جامعة بيرزيت
BIRZEIT UNIVERSITY

COMP4388: Machine Learning
Fall 2022/2023

Deadline: Monday 2 January 2023

In this project you will deal with a dataset related to customer churn. You are required to perform a series of tasks that are explained in this document.

The dataset can be found under this link:
https://www.dropbox.com/s/0su82yficgfuabn/Customer%20Churn.csv?dl=0

Details of the dataset can be found here:
https://www.dropbox.com/s/0yienihzrs0axlf/COMP4388-Data%20Set%20Information.docx?dl=0

You have to perform the following tasks:

1. Following the steps we have learnt in Exploratory Data Analysis (EDA), print the summary statistics of all attributes in the dataset.
2. Show the distribution of the class label (churn) and indicate any highlights in the distribution of the class label.
3. For each age group, draw a histogram detailing the amount of churn in each sub-group.
4. For each charge amount, draw a histogram detailing the amount of churn in each sub-group.
5. Show the details of the charge amount of customers.
6. Visualise the correlation between all features and explain them in your own words.
7. Split the dataset into training (80%) and test (20%).

As for the correlation, please detail the correlation between features and make sure to have your features that will be input of the machine learning models being clean. Based on the correlation, you have to decide which features to stay for the learning stage and which can be deleted.

Regression tasks:

1. Apply linear regression to learn the attribute "Customer Value" using all independent attributes (call this model LRM1).
2. Apply linear regression using the set of 3-most important features (from your point of view); and explain why did you use these 3 attributes (call this model LRM2).
3. Apply linear regression using the set of the most important features (based on the correlation coefficient matrix) and explain why did you use these 3 attributes (call this model LRM3).
4. Compare the performance of these models using adequate performance metrics

Classification tasks:

1. Run k-Nearest Neighbours classifier to predict churn of customers (the "Churn" feature) using the test set
2. Run Naive Bayes classifier to predict churn of customers (the "Churn" feature) using the test set
3. Run Logistic Regression classifier to predict churn of customers (the "Churn" feature) using the test set
4. Compare the performance of Logistic regression, Naive Bayes, and kNN classifiers in an appropriate results section. Compare the classification performance of the generated classification models and make sure to use the appropriate performance metrics. You should include at least the ROC/AUC score and the Confusion Matrix. Report the results in an appropriate table and explain in your own words why one model outperforms the other.

You have to turn in a softcopy of your Python code and a Word document containing the information required as specified above. The document should be on a paper-format. Please send your submissions as a reply to the message sent on Ritaj only with the files named "COMP4388-XXXXX.docx/pdf" and "COMP4388-XXXXX.py" where XXXXX is your BZU-student ID number.

If you have any questions, please feel free to contact me via Ritaj or email:

rjarrar@birzeit.edu