Hello Danielle,

I just wanted to report back on my findings for the customer preferences, stemming from the survey results we received. For this project I used a couple of different classifiers in RStudio to find the projected results we needed from the incomplete questionnaire surveys. For this report I will go into some detail about the quality of the results I found, my preference for the classifier to be used, and a brief overview of the brand preferences based on customer feedback.

With this project we were trying to determine which brand of laptops customers would prefer: Acer or Sony. The initial problem we had, as you know, was that customer questionnaires were incomplete; meaning the best way for us in a data analytics perspective to find the answers was to use a classifier, RandomForest and kNN in this case, to predict the most likely answers our customers may have. Starting off I reviewed the data that would be used for the training and test sets and analyzed it to see if there was any redundant data that could cause overfitting. While I did not find anything wrong with the data I did notice that the correlation for every attribute was very low and so decided not to remove any attributes as I thought it would harm the training and testing sets of data.

Now the kNN classifier we have used previously in a project in Weka and was very useful to us at the time, but I did not find it as helpful in R for this work. For one the accuracy and kappa scores were good, but not great with the top results being: 0.87 and 0.72 for Accuracy and Kappa respectively. I tried the resampling function in R and found the results to be worse with the accuracy at 0.19 and Kappa of 0. So based on these results I went back and reviewed the data and tried rethinking how I created the kNN classifier for this project just in case something had been missed that could be causing an issue, but the results did not improve.

Next I tried using the RandomForest classifier in R. What I found was that besides the results being better for me that the RandomForest ran easier than kNN by leaps and bounds. For instance the syntax of the code was easier to follow and implement. The resulting performance metrics I found at their best Accuracy was 0.92 and Kappa 0.83 with 1,000 trees, having also tried 200, 500, and 700 trees. The resampling function found an Accuracy of 0.91 and Kappa of 0.82. I cannot say why the resampling of the RandomForest classifier worked so well over the kNN other than the basic differences in how the two classifiers work, but based on the ease of use of the code and the performance metric results it is not really a contest that the RandomForest is the classifier I decided to go with for this project.

With the RandomForest classifier chosen I used it to create the prediction data that used the incomplete customer questionnaire to find the projected answers we needed. Based on those results for Acer and Sony customers I copied them into the incomplete customer survey to find demographic data. This resulted in some interesting and surprising observations. One example of something that surprised me was that the salary distribution was so different between Acer and Sony customers. The maximum salary for Acer was $125,839 with an average of $74,689 and Sony customers had a maximum of $150,000 with an average of $92,251. And I found that Sony customers being more financially well off than Acer customers was a trend that continued with their credit levels on average being higher, they drive nicer cars, and are more likely to have a higher education (college and professional degrees). Something that I found interesting was the overall breakdown of survey data broken down by whether it came from the complete or incomplete predicted survey answers and whether it was for a customer that favored Acer or Sony laptops, which all this information could not be found until the predicted data was generated from the incomplete survey. I found that altogether between the completed survey that had the actual data and the incomplete survey with its predicted data there were 15,000 instances of choices made for Acer and Sony branded laptops. For Acer the combined choices were 5,688 or 38% of total instances. While Sony branded laptops were chosen 9,312 or 62% of the time. See graph 1 for the compiled survey findings.

|  |  |  |
| --- | --- | --- |
|  | **Acer** | **Sony** |
| **Complete Survey (Actual)** | 3783 | 6217 |
| **Incomplete Survey (Predicted)** | 1905 | 3095 |
| **Total** | 5688 | 9312 |
| **Percentage** | 38% | 62% |

Graph 1

All of this information and more may be found in an Excel workbook that I will include with this email that details all the findings from my research as well as the code created in R. And if you have any questions or comments just let me know and I will help as best I can.

Regards,

Edward Tanzer