**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  Sba22382 |
| **Assessment Due Date:** | 11/Oct/2023 |
| **Date of Submission:** | 26/Nov/2023 |

**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**

Introduction

1. Selection of the dataset

2. Exploratory data analysis

3. Data Preparation

4. Machine Learning Models

5. Discussion and interpretation of the results

6. Reflective Journal by Miroslava Slavikova

7. Reflective Journal by Zygimantas Jakubauskas

Conclusion

References

**Introduction**

The aviation industry is undoubtedly a challenging business, airlines are not only facing fierce competitors but complex regulations by the government regulatory agencies, with increasing environmental responsibility they are under pressure to adopt sustainable practices like carbon offset programs, managing the fluctuating fuel prices and there are the ever-rising passenger’s expectations just to mention to few. Airlines are constantly inventing new ways of keeping customers happy while trying to foster their loyalty. We are going to look at some strategies and considerations to analyse the key factors impacting passenger’s satisfaction during their air travel.

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. **Exploratory data analysis**

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 variables don’t contribute to the dataset and we made the decision to remove them.

With these columns removed there are 4 important continuous variables left – “Age”, “Flight Distance”, “Arrival Delay in Minutes” and “Departure Delay in Minutes”. We analysed closer these values to make the decision regarding scaling.

A graph of a number of people

Description automatically generated

*Figure1.Displot of variable “Age”*

Distribution of variable “Age” is not exactly normal but it’s not skewed either. Skewness is -0.003606211745335888 meaning that “Age” doesn’t require scaling.

Skewness of “Arrival Delay in Minutes” and “Departure Delay in Minutes” are respectively 6.670124610533305 and 6.82198031017346 showing large positive skews. These variables require scaling.

Variable’s “Flight Distance” analysis shows big gap between minimum and maximum values and large standard deviation. Boxplot of “Flight Distance” clearly shows few outliers. “Flight Distance” also requires scaling.

A blue rectangular object with black lines

Description automatically generated

*Figure2. Boxplot of variable “Flight Distance”*

1. **Data Preparation**

**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, that 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

*Figure3. Pie chart of Customers satisfaction distribution*

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

**Imputing missing values.** There are 393 missing values in dataset represented as “nan”. We imputed them using SimpleImputer from sklearn library using “median” as the strategy.

**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 close to normal distribution, skewness is very close to zero and doesn’t need to be scaled.

1. **Machine Learning Models**

To find the best performing machine learning models we run and compared the results of 5 models – Naïve Bayes, Random Forest, Decision Tree, Linear Regression and Support Vector Machine, using test size of 20% for all of them. SVM produced results with accuracy of 92.4% but is very slow to run and we dropped it for further analysis.

A close-up of a number

Description automatically generated

The best results, even including SVM were produced by Random Forest and Decision Tree. The worst results are produced by Naïve Bayes model.

A screenshot of a white background

Description automatically generated

*Figure4. Table of ML models Accuracies and Predictions*

For further analysis which include cross validation and 10%, 20%, 30% training splits we are using these two best performing models.

Random Forest is deeper analysed (using cross validation techniques and three different training splits) in Miroslava’s reflective journal and respectively Decision Tree is deeper analysed in Zygimantas’s reflective journal. Cross validation is especially useful to analyse smaller datasets where may not be enough data to make accurate predictions. Our dataset has almost 130000 observations and cross validation results vary little.

