Final Proposal

Jack Wright

10/22/2021

GStore, Revenue per Customer Prediction

For our final project, we will be looking at a kaggle competition `Google Analytics Customer Review Prediction` where we will *predict revenue per customer* for a Google store based upon a dataset provided from the online outlet.

In this competition, you're challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer. Hopefully, the outcome will be more actionable operational changes and a better use of marketing budgets for those companies who choose to use data analysis on top of their data.

Why is this important?

It is a widely held belief that 80% of a business' revenue comes from 20% of their customers, also known as the *80/20 rule*. If we are able to predict revenue per customer, then we should be able to identify this most important 20% of a stores customers. We selected a kaggle competition because of the benchmarks set that will allow us to gauge our progress as analysts.

Background

Papers on predicting revenue per customer, or *customer lifetime value* prescribe a lot of the same techniques that we have become accustomed to. Tidying, Exploration, model building and analysis. But there are a few features unique to predicting future spending that we may choose to look at.

activity analysis:

As highlighted in the paper Predictive Modeling Using Transactional Data, activity analysis sets a time frame for which a customer can be defined as inactive and can be used to model attrition of a business. Maybe we can leverage this into a dummy variable to predict future spending.

Cohort and Trend Analysis:

There seems to be some trend in future revenue prediction between *attriters* and *high transactors*. This makes intuitive sense that people who have spent more, will spend more in the future. It might be interesting to give this a time dimension as well as just a scalar

quality.

Predictions on groups with varying amounts of data:

In the paper Predictive Profiles for Transaction Data using Finite Mixture Models, The authors bring up the problem of making predictions on transactors with low amounts of data (say one purchase only). They recommend using *Bayesian estimation for learning predictive profiles*. In essence, we can create different predictive profiles, or models, depending on how much data we have about the transactor. For example, a transactor with a low amount of information (one purchase) would have a more general model to the data applied, while as a frequent transactor could have a much more data driven model.

This technique is generally used in machine learning, but I don't see why we couldn't apply it to logistic regression.

Methodology:

We will generate both linear regression models and Random Forest decision trees to model the data and the source of our data is the Google 'Gstore' where you can purchase google branded products online.