Final Project

Due Date: 11.59 pm Dec 20, 2020

Submit a single notebook file (.ipynb) on Quercus

Background and Introduction

In this project you will be working with Black Friday sales data of a specific store. The store has already collected data that entail buyer's info such as age, occupation, marital status and others. The dataset that you need to use in this project includes around 500,000 observations.

The ultimate purpose of this project is to make a regression model to estimate purchase amount by each person depending on their features like gender, occupation, and etc. However, firstly you need to explore the dataset and extract useful information that will answer questions like what gender shops more on Black Friday? Do the occupations of people have any impact on sales? Which age group is the highest spender? And etc.

This project follows the process below in short:

A. Data Cleaning: you need to first clean the data before any analysis.
B. Data Analysis: you will answer some questions about different features and customers purchase behaviors to extract insight and useful information.
C. Data Modeling:
    1- Feature Engineering: you need to engineer features to prepare them for regression model
    2- Linear Regression Model: train and fit a linear regression model to data.
    3- Final Model Evaluation: evaluate the model using test dataset.

Learning Objectives

- How to clean data
- How to visualize data
- How to do exploratory data analysis
- How to do feature engineering
- How to make machine learning models
- How to evaluate regression models

Tool Required

- You have to use Python as the only tool in this project. You are free to use any packages in Python, but it is encouraged to use these packages: Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn
- Data Files
    - **Black_Friday.csv**: the main csv file that contains dataset to explore and create model
    - **Black_Friday_Final_Test.csv:** a similar dataset to Black_Friday.csv containing different transactions information, which is only for the final testing, and evaluation phase.

To Do

A. Data Cleaning (10%)

First, you need to clean the data. The existing dataset has missing values in some columns, and also some continuous or discrete ordinal columns are treated as string (object) because they include special characters like +.

Write a function clean_data that takes in a DataFrame as an input, and returns a clean DataFrame. The clean DataFrame should follow these rules: `clean_data(df) -> clean_df`

1- This function needs to turn ordinal columns into column with a proper data type. If a column like *age* or *stay_in_current_city_years* represents an ordinal feature (discrete numeric), it should be a number instead of a string object.
2- This functions needs to turn any categorical column containing numbers into strings. For example, a column including IDs should be a string data type instead of numeric data type (because there is no order for IDs).
3- This function needs to handle missing values. It should either fill missing values with proper value, or drop the entire column containing too many missing values.

B. Data Analysis (30% + 10%)

Using clean DataFrame from previous section answer questions below by visualizing proper plots:

1- Compare number of customers, and total amount of purchase for each gender? Is this a balance dataset with respect to gender? (5%)
2- Compare total amount of purchase for different cities. (5%)
3- Compare total amount of purchase for combinations of gender and age. Which gender and age is better target for marketing campaign? (5%)
4- Using box and whisker plot, compare distribution of purchase amounts among combinations of genders and age. (5%)
5- Using box and whisker plot, compare distribution of purchase amounts among different occupations. (5%)
6- (Optional) Create a heatmap or scatter plot showing correlation matrix of all of two numerical features. Which features are positively correlated with purchase amount? (+10%)

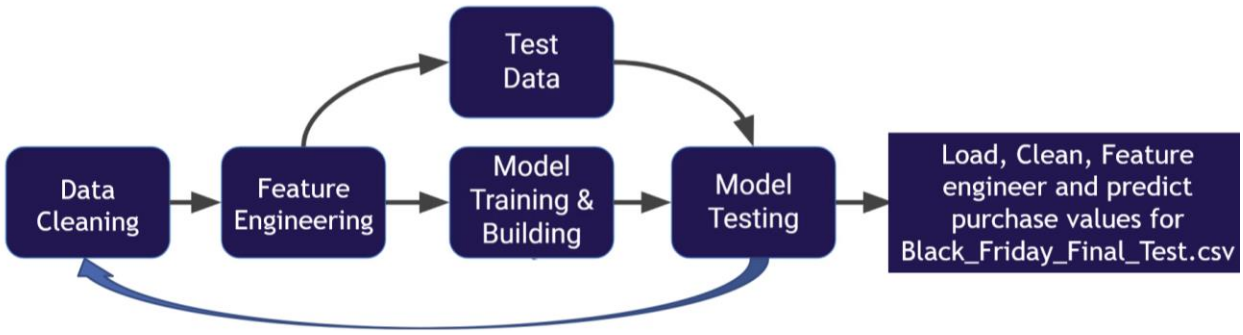
C. Data Modeling (60% + 5%)

In this section, you need to create an estimator to predict amount of purchase for testing data points.

The first step is to preprocess the features, so you can feed it into your regression model. Then, you need to split your data into training and testing, and fit a multivariant linear regression to the training dataset, and then test it on testing dataset.

By evaluating the regression model (MSE and MAE), you can go back and modify any feature engineering or cleaning process you have applied on dataset and try to minimize the error by iterating through this process.

At the end, when you finalize your regression model, load the *Black_Friday_Final_Test.csv* dataset and apply same cleaning and feature engineering you have applied to *Black_Friday.csv* so far. Then, run the regression model to predict purchase amount. Report the mean squared error (MSE) and mean absolute error (MAE) for this dataset (*Black_Friday_Final_Test.csv*). This part assures that final testing process would be consistent for all students, because there is no randomness involve.



1- Feature Engineering: In this part, you need to preprocess and engineer features (columns) to be able to feed it to regression model and predict purchase amount better.
   I) For categorical columns, use dummy variable to turn categorical column into columns of binary values (one-hot encoded vector)[1]. (15%)
   II) (Optional) Normalize all numerical columns using mean normalization. (+5%)
2- Linear Regression Model: split the dataset into training and testing dataset (80% for training and 20% for testing). Next, create and fit a multivariant linear regression model to training dataset. Then, test the trained model over testing dataset. What are the MAE and MSE? Feel free to modify or tune anything you want in previous parts to reduce MAE and MSE. (30%)
3- Final Model Evaluation: Finally, after reaching the optimum model, load *Black_Friday_Final_Test.csv* dataset into a DataFrame. First apply all cleaning and feature engineering to this DataFrame as well, and then apply the trained regression model from part 2 to get predicted values. Finally, calculate and report (print) MAE and MSE. (15%)

Submission:

Submit a single notebook file (.ipynb) via Quercus with the following naming convention: lastname_firstname_final_project.ipynb

Make sure that you comment your code appropriately and describe your approach in short. Your module should be self contained, i.e., the functions you submit cannot call functions you defined in other Python modules or Python codes.

Keep in mind, in each step of this project, there is no absolute correct answer. Your approach should be based on your rationale and reasoning, so try to shortly explain what you are doing and why.

The total mark for this project is 115% (35% of the final mark). The additional 15% (5% of final mark) is optional, and it can make up for the fifth assignment.

[1] https://towardsdatascience.com/the-dummys-guide-to-creating-dummy-variables-f21faddb1d40