Course 3, Task 2: Report to Blackwell

Predicting Customer Brand Preferences

Anders Dowd - Spring 2022 Data Analytics Certification Program

(A) Overview

The Blackwell Electronics' sales division enlisted the help of a market research firm to conduct a consumer survey regarding computer brand preference. The survey consisted of a handful of demographic questions as well as a choice between two well-known computer brands. Unfortunately, about one third of the responses were incomplete or corrupted somehow. By employing machine learning in RStudio, we shall attempt to deduce the likely responses of the customers with missing data using the completed surveys as a training set.

(B) Data

We were provided with two data sets. One consisted of customer info from approximately 10,000 completed surveys, while the other contained completed demographic data but corrupted responses to the brand preference question. The features available were salary, age, education level, car model, zip code, and credit limit. Preprocessing the data consisted of ensuring all features and the target variable were the proper data type to work with the algorithms in both data sets, then splitting the complete set into training and test sets. Ultimately, the final model would be utilized on the incomplete data set to predict the missing values.

(C) Modeling

Our goal is to determine whether a group of customers is more likely to prefer Sony or Acer computers, making this a straight-forward classification problem. There is no shortage of available classification algorithms in R, but for our purposes we chose three to build our models with: Gradient Boosting, Random Forest, and C5.0.

I proceeded by running out-of-the-box models with the three algorithms, then checking variable importance on each one using the varImp() function. There was concurrence regarding the top three features, but the fourth and fifth varied by algorithm. I created unique feature groups for each algorithm based on the results of the varImp(). Once features were selected, I tuned each model and ran cross-validation to determine the best parameters and then compared the results to select the best fit model.

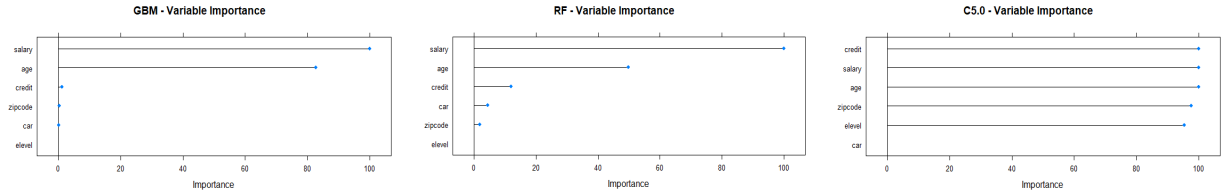

Figure 1: Comparative Variable Importance Between Models

(D) Method 1: Gradient Boosting Classifier

I first built GBM models using both automatic and manual tuning grids and a 10-fold cross-validation. Ultimately the best model I produced consisted of a tune length of 10 with 50 trees set to an interaction depth of 6 and a minimum observations per node of 5. This model returned an **Accuracy Score** of 0.9247059 and a **Kappa Score** of 0.8405758.

(E) Method 2: Random Forest Classifier

I then built random forest models using both automatic and manual tuning grids and a 10-fold cross-validation. Ultimately the best model I produced used mtry=2 and returned an **Accuracy Score** of 0.9216103 and a **Kappa Score** of 0.8334943.

(F) Method 3: C5.0

The first two algorithms produced highly successful models, but for the sake of thoroughness I also built a C5.0 classification model. The best C5.0 model I produced used a tune length of 10 with 10 trials on a tree model and winnow=TRUE. It returned an **Accuracy Score** of 0.9230882 and a **Kappa Score** of 0.8363281.

(G) Model Selection and Generating Predictions

All three models were highly successful, but the Gradient Boosting Classifier had slightly better numbers and, once tuned, significantly faster run time. Just to make sure it was all it appeared to be, I checked its ROC, sensitivity, and specificity scores. They returned 0.9755921, 0.9109786, and 0.9328380 respectively, making me confident in the model's performance. The next step was to test it against my test set. The model proved to hold up on the test data, returning an **Accuracy Score** of 0.9316896 and a **Kappa Score** of 0.8548718.

Finally, now confident in the model's ability, I implemented it on the incomplete data set to predict the missing brand preferences. The proportion of Acer to Sony was similar in the predictions to the ground truth in the complete data set, which bolsters confidence that the model is reliable. In conclusion, Sony is the preferred brand by more than a 3:2 ratio. Of the 5000 missing data points, the model predicts 3090 would prefer Sony to 1910 for Acer.

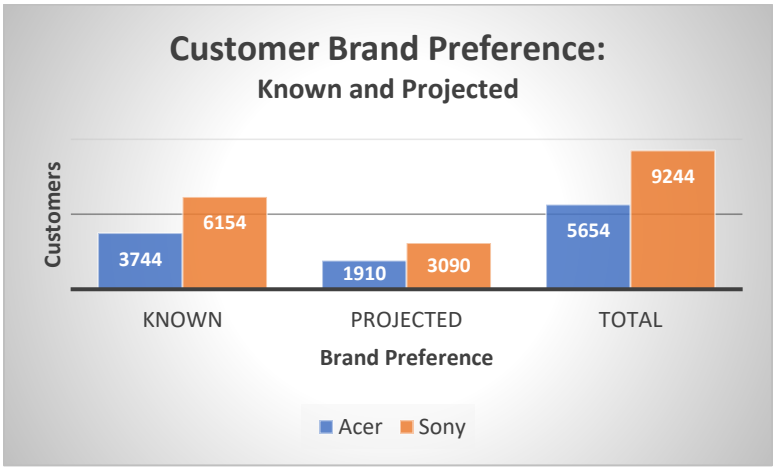

Figure 2: Customer Brand Preference