Brand Preference Prediction

Summary

- The predictive exercise produces a model with high accuracy to predict the brand of computer a customer prefers.

- In order to improve accuracy, we customized the models from its default mode, applied feature selection to optimize the data for the model, and compare the results.

Problem Statement

Our task is to investigate and create a highly accurate model using survey information to predict which computer brand customers would prefer. This model would be an extra tool for Our client's sales force to understand customer's preferences in computer choices.

Data Sets Description

For the models' training and testing we are using the data set provided by Our client via csv file that includes 9.9K records (observations), each related to a survey participant. Each record contains demographic information (education, age, us region they live at), brand of main car owned, total credit available, and computer brand preferred.
A second set with similar information, but for 5K respondents and missing accurate answers for the computer brand portion, was also provided. This data set will be the first real case we will apply the computer brand predictive model we will select.

Modeling – Data Preparation

Based on the EDA, and thru the use Recursive Feature Elimination (RFE), we created a separate dataset where the **car** feature was not included. In addition, there is no strong correlation among the features, so there is no need to remove any other feature due to this reason. We also converted all categorial features to factor, so they match the classification algorithms targeted to predict the computer brand (a categorical variable).

We did not normalize data, nor did we do any feature engineering on the data at this point since most is categorical and the numeric data did not exhibit much in terms of outliers.

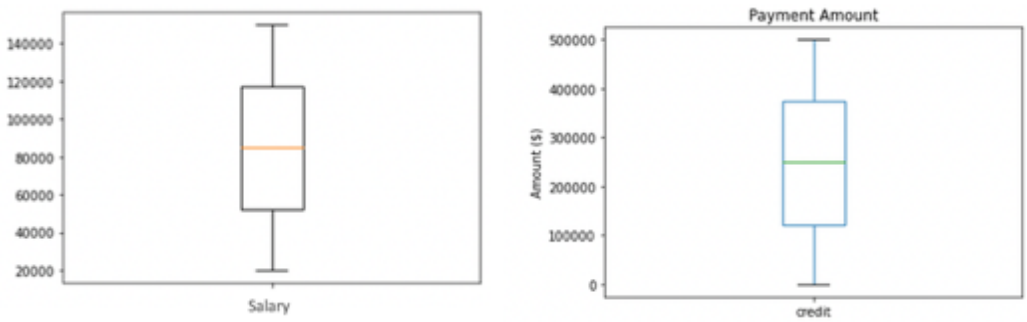


Figure 1: Boxplots – Salary & Credit limit

Modeling – Choosing the model

We ran three models: **Gradient Boosting (GB), C5 & Random Forest (RF)** using the caret package's train function. Each model was run first in its out-of-box mode, and then we changed parameters aimed to improved accuracy (see "modified" or "mod." In Tables 1 & 2). For all models, we set the train control for resampling to be a 10-fold cross validation, repeated 3 times.

We ran the models using the normal data set as received from the csv file, and then we ran the same models using the RFE dataset described earlier. The results are shown in tables 1 & 2, respectively. In the case of the parameter changes, and for the **RF model**, we added **mtry** values (number of variables used per tree in the forest) around the space where the OOB version showed as best accuracy. In a similar manner for **GB,** and after the OOB run, we adjusted the model by modifying the **max. tree depth**, the **number of boosting interactions**, **shrinkage**, and **min. number of observations in the trees' terminal node**s. Finally, in the case of **C5**, we again look at the OOB results and we adjusted the **number of trees to generate**, and **model type** to be used. The actual parameter values are in the accompanying R pipeline.

Classifier Model (RFE)	Accuracy (%)	Kappa (%)
Gradient Boosting Mod.	**92.75%**	**84.66%**
C5 Modified	**92.66%**	**84.39%**
Gradient Boosting OOB	92.39%	83.92%
C5 OOB	92.35%	83.72%
Random Forest Mod.	92.24%	83.55%
Random Forest OOB	92.11%	83.25%

Table 1: RFE (no "car" feature) test sample

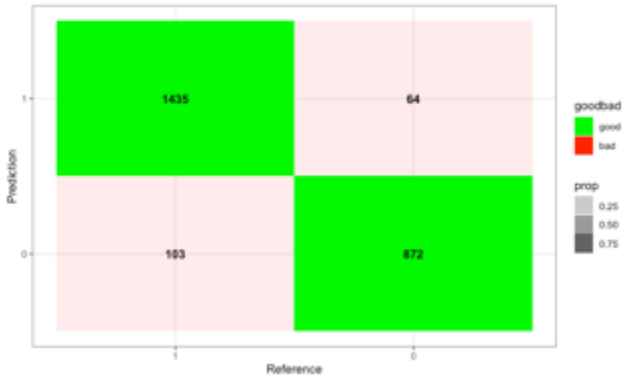
Classifier Model (non-RFE)	Accuracy (%)	Kappa (%)
Gradient Boosting Mod.	**92.470%**	**84.050%**
Random Forest OOB	**92.187%**	**83.390%**
Random Forest Mod.	92.121%	83.271%
C5 OOB	91.972%	82.926%
Gradient Boosting OOB	91.932%	82.968%
C5 Modified	91.729%	82.387%

