SPRINGBOARD DSC

CAPSTONE PROJECT 1 PROPOSAL

Sheema Murugesh Babu

March 2019

1. What is the problem you want to solve?

The challenge here is to predict purchase prices of various products purchased by customers based on historical purchase patterns. The data contains features like age, gender, marital status, categories of products purchased, city demographics etc.

2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have done otherwise?

Retailers would greatly value these predictions which will help them to know their customer interests better. Such a predictor has a clear commercial value to the store owners as it would assist with their financial planning, inventory management, marketing, and advertising. This will also help them to create personalized offer for customers against different products. This helps in getting the peek sales and increased revenue, in turn increasing the company's turn over and preserve their successful business.

3. What data are you going to use for this? How will you acquire this data?

The data was found from the "Black Friday" dataset provided by Kaggle's website. https://www.kaggle.com/mehdidag/black-friday

4. In brief, outline your approach to solving this problem (knowing that this might change later).

I'll be following a typical data science pipeline, which is "OSEMN" (pronounced awesome).

Obtaining the required data is the first approach in solving the problem. I would have to download the dataset from Kaggle's website and import it as a "csv" file to my working environment.

Scrubbing or cleaning the data is the next step. This includes data imputation of missing or invalid data and fixing column names.

Exploratory data analysis will follow right after and allow further insight of what our dataset contains, looking for potential data quality issues. Understanding the relationship each explanatory variable has with the response variable resides here and we can do this with a correlation matrix. The creation or removing of features using feature engineering is a possibility. The use of various graphs plays a significant role here as well, because it will give us a visual representation of how the variables interact with one another. We will get to see whether some variables have a linear or non-linear relationship. Taking the time to examine and understand our dataset will then give us the suggestions on what type of predictive model to use.

Modeling the data will give us our predictive power on the purchase patterns of various customers. Types of models to use could be RF, SVM, LM, GBM, etc. Cross validation is used here, which will allow us to examine our model’s accuracy and tune our model’s hyperparameters if necessary. We can also use some feature importance analysis with information extracted from Random Forest models.

Interpreting the data is last step. With all the results and analysis of the data, we can find explanations to questions like, which category of products were sold the most or what population of gender helped in revenue generation? What relationship of variables were found? If our model’s training performance greatly differs from its testing performance, a chance of overfitting is likely. Ways to prevent overfitting include: collecting more data, choosing simpler models, cross validation, regularization, use of ensemble methods, or better parameter tuning. This analysis also gives a brief overview of the feature importance that affected our model and how we can improve our model in the future.

5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

As required, my deliverables will be all the Jupyter notebook I will develop, a final report, and a presentation slide deck.