ML ASSIGNMENT 1 REPORT

SOLVING THE TASK OF FLIGHT DELAY ESTIMATION USING MACHINE LEARNING

Umaraliev Odilbek
Data Science
Innopolis University
o.umaraliev@innopolis.university

MOTIVATION

I fell in love with Machine Learning during my Master degree in Innopolis University. This is my first assignment from an ML course. It was very interesting and very effective to apply studied algorithms to real task. I already got this course in my 3rd year bachelor's degree. We got this course only one time for a week. But despite this I decided to make a Diploma project exactly in this field and I made a 'Machine Learning model to predict the price of the apartments in Kyrgyzstan and building web application'. ML is very popular and interesting field. It includes AI, which is in demand. Besides, ML and AI jobs have jumped by almost 75% over the past four years and are poised to keep growing.

BRIEF TASK DEFINITION

In this task, I used some machine learning algorithms from the course of Machine Learning to solve the task of flight delay estimation. The purpose is not to obtain the best possible prediction but rather to emphasize on the various steps needed to build a model. I show how to build linear and polynomial models for univariate or multivariate regressions and also, I give some insight on the reason why regularisation helps us in developing models that generalize well.

This task consists of:

- Data Preprocessing
- Visualization
- Outlier Detection & Removal
- Applying ML Algorithms (models)
    - Simple Linear Regression (Regularization)

    - Multiple Linear Regression
    - Polynomial Regression
- Performance Measurement

DATA DESCRIPTION

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. The variables are: Departure Airport, Scheduled departure time, Destination Airport, Scheduled arrival Time and Delay.

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

Figure 1: Dataset

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	Flight delay in minutes

Table 1: Variable Description

In future, from Scheduled departure time and Scheduled departure time columns extracted timestamps. These features are splitted to: Year, Month, Day, Weekday and Time.

DEP Year	DEP Month	DEP Day	DEP weekday	DEP Time	ARR Date	ARR Year	ARR Month	ARR Day	ARR weekday	ARR Time
2015	10	27	1	07:40:00	2015-10-27	2015	10	27	1	20:45:00
2015	10	27	1	09:50:00	2015-10-27	2015	10	27	1	20:35:00
2015	10	27	1	10:45:00	2015-10-27	2015	10	27	1	23:35:00
2015	10	27	1	12:30:00	2015-10-28	2015	10	28	2	01:20:00
2015	10	27	1	14:15:00	2015-10-27	2015	10	27	1	16:40:00

Figure 2: Splitted Timestamps

Variable Flight Duration Time is calculated and assigned to a new variable.

COMPARISON OF MODELS

Simple Linear Regression

Wikipedia: In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

Before using Linear Regression model, we need to split our data to the train data (all the data from year 2015 till 2017) and to the testing set (all the data samples collected in year 2018)

```
In [40]: # IN OUR DATASET THERE IS FLIGHT DATES IN THESE YEARS 2015,2016,2017,2018
         # WE TAKE ALL YEARS WHICH ARE SMALLER THAT 2018 AS train_data
         # AND 2018 YEAR AS test_data

         train_data = new_data[new_data['DEP Year']<=2017]
         test_data = new_data[new_data['DEP Year']>2017]
```

Figure 3: Splitting data to train and test.

After splitting data to the train and test data, we split all columns. But, to apply Simple Linear Regression we need for only 2 columns. In our case, 'Flight Duration' column used as predictor and the Delay column as target.

Performance Measuring of Simple Linear Regression:

Mean Absolute Error (MAE)	10.348050804475623
Mean Squared Error (MSE)	354.4951688931716
Root Mean Squared Error (RMSE)	18.828042088681755
R2 Score	-0.042731738870564806

Table 2: Performance Measuring of Simple Linear Regression

Plotting:

