**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning |
| **Assessment Title:** | CA1 Project |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Zygimantas Jakubauskas  Miroslava Slavikova |
| **Student Number:** | Sba22342 |
| **Assessment Due Date:** | 11/Oct/2023 |
| **Date of Submission:** |  |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Machine Learning CA1 Project**

**Table of Contents**

1. Introduction
2. Selection of the dataset
3. Data Exploration and Preparation
   1. Exploratory data analysis
   2. Data Preparation Tasks
4. Feature Engineering

4.1 Scaling

1. Visualisations
2. Conclusion
3. References
4. **Introduction**

Aviation industry is a difficult business. Airlines constantly inventing new ways of keeping customers happy trying to make them loyal …… to be continued.

1. **Selection of the dataset**

The theme we chose for our analysis is Transport and dataset is air\_data.csv. Dataset contains data collected from passengers sharing their experiences after the flight. Dataset contains personal data, such as Age, Loyalty status or Gender, and grades given by each passenger evaluating aspects like Onboard Wi-Fi, Onboard Food, Ease of Online Booking, and others. Original dataset has 25 variables and 129880 observations.

The last variable of the dataset is Satisfaction and it contains 2 values – Satisfied and Neutral/Dissatisfied. Our task is to build Machine Learning model, thar could predict customers overall satisfaction after completing the journey. We will analyse what aspects of flight experience have most influence on the final decision, and what sectors of the service airlines must improve to keep loyal customer base and increase it.

1. **Data Exploration and Preparation**

We started by checking duplicates and missing values. Dataset contains no duplicates, but variable Arrival Delay In Minutes contains 393 missing values. We handled these missing values by using SimpleImputer from sklearn and filling in with median values.

Dataset contains two insignificant variables – Unnamed (the row number) and Id (customer’s identifier). These variabes don’t contribute to the dataset and we made the decision to remove them.

1. **Feature Engineering**

**4.1. Scaling**

Once all the categorical data is encoded, there are 4 columns of continuous variables left. “Age”, “Flight Distance”, “Departure Delay in Minutes” and “Arrival Delay in Minutes”. Boxplots show that the two Delay columns contain scarse data. For this reason to scale them we are using L2 normaliser. “Flight Distance” is skewed, skewness is greater than 1. Therefore, to scale “Flight Distance” we are using MinMax scaler. “Age” is distributed normally, skewness is very close to zero and doesn’t need to be scaled.

**“Draft - (Intro/ Body/ Conclusion)”**

**Personal Reflective Report of Miroslava Slavikova**

Intro ???

**Encoding**

Once we explored the data, I started to replace text with numerical values where “satisfied” was assigned 1 and “dissatisfied” was assigned 0. Additionally, we needed to remove columns with “Unnamed” and “ID” columns. We won’t be able to analyse un-named values, if we don’t know what they are and personal IDs are sensitive personal data and subject to an additional protection under GDPR act. In summary, this helped to remove those columns in order to clean up our data and focus only on relevant information.

**Checking for missing values** … “Arrival Delay in Minutes”

To analyse the most important features in data set, I’ve created a code that assigns each feature importance score. I assigned the column names of the training data to the “importance” column and with **fi.head(20)** command I can request to display the top 20 rows that represent the features with highest score of importance. This command sorts the Data Frame based on the “importance” column in descending order where the most important features come first and on top of the graph.

**Train/Test**

**ML models**

**10,20,30 splits**

Timeline and/or Conclusion

With a full time job and family responsibilities, time is always scarce. Dedication, weekends, communication via messenger, GitHub, in person

**References:**

<https://scikit-learn.org/>

<https://www.kaggle.com/>

<https://realpython.com/>

<https://towardsdatascience.com/predictive-analytics-on-customer-behavior-with-support-vector-machines-svm-7e68fd2be610>