Dear Danielle,

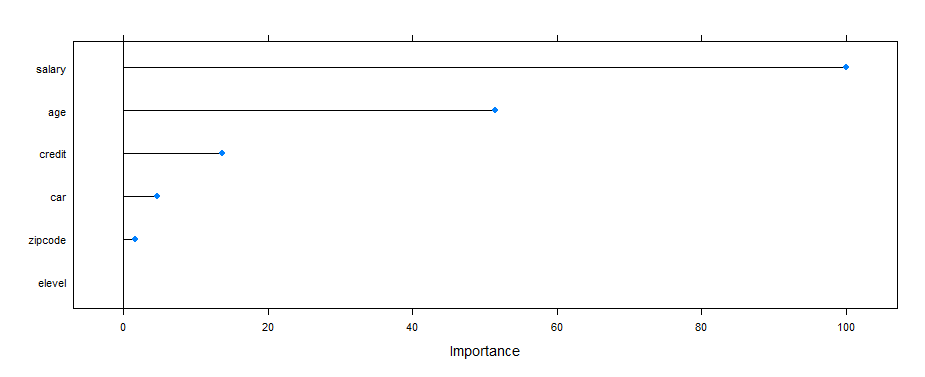
Per your request, attached is my report on Brand Preference Prediction

R and R-studio was used to predict the customer brand preferences from an “incomplete survey” data set sent from the marketing department. This data set has 5000 records and 6 features (Salary, Age, Credit, Car, Zipcode, Education) along with the target variable (Brand) which were somehow corrupted. The model was trained on a similar “complete survey” data set about twice the size in records and target variable was not “corrupted”. From the complete survey we were able to train a strong model to predict the brand outcomes for the incomplete survey. Models from four classification algorithms were developed and the best model was used for the predictions. The four model algorithms looked at were:

* Stochastic Gradient Boosting Classification (gbm)
* Random Forest Classification Model (rf)
* Random Forest Classification features reduction (rfeControl)
* C5.0 Decision Trees and Rule-Based classification (C50)

Modeling Methods:

10-Fold cross validation was used in all four models’ training. The Carat package was also used to optimize each algorithm’s parameter values by adjusting the tunelength up to 5 times. Additionally, a review of feature importance was done. Below is a graph showing the feature importance of the C5.0 model.



All models aligned that education level, zip code, or car have the least importance in predicting Brand and most likely one of these could be dropped from the feature inputs of the models without significantly altering performance. The Random Forrest features reduction (rfe) model did drop Education Level from its features. We however chose to move forward with all features in all other models since dropping a feature does not significantly reduce computation costs.

Model Selection Criteria:

For our situation, there is no sensitivity to a model’s bias to predicting one outcome over the other; therefore, the model’s precision or recall is not relevant and only maximizing accuracy was used in our decision for which model to use while keeping kappa in mind. Below is a table showing each model’s accuracy and kappa.

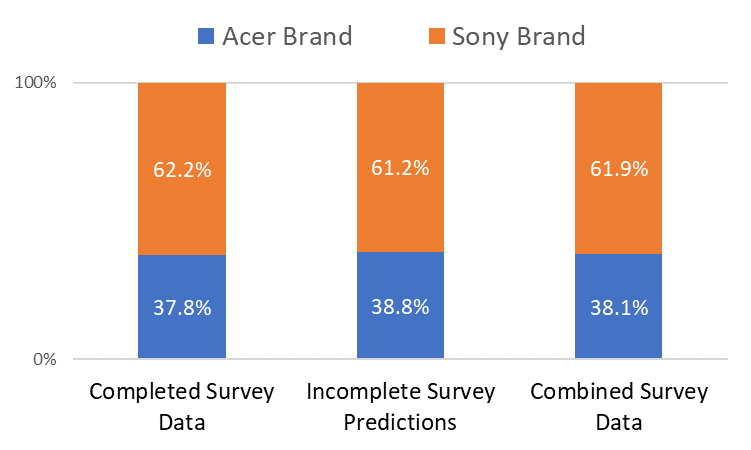


As shown in the table, we used the Stochastic Gradient Boosting Classification (gbm) model to run our predictions on the incomplete survey because it had the highest accuracy score (0.921) with a kappa value in line with the other models.

Results:

Applying the gbm model to predict the incomplete survey brand results, our predictions have 38.8% Acer models and 61.2% Sony models. This compares with the complete data set of 37.8% Acer and 62.2% Sony. See the graph below:

Predicted Results vs. Completed Data Results



Given that our prediction model has a 92.1% accuracy, it can be said it has an error range of ±3.85%. As such we would expect the percentage of Acer Brands in the completed survey (37.8%) to fall within the range of 37.4% and 40.0% of the Acer Brands in the predicted survey: which it does.

Assumptions:

* Data in both surveys are randomly selected from the same population