Report
Flight Delay Forecasting

Machine Learning (F21)
Innopolis University, 2021
Assignment 1

Abstract

Accurate flight delay prediction is fundamental to establish the more efficient airline business. Recent studies have been focused on applying machine learning methods to predict the flight delay. Most of the previous prediction methods are conducted in a single route or airport. The aim of this study project is to utilize different machine learning algorithms on real world data to be able to predict flight delays for all causes, in order to create more efficient flight schedules. We will analyse different algorithms from the accuracy perspective and propose a combined method in order to optimize our prediction results.

preprocessing

Through this work, I implement many preprocessing steps such as cleaning and visualization of the data before modeling in order to have a good performance.

1. **Setup and import data**

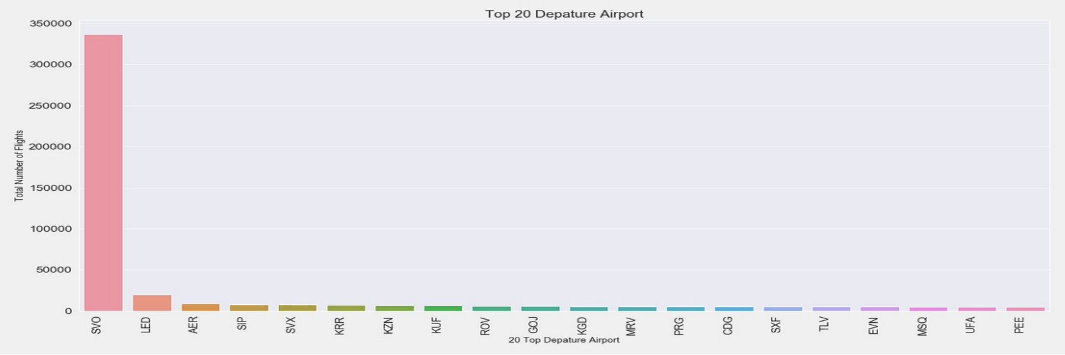
2. **Exploratory data analysis (EDA) :**

   Through this step, we got important statistics and information about data like shape, features type, missing values, and duplicates.

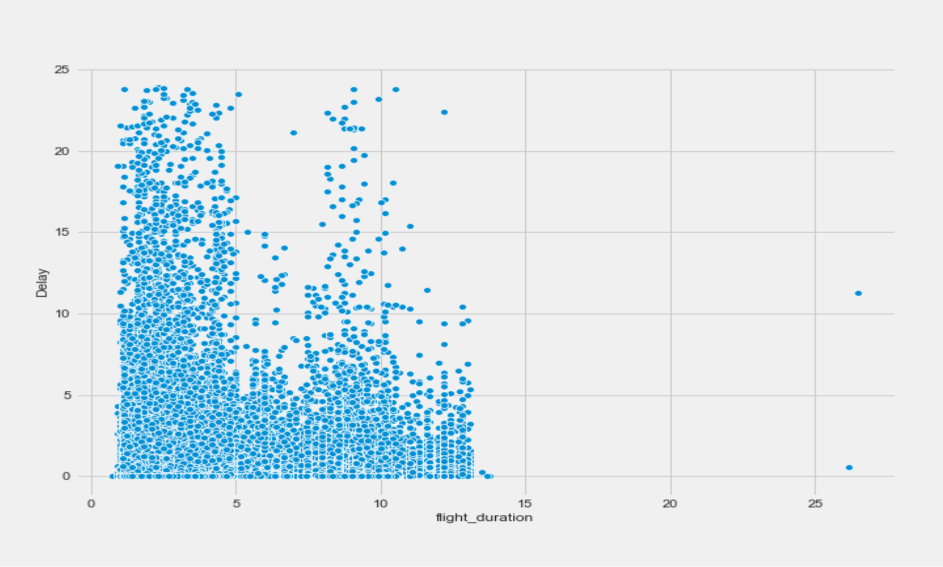
3. **Data preprocessing and visualization :**

   In this step, we did some cleaning and feature extraction such as convert DateTime from an object into an integer and extract a new feature [flight duration].

- Draw other graphs to build strong insight from data such as show top 20 departure and destination airports and found that 'SVO'.