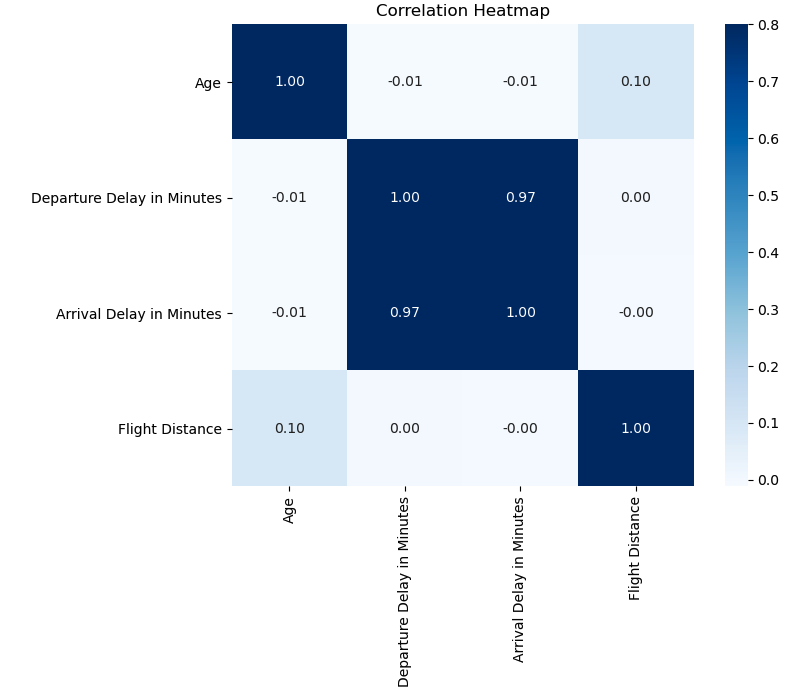
1. **Reflective Journal by Miroslava Slavikova**

At our initial stage of planning, we started our project by discussions, brainstorming and asking various questions. What would be a good topic, how to find relevant dataset, what results 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 used 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. It turned out to be a very challenging part of the work, from occasional delays due to errors on the system when pushing/pulling files to endless troubleshooting of errors. But I have tried to embrace the challenges in order to continue this learning journey knowing every obstacle is an opportunity for personal growth.

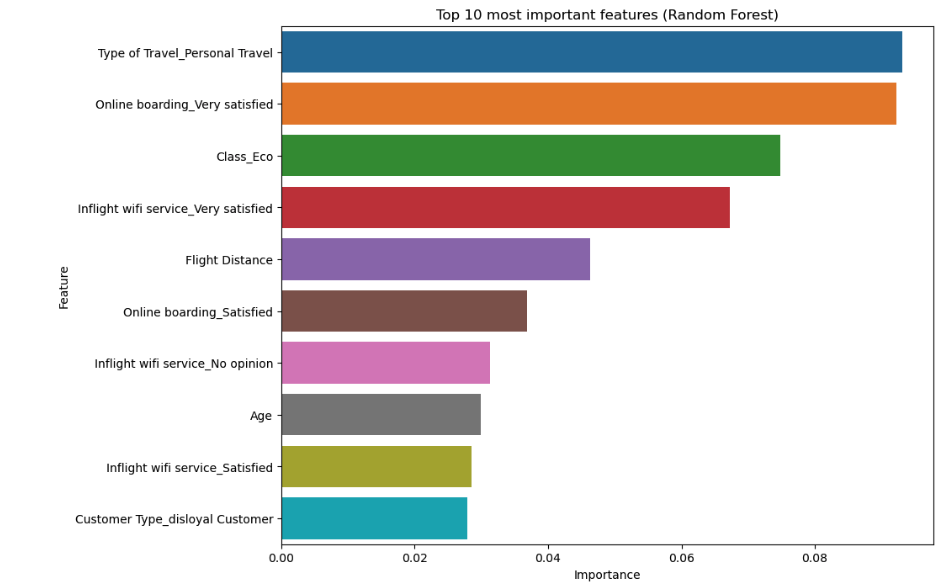
Once found and agreed on dataset, we explored the data and performed EDA. I started to replace text with numerical values where “satisfied” was assigned 1 and “dissatisfied” was assigned 0. Additionally, we needed to remove columns with “Unnamed” and “ID” columns. We won’t be able to analyse un-named 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.

