**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Higher Diploma Data Analytics for Business   * Data Preparation & Visualisation * Statistical Techniques for Data Analytics * Machine Learning |
| **Assessment Title:** | Individual / Practical |
| **Lecturer Name:** | Dr. Muhammad Iqbal  David McQuaid  Marina Soledad Iantorno |
| **Student Full Name:** | Miroslava Slavikova |
| **Student Number:** | sba22382 |
| **Assessment Due Date:** | 05 Jan 2024 |
| **Date of Submission:** | 02 Jan 2024 |

**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. |

Attrition – title……

**Subject area:**

Table of Contents

**Introduction3**

**Business Analysis and Project Plan3**

**Machine Learning Models3**

**Libraries 4**

**Dataset and Data Understanding4**

**Data Preparation4**

**Data Visualization5**

**Hotel Revenue Management7**

**Analysis of Results14**

**Future recommendations** 14

**Challenges15**

**Milestones**15

**Conclusion15**

**GitHub link16**

**Timeline link16**

**Bibliography & References16**

**Introduction**

**Business Analysis and Project Plan**

**Machine Learning Models**

I have used three different supervised machine learning models:

* Decision Tree with accuracy score of 0.810
* KNN with accuracy score of 0.800
* Logistic Regression with accuracy score of 0.775

The results provided reasonable values, albeit not outstanding. Therefore, I have tried to perform PCA that not only compresses data, but aims to improve the accuracy score. Running the KNN model thereafter improved the accuracy score to 0.842.

**Libraries**

I have downloaded all the necessary libraries, panda for data manipulation and analysis library, matplotlib.pyplot and seaborn for data visualisation, numpy for numerical computing, sklearn libraries for future scaling for ML models, PCA, encoding, training and testing sets, linear regression model for predictive modelling. I’ve also included a Folium, very useful Python Library used for visualizing geospatial data. Libraries have been imported and assigned the abbreviated formats. The abbreviated format makes recalling and use of these libraries more efficient. Lastly, I have uploaded a csv file that we renamed as a hotel.df for easy reference.

**Dataset and Data Understanding**

**Data Preparation**

* Counting the sum of missing values, the dataset shows missing data for country, agent and company
* Filling the missing values with zero
* Deleting rows with zero values for adults, children and babies in the same row, total of 180
* Identifying the target value ADR, vs is-cancelled

**Data Visualisation**

I performed several visualizations to review the data in order to have better understanding of the dataset.

1. pair plot of the first ten variables that shows us the relationships between pairs of variables (ref. to Jupyter file, code block: 16 )
2. line plot shows lead time for all bookings and there’re some outliers booking as far as 2 years in advance
3. Heatmap of correlations shows us relationships between variables and how they are correlated.

**Attrition …….**

**Let’s have a look at few categories:**

there’ll be a higher demand for the rooms due to the large number of reservations. For the off-peak season, when there’s a lower demand, hotel can create special deals to attract more traffic and foster consumer’s loyalty. creating a hotel package will additionally diversify the revenue

**Analysis of Results**

**Future Recommendations**

## **Challenges encountered**

I have encountered numerous challenges while working on this dataset. These may include looking for the suitable dataset, they were either too large or too small, I went with the bigger size of dataset. Larger volume of data provides more information for the training and might lead to better performance.

Additionally, I’ve identified a lot of missing vales, had a challenge with identifying the target value - (ADR, vs is-cancelled) and identifying the best performing model with better accuracy score.

Admittedly, the project overall was a challenging task, but generally speaking, I really enjoyed learning and attempting to understand such a complex subject of study as Data Analysis and Machine Learning.

**Milestones**

Researching while working on this project helped me to gain better understanding of visualisation techniques and how to interpret the various graphs and charts, gradually gaining confidence in this subject. I've also developed a better understanding of GitHub, including version control and how to create repositories.

**Conclusion**

**GitHub link:**

**Timeline:**

**References & Bibliography:**

<https://www.sciencedirect.com/science/article/pii/S2352340918315191>

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

https://pypi.org/project/fasteda/

<https://towardsdatascience.com/how-to-clean-your-data-in-python-8f178638b98d>

<https://www.analyticsvidhya.com/blog/2020/06/guide-geospatial-analysis-folium-python/#:~:text=Folium%20is%20a%20Python%20library,library%20for%20plotting%20interactive%20maps>.

<https://realpython.com/>

<https://www.geeksforgeeks.org/>

<https://pandas.pydata.org/>