Capstone Project: Analysis of Chinese Mobile Phone User Population vs Phone Specifications

The project conducted was focused on using a user's personal information to determine the type of phone that user has using a classification model. The hopes of this model is to allow for a marketer to give better and more reliable suggestions to potential consumers. This idea was sparked through a public dataset found on Kaggle documenting thousands of phone users over the last recent years.

The first step of my project was to collect and load the data. The first portion of the data was quite simple to collect, just download the dataset through the Kaggle website. The second portion of the data was gotten by scrubbing the gsmarena website. I pulled the data through a C# program given on GitHub after making minor modifications to update the program to deal with minor inconsistencies.

The second step was cleaning the data. The two datasets loaded into my Jupyter notebook without any issues. But it was impossible to match up the datasets because the former dataset was recorded in Chinese. A good amount of effort was made to translate the dataset from Chinese to English. In addition to the fact that misspelling were common with the gsmarena dataset due to the fact that the data was pulled from the web.

During the cleaning of the data, I used exploratory data analysis to gain a general understanding of what my datasets are. I displayed subsets of each datasets and information about the observations. The collected data comes from TalkingData SDK and only a portion of the device_ids are matched with the event data. Also, the device_id to device_brand/model has translation problems and foreign-only releases. A lot of brands and devices cannot be matched with the gsmarena data so I had to eliminate a good portion of the available data. This analysis surprisingly does not include "Apple" as a brand which can be a result of the TalkingData SDK being unable to collect data from Apple devices or being an Android-only SDK. The majority of data came from 2013 through 2015 and the basis of this range is on the release dates of the phones.

I was able to display graphs showing the market share that each phone manufacturer had by age group and gender. I calculated that there is a small positive correlation between the number of models released and the number of users but the p-value is too high to be considered a valid conclusion. You can also see the outliers of the data. For example Xiaomi has only release ~16 models but has sold a ton of phones, whereas Lenovo released ~48 models but sold less than a thousand. I displayed the top 10 most popular phone models being used which also gave the conclusion that the majority of my data comes from Xiaomi phone users. This quickly allowed me to realize that the data has a severe class imbalance problem, where a lot of the smaller brands had little to no market share and were eliminated from my model.

The next portion of my project was to merge the two datasets. The Kaggle dataset had to be internally merged because it was already split into five separate parts with each part holding a section of the information I needed. The features that I selected for my model are: label id, gender, age, and year. The label id represents how the phone is used and the year is the year the current phone that the user has was released. This year can be assumed as when the user purchased the phone as most android phone users don't purchase used or older models. These features should be easy to determine from a potential consumer through a few targeted questions.

The final step was the create a model. The first classifier I decided to use was logistic regression, mainly because of its simplicity and quick training time. The log loss of this model was not particularly great, so I investigated the cause of the issue by graphing the classification using PCA. The main issue was logistic regression even using class weights to balance out the data was that the classifier was unable to classify the lesser used phone brands.

I quickly changed my model to XGBoost which is a popular parallel tree boosting classifier. This classifier performed significantly better than the logistic regression classifier and had a score of .44. I plotted the multiclass roc-auc curve with the area values under each curve. The model, while far from being perfect, gives direction to potential marketers in deciding on the course of action per consumer based off of limited information.