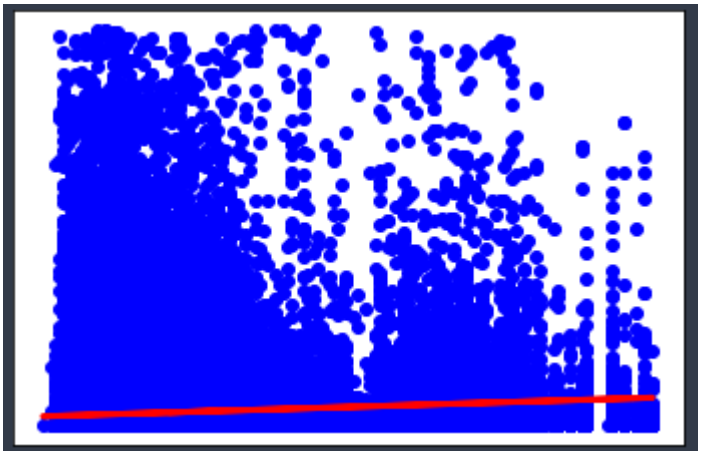

Figure 4: Plotting (Simple Linear Regression)

Multiple Linear Regression

The difference between Simple and Multiple Linear Regression is, in Multiple LR we use more then one predictor variables. So, to apply this model we can take all variables as predictors expect 'Delay' column, this columns will be our target variable.

Performance Measuring of Multiple Linear Regression:

Mean Absolute Error (MAE)	10.50387210597826
Mean Squared Error (MSE)	357.36859842404186
Root Mean Squared Error (RMSE)	18.90419525988985
R2 Score	-0.05118380376217213

Table 3: Performance Measuring of Multiple Linear Regression

Remark: To reduce dimensionality ***PCA*** can be used.

Principal Component Analysis (PCA)

Wikipedia: Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

Polynomial Regression

Wikipedia: Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x).

We use only one column as a predictor variable to apply the Polynomial Regression model. This predictor is the same as the previous model. So, our independent variable is 'Flight Duration Time' and the dependent variable is 'Delay'. For choosing the best model, we can change the degree of Polynomial Regression model. I tried to use [1,3,15] degrees and compare performance.

Degree 1:

Mean Absolute Error (MAE)	10.422671156964151
Mean Squared Error (MSE)	476.16694217848874
Root Mean Squared Error (RMSE)	21.82124978497998
R2 Score	0.005691420827001692

Table 4: Performance Polynomial Regression (Degree 1)

Plotting Degree 1:

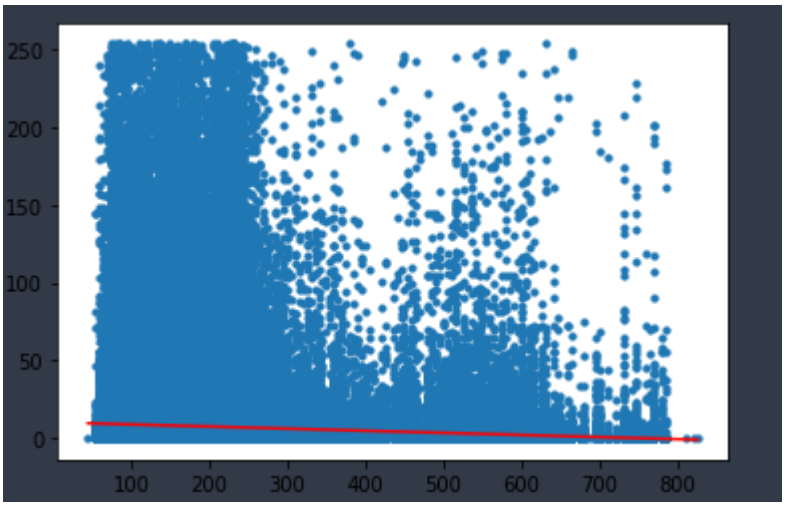

Figure 5: Plotting with Degree 1

Degree 3:

Mean Absolute Error (MAE)	10.394214810573805
Mean Squared Error (MSE)	475.46024266295694
Root Mean Squared Error (RMSE)	21.80505085210665
R2 Score	0.007167116279474417

Table 5: Performance Polynomial Regression (Degree 3)

Plotting Degree 3:

