SYS 6018: Data Mining

11/10/2018

Google Analytics Customer Revenue Prediction: Reflection Paper

Catherine Beazley, Saurav Sengupta, Aman Shrivastava

This problem would help any online store that seeks to predict its online sales over different periods in the year. Advertising companies would find this model particularly useful, since, this gives us a way to model consumer behavior on different products at different times over the year. Knowing this behavior, advertising companies could better optimize advertising certain products to certain consumers at certain times.

The problem was made challenging by the huge size of the training data. A number of parameters had same values in all rows, were highly skewed or had null values in most of the rows which made it harder to derive correlation and create new features. Some of the data were given in form of JSON objects which had to be flattened out to get values into columns of the dataframe.

This problem could be applied to a brick and mortar store as well. We found that some of the time variables were the most important in our models. A store in a mall for example, could use a dataset similar to this one to see when people are buying from their stores and how much people are buying to optimize when they should open their stores, how many staff need to be working at specific times to attend to the customers, and how much inventory to stock in the store at certain times.

Something that could account for the differing performance is the number of regressors. The data contained a large number of regressors and some models perform better or worse with a large numbers of regressors. Random forests handle a large number of regressors better than some other models because it can compare accuracy at each node among a subset of regressors, and ultimately use the many options of regressors to optimize each split of the data. In this case, a large number of regressors can help performance without overfitting the training data.

In contrast, using a large number of regressors in OLS and Regression Splines often leads to overfitting of the training data which worsens performance of the model with the testing set. It is up to the data scientist to find to optimal subset of regressors for these models to maximize testing error, which can lead to a worse performance in these models since performance will depend on this choice of regressors.