***Data Analytics Report***

**Customer Brand Preferences – Sony or Acer?**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1.0 | June 12th | Steffen Adolf | Initial write |
|  |  |  |  |
|  |  |  |  |

**Task**

Based on the survey of our customers conducted by a market research firm, we want to find out which of the brands “Sony” and “Acer” is preferred by our customers.

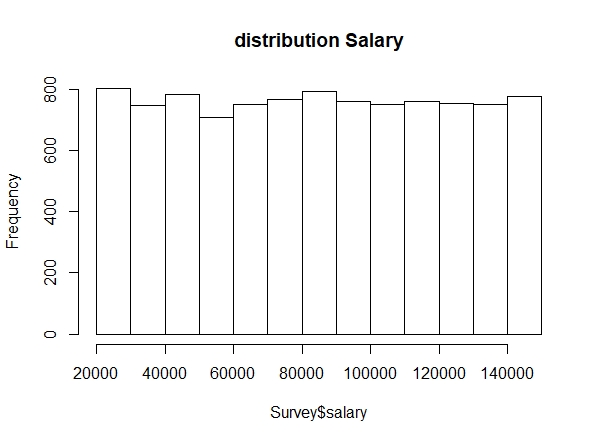
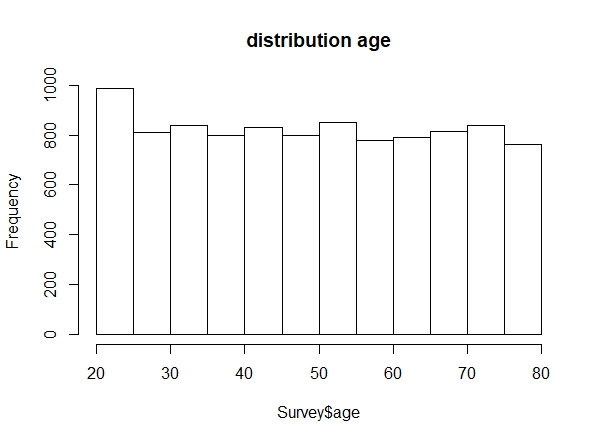
There are 10.000 complete answered observations and 5.000 incomplete ones. Therefore, with the help of the complete observations, predictions about the missing responses can be made.

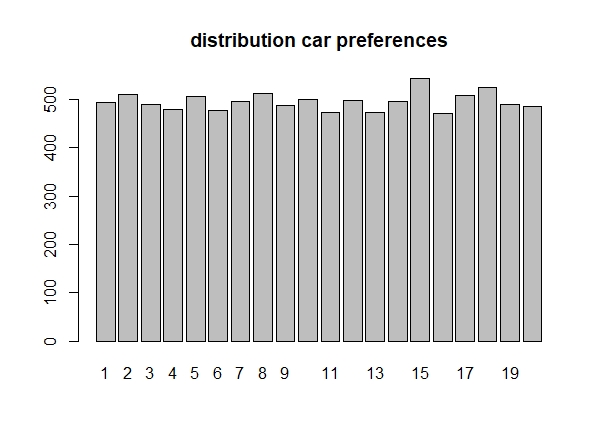
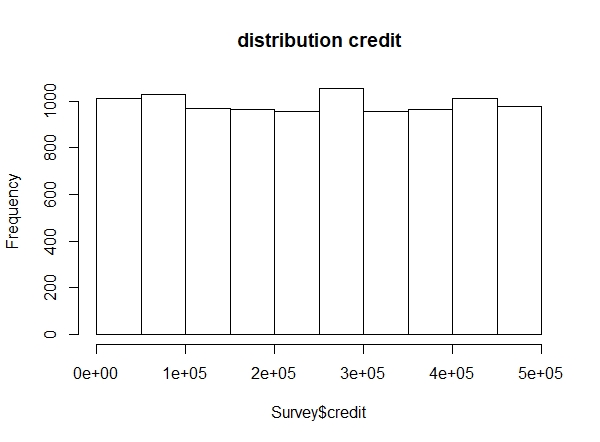
Then the full survey of 15.000 can give a clear picture of customers preference.

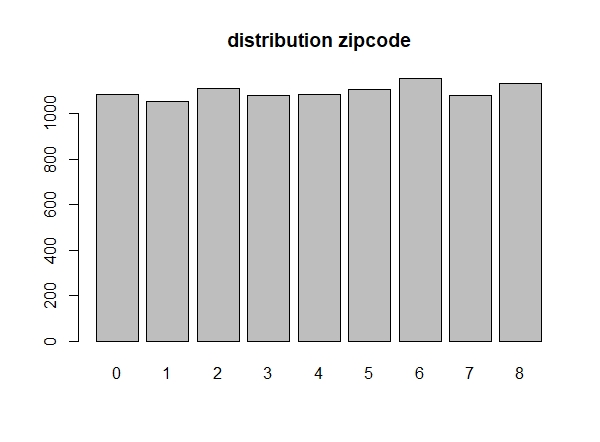
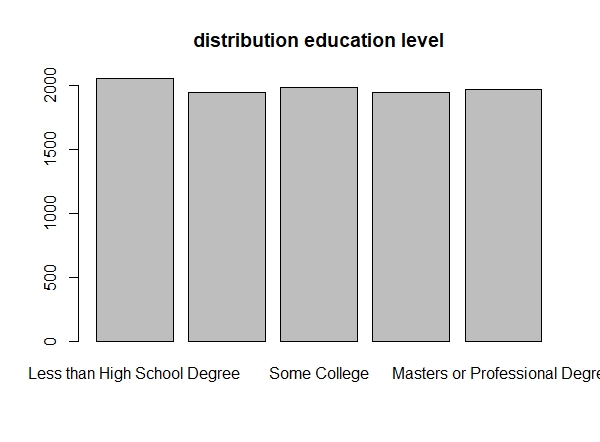
**1. Data Exploration**