The heatmap gives us visual representation of the correlation between the selected variable and colours show the strength and direction of this corelation where dark colour suggests stronger positive correlation and the nearly white colour shows us there is no correlation at all. In our graph we can see there is a strong relationship between the "Arrival Delay in Minutes" and "Departure Delay in Minutes".



*Figure5. Correlation Heatmap*

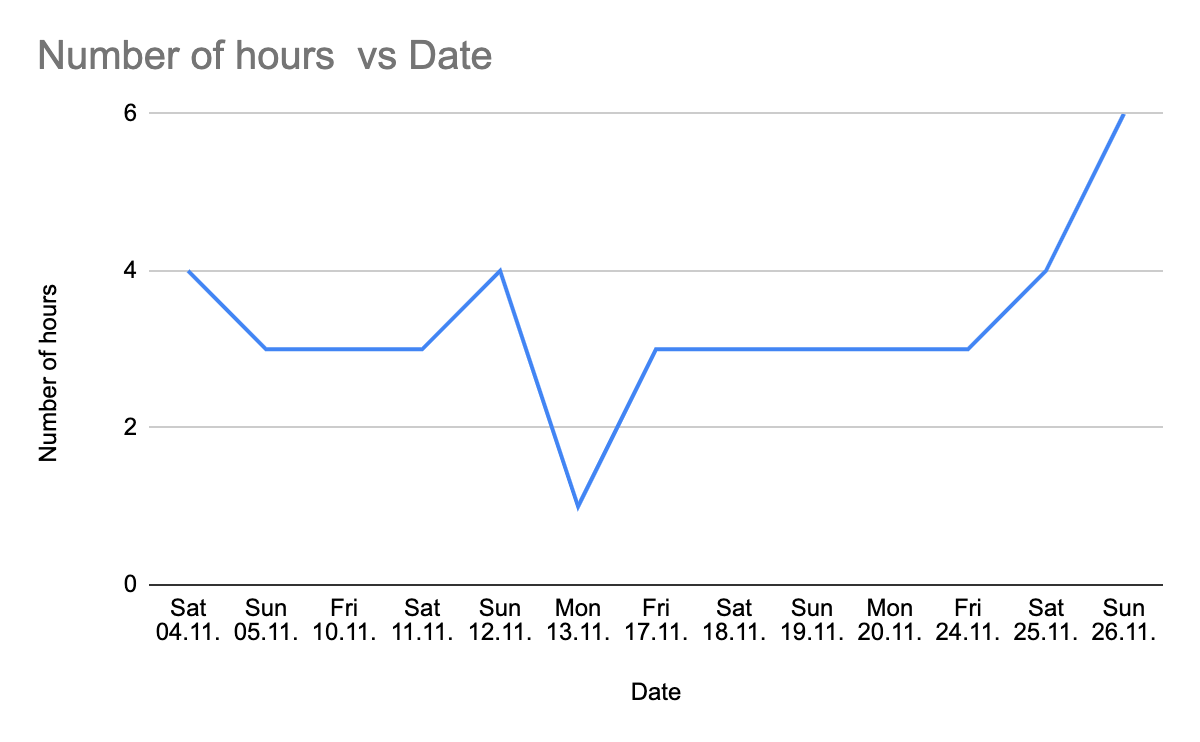
 To obtain the most important features in data set, I’ve created a code that assigns each feature importance score. I assign the column names of the training data to the “importance” column and with the command **fi. head(10)** I can request to display the top 10, or 20 rows for example, that represent 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.



*Figure6. Top 10 most important features using Random Forest model*

Admittedly, the project 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. Unfortunately, we’ve constantly faced time constraints, having busy and rather different schedules, it was a rocky start but eventually found a suitable times to work together when needed. We’ve communicated frequently and we even arranged a time to meet in person twice on Friday 17 Nov and 24 Nov. The meetings in person were insightful and helped me to stay focused and cleared any pending matters or questions.

