**C3T2 Brand Reference Report for Blackwell Electronics**

For this project we were working with two datasets. We were using data mining techniques and machine learning predictive models to answer the question about brand preference of the customers for Blackwell electronics.  To have a complete report we had to answer the following question:

* What brand of electronics did Blackwell customers prefer in their cars?

 We were working with two datasets. The first complete dataset was used to build predictive tuned up models about customer brand preference.  After we optimized our model with the first dataset, we turn our attention to the second, incomplete survey dataset, to predict what brand (0-Acer or 1-Sony) customers preferred in their cars. Then we added the numbers from both data sets for each brand preferred and presented that information in the table format.  Both datasets were clean and did not have any duplicates.  We explored the first dataset and we discovered that based on salary and age distribution, the data was evenly distributed. We also ran different redistributions and we did not notice anything specific that will help us answer our final questions.

            When we turned our attention to building models, we were enthusiastic to build as many models as we could, tune them up, and choose the best fitting one.  Then, our plan was to apply that model to the dataset with incomplete responses so that we would be able to see what brand of electronics customers prefer.

We started our prediction by building the Random Forest classification predictive model with tuning length 1. We achieved the accuracy of the model at .91, which means that out of 10 observations, if we look at salary, age, education level, car, zip code, and credit, we can correctly predict the brand of electronics that customer preferred 9 times out of 10. For the same model we achieved a Kappa score of .829

We were working with the model of Random Forest and we tried to tune the parameters of the model in different ways to see if we could reach a better result. We tried a different Tune Length of the model, we used an automatic grid, and we tried manually adjusting the grid.  We also did a random search of the best fitted model from all of the tuning that we did.  After all of this our model did not improve that much.  However, we can say that our accuracy of .91 is a pretty good accuracy, and we can use this model on our incomplete dataset.

While working with the same dataset and looking for the best model, we also tried Gradient Boosting and found that the accuracy for this model was lower when the number of trees was low and the accuracy improved when the number if trees was higher. The lowest accuracy for the GBM model was. 73 and the highest was. 91.

We did not stop there, but rather we experimented with the c5.0 model and we got an accuracy of .84, which is lower than the previous models.

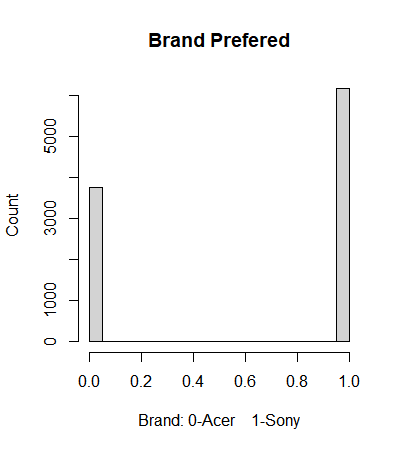
After running all of the models, we decided to use the random forest model to predict the brand of electronics for the incomplete dataset. As mentioned earlier, the decision tree model had the highest accuracy of any tuning that we experimented with.

After running our selected model on both datasets, we came up with the information on brand preferences for both datasets and got the following results (see table below):

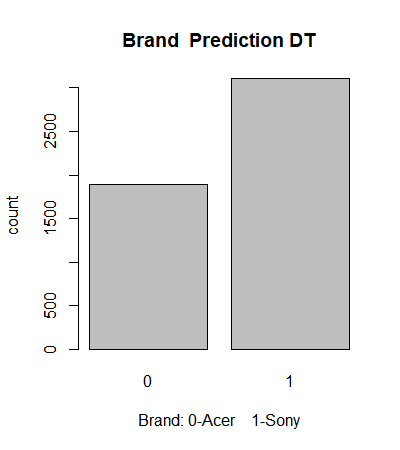
|  |  |  |  |
| --- | --- | --- | --- |
| **Brand Preference** | **Complete Data** | **Incomplete Data** | **Total** |
| Acer | 3744 | 1890 | 5634 |
| Sony | 6154 | 3110 | 9264 |

It would be short of a full report not to include the importance of the attributes given in the dataset. We looked at the importance of the given attributes and we saw that salary, age and credit amount available were the most important attributes in building the predictive models, and where car brand, region (zip code) and education level were the least important variables in building the models. Looking at the importance of variables in R, it is an especially useful function that we can definitely use for the other projects to come.

In conclusion, we can tell from running our models and from fitting the incomplete dataset to our chosen model that the Sony brand is preferred to the Acer brand.  In the table below you can see brand preference based on the complete dataset response.



In the table below you can see predicted brand for data incomplete.



This conclusion was accomplished by running different predictive models and choosing the best fitted one, with the highest accuracy.