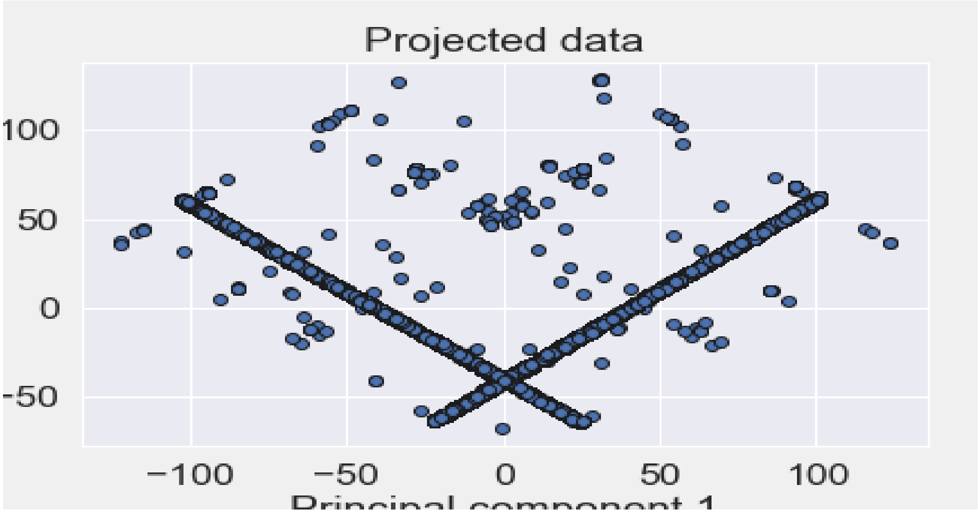


- show relationship between flight duration and delay.



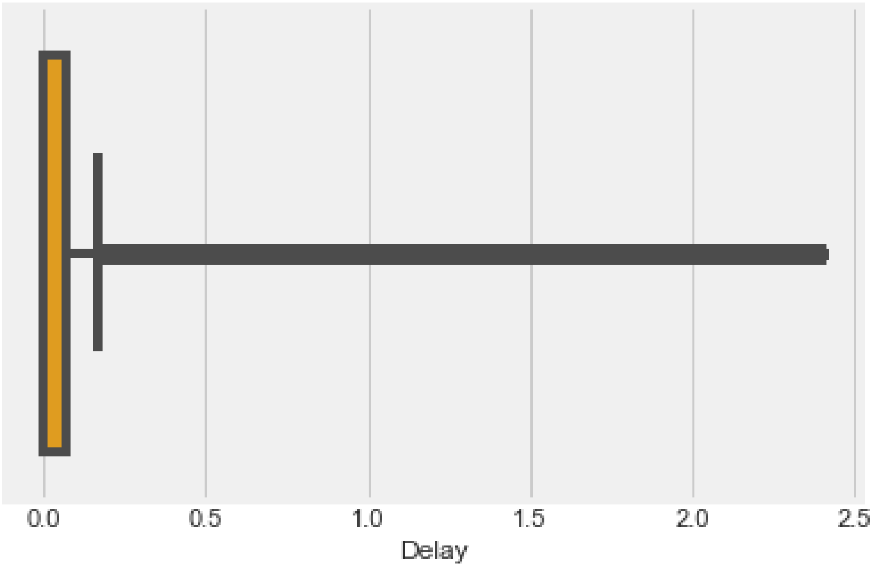
- **Encoding**: encode Departure Airport and Destination Airport columns with LabelEncoder.

