Competition 4: Predicting how much GStore customers will spend

Andrew Evans (ace8p), Ning Han (nh4mq), Sameer Singh (ss8gc)

The Google Analytics Customer Revenue Prediction competition has a self-evident problem statement. Accurately predicting revenue for an online retail site has clear benefit to making informed business decisions that could increase revenue, reduce expenses, or optimize a supply chain. This competition provides detailed information describing characteristics of customers themselves as well as their unique visits to the site. Having the ability to predict revenue with that information can make it possible for a marketing team to reallocate ad spending geographically, or across access devices to increase customer traffic from sources that drive revenue. It would also be possible to see platforms where there is little revenue potential and limit development resources on keeping the store operational where there is little upside to doing so. Aside from changing marketing and development resources, an online retailer could derive insights about customers from training data, and then adjust production or inventory as customer traffic trends in new directions over time.

This problem is challenging because of the nature of sales being driven by a minority of customers or transactions. As stated in the competition's description, "The 80/20 rule has proven true for many businesses–only a small percentage of customers produce most of the revenue." When encountered in prediction, this Pareto principle means that an appropriate estimate in the majority of cases should be zero or very small in value. Unless the predictors themselves align with the divide between those who are making purchases and those who are not (e.g. if iOS accounted for 80% of traffic and accounted for less than 20% of sales revenue while other platforms made up 80% of the revenue with only 20% of the traffic) then predicting the right value could be difficult without knowing underlying motivations not present in the predictors' values or categories.

This *challenge* is similar to the challenge discussed by Dr. Thomas Plöetz when he delivered a seminar regarding Computational Behavior Analysis to students and faculty at UVa. With data collected from wearables or cameras attempting to analyze behaviors of people or animals, there is an overwhelming portion of the data where the appropriate label or classification of the null or benign case.

The competition's problem (predicting revenue per customer via a web store) is similar to a plethora of other situations across online retailers like Amazon.com, Walmart, or Apple's online store, but also similar to revenue prediction in scenarios like mobile game applications. Similarly, a small portion of customers drive the majority of revenue in mobile games. To predict the revenue per customer, a company might instead need to look at predictors like user behavior sequences or interaction between one user and others which could be indicators of purchase likelihood or revenue quantity.

Our team saw Random Forest make the best prediction for this competition. There are many reasons this might be the case. Often, Random Forest performs well due to its strengths such as its inherent feature selection as part of the tree building, and its ability to handle many datatypes (e.g. numeric, categorical, logical, and chronological) without manipulation. This competition had a wide variety of predictors and RF might have found predictors in its tree splits that align well with the divide between the 20% of customers that drive revenue and the others who do not. By comparison, our Lasso Regression model may have performed worse due to its nature to shrink coefficients to zero. If two predictors are correlated, a Lasso model can wind up dropping one altogether. If that happens, but the nature of the prediction is dependent on an interaction between those predictors, the Lasso will perform worse where the RF might instead produce splits in trees at levels of those predictors.