|  |  |  |
| --- | --- | --- |
|  |  |  |
| Ramses Meza  May 2021  <https://github.com/lordtable/UTDSGC_C3T2> |  | Data Analytics Course 3.2: Predicting Customer Brand Preferences |

Business Acumen

Blackwell Electronics wishes to pursue a deeper strategic relationship with one of two preferred major computer manufacturers: Acer and Sony. In order to make an informed decision on which manufacturer to choose, the sales team commissioned an outsourced survey of existing customers to know which is their preferred brand, along with other demographics information. However, a significant portion of the survey was found to have the brand preference question incorrectly captured. The goal of the current Data Science project (using RStudio) is to use the reliable portion of the survey to build a model able to predict the customers preferred brand, and if successful, use this model to create synthetic brand preferences responses on the portion of the survey that contains a corrupted brand preference information. Finally, provide a global outcome of the preferred computer brand.

**Data Management, Cleaning & Pre-processing**

The datasets were provided on a .csv file format and locally stored. Two files were provided: one containing the uncorrupted portion of the survey (called *CompleteResponses*) and other file containing the survey with the corrupted ***brand*** feature (called *SurveyIncomplete*). Both files are structured in a table format: each row represents an undisclosed surveyed customer. Each column represents an attribute or feature surveyed, such as:

* **salary**: Customer’s yearly salary in US$ [numeric]
* **age**: Customer’s age [numeric]
* **elevel**: customer’s highest education level obtained [categorical]
* **car**: Make of customer’s primary car [categorical]
* **zipcode**: Customer’s zipcode based on US regions [categorical]
* **credit**: Amount of credit available in US$ [numeric]
* **brand**: Preferred computer brand. This is the target feature: either Acer or Sony [categorical]

Each survey file was managed as a RStudio Dataframe. The only data cleaning/pre-processing needed consisted on the data type transformations to factor of the ***brand*** feature. Early screening of the data using the *summarytools* library indicated that no missing values, no duplicated instances were detected. The *CompleteResponses* dataset to be used for model training and testing consists of 9,898 observations/instances. Of these, 75% were retained as instances used for training all the predictive models to be later described, while 25% of the instances were kept for models’ testing and evaluation. The *SurveyIncomplete* dataset has the same overall data quality description as *CompleteResponses*, but having the ill-captured ***brand*** feature. It contains 5,000 observations/instances. The brand feature values of this file will be replaced by synthetic or model-generated values generated by a machine learning classifier, subject to the success on designing one based on the *CompleteResponses* dataset.

**Exploratory Data Analysis (EDA)**

EDA was performed on the *CompleteResponses* dataset using combinations of R libraries, such as *summarytools*, *explore* and *dlookr*. The first goal was to screen for data errors. Secondly, to get familiarity with each feature characteristics, such as ranges and normality checks. Then, attention was focused on features inter-dependence, aiming for feature selection and/or dimensionality reduction. It was found that variables are poorly correlated to each other. Finally, using the *explore* library, it is possible to set the ***brand*** feature as target variable and build a very quick-look Decision Tree (DT) in order to highlight features most likely to influence a potential classification. Figure 1 shows such quick DT, where the ***salary*** and ***age*** features could be the main drivers on a subsequent predictive model.

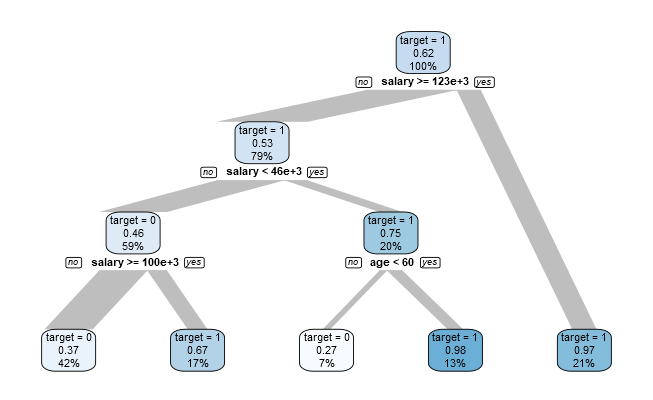


Figure 1: Quick-look Decision Tree from the *dlookr* library

**Models Tuning, Assessment & Final Model Selection**

Two classification modelling algorithms were chosen to be built using the same selected features to predict the ***brand*** variable: Stochastic Gradient Boosting (GBM) and Random Forest (RF). I took advantage of the fact that most R models have a built-in capability for tuning parameters, that allows parameters selection based on a systematic testing and performance based on the accuracy metric for each parameter set tested. The tuning was executed using the training data as follows:

* **Stochastic Gradient Boosting (GBM)**: Using a 10-fold cross-validation and an Automatic Grid parameter testing, which tested for the *number of trees* and *interaction depth* best permutations based on the accuracy and kappa values for each combination. The best performing (highest accuracy and kappa values) tuned GBM model resulted from an *interaction depth*=3 and *number of tress*=150. The corresponding accuracy= 0.9193140 and kappa=0.8294875.
* **Random Forest (RF)**: Using a 10-fold cross-validation and manually tune for 5 different *mtry* values (1 to 5), the best performing (highest accuracy and kappa values) tuned RF model resulted from *mtry*=3. The corresponding accuracy= 0.9185089 and kappa=0.8269785.

I validated each tuned model by applying them to the test dataset, and obtained each model prediction metrics yielded from each model confusion matrix. These metrics are shown on Table 1. Even though both models perform very well and are very comparable in terms of predictive metrics, I decided to perform the final classification using the Stochastic Gradient Boosting (GBM) tuned model, mainly because of its much faster execution time.

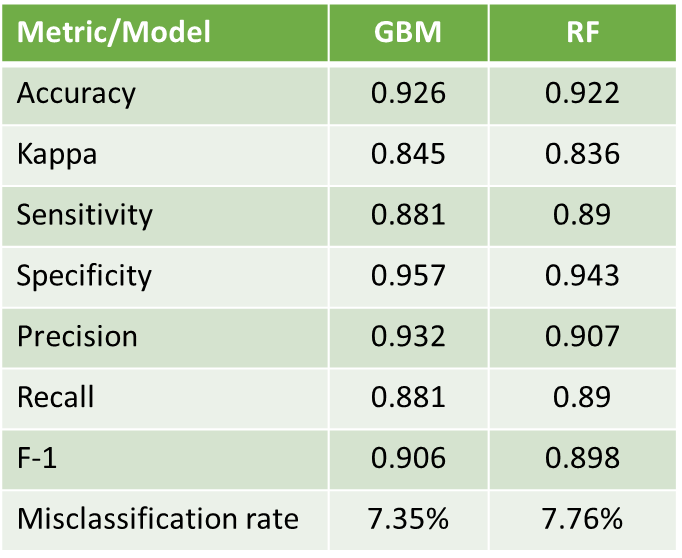


Table 1: Tuned-models metrics performance based on validation/testing data

For this model, the most important predictors or features are ***salary*** and ***age***, as shown on Figure 2, in agreement with the insights drawn during the EDA. Feature ***credit*** plays a very minor role while ***car***, ***zipcode*** and ***elevel*** play insignificant roles as predictors.

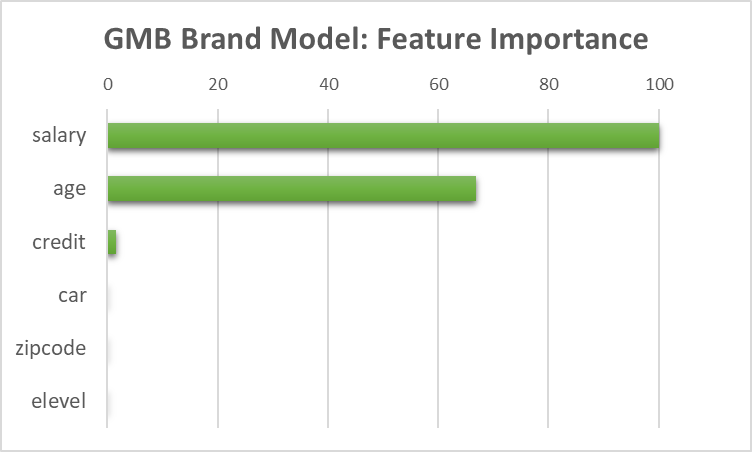


Figure 2: Variable Importance based on the final, tuned GBM model.

