SPRINGBOARD DSC
CAPSTONE PROJECT 2 PROPOSAL
By Sheema Murugesh Babu
September 2019

1. What is the problem you want to solve?

The challenge here is to build a model that identifies customers with the intention to leave a service in the near future. The data contains Demographic information like gender, age range, and whether they have partners and dependents, contains customer account information, services that each customer has signed up for and lastly customers who left within the last month – the column is called Churn.

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?

Many business establishments would greatly value these predictions which will help them to know if a customer stops using their service or stops doing business with them. Customers turning their back to a service or a product are no fun for any business. Such a predictor has a clear commercial value to business owners. Keeping the right customers can be quite valuable for a company. Not only because customer acquisition is much more expensive, but as more and more business models are shifting towards subscription plans, a customer can be worth thousands of dollars in future. Reducing churn ultimately leads to a sustainable growing business.

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

The data was found from the "Telcom Customer Churn" dataset provided by Kaggle's website. https://www.kaggle.com/blastchar/telco-customer-churn

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

Below points are referenced from the site https://medium.com/breathe-publication/life-of-data-data-science-is-osemn-f453e1febc10. 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 the explanatory variable has with the response variable resides here. The creation or removing 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. Taking the time to examine and understand our dataset will then give us suggestions on what type of predictive model to use.

Modeling the data will be the next step to follow. Types of models to use could be Random Forest Classifiers, Logistic Regression, Naive Bayes, etc. Cross validation could be 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. If our model's training performance greatly differs from its testing performance, the 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.

Interpreting the results is the last step. With all the results and analysis of the data, we can find explanations to questions like, which customers are likely to churn and how to stop them from leaving. 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.
