Customer Brand Preferences - Course 3, Task 2

Blackwell Electronics Investigation

Primary Objective:

- Build and assess models to predict customers' computer brand preference based on complete survey data using R Studio
- Use the best model to predict brand preference for an incomplete customer survey data set

Business Question:

Can we use Data Science to better predict what brand of computer a customer will prefer?

Data Set

For this analysis and modeling we used a data set of 9898 rows with customer information including salary, age, education level, primary vehicle, region, credit score (independent variables), and preferred computer brand (dependent variable). Using this data set, we'll build and select an appropriate model for predicting outcomes in a data set of 5000 rows that lacks the preferred computer brand.

Feature Engineering

Before building models, we investigate whether any of the independent variables show correlations higher than 0.75. In such a case, one of the highly correlated features is eliminated prior to model building. As shown in Figure 1, none of the independent variables are highly correlated, with the highest correlation being 0.06.

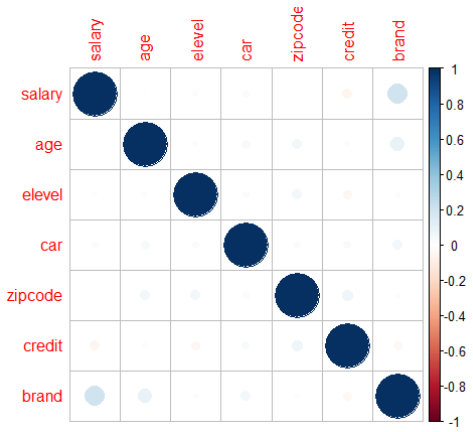

Figure 1: Correlation plot of features

Using Caret Recursive Function Elimination, the Random Forest model was applied to see if the feature set could be reduced for creating a more efficient model. The tool showed that with just two features – customer salary and age, the model actually increased (from 0.882 to 0.909) versus using all 6 available features.

Predictive Modelling

The following 5 models were built and investigated to find a good predictor to the dependent variable using a 10-fold cross-validation:

- Random Forest with the out-of-the-box data
- Random Forest with the out-of-the-box data, with a manual grid
- Stochastic Gradient Boosting with the out-of-the-box data
- Random Forest with the reduced feature set
- Stochastic Gradient Boosting with the reduced feature set

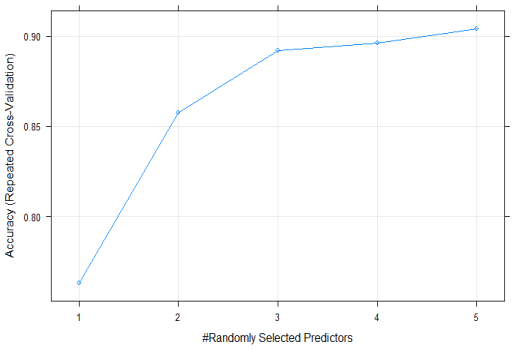

Figure 2: Random Forest accuracy versus mtry

For the Random Forest with manual grid, Figure 2 shows that mtry=5 shows a valuable increase in model accuracy. This accuracy increase, along with the associated high kappa coefficient, results in this Random Forest manual grid model performing the best on this data set shown in Table 1. Surprising in this result was how close all of the models were in terms of accuracy for this training data, with the entire set ranging between 88% and 91% accuracy.

Table 1: Model accuracy and kappa

Model	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Random Forest (RF)	0.813	0.893	0.905	0.899	0.917	0.934	0
RF manual grid	0.827	0.896	0.913	0.904	0.920	0.934	0
Gradient Boosting (GBM)	0.760	0.856	0.880	0.871	0.906	0.908	0
RF with reduced variables (FE)	0.867	0.893	0.899	0.899	0.906	0.934	0
GBM with reduced variables (FE)	0.840	0.871	0.887	0.895	0.926	0.960	0
Kappa							
Model	Min	1st Quartile	Median	Mean	3rd Quartile	Max	NA's
Random Forest (RF)	0.606	0.775	0.800	0.787	0.826	0.862	0
RF manual grid	0.632	0.788	0.814	0.798	0.833	0.862	0
Gradient Boosting (GBM)	0.487	0.700	0.744	0.726	0.803	0.809	0
RF with reduced variables (FE)	0.719	0.771	0.789	0.786	0.801	0.858	0
GBM with reduced variables (FE)	0.654	0.730	0.759	0.778	0.843	0.915	0

When this model is applied to the Test data set from the complete survey data set (Ground Truth), we see a slight increase in accuracy and kappa to 0.923 and 0.839 respectively, further cementing this as a good model for this data set.

When this model is applied to the incomplete customer survey dataset we find that customers prefer Sony computers significantly over Acer computers. In Figure 3 the predicted brand preferences (orange) are stacked on top of the existing known brand preferences (blue) from the complete survey data.

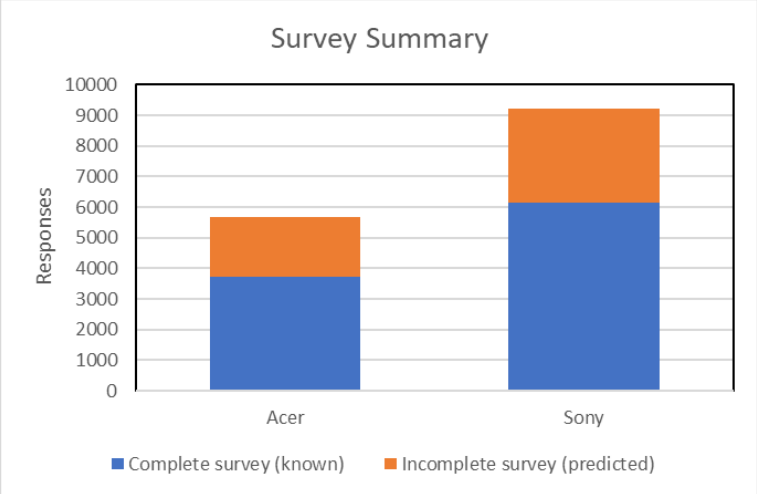

Figure 3: Brand preference for complete survey and incomplete survey data