Message

From: Danielle Sherman

Brand Preference Prediction

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
Attachments

POA

Project Roadmap

Your Task

You have been asked by Danielle Sherman, CTO of Blackwell Electronics, to predict the customers' brand preferences that are missing from the incomplete surveys by conducting two classification methods in R. Once you have determined which classifier — C5.0 or RandomForest —works better on the provided data set, she would like you to predict the brand preferences for the incomplete survey responses and prepare a report of your findings.