**Predicting Flight Delays**

**By:**

Faisal Saimeh, Data Science Trainee

**September 10,2021**

**Prepared For:**

National ICT Upskilling Program, Al-Hussain Technical University

A picture containing text, outdoor, scoreboard

Description automatically generated

Table of Contents

Summery ................................................................................1

Introduction.............................................................................1

Methodology...........................................................................2

Result......................................................................................3

Discussion...............................................................................5

Conclusion..............................................................................5

Project Code...........................................................................5

References..............................................................................6

**Summery**

The main goal of the project is to predict flight delays caused by various factors. Flight delays will have a negative impact on passengers, the aviation industry, and airport authorities, mainly economic. Therefore, these factors show how necessary and relevant it is to predict delays regardless of the airline's grid range. Perform predictive analysis, which includes a series of statistical techniques from supervised machine learning and data to make predictions. With the help of regression analysis used by Python, this prediction will be useful It helps to analyze the performance of various airlines and airports in detail, and then make an informed decision. Additionally, in addition to passenger-related assessments.

1. **Introduction**

This project was created by Fabien Daniel and shared at Kaggle. I had my contribution in this project by achieving the best result in data prediction, to have a better result the coefficient of determination denoted by R².

* 1. **Purpose**

This project develops a model aimed at predicting flight delays at take-off. The purpose is to obtain the best possible prediction by emphasizing the various steps needed to build such a model.

* 1. **Scope**

This report details the flight data problem and shows who are the most 14 companies have a delay in their flight time from their origin airports, I am going to focus on the max delays company, who have the highest percentage delay comparing with other companies.

* 1. **Problem**

The elapsed time will be predicted to know the flight delay. In data science, it is a Regression problem. It is a type of problem in which the output variable is a real or continuous value.

1. **Methodology**
   1. **The main aspects of Python covered throughout the project are:**

Python is the language used in this project, and some libraries, have been used belong to Python. First, for data manipulation, I used Pandas and NumPy. Then, for modeling data, I called Sklearn. It is a machine learning library. Finally, for visualization, I used Matplotlib and Seaborn libraries.

* 1. **This project consists of six main steps:**

Firstly, we need to load the libraries that we need for this project. Secondly, uploading the flights’ file that contains the data and views the data. Thirdly, cleaning the data by re-organizing the data frame and dealing with missing data. Fourthly, uploading the airlines' files to know the max, min, and mean delays in flights by doing basic statistical descriptions of the airlines. Fifthly, splitting data into training data and testing data, then encoding categorical data. Finally, predicting flights delay using six machine learning models, Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, KNN Regression, and Multiple Linear Regression.

1. **Results**

I added three machine models to the project Lasso, KNN, and Multiple Linear Regression. The result split into two sections before my editing on code and after, to clarify more, these tables and graphs will clarify the improvements. These results tests were taken on American Airlines.

* 1. **Tables**

Table 3.1.1: Before Editing Table 3.1.2: After Editing

|  |  |
| --- | --- |
| Machine Learning Model | R² Score |
| Linear Regression | 0.04396844305486225 |
| Polynomial Regression | 0.4260746917181413 |
| Ridge Regression | 0.041546895480597534 |
| Lasso Regression | 0.09030034320259406 |
| KNN Regression | -0.46432775846604235 |
| Multiple Linear Regression | 0.04396844305486225 |

|  |  |
| --- | --- |
| Machine Learning Model | R² Score |
| Linear Regression | 0.9999999999996676 |
| Polynomial Regression | 0.9995335553782245 |
| Ridge Regression | 0.9865625116926238 |
| Lasso Regression | 0.9999955060268179 |
| KNN Regression | 0.9932590777105758 |
| Multiple Linear Regression | 0.9999999999996676 |

* 1. **Graphs**

In graphs, I focused on the worst result of two machine models. First, the Ridge model already existed. Secondly, the KNN model I had added.

**Ridge Model KNN Model**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 3.2.3: KNN Model Before Editing

Figure 3.2.1: Ridge Model Before Editing

Chart, scatter chart

Description automatically generatedChart

Description automatically generated

Figure 3.2.4: KNN Model After Editing

Figure 3.2.2: Ridge Model After Editing

1. **Discussion**

We see that there are excellent improvements in results from the previous tests. Firstly, I kept the format of the data. I didn't change it like the last test. Secondly, using the One Hot Encoder procedure appropriately gives a cleaner and understood data for the machine models. Because of these changes, the results of the models have increased very high, R² score has better results from the previous tests.

1. **Conclusion**

In the results, before editing the data procedure, the Polynomial Regression model was the best result with a 0.4260 score. The rest were so bad, particularly KNN Regression with a -0.4643 score. After editing data procedure models' scores improve, the Linear and Multi Linear Regression models were the best results with 0.9999 score and the lowest score was to Ridge Regression model with 0.9865. The improvements were magnificent, on the KNN Regression jump from -0.4643 to 0.9932 score. The conclusion, before judging the models. We need to make sure of the way of data how was treated before modeling.

1. **Project Code**

In this section, you will find my GitHub links, repository and project links.

Capstone Project repository:

<https://github.com/FaisalSaymeh/Capstone-Project-Predicting-Flight-Delays>

Project Code:

<https://github.com/FaisalSaymeh/Capstone-Project-Predicting-Flight-Delays/blob/master/Capstone_Project_Predicting_Flight_Delays_January_.ipynb>

# **References**

|  |  |
| --- | --- |
| [1] | F. Daniel, "Kaggle.com," September 2017. [Online]. Available: https://www.kaggle.com/fabiendaniel/predicting-flight-delays-tutoria. |
| [2] | F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, . J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: Machine Learning in Python, M. Braun, Ed., 2011, pp. 2825--2830. |
| [3] | M. Badhe, "Multiple Linear Regression Using Python and Scikit-learn," May 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/. |