**Synthetic Brand Survey Responses & Overall Final Brand Preferences**

The final GBM classifier model was applied to a dataframe copy of the *SurveyIncomplete* dataset, effectively replacing the corrupted or ill-captured ***brand*** preferences responses by synthetic responses generated by the calibrated and tuned GBM model. A before and after comparison of the feature for this portion of the survey can be made based on Table 2: the (***brand***) corrupted or ill-captured survey is strongly unbalanced (~99% of instances correspond to Acer). After the application of the GBM model, the synthetic brand responses are now more balanced, now with a significant majority preferring Sony (~61%).

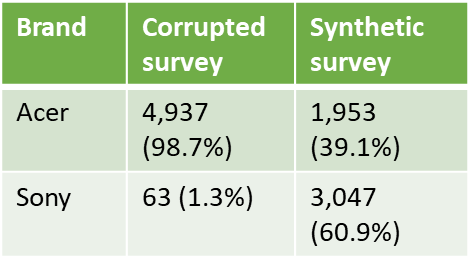


Table 2: Brand preferences instances before and after application of the GBM model to the corrupted or ill-captured portion of the survey.

This portion of the survey containing the synthetic brand preferences responses can now be merged with the *CompleteResponses* dataset in order to yield a broader and more reliable evaluation of customers brand preferences. Figure 3 shows a stacked bar chart that clearly shows that about 62% of the total surveyed customers prefer Sony. The reliable predictive GBM model allowed to yield synthetic responses that permitted the incorporation of all the surveyed customers, **and unmistakably point towards Sony as the brand to engage with on deeper strategic partnerships**.

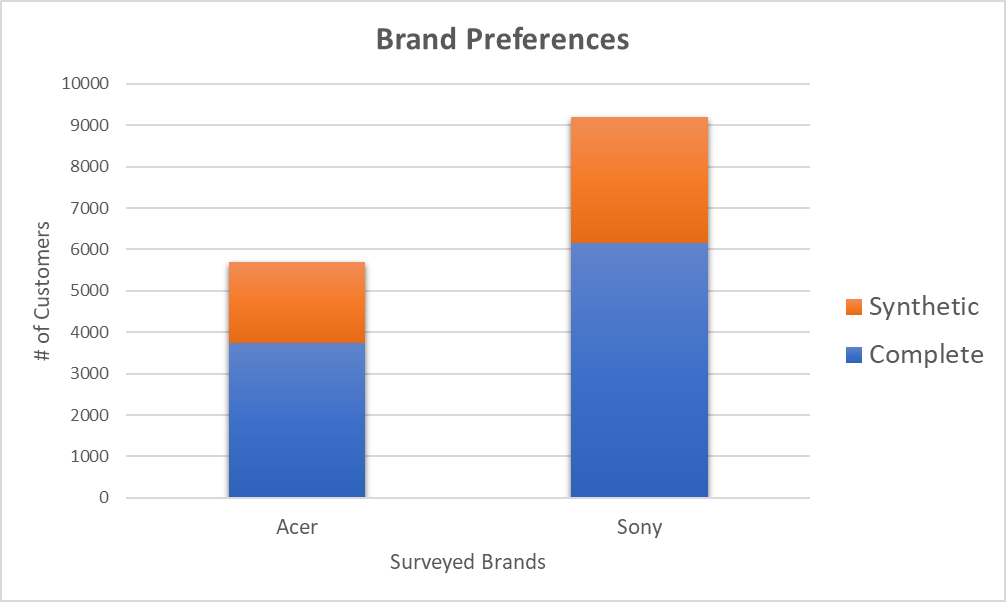


Figure 3: Final Customers Brand Preferences from all surveyed customers.