Trial to Paid Conversion Prediction

Shawn Kim

**Capstone Table of Contents:**

1. Client and Background
2. Feature Hypothesis Generation and Data Exploration:
   * Identification of variables that can impact conversion
   * Analysis based on the scoping in previous step to identify final data set
3. Data Preparation & Data Cleansing
   * Processing of features for model intake
   * ML feature selection for feature set validation and discovery
   * Outlier treatment
4. Modeling
   * Testing of different ML models (XGBoost and RandomForest)
   * Fit best model

**1. Client and background**

The client is my current company that is in the media industry. The project deals with their online streaming platform that they have built into an subscription business model along with ad revenue model for certain subscription tiers. Their competitors include other network streaming portals, and the more established Netflix/Hulu/Sling/Amazon.

A key problem that they have to solve is being able to predict which trial subscriber are most likely to become paying subscribers. A model that enables the client to predict subscription likelihood gives them the ability to become smarter about how they spend their marketing budget. Targeted campaigns can become smarter.

My capstone project’s goal is to predict the likelihood of conversion for trial users.

**2. Feature Hypothesis Generation and Data Exploration:**

The first step was to generate a list of feature hypotheses. There is a lot of data available so it was important to first create a list of hypotheses to guide the EDA phase. After several brainstorming sessions with the product teams, marketing teams, and my own team, I was able to generate a list of 40 possible hypotheses. Once I had this list, I went through and performed EDA using Tableau to see how many days of the trial were necessary to effectively predict conversion. Using histograms, I was able to determine that large differences in conversion rate appeared as soon as the first day.

In the next phase, I performed EDA in Python in order to identify any outliers, check for normal distributions/if transformations were necessary, address skewness, and address any zeros and null values.

Through this process identified the scope of the data. Resolved a final data set with several conditions based on the EDA.

**3. Data Preparation & Data Cleansing**

I used pandas-profiling function to quickly identify each of these issues. I had to drop some of the features that were too low in sample or only affected a small subset of the data. For high levels of skewness and outliers, I performed transformations using a function called power transform to normalize distributions and also remove outliers. For values with too many zeros or null values, I used imputation functions to add the mean for continuous features or the most common value for categorical values.

With the final cleaned up data set, I created train and test sets (70/30 split). I then used randomforest to perform a feature importance ranking. By graphing the results, I was able to quickly see which features were the most important to conversion rate.



**4. Modeling**

Tested 3 different models: Decision tree, random forest, and XGBoost. The decision tree model gave me an accuracy rate of 63.1%. The random forest model gave me an accuracy rate of 60.3%. The XGBoost model gave me an accuracy rate of 63.4%. The winner was the XGBoost model that gave me the highest level of prediction accuracy.

https://github.com/shawnyskim/firstproject/blob/master/Model%20for%20SB%20Capstone%20Project%20-%20Predicting%20Trial%20to%20Paid%20Subscribers.ipynb