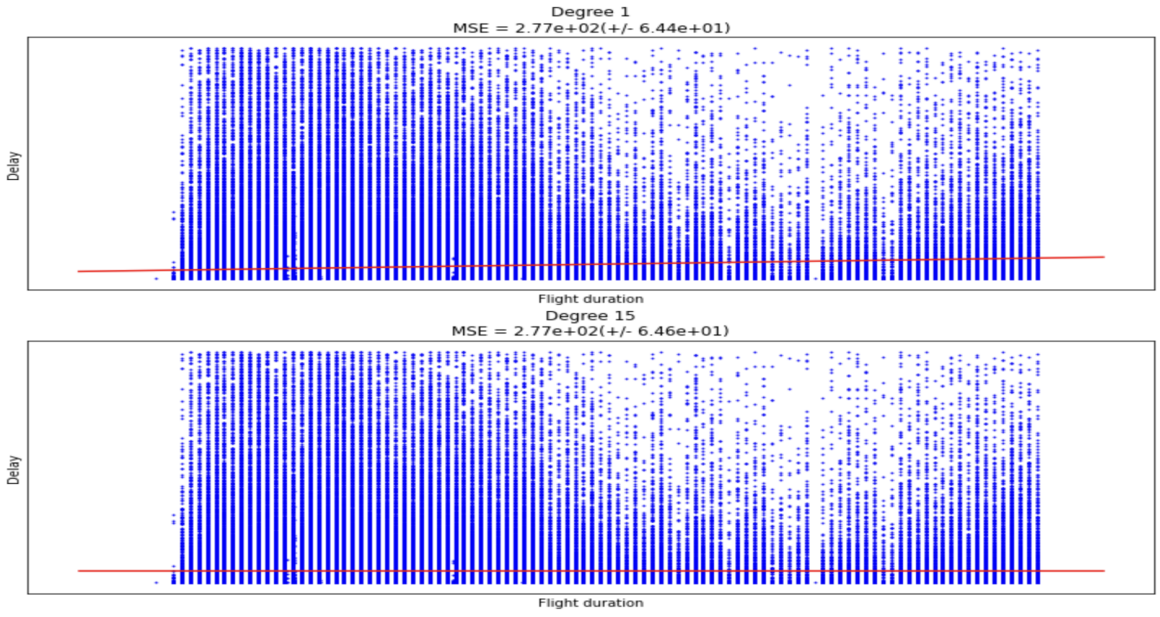


Even considering multiple features and PCA dimension reduction, it did not affect the model significantly, I would say at all.

```
Model intercept : [3.42129497]

Model coefficients : [[-0.00590683
0.01256019  0.01705326  0.01705326
0.01533516]]
```

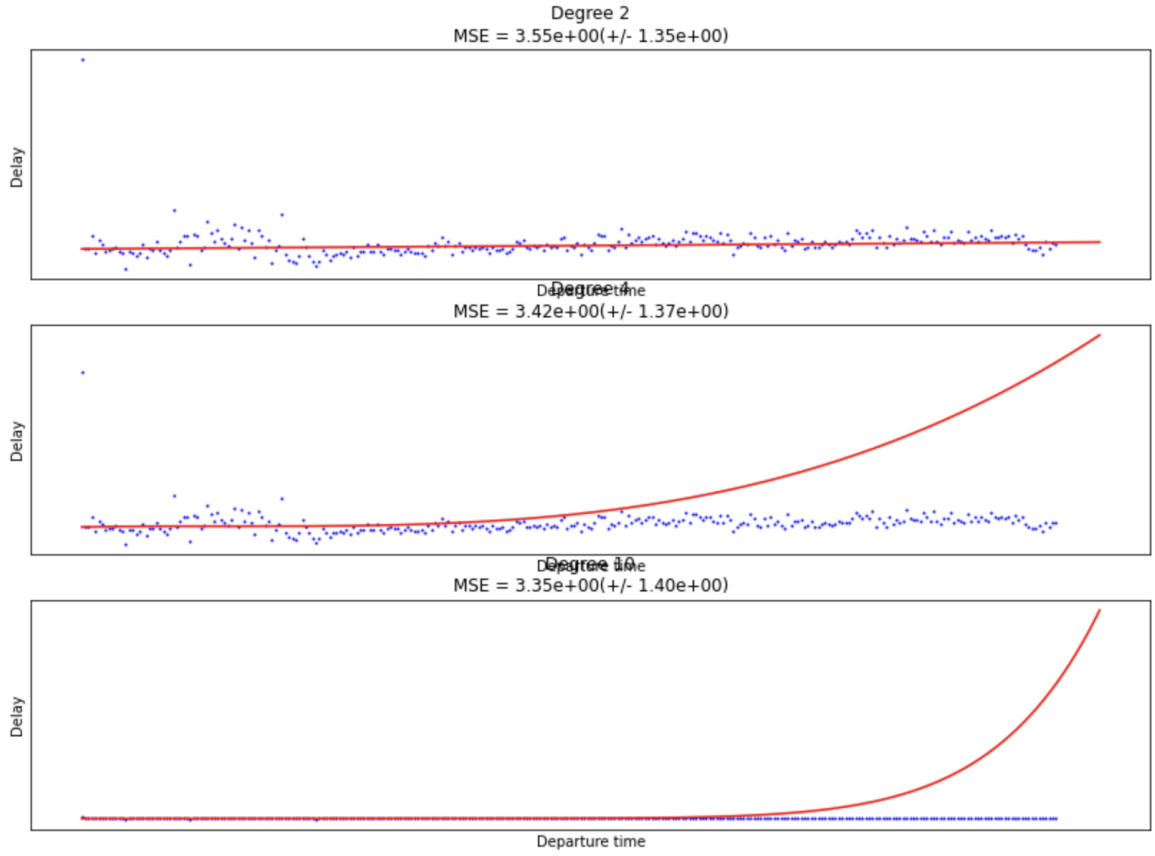
Polynomial Regression

And the following two polynomial regression models predict depending on fligh_duration feature(with degree 1 and 15) and departure_time.



df

Degree 2, 4 and 10



CONCLUSION

Even though it did not fit even nearly to perfect I had experience of working with data near million and multiple features. I learned that by examining data from different perspectives can result in various correlations.

REFERENCES

1. https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
2. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html#sklearn.preprocessing.LabelEncoder.fit_transform
3. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.astype.html