Google Analytics Customer Revenue Prediction Report

Team member: Jiangxue Han, Shaoran Li, Sean Mullane

- **Who might care about this problem and why?**
  Customer analytics could be useful to many, including marketing campaigns, sales and product teams, upper management and even the board of directors. Customer analytics refers to the processes and technologies that give organizations the customer insight necessary to match customers to product offerings that are attractive, relevant and timely. Nowadays, customers are more connected and generate large amounts of data. This provides great opportunities for customer data analytics since data is easier to collect from information platforms. If this data could be interpreted better, then purchasing behaviour could be predicted with more accuracy and this would enable business strategy to be tailored for increased revenue generation.

- **Why was this problem challenging?**
  This problem is challenging due to the size of data (25 G) and data structure in the fields (dictionary). There are huge chunks of missing data; we took several different approaches to resolve this. For the *revenue*, *totals.bounces*, *totals.newVisitis*, *totals.timeOnsite* and *totals.sessionQualityDim* columns we used value 0 for "NA". For any columns that have more than 90% of "NA" we dropped them. For *pageviews* column we used median imputation.

  Also, the dataset was updated 3 days before deadline, and although it was not entirely different, due to the increased size our group has tried various methods such as splitting data into smaller chunks, uploading into AWS server, selecting random sample, etc. The response variable was also highly imbalanced, with the large majority of visitors spending nothing at the store and a large portion of the revenue coming from a small number of users.
- **What other problems resemble this problem?**
  Credit card rate segregation seems similar to this problem. The example of Capital One's use of customer analytics for market segmentation is useful to illustrate the potential benefits. A long time ago Capital One was only a small bank. They became big player because two engineers in the company proposed the idea that credit card rates should be different for different customers. The customers that are most active and loyal should be rewarded with better rates and be focused on by the bank. At that time most credit card rates were the same across different customer groups. They conducted this experiment for five years and turned a loss into profit by more than 50%. By conducting analysis and studying customer behaviour, Capital One was rewarded greatly. We expect similar results for this competition, that Google will exploit these analysis and result in tremendous increase in their revenue.

- **What might account for the differing performance levels of the mandatory models?**
  Random forest performs the best since many of the variables in this dataset are categorical variables. Random forest is also well-suited to data sets with complex interactions between variables due to the large number of trees using a variety of splits. For contrast, OLS and spline models must have these interactions defined explicitly which is burdensome in the presence of a large number of variables, especially those created by one-hot encoding categorical variables. Adding these interaction terms would also cause issues with multicollinearity in an OLS or spline model. The same goes for GBM versus OLS and splines.