**Course 3 Task 3: Multiple Regression in R – Michelle Giniewicz**

**Intro and Model Selection**

I began my analysis by reviewing the Existing Product Attributes dataset of 81 different products with attributes like Price, 1-5 Star Reviews, Positive/Negative Service Reviews, Volume, etc. The first business question I needed to answer was: What are the predicted sales of four different product types: PC, Laptops, Netbooks, and Smartphones. I first analyzed the data to see if any variables were highly correlated and therefore would have an impact on the sales volume. I found that there was a direct correlation between the Volume and x5StarReviews. Additionally, there was high correlation between x1StarReviews and x2StarReviews, and between x3StarReviews and x4StarReviews. Based on this, I created a new dataset that removed the x5Star Reviews, x3StarReviews, and x1StarReviews.

Chart, bubble chart

Description automatically generated

Additionally, I noticed that there were two outliers within the data with much higher volumes than every other product. Since I wasn’t sure if that may impact the accuracy of my machine learning models, I created another new dataset that removed those two products.

Next, I created three machine learning models using each of my first two datasets (the complete Existing Product Attributes dataset and the dataset with the highly correlated variables removed). The three machine learning models I created were: 1) Support Vector Model, 2) Random Forest Model, and 3) Gradient Boosting Model. I then tested each of these six machine learning models with a test dataset, to see how accurate they would be at predicting the volume of the new incomplete dataset. After training and testing the models, I found that the Support Vector Model and Gradient Boosting Model for both datasets were not a good fit for this data, since they predicted negative amounts for the sales volumes. Based on this, I then knew that Random Forest would be my best model. However, I also decided to run my third dataset (with the outliers removed) using these models, to see which would be the best fit. At the end of my testing, the Random Forest model with the outliers removed and the highly correlated variables removed was the most precise model. Below is the resample output that I received for this model:



**Predicting Sales of Four Product Types**

After choosing the Random Forest Model with both outliers and highly correlated variables removed, I then used this model to accurately predict the sales volume in the New Product Attributes dataset for the four target product types: PCs, Laptops, Netbooks, and Smartphones. Below are the sales volumes that my model predicted:

Table

Description automatically generated with medium confidenceChart, bar chart

Description automatically generated

**Lessons Learned and Recommendations**

One of the main lessons that I learned from this analysis was just because a machine learning model seems precise, you should also look at the predicted output to see if it makes sense for your dataset. For example, the Support Vector Model had pretty high R-squared values with my OOB dataset. However, once I looked at the predicted data, I realized that it was predicting negative sales volume, which proved it was not a good fit for this dataset.

Another lesson I learned is that having outliers like we did in the Existing Product Attributes dataset can make a significant difference in the accuracy of your model. For example, prior to removing the outliers in the dataset with the removed highly correlated variables, my Random Forest model had an R-squared value of 0.84 with the predictions against my test data. Once I removed the outliers and re-ran the model, I then got an R-square value of 0.98.

Based on the correlation analysis, I would recommend the sales department focus their efforts on products that have a lot of 5-star reviews. There was a perfect positive correlation between the sales volume and x5StarReview attribute, meaning that products with a lot of 5-star reviews will also have a lot of sales.