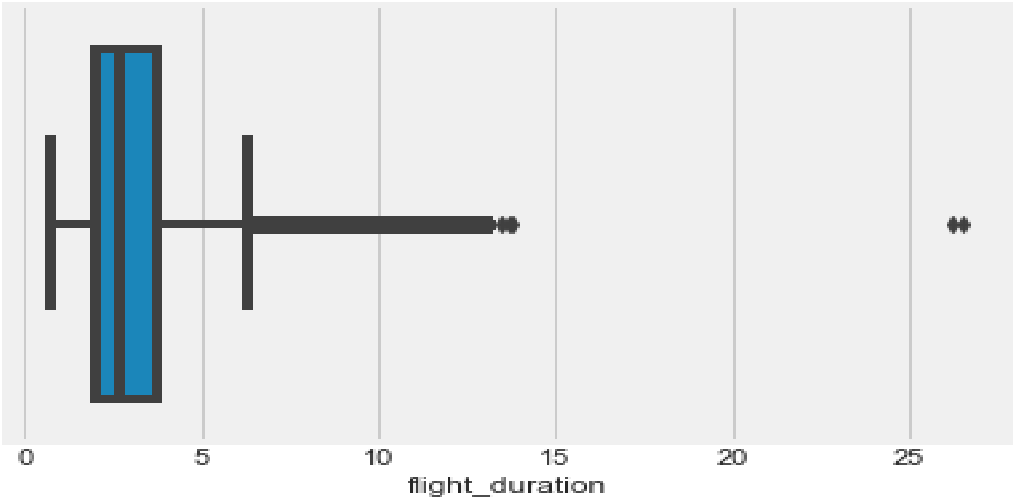
- **PCA**



4. Outlier Detection & Removal :

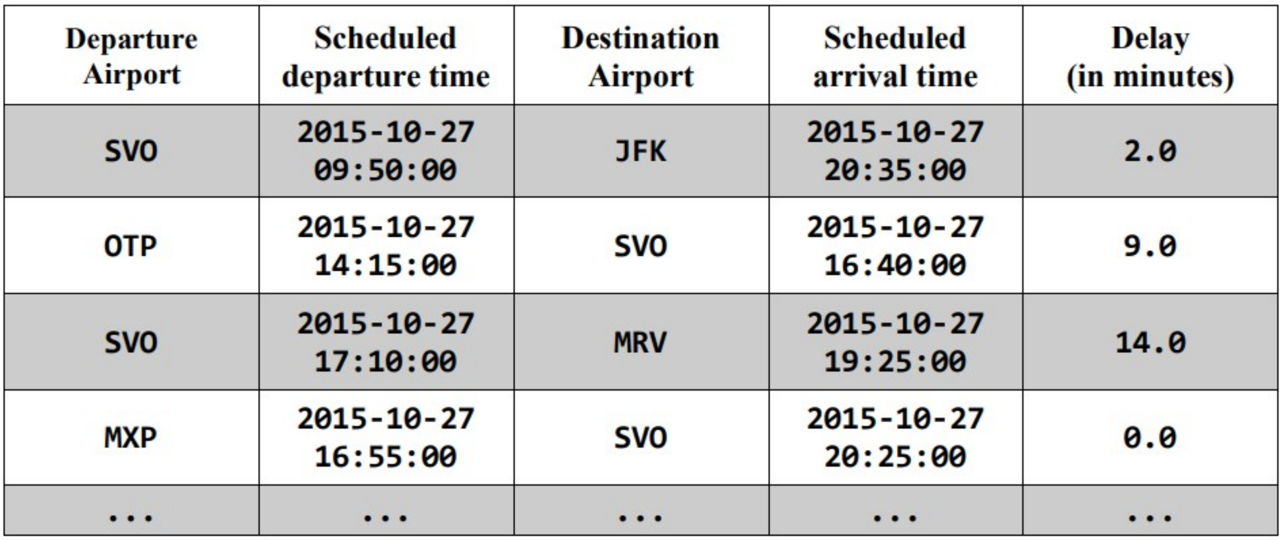
After analyzing, we found that flight duration and delay have some outlier values as the following graphs show. After that we removed outliers.





Dataset

Obtained the data from Innopolis University partner company analyzing flights delays. Each entry in the dataset file corresponds to a flight and the data was recorded over a period of 4 years. These flights are described according to 5 variables. A sneck peek of the dataset can be seen in the table below:



Departure Airport	Scheduled departure time	Destination Airport	Scheduled arrival time	Delay (in minutes)
SVO	2015-10-27 09:50:00	JFK	2015-10-27 20:35:00	2.0
OTP	2015-10-27 14:15:00	SVO	2015-10-27 16:40:00	9.0
SVO	2015-10-27 17:10:00	MRV	2015-10-27 19:25:00	14.0
MXP	2015-10-27 16:55:00	SVO	2015-10-27 20:25:00	0.0
...	...	...	...	...

