CS3120 Introduction to Data Science

Project Description

Deadlines Pre-processed dataset - April 10, 2022
Final Project Report - May 1, 2022

Overview

Customer churn is the loss of customers by a business for different reasons such as poor service and better price somewhere else. It is one of the most critical and challenging problems for telecommunication companies, credit card companies, cable service providers, etc. Since acquiring new customers costs more than retaining existing ones, analysing customer churn is vital for businesses.

CEO of Chatterbox Telco Pvt Ltd in the Banana Republic, Mr. William wants to analyze the customer churn in his company and decides to bring a data science engineer on board. Suppose he hired you for this job. You were provided with a dataset that contains the package type and usage details of a customer and whether they left Chatterbox or not. In particular, the dataset consists of 19 predictor variables and one target variable. Table 1 shows the description of each variable.

Variable Name	**Description**
customer_id	Customer identification number
account_length	Number of months the customer has been with the current telco provider
location_code	Customer location code
international_plan	If the customer has international plan or not
voice_mail_plan	If the customer has voice mail plan or not
number_vm_messages	Number of voice-mail messages
total_day_min	Total minutes of day calls
total_day_calls	Total number of day calls
total_day_charge	Total charge of day calls

total_eve_min	Total minutes of evening calls
total_eve_calls	Total number of evening calls
total_eve_charge	Total charge of evening calls
total_night_minutes	Total minutes of night calls
total_night_calls	Total number of night calls
total_night_charge	Total charge of night calls
total_intl_minutes	Total minutes of international calls
total_intl_calls	Total number of international calls
total_intl_charge	Total charge of international calls
customer_service_calls	Number of calls to customer service
Churn	If the customer left or not (target variable)

Your Tasks

Your goal in this project is to analyse the given dataset and derive valuable insights that would be useful for Mr. Williams to make strategic decisions to improve customer retention. For this purpose, analyse the given dataset with the knowledge you gathered in the class. Before analysing the data, as usual, pre-process the dataset.

For this project use Python 3.6 or later. Also, you can use any Python library required for data pre-processing and analysis, e.g., Pandas, Scikit-learn, Numpy, Matplotlib, Seaborn, etc.

1. Data Pre-processing

1. Load the provided CSV file into a Data Frame object named *chatterbox*.
2. Identify the data type of each variable.
3. Identify if there are data quality issues evident in the dataset and handle them.
4. Conduct necessary data transformations.

2. Data Analysis

Carry out descriptive, exploratory, and predictive data analysis on the provided dataset. You may use predictive analytics to validate the pre-processing quality.

Submissions

Data Cleaning

Create a Python script renamed as `<<Your_Student_Id>>.py` to clean the given datasets. Use Python 3.6 for this project and you can use the following libraries for data cleaning (Please use the specified versions).

- Pandas - 0.24
- Numpy - 1.17
- Scikit-learn - 0.22

The data cleaning section of the project will be evaluated automatically. When your submitted Python script is run, it should read the Train_Dataset.csv and Test_Dataset.csv files located in the same folder and write the cleaned datasets as CSV files into the same folder. The cleaned datasets should be renamed as follows: the cleaned training dataset as `Train_Dataset_<<Your_Student_Id>>.csv` and the cleaned test dataset as `Test_Dataset<<Your_Student_Id>>.csv`. Prior to submission, run and test your script properly and make sure it is free of bugs.

Final report

The final report should describe and justify the pre-processing steps you employed. Further, report the five most significant insights you derived from the data analysis. For each insight describe how you arrived at it with supporting visualizations and analysis.

The suggested report format is as follows.

- Problem overview
- Dataset description
- Data pre-processing
- Insights from data analysis
    - Insight 1
        - Description of the approach with supporting visualizations and analysis.
    - Insight 2
        - Description of the approach with supporting visualizations and analysis.
    - **…**

The report should be in PDF format and **no longer than 8 pages**. It should be named using your student ID (e.g., 1234567U.pdf). Upload the report to Moodle before **01 May 2022 11.59 p.m.**