Table 2: Full-feature test sample

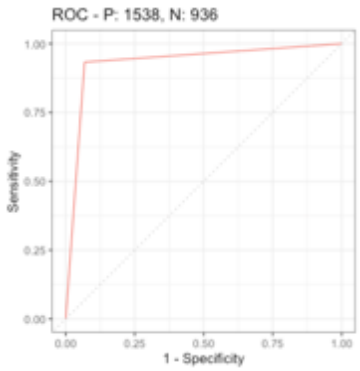
Looking at Tables 1 & 2 above, the first result is that regardless of data set used, and for the most part, the modified models performed better than the OOB version in terms of both **accuracy** and the positive and significant concordance of the prediction with the actual model output (**Kappa**). Second, the models using the data set with reduced features (RFE) were more accurate than the models using the original data set. Last, the models in blue in Tables 1 & 2 are the selected models to run thru validation in order to choose a preferred one.

Modeling – Validation

The formal comparison of both pairs of models (RFE, and Non-RFE) using caret's resample function resulted in that **both modified Gradient Boosting** are the most accurate models (92.2% and 92.6% for RFE and non-RFE respectively). The next step was to test both modified GB winning models using the testing set. The result was that the modified GB model, using the full-feature data set, had a higher accuracy (93.25% vs. 92.52%) and kappa values (85.77% vs. 84.2%).
Hence, **we proceed to the prediction step using the full-feature set GB modified model**.



Accuracy = 93.25% Kappa = 85.77%



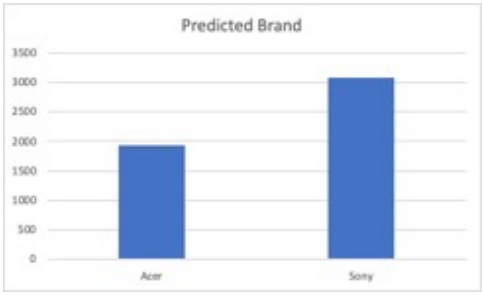
AUC = 0.932

Figure 2: Full-feature GB Mod. Confusion matrix **Figure 3**: Full-feature GB Mod. ROC Curve

The ROC curve has a healthy separation from the random-choice line, and the area under the curve (AUC) total is 0.932. Both characteristics are very good signs of the predictability strength of the model, and it aligns with the 93.25% accuracy discussed before.

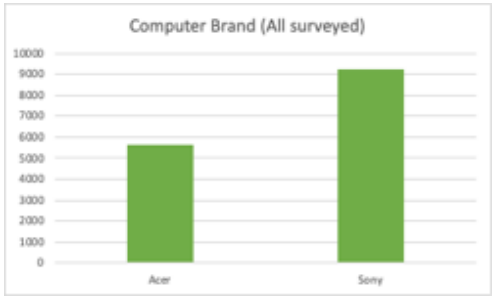
Prediction

The chosen **full-feature-data-set based modified GB** model was used to predict the computer brand for the data set given which did not have this information accurately recorded. The results show that the **proportion of brand preference is the same** for either the incomplete survey or the full survey (which includes the complete data set).



Brand	Count	% Total
Acer	1,922.00	38.44%
Sony	3,078.00	61.56%
TOTAL	**5,000.00**	100.00%

Figure 4: Predicted brand totals – Incomplete Survey



Brand	Count	% Total
Acer	5,666.00	38.03%
Sony	9,232.00	61.97%
TOTAL	**14,898.00**	100%

Figure 5: Brand totals – Incomplete Survey with predicted, plus the Complete Survey data

Additional Technical Details

The provided training/testing data set has the following characteristics:

Number of observations	9,898 (each a single person credit behavior)
Attributes	7 [edu. Level, salary, age, car owned, home zip code, credit limit, computer brand preferred]
Duplicates	0
Bad/Missing data	0
Unused Attributes	Car in the case of Feature selection

Table 3: Correlation Table

The provided data with incomplete Brand feature to apply the model to has the following characteristics:

Number of observations	5,000 (each a single person credit behavior)
Attributes	6 [edu. Level, salary, age, car owned, home zip code, credit limit] Computer brand preferred is not valid.
Duplicates	0
Bad/Missing data	0
Unused Attributes	Car in the case of Feature selection

Table 4: Correlation Table

Correlation result

The strongest correlation was{Salary, Brand} = 0.206 shown in the figure below.