To conclude, 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.



*Figure7. Estimated time spent on the 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 by them, but with this data we also can predict what can make customers loyal (although we didn’t analyse this on this occasion), something that is crucial for every airline. Very interesting piece of information is Importance’s Graphs, that could give deep insides into areas where airlines should focus, what aspects of their service to improve, or even what segments of passengers to target to offer their products. I could see for myself how machine learning techniques can influence marketing decisions.

This was a vast time-consuming project, with big chunk of time dedicated to trying to understand GitHub. I wish we had in-depth lesson on GitHub prior to the project as various errors, such as conflicting errors were a constant feature. Overall, all of my evenings and weekends for the last two weeks were dedicated to the assignment either working on my own or together with the partner.

A graph showing time and date

Description automatically generated

*Figure8. Estimated time spent on the project*

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 aiming to understand Machine Learning inside out. I definitely improved my Data Preparation and Machine Learning skills during this project.

As mentioned above, I’ll take a deeper look at Decision Tree model, as this was one of the better performing models outperformed only by Random Forest. Running model at three different test sizes – 0.1, 0.2 and 0.3 following results were produced:

A screenshot of a graph

Description automatically generated

Model performed best at test size 0.2, and at this test size confusion matrix’s accuracy was the highest 0.948%. This is the plot of Confusion Matrix.

A blue squares with white text

Description automatically generated

*Figure9. Decision Tree confusion matrix at test size 0.1*

As 0.2 is the best performing test size, I’m using it to calculate cross validation scores. I’m using two different folds – 10 folds and 20 folds, and scores are better at 20 folds, giving the Mean Accuracy of 0.947. The difference compared to original testing is 0.947 – 0.948 – 0.001%. This is a big dataset and test size 0.2 is almost 26000 observations and only two classes to predict and this small difference in this case is expected. Cross validation technique is more relevant and necessary for smaller datasets.

We were also looking at reducing dimensionality using PCA. However, this dataset, even when encoded only would increase to 80 columns, and variance isn’t big. Many encoded variables would only have 4 or 5 values. Scree plot looks flat, and we didn’t see much benefit of adding it to the analysis. Therefore, we took the decision not to PCA at all for further ML analysis.

There is much more that could be done with this dataset, we only scratched the surface.

1. **Conclusion**

By combining our experience and diverse skill, we were able to analyse the dataset air\_data.csv dataset and gain valuable insights about passenger’s priorities, ever-increasing expectations and how each aspect of the flight reflects on the client’s satisfaction levels. Such data are a valuable information for airlines to remain competitive, to understand passenger’s behaviour and how to improve the traveller’s experience.

Data-driven decision making has become increasingly important to understanding how to boost customer’s satisfaction, reduce operation costs and in return increase revenue in already a very thin margin industry. For example, rewards program offers very valuable data, and by analysing travellers’ behaviour, they help to create more personalised offers to loyal customers such as free tickets, business or first-class updates on certain routes, priority bookings or extra baggage allowance. Such programs can further foster loyalty and support client’s retention.

In efforts to reduce cost, machine learning can be also applied to optimise the operations aspects of the industry such as maintenance, inventory management, weather patterns or flight scheduling. Therefore, the airlines must embrace the data driven decision making process and leverage data effectively in order to meet the evolving needs of their clients and boost their bottom line at the same time.

**References and libraries:**

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

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

<https://realpython.com/>

<https://towardsdatascience.com/>

<https://www.lhsystems.com/blog-entry/why-it-critical-aviation-industry-be-data-driven-digital-age>

<https://www.analyticsvidhya.com/blog/2021/04/how-aviation-industry-uses-data-science/>

**GitHub link:**

https://github.com/ZygimantasJakubauskas/Mashine-Learning-CA1/tree/main/ML

***Word Count***

*Word count total 2129*

*of which*

*Word count Miroslava reflection 581*

*Word count Zygimantas reflection 528*