The description of the 5 variables describing each flight are:

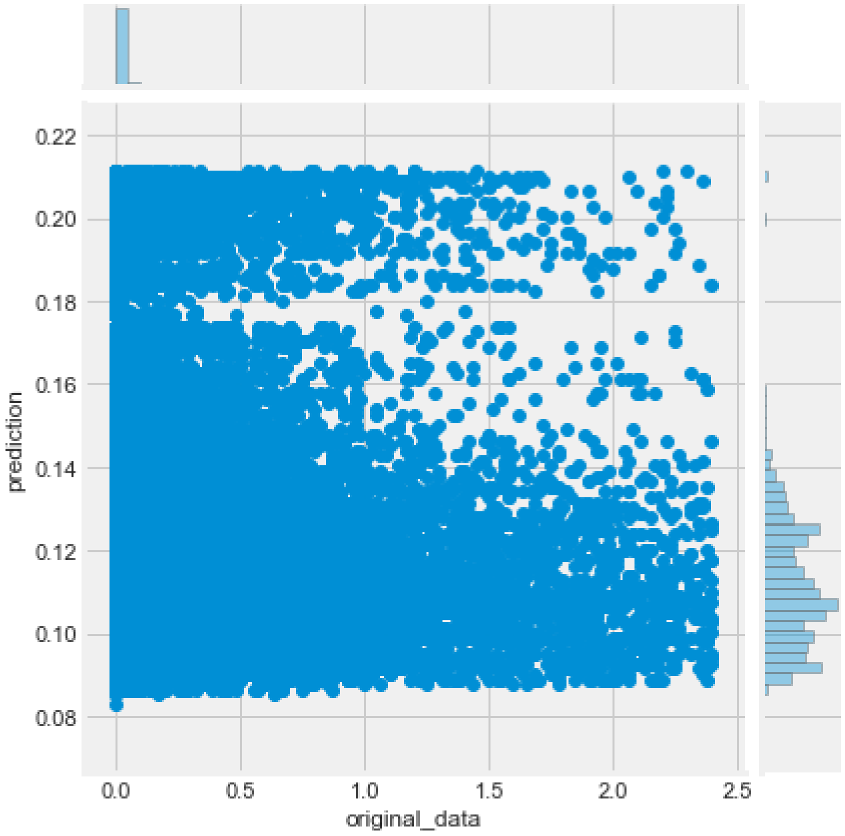
- Departure Airport : Name of the airport where the flight departed. The name is given as airport international code.
- Scheduled departure time : Time scheduled for the flight take-off from origin airport;
- Destination Airport: Flight destination airport. The name is given airport international code;
- Scheduled arrival time : Time scheduled for the flight touch-down at the destination airport;
- Delay (in minutes) : Flight delay in minutes;

Modeling

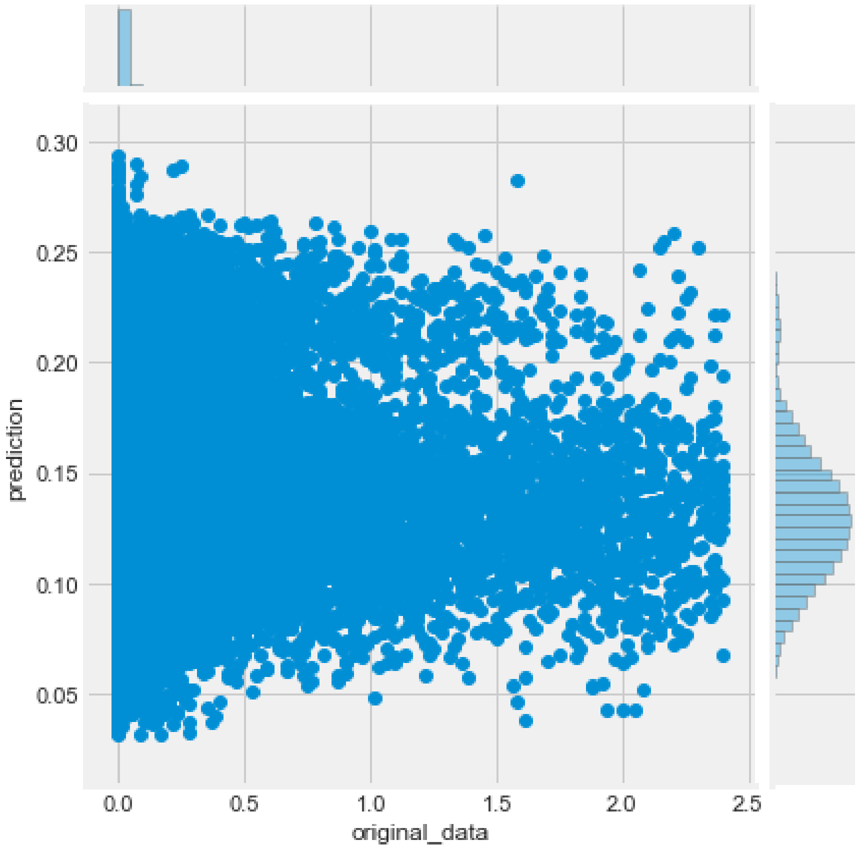
We used many various models to predict flight delays and have a high performance. In the following table the performance for each model.

