1. Task Description

Hello,

Let me begin by thanking you for your investigation into the use of R in our day-to-day data analytics activities. I believe it's going to be a useful addition, and your investigation was integral in my decision to bring on board as one of our main tools. We will not be replacing RapidMiner completely, since RapidMiner is a very useful tool for visualization and analytics, but we will be using R and R Studio going forward in the next project since we have some deep analytics work to do. Speaking of that…

The sales team engaged a market research firm to conduct a survey of our existing customers. One of the objectives of the survey was to find out which of two brands of computers our customers prefer. This information will help us decide with which manufacturer we should pursue a deeper strategic relationship. Unfortunately, the answer to the brand preference question was not properly captured for all of the respondents.

That is where you come in: I want you to investigate if customer responses to some survey questions (e.g. income, age, etc.) enable us to predict the answer to the brand preference question. If we can do this with confidence, I would like you to make those predictions and provide the sales team with a complete view of what brand our customers prefer.

To do this, I would like you to run and optimize at least two different decision tree classification methods in R - C5.0 and RandomForest - and compare which one works better for this data set.

I have already set up the data for you in the attached CSV files: the file labelled CompleteResponses.csv is the data set you will use to train your model and build your predictive model. It includes ~10,000 fully-answered surveys and the key to the survey can be found in survey_key.csv. The file labelled SurveyIncomplete.csv will be your main test set (the data you will apply your optimized model to predict the brand preference). You'll be applying your trained and tested model to this data to prepare the model for production.

When you have completed your analysis, please submit a brief report that includes the methods you tried and your results. I would also like to see the results exported from R for each of the classifiers you tried.

Thanks,
Danielle

Danielle Sherman
Chief Technology Officer
Blackwell Electronics
www.blackwellelectronics.com

2. Task Solution

Two data sets (SurveyIncomplete.csv and CompleteResponses.csv) were provided by the CTO, to conduct a classification analysis to predict customer preferences over a computer brand.

This report contains the results and results discussion section as well the recommendations section, the model will be analyzed in these sections.

A dataset named "CompleteResponses" contains 9,898 survey responses, this file will be used to develop a classification model to predict the incomplete surveys provided in the dataset "SurveyIncomplete", this file contains 5,000 incomplete surveys.

The solution for this task will be conducted using two different machine learning algorithms and then will be compared to identify variances between both models.

Model 1 – Method C5.0 on the training set with 10-fold cross validation and an Automatic Tuning Grid
Model 2 – Random Forrest with 10-fold cross validation and manually tune 5 different MTRY values

2.1 Models Configuration

Model 1 – Decision Tree with method C5.0

This model contains the following processes and parameters:

- Create Data Partition (75/25 split)
- Pre-processing Steps (RFE, normalization)
- Train Control (method = "repeatedcv", number = 10, repeats = 1)
- Train (method='C5.0', tuneLength = 7)
- Quality Metrics (Model metrics, VarImp, Confusion Matrix)
- Predictions (Test Data and Survey imcomplete)
- Post Resample

Model 2 – Random Forrest

This model contains the following processes and parameters:

- Create Data Partition (75/25 split)
- Pre-processing Steps (RFE, normalization)
- Train Control (method = "repeatedcv", number = 10, repeats = 1)
- Train (mtry=c(1,2,3,4,5), method='RF')
- Quality Metrics (Model metrics, VarImp, Confusion Matrix)
- Predictions (Test Data and Survey imcomplete)
- Post Resample

3. Results

Both models were run under the specified conditions, the customer preferences are plotted in the image below, there is a variance in terms of customer predictions of 0,6% (59 observations) between the models.

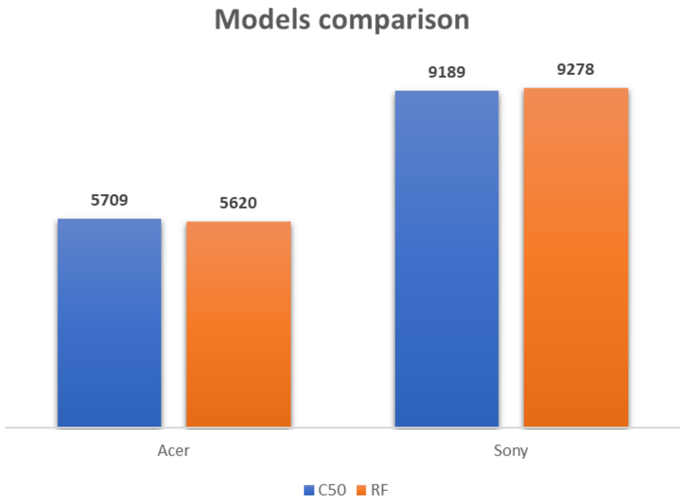


Figure 1 – Customer Preferences (Models Comparison)

