Customer Preference

Predict customer brand preference

Rayon Susan Koshy

**OBJECTIVE**

* Analyze the most preffered brand by the customers
* Build model to predict the brand preference to fill the incomplete survey data set

Contents

[**I.** **DATA SET & FEATURES** 4](#_Toc43834256)

[**II.** **CORRELATION BETWEEN THE VARIABLE** 6](#_Toc43834257)

[**III.** **BRAND Vs SALARY** 7](#_Toc43834258)

[**IV.** **MODELLING** 8](#_Toc43834259)

[**V.** **FEATURE IMPORTANCE** 9](#_Toc43834260)

[**VI.** **SUMMARY** 10](#_Toc43834261)

1. **DATA SET & FEATURES**

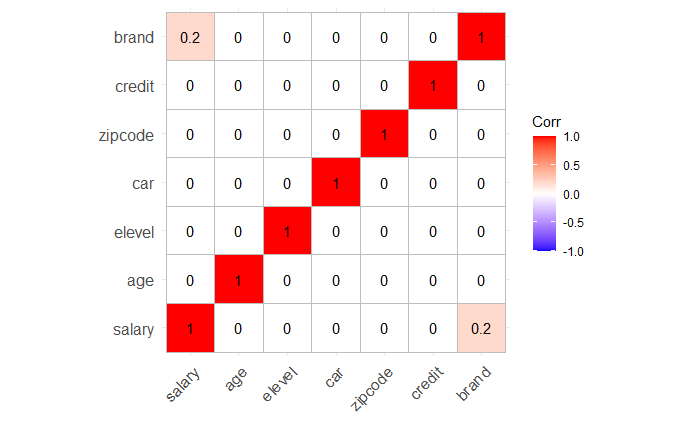
The below table gives attribute information:

|  |  |  |
| --- | --- | --- |
| Variable | Description | Category |
| SALARY | Yearly salary of the customer |  |
| AGE | Age in year |  |
| SEX | Gender | 1 = male  2 = female |
| ELEVEL | Educational qualification of the customers | 0 = Less than High School Degree  1 = High School Degree  2 = Some College  3 = 4-Year College Degree  4 = Master's, Doctoral or Professional Degree |
| CAR | Primary car of the customer | 1 = BMW  2 = Buick  3 = Cadillac  4 = Chevrolet  5 = Chrysler  6 = Dodge  7 = Ford  8 = Honda  9 = Hyundai  10 = Jeep  11 = Kia  12 = Lincoln  13 = Mazda  14 = Mercedes Benz  15 = Mitsubishi  16 = Nissan  17 = Ram  18 = Subaru  19 = Toyota  20 = None of the above |
| ZIPCODE | Zip code corresponding to the region | 0 = New England  1 = Mid-Atlantic  2 = East North Central  3 = West North Central  4 = South Atlantic  5 = East South Central  6 = West South Central  7 = Mountain  8 = Pacific |
| CREDIT | Amount of credit is available to the customer |  |
| BRAND | Brand which the customer prefer | 0 = Acer  1 = Sony |

**TABLE1**: Attribute information

1. **CORRELATION BETWEEN THE VARIABLE**

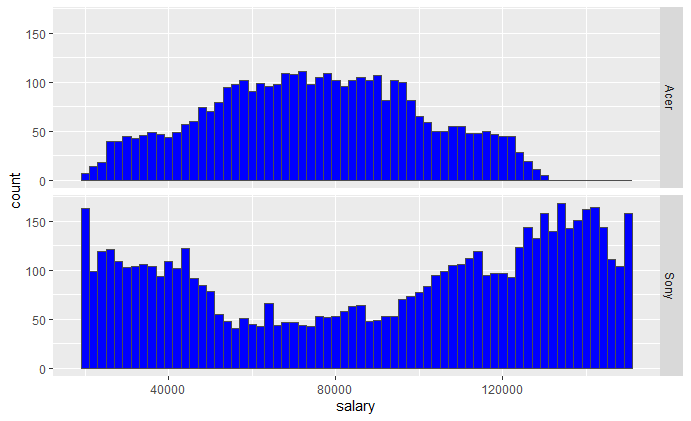
The below shows correlation matrix, which in turns shows up the correlation between each variable. From Fig1, it is clear there is a relation between the variable brand & salary.



**FIG1**: Correlation matrix

1. **BRAND Vs SALARY**

The below figure has two graphs. One on the top shows the number customers in each salary group who prefer Acer & the one in bottom shows the customers who prefer Sony.



**FIG2**: Customers in each salary group

Inference from the graph:

* Customers who prefer Sony are spread in all ranges of salaries even though there is a smaller number of customers with salary around 50,000 to 1,00,000.
* Customers preferring Acer are mostly in the range of 50,000 to 1,00,000
* There is very minimal number of customers of salary above 1,30,000 who prefer Acer
* The main brand preferred by most of the customers is **Sony** when compared to Acer.

