**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

2.1 Analysis of Target variable

1. Data Exploration and Preparation
   1. Exploratory data analysis
   2. Data Preparation Tasks
2. Feature Engineering

4.1. Encoding

4.1. Scaling

5. Discussion and interpretation of the results

6. Reflective Journal by Miroslava Slavikova

7.Reflective Journal by Zygimantas Jakubauskas

Conclusion

References

1. **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.

* 1. **Analysis of Target variable**

The last variable of the dataset is “Satisfaction” and it contains 2 values – Satisfied and Neutral/Dissatisfied. This variable is our Target variable. 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.

“Satisfaction” has two values – Satisfied and Neutral and Dissatisfied, which for machine learning purposes we encoded as 0 and 1.

A blue and orange pie chart

Description automatically generated

1. *Figure1. Pie chart of Customers satisfaction distribution*
2. **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. Encoding**

Reduced dataset contains 4 continuous and 19 categorical variables (1 of these is our target variable). Target variable is is encoded by giving labels 0 and 1. The rest of categorical variables is encoded using pandas get\_dummies function.

**4.2. 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 sparse 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.

1. **Discussion and interpretation of the results**
2. **Reflective Journal by Miroslava Slavikova**

Draft - (Intro/ Body/ Conclusion)

Personal Reflective Report of Miroslava Slavikova

At our initial stage of planning, we started our project by discussing it and asking various questions. What would be a good topic, how to find relevant dataset, what are we going to deliver and what would be the goals and expectations. We were in contact couple of times a week to check our progress and discuss improvements where needed and we aligned on the next steps to move forward. Personally, I am new to this subject and it was challenging to stay focused on the key objectives. I did several visualisations that we decided to drop as they were not in line with our main objective.

Additionally, with introduction of GitHub version-controlled system that we use to record our progress and track our collaboration timeline, I’ve also taken some time to understand how to use the system effectively and how to share files correctly.

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 ID is a 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 assign the column names of the training data to the “importance” column and with **fi.head(20)** I can request to display the top 20 rows that represent the features with highest score of importance. This command sorts the dafaframe 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

Conclusion

Admittedly, it was a challenging task for me given my current level of experience, but generally speaking, I really enjoyed learning and attempting to understand such a complex subject of study as Machine Learning. Furthermore, Zygimantas is an excellent team player, very focused and reliable. He was very supportive and understanding during this journey and I could not ask for a better partner on this project.

1. **Reflective Journal by Zygimantas Jakubauskas**

I personally enjoyed working with this air\_data.csv dataset. Plane travelling experience is something I can relate to, and I can see this type of data being used in real world. It gave me knowledge what airlines do to improve their customer service, and how little details can affect overall picture. We were predicting satisfaction of customers after taking the journey, but with this data although we can predict what can make customer loyal (although we didn’t analyse this on this occasion), something that is crucial for every airline.

As for practical part of the assignment it’s difficult to highlight any particular area, because a lot of work was done in collaboration with my partner, sharing the ideas, trying different approaches and solutions to the problems. I personally tried to analyse deeply every aspect of the assessment.

1. **Conclusion**