The quality metrics of each model are:

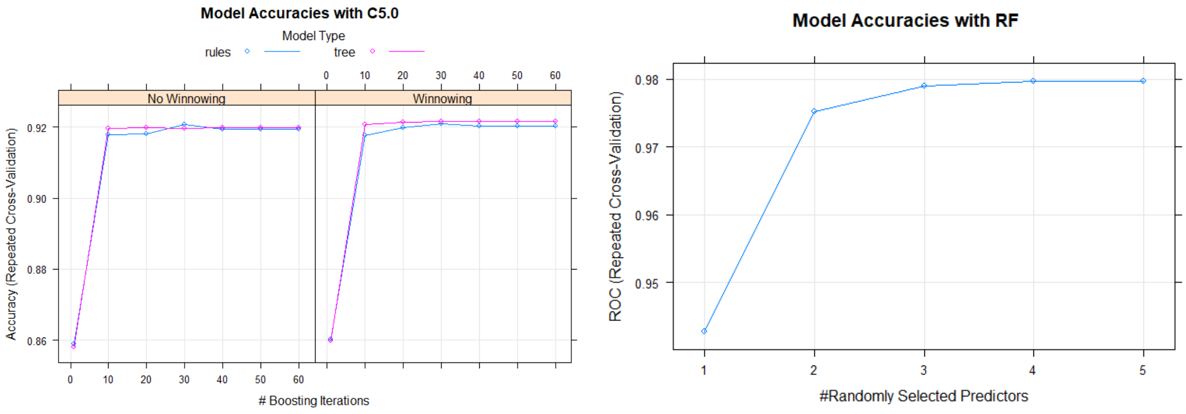


Figure 2 – Model accuracies for C5.0 and RF

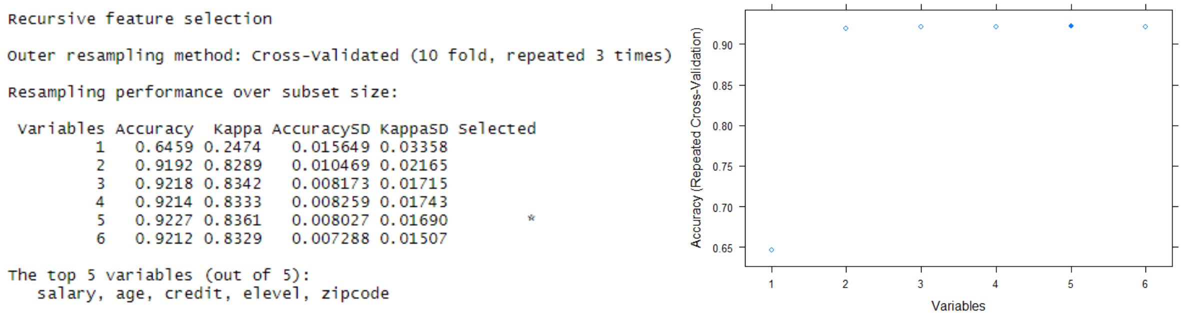
Figure 2 contains the model accuracies for each model, the model RF has the higher accuracy with a value of 0,97 in MTRY= 4 while the model C5.0 has a value of 0,92 in trial 30, tree and winnowing.

In the post resample process the C50 model was the one with the highest accuracy 0.9244 and highest Kappa value as is show in the figure 3. Meaning the most accurate model between these two models.

	C5.0	**RF**
Accuracy	0.9244139	0.9223929
Kappa	0.8408114	0.8350861

Figure 3 – Post Resample

The recursive feature selection (RFE) showed the top variables accuracy (Figure 4), in both model 5 out of 6 variables were important (Salary, Age, Credit, Elevel and Zipcode) the variable with the lowest accuracy is Car with an accuracy of 0.64 and a Kappa of 0.24.



```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 3 times)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.6459 0.2474   0.015649 0.03358
         2   0.9192 0.8289   0.010469 0.02165
         3   0.9218 0.8342   0.008173 0.01715
         4   0.9214 0.8333   0.008259 0.01743
         5   0.9227 0.8361   0.008027 0.01690        *
         6   0.9212 0.8329   0.007288 0.01507

The top 5 variables (out of 5):
   salary, age, credit, elevel, zipcode
```

Figure 4 – Recursive Feature Selection Results

The variable importance for model C5.0 shows that Age, Salary, car and zipcode are the most important variables; while the model RF shows that Salary, Age, Credit, Car and Zipcode are the most important variables. As is show in figure 5.

