Data Science pipeline on-premises and on-the-cloud

Sem 2, 2022

Marking	Student Id	Student Name
Ibrahim Radwan		

Part Number	Sub Mark	Out of
Part A		**60**
Part B		**40**
Total Mark	0	**100**

Part A

Items	Done Completely	Done Partially	Mark	Out of
Does the code of this part work correctly?				
Step 1: Problem formulation - The reason of using ML in this problem is defined - The business problem is defined successfully - Type of ML problem is defined	☐	☐	0	**5**
Step 2: Data Processing and visualization - Using relative paths to ensure the code is working on other machines - The code placeholders have been completed and working fine - The CSV files are combined into a single CSV - The data in the combined csv file is explored successfully - Answers to the mentioned questions are provided and the cell is turned into markdown format. - Features for the ML have been defined and encoded to pass them to the ML approaches. - The first combined CSV file is saved to the local machine.	☐	☐	0	**15**
Step 3: Model training and evaluation - Data is split into train and test sets - The baseline for the classification task is built - The baseline for the classification task is evaluated on the test data using different evaluation metrics - Answers to the mentioned questions are provided and the cell is turned into markdown format.	☐	☐	0	**10**
Step 4: Deployment - Link of the pushed repository to the GitLab is provided and is working fine - A readme.md file is created with the instructions of running this code on other machines	☐	☐	0	**10**

Step 5: Feature Engineering - Indicator variables for the holidays are defined - Weather data is downloaded and loaded into the notebook - The weather stations are mapped to airport codes - Other code parts for imputing and encoding the features are done successfully - A new baseline classifier is created successfully - The difference in performance between the two baselines are documented	☐	☐	0	**10**
Step 6: Using Tableau A dashboard is provided either as a link, ppt or as tableau notebook	☐	☐	0	**5**
Conclusion: Conclusion is deduced highlighting the model performance, challenges and the important things learned in this project.	☐	☐	0	**5**
Part A – total			0	**60**

Part B

Items	Done completely	Done Partially	Mark	Out of
Does the code of this part work correctly?				
Step 1: Prepare the environment - The notebook is created and downloaded from the AWS environment.	☐	☐	0	**5**
Step 2: Build and evaluate simple models - The data is split into train, validate and test sets - The linear estimator model is created - The model is hosted on another machine - The evaluation is done with suitable metrics	☐	☐	0	**15**
Step 3: Build and evaluate ensemble models - An ensemble model is created using the xgboost estimator - The model is hosted on another machine - The evaluation is done	☐	☐	0	**15**
Final Comments: - The final comments of using the two approaches is conducted.	☐	☐	0	**5**
Part A – total			0	**40**