Customer Brand Preferences Report

Summary (problem scope)

This document is a report on Blackwell’s problem regarding customer preference of computer between the options of Sony and Acer. This problem was solved using predictive analytic tools. Commentary will involve what models were used, their predictive power, how they were improved, and what Blackwell customers actually may have preferred.

Before I go into the analysis, it is important to make clear what the problem was. Blackwell had two customer survey datasets, one of which was accurate and fully complete whereas the other one was incomplete and therefore inaccurate.

Blackwell wanted to know if it would be possible to approximate what the actual computer preferences could have been from the incomplete customer survey dataset by using the complete customer survey dataset as a guide.

By using the tools of R, RStudio, and the caret package, which contains several relevant predictive analytic tools for this problem, it was possible to determine what the computer preferences outcomes could have been as if the customers had entered that information correctly. More will be said about this at the end of this report.

Analysis

There were three models tested. They were Stochastic Gradient Boosting (GBM), Random Forest (RF) and C5.0 (c5). These models were selected for testing because they could be used for classification problems, which summarizes Blackwell’s problem. Blackwell wants to predict between two options, Sony and Acer, which are both textual information.

In order to train and test these models for their ability to predict customer computer preferences, a subset of the complete dataset (CompleteResponses.csv) was used to form the primary testing set.

The primary metrics used to evaluate the predictive power of the models involved accuracy and kappa scores. This is how each model performed:

GBM

|  |  |
| --- | --- |
| Accuracy Score | 0.921 |
| Kapp Score | 0.832 |

RF

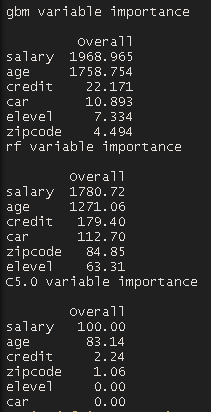
|  |  |  |
| --- | --- | --- |
|  | Accuracy Score | Kappa Score |
| mtry1 | 0.864 | 0.703 |
| mtry2 | 0.923 | 0.838 |
| mtry3 | 0.925 | 0.841 |
| mtry4 | 0.925 | 0.842 |
| mtry5 | 0.923 | 0.836 |
|  |  |  |
| average | 0.912 | 0.812 |

C5.0

|  |  |
| --- | --- |
| Accuracy Score | 0.916 |
| Kapp Score | 0.821 |

Given these results, I decided to choose the GBM model in order to predict the customer computer preferences. While all of the models performed at a very similar level, the GBM model had better predictive performance. This better performance is important because this translates to a relatively clearer idea as to the computer brand preferences for the customers.

Besides the aforementioned prediction metrics, it was possible to see how each of the models prioritized different features of the dataset. This is how each model prioritized each of the features:



At a glance, we can clearly see that the “salary” and “age” features were the important variable for each of these models. This means that, the salary and age of a customer may have the highest relative effect on whether or not they will prefer a Sony or Acer computer.

Given that I have chosen the GBM model to predict the customer computer brand preferences, here are the results:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Acer | Sony | Total |
|  |  |  |  |
| Complete customer survey (actual) | 3744 | 6154 | 9898 |
|  |  |  |  |
| Incomplete customer survey (predicted) | 1921 | 3079 | 5000 |
|  |  |  |  |
| Total | 5665 | 9233 | 14898 |
|  |  |  |  |

As we can see from the customer preferences table above, Sony ended up being the most popular computer brand. What is particularly interesting, and perhaps even reassuring as far as the performance of the GBM model is concerned, is that the ratios of preference between the two datasets are similar. That is, for the complete dataset, 62% of customers (out of 9,898) preferred Sony while only 38% of customers preferred Acer. When using the GBM model to predict the customer preferences between the two computers, we get a similar outcome. That is, for the incomplete dataset, approximated (predicted) preferences were as follows: 61% of customers (out of 5,000) preferred Sony while only 39% preferred Acer.

That being said, even though the GBM model did not predict wildly divergent or unreasonable results, it is necessary to remind ourselves that no prediction --- and no model or analytics process --- is absolute and 100% guaranteed to result in perfect, totally reasonable or always desired outcomes. So, it is best for further research and analytics to be done to further clarify and verify the usefulness of predictions.