Machine Learning (F21)
Innopolis University, 2021
Assignment 1

Task: Flight Delay Forecasting
Due Date: 24th September 2021
Submission Format: GitHub repository link and report (.pdf)
Data: Dataset

1. Task Description

In this assignment, you are going to be solving the task of flight delay estimation using machine learning. The goals are:

- Preprocess, visualize and split the dataset
- Select 2 or more appropriate machine learning models to estimate the flight delays (i.e. Linear regression, polynomial regression, etc.)
- Use at least 1 machine learning model with regularization to estimate flight delay.
- Compare the selected machine learning models performance using the appropriate evaluation metrics.
- Describe which model is better based on the test and training set performance. Does the model overfit? Underfit?
- Outlier detection and removal

2. Dataset

The Dataset comes 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:

Variable name	Description
`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 as airport international code
`Scheduled arrival time`	Time scheduled for the flight touch-down at the destination airport
`Delay (in minutes)`	Flight delay in minutes

3. Data Preprocessing and Visualization

The simplest way to convert the string representation into the machine-readable format is to substitute the characters with a unique integer identifier. This can be easily achieved by using Label encoder `LabelEncoder` from sklearn. You are free to apply other ways for handling categorical string data and missing values. **Use encoder of your choice.**

For feature engineering, more features can be extracted from the timestamps (i.e year, month, day, day of the week). These time features can be easily be extracted using pandas `pandas.Series.dt`. For data visualization on 2D plane dimension reduction methods such as **PCA** can be used. The simplest way is to select one meaningful feature and plot it against the target variable `Delay (in minutes)`. We suggest plotting the flight duration which is the time difference between departure and arrival.

The data should be split to train and test. The data is split based on Scheduled departure time. The train data is all the data from year **2015** till **2017**. All the data samples collected in year **2018** are to be used as testing set.

4. Outlier Detection & Removal

when preparing datasets for machine learning models, it is really important to detect all the outliers and either get rid of them or analyze them to know why you had them there in the first place. In training machine learning models (especially supervised models), outliers can deceive the training process resulting in prolonged training times, or lead to the development of less precise models.

Outliers are not easily recognizable in the data collection stage however they can be detected in the analysis stage. There exist several ways for detecting outliers (i.e visualizing the data and spotting the data points diverting from the majority of data). To test if there exist outliers in the flight delay dataset, sample one-month data and apply the outlier detection method of your choice. For more information about the different approaches for outlier detection see the references section.

5. Machine learning models

For estimating flight delay time, you will need to select the appropriate machine learning algorithm. From the Course of machine learning, you have studied several machine learning algorithms (i.e linear regression, logistic regression, polynomial regression, etc.). If you decide to use an algorithm taking one variable as input, `flight duration` should be used as an independent (predictor) variable. A minimum of 3 machine learning algorithms should be used for the flight delay estimation. One of the algorithms should have regularization.

6. Performance Measurement

To measure the performance of the selected models there, exist a number of metrics studied in the course of machine learning (i.e. MSE, precision, recall, RMSE, F1-score, R^2 and weighted F1-score).

7. Report & Source Code

After performing the comparison of the machine learning models, the results should be presented in a form of a report. The implementation should be in python. The implementation repository should be available in GitHub or GitLab. Your repository should contain the following:

1. Main python script (jupyter notebook files can be included in a separate repository folder)
2. Readme file (i.e. how to run the main script)
3. Documentation (code documentation and Readme)

Your report should contain the following:

1. Motivation, explanation of what a reader should expect from your report
2. Brief task definition and data description
3. If you use an alternative data input format, explain it
4. Comparison of 3 selected models. Describe which model is better based on the test and training set performance. Does the model overfit? Underfit?
5. Use graphs and tables to document the results of your experiments

The report should be submitted in PDF format.

Resources

- Ways to Detect and Remove the Outliers
- How to Remove Outliers for Machine Learning
- 5 Ways to Detect Outliers/Anomalies That Every Data Scientist Should Know)
- Sklearn: Novelty and Outlier Detection