Dear Danielle,

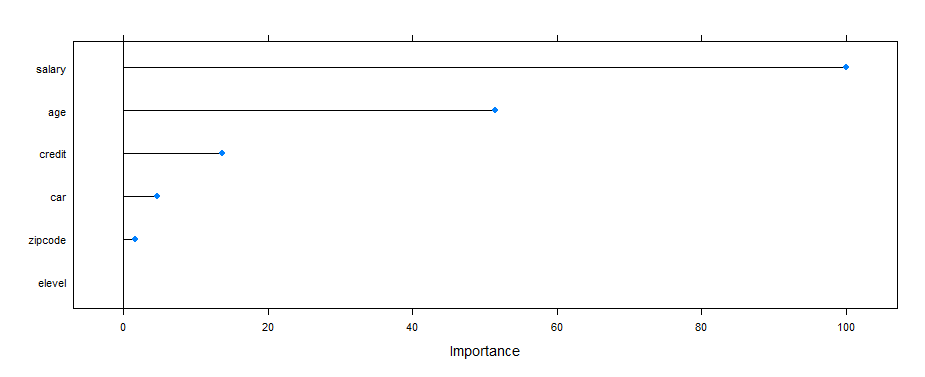
Per your request, attached is my report on Brand Preference Prediction

R and R-studio was used to predict the customer brand preferences from a “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 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 (rfe)
* 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 agreed that Education Level has the least importance in predicting Brand and most likely could be dropped from the features list without significantly altering each model’s performance. The Random Forrest features reduction model did drop the Education Level feature. We however chose to move forward with all features in all other models since dropping the feature was not a significant reduction in computation costs.

Decision 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. Below is a table of each model’s accuracy and kappa.

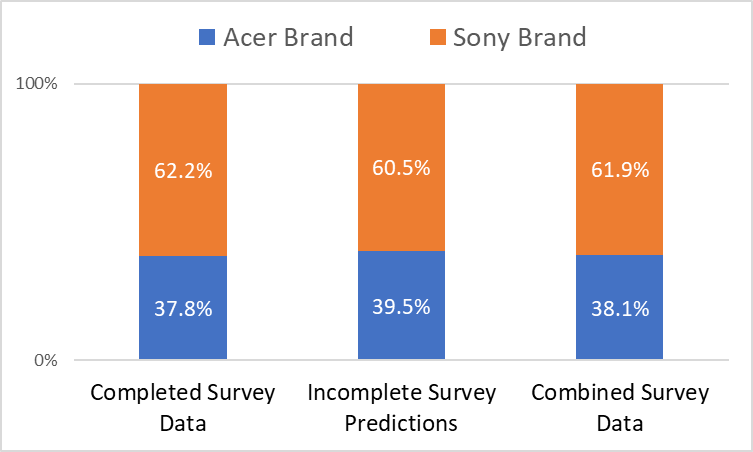


As shown in the table we used the Random Forrest Classifier model to run our predictions on the incomplete survey because it had the highest accuracy score (0.925) and the lowest kappa (0.818).

Results:

With that, our predictions on the incomplete data set have 39.5% Acer models and 60.5% 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 model is said to have a 92.5 accuracy, it can be said that it has an error rate of ± 3.5%. As such we would expect our predicted results to come within in a range of 37.8% ± 3.5% of the completed data survey which it does.

Assumptions:

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