Data Bootcamp Problem Set 6 Spring 2020

Instructions:

- **After completing the assignment, please submit your .ipnyb file to NYU Classes with the following naming convention: Lastname_Firstname_NetID_ProblemSet# (ex. Smith_John_js123_ProblemSet2)**
- Submit your answers in a **Jupyter notebook** with **proper markdowns to indicate problem numbers.**
- **Write the questions in markdown before you provide your answers.**
- **When copying the dictionary or any values directly from this file, make sure that all the quotations and brackets are in the right form in Jupyter Notebook. (Especially for string quotations – sometimes if you copy directly from a pdf file, the quotation breaks and it won't show up properly as a string in Jupyter Notebook)**
- See Grading Guidelines under Announcements on NYU Classes.
- For problem 1 import Ecommerce_data.csv.
- For problem 2 import credit_default_dataset.csv. Also see attached data dictionary if you want to know more about the customer features
- In-line comments are preferred for this assignment but not mandatory
- No explanations are expected at the end of answers, unless requested

Problems:

1. Regression Task. You are a data scientist with an e-commerce firm. The firm wants you to analyze their customer data and predict the yearly amount spent.

   a) Explore correlations between customer features and the variable you are trying to predict. Show your exploration by displaying scatter charts or other visualization techniques you used.

   b) Use numerical features of the customers as your customer features X and variable y (to predict) as the "Yearly Amount Spent" column.

2. Given customer characteristics like age, income, education, payment history, etc. your task is to predict whether they will default or not (Choose any variable or combination of variables from the dataset that you think may work best to predict the likeliness of default).

   a) Apply logictic regression to perform classification task. Use suitable method to evaluate the performance of your approach.

   b) Apply K-nearest neighbors to perform classification. Use a suitable method to evaluate the performance of your approach. Fine the most suitable value of 'K'.

   c) Apply Random Forest to achieve classification. Use a suitable method to evaluate the performance of your approach. Tune the parameters of this model- the max_depth and max_features.

   d) Choose a model based on your results from a, b and c. Predict y-hat using predict() with the chosen model, then display the predicted default values using the .value_counts() method.