1. **MODELLING**

In this project two data set is given.

1. CompleteResponses.csv
2. SurveyIncomplete.csv

CompleteResponse.csv has the complete response of the customers where SurveyIncomplete.csv is the incomplete data set. Our task here is to find an accurate model to predict & complete the incomplete data set.

Here, we review the different models that we tried for the case study problem and results from Accuracy score. The below shows the Accuracy score on using different models:

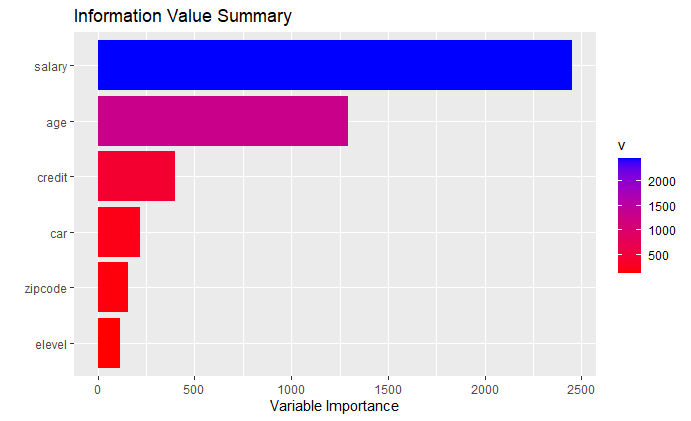
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Train Control parameter | Best model tuning parameter | Model Accuracy score | Model Kappa | Prediction Accuracy score | Prediction Kappa score |
| Random Forest Classifier | * 10-fold cross validation * Automatic Tuning Grid | * mtry = 18 | 92.03% | 83.08% | 92.5% | 84.1% |
| Random Forest Classifier | * 10-fold cross validation * Manual Tuning Grid(mtry = 1,10,20) | * mtry = 20 | 92.2% | 83.4% | 92.5% | 84.3% |
| BoostedTree(C5.0) | * 10-fold cross validation * Automatic Tuning Grid | * trials = 40 * model = tree * winnow = FALSE | 91.9% | 82.78% | 91.95% | 82.7% |
| BoostedTree(C5.0) | * 10-fold cross validation * Random search | * trials = 52 * model = rules * winnow = TRUE | 91.8% | 82.7% | 92.36% | 83.91% |

Inference:

From the above random forest by manual tuning & mtry = 20, gives the high performance with the accuracy around 92.5%

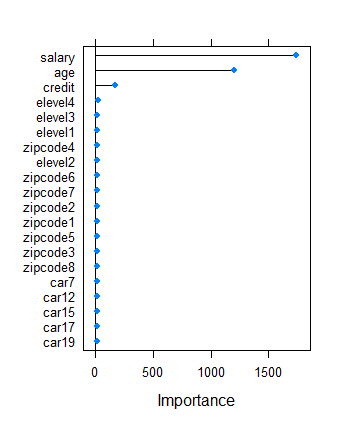
1. **FEATURE IMPORTANCE**

The influence of features is a parameter is important as they increase & decrease the accuracy. Some feature has a positive impact while some feature has negative. The below gives you an over importance of feature



**FIG3**: Feature Importance

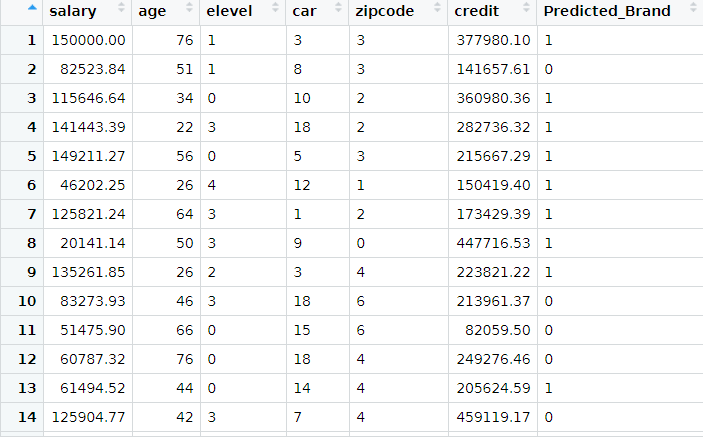
The below shows the feature importance as per the random forest algorithm:



**FIG3**: Feature Importance

# **SUMMARY**

One of objective of the project is to analyze the best fit algorithm & use the algorithm to fill the Incomplete survey. From the analysis, Random forest by tuning mtry as 20, gives the high performance with the accuracy around 92.5%. Using the model, the brand preference of the customers in the incomplete data set was predicted as shown below:



The main brand preferred by most of the customers is **Sony** when compared to Acer.