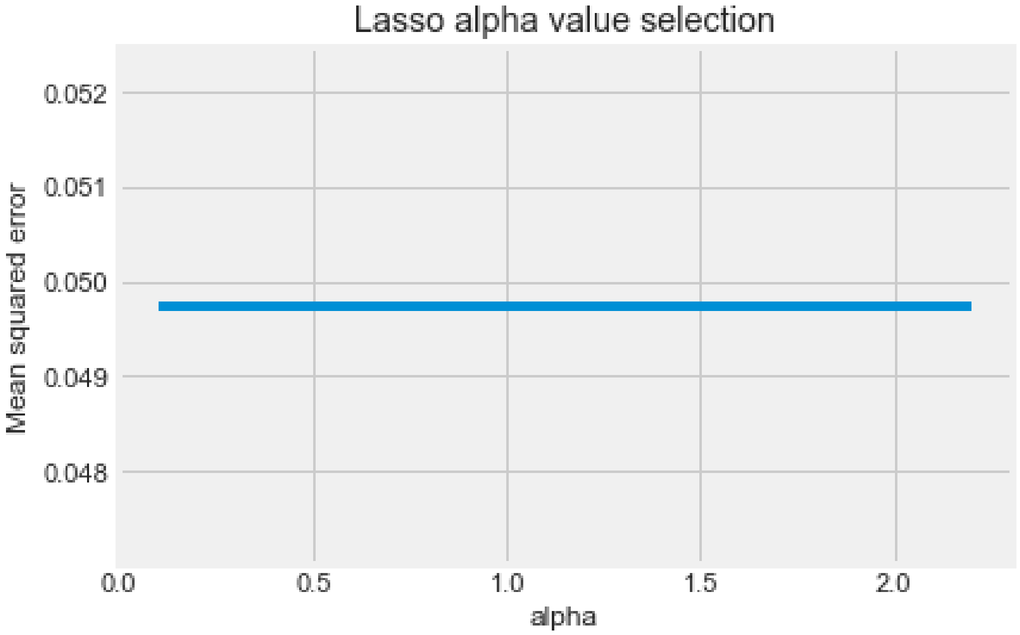
Simple linear regression	**Multiple linear regression**	**Simple polynomial regression**	**Multiple polynomial regression**	**Lasso**
Test MSE : 0.04947	Test MSE : 0.051625	Test MSE : 0.139052 Train MSE: 0.076667	Test MSE : 0.04528 Train MSE: 0.07544	Test MSE : 0.04975 Train MSE: 0.0772
Test MAE : 0.13913	Test MAE : 0.150595	Test MAE : 0.13913 Train MAE: 0.153534	Test MAE : 0.090147 Train MAE: 0.15034	Test MAE : 0.14059 Train MAE: 0.15523
Test RMSE : 0.2224	Test RMSE : 0.227212	Test RMSE : 0.22231 Train RMSE: 0.27688	Test RMSE : 0.21281 Train RMSE: 0.27467	Test RMSE : 0.22305 Train RMSE: 0.2779

Simple linear Regression



Multiple linear Regression





Conclusion:

Found that simple linear regression is better than simple ploynomial regression, but in other hand multiple ploynomial regression is better than multiple linear regression. I see all models have no overfitting but have some sort of underfitting. I recommend to grab more data to have high performance.