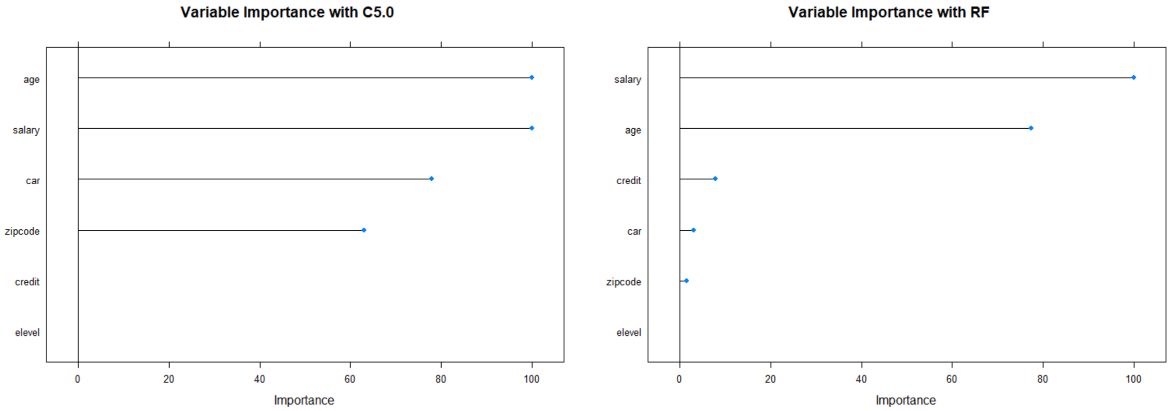
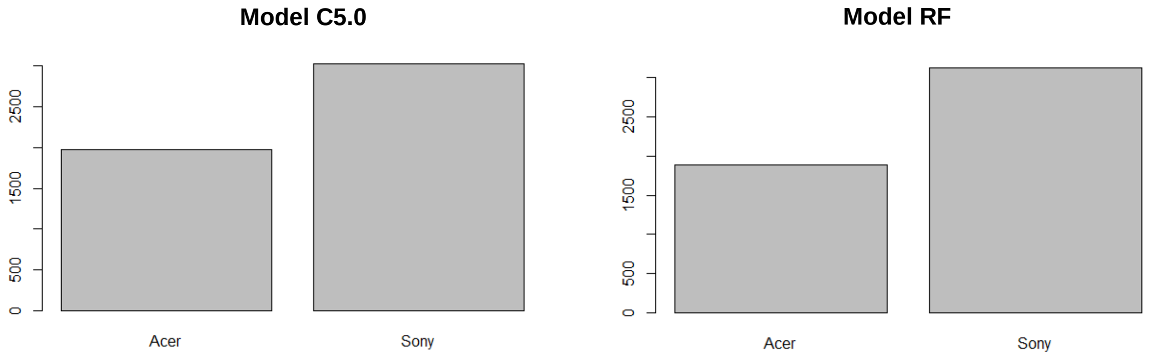


Figure 5 – VarImp for Model C5.0 and Model RF



	Model C5.0	Model RF	Variance
Sony	3035	3124	2.85%
Acer	1965	1876	4.53%

Figure 6 – Prediction distribution for model C5.0 and model RF

The final prediction in the “Surveyimcomplete” file shows that prediction for Acer in Model C5.0 are 1,965 units while in Model RF are 1,876, this is a variance of 89 units (~ 4,53%) between RF and C5.0 model. In the other hand the predictions for Sony are 3,035 and 3,124 respectively, this is a variance of 89 units (~2,85%).

The final predictions shows a net variance of ~0,97% in the customer preferences distribution of Sony (89 units ~0,97%) and Acer (- 89 units ~0,97%) between both Model C5.0 and Model RF.

	Model C5.0	Model RF	Variance
Sony	9,189	9,278	0,97%
Acer	5,709	5,620	0,97%

Figure 7 – Customer preferences for model C5.0 and model RF

4. Results Discussion

The models developed to conduct the prediction of the incomplete survey shows a variance of ~0,6% between the model accuracy and ~0,2% in the kappa value, depending of the business context these values could be negligible.

In the context of this exercise these values show a variance of 89 units, which is low. However, for a larger dataset and a different business context this could represent and inconvenience.

5. Recommendations

Use the model RF to conduct the predictions since the Accuracy and Kappa values were the best values [higher Accuracy and agreement between the appraisers (Kappa)].

6. R scripts

Provided in a zip file