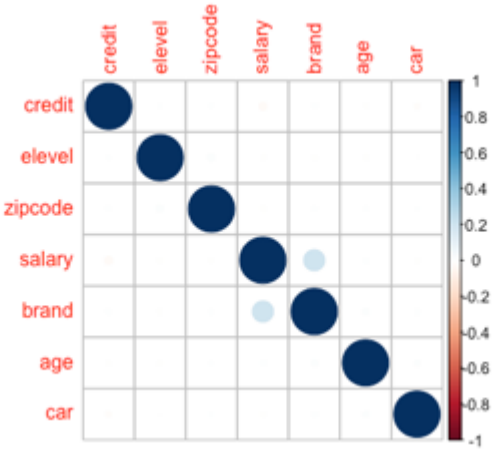


Figure 6: Correlation

	salary	age	elevel	car	zipcode	credit	brand
salary	1.000000000	0.007978566	-6.620234e-03	-6.090575e-03	-0.005471132	-0.025126808	0.206489883
age	0.007978566	1.000000000	-5.830340e-03	1.024607e-02	0.003681375	-0.004400692	0.013713286
elevel	-0.006620234	-0.005830340	1.000000e+00	-4.676852e-05	0.018095400	0.002720642	-0.004828912
car	-0.006090575	0.010246067	-4.676852e-05	1.000000e+00	0.001526528	-0.010329137	0.005923147
zipcode	-0.005471132	0.003681375	1.809540e-02	1.526528e-03	1.000000000	0.004962011	0.004665088
credit	-0.025126808	-0.004400692	2.720642e-03	-1.032914e-02	0.004962011	1.000000000	0.005688438
brand	0.206489883	0.013713286	-4.828912e-03	5.923147e-03	0.004665088	0.005688438	1.000000000

Table 5: Correlation Table

Model Comparison results

```
Accuracy
       Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
gb 0.9204852 0.9222746 0.9252529 0.9268576 0.9283311 0.9434724
c5 0.9044415 0.9177898 0.9259268 0.9243006 0.9286194 0.9461642

Kappa
       Min.   1st Qu.    Median      Mean   3rd Qu.      Max. I
gb 0.830909 0.8356892 0.8409393 0.8449626 0.8478073 0.8798124
c5 0.796115 0.8252986 0.8410063 0.8384658 0.8479524 0.8848936
```

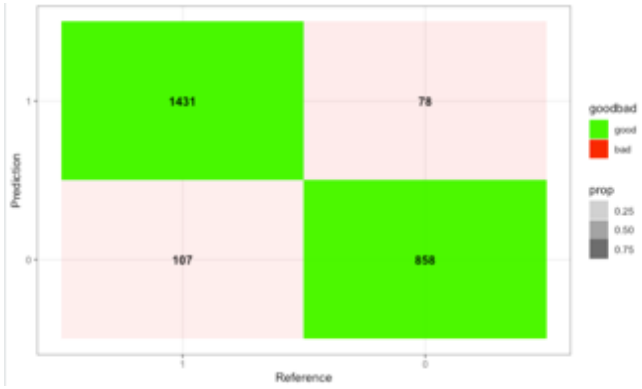
Figure 7: RFE Model Comparison

```
Accuracy
       Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
gb 0.9083558 0.9195011 0.9212125 0.9225477 0.9255884 0.935397
rf 0.9083558 0.9134378 0.9192463 0.9202574 0.9258760 0.935397

Kappa
       Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
gb 0.8057906 0.8291816 0.8335149 0.835822 0.8416918 0.8635936
rf 0.8044280 0.8178265 0.8283053 0.830916 0.8427723 0.8624509
```

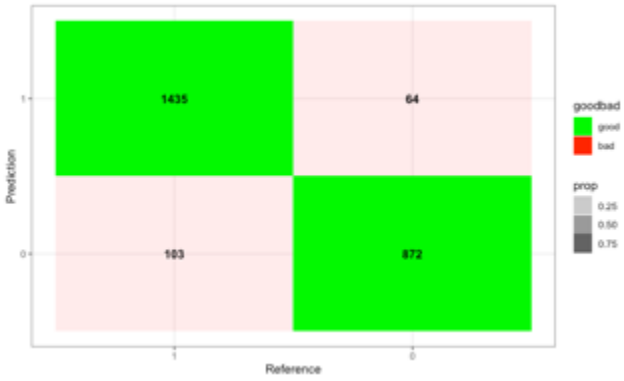
Figure 8: Full-feature Model Comparison

Test Confusion Matrix



Accuracy = 92.52% Kappa = 84.20%

Figure 9: RFE - GB Modified



Accuracy = 93.25% Kappa = 85.77%

Figure 10: Full-feature - GB Modified