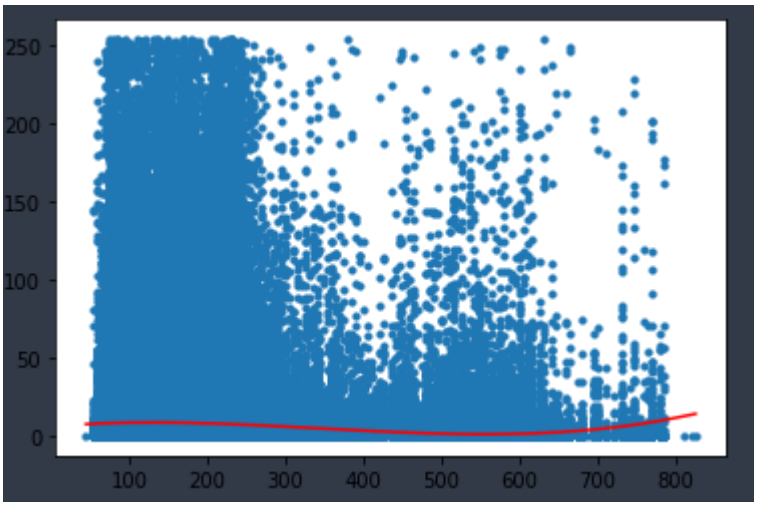

Figure 6: Plotting with Degree 3

Degree 15:

Mean Absolute Error (MAE)	10.4521098915151
Mean Squared Error (MSE)	476.7505407247248
Root Mean Squared Error (RMSE)	21.834617943181986
R2 Score	0.004472778813214395

Table 6: Performance Polynomial Regression (Degree 15)

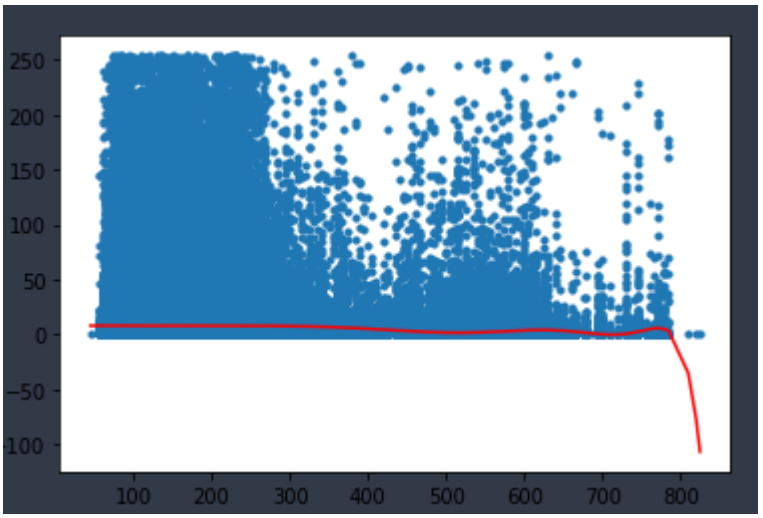

Figure 7: Plotting with Degree 15

Regularization

Wikipedia. In mathematics, statistics, finance, computer science, particularly in machine learning and inverse problems, regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting. Regularization can be applied to objective functions in ill-posed optimization problems.
***Lasso regression** is a regularization technique. It is **used over regression methods for a more accurate prediction**. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).*

After measuring performance lasso model I get:

Mean Absolute Error (MAE)	10.401714573194239
Mean Squared Error (MSE)	475.2080869321574
Root Mean Squared Error (RMSE)	21.799268036614382
R2 Score	0.006946877548726693

Comparing all model performances

	Simple LR	*Multiple LR*	*Poly LR (deg=1)*	*Poly LR (deg=3)*	*Poly LR (deg=15)*	*Regularization (Lasso)*
MAE	10.3480	10.5038	10.4226	10.3942	10.45210	10.4017
MSE	354.4951	357.3685	476.166	475.460	476.7505	475.208
RMSE	18.8280	18.9041	21.8212	21.8050	21.8346	21.7992
R2	-0.0427	-0.0511	0.0056	0.0071	0.0044	0.0069

Link to GitHub Repository: https://github.com/Odilbek99/ML_Assignment_1/