**1.1 Distribution**

The distribution of each of the features of the data set is in a very unusual form. The analysis shows that all of the features are more or less uniform distributed.

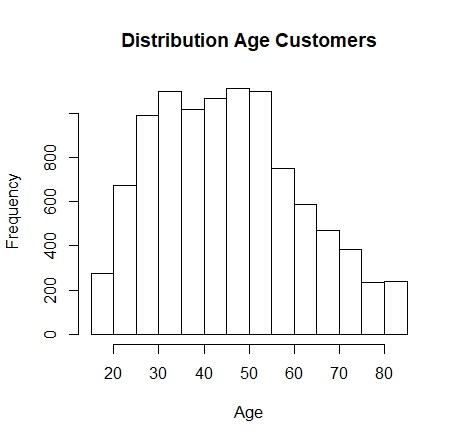






This kind of distribution is untypical, especially for such a set of features. It is possible of course, that the research company has deliberately chosen a survey sample which represents customers from all regions or age groups in a uniform way.

But in combination with all the other features also distributed uniform, like customers salary and their education level, the whole data set seems to be in a not very plausible shape and therefore the quality of the data set itself is in doubt.

A recent study about ***customer buying patterns*** gave us clear information about the true age distribution of our customers.

-right skewed distribution of age of our customers, a typical and realistic distribution form for this kind of attribute.

Why should an experienced market research company deliberately choose a sample that represents a different form?

**1.2 Correlation and Causality**

A closer look on the correlations between the features just confirms the impression that the quality of the data set is doubtful.

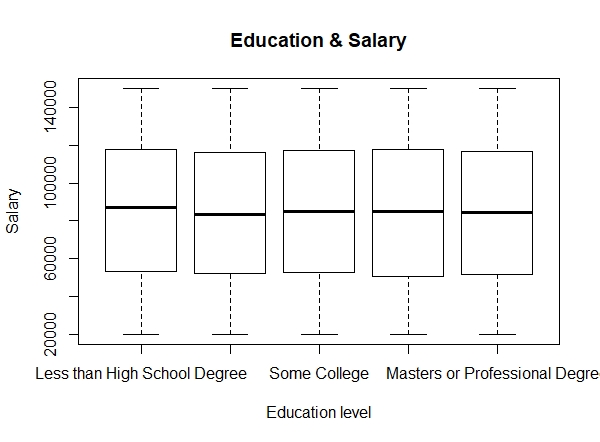
There is no correlation between the numerical attributes at all.



No correlation between age and salary of a customer and no correlation between salary and given credit seems implausible.

I

A look on the relation between the education level of interviewees and its salaries is especially incomprehensible.



|  |  |  |
| --- | --- | --- |
| Education Level | mean salary | median salary |
| Less High School | 85827.51 | 87022.03 |
| High School | 84004.49 | 83453.12 |
| Some College | 85003.14 | 84812.22 |
| College Degree | 84971.76 | 84916.92 |
| Master’s Degree | 84497.85 | 84324.41 |

This data set shows no increase in average salaries with rising educational level. Even without deeper research it is intuitively clear that normally there would be a clear connection between these two attributes.

**1.3 Conclusion**

The explorative analysis of this survey shows a clear lack of quality of the data. The distribution of each of the features is unnatural. The absence of any correlation between features, which under normal conditions would clearly have one, just confirms that the data set can’t be in a correct form.

That’s why the following modelling process and the predictions should be interpreted cautiously.

**2. Modelling and Predictions**

**2.1 Data Transformation and Feature Selection**

The survey provides a lot of observations from regions in the US where we don’t have any customers. To get a clear picture of our customers preferences, I concentrated the analysis on the observations from the “South East” (regions: “South Atlantic”, “East South Central”) and left out all the other regions. The number of observations has been reduced from 10.000 to 2.195 then.

The C5.0 model has an in-built feature selection function so that a manual selection is unnecessary.

**2.2 Model Performance**

**C5.0**

1647 samples

6 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 1482, 1482, 1482, 1483, 1483, 1482, ...

Resampling results across tuning parameters:

model winnow trials Accuracy Kappa

rules FALSE 1 0.8587073 0.7107636

rules FALSE 10 0.8795805 0.7489365

rules FALSE 20 0.8797849 0.7494587

rules FALSE 30 0.8797849 0.7494587

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were trials = 20, model = rules and winnow = FALSE.

Confusion Matrix

Reference Accuracy: 0.8686

Prediction 0 1 Kappa: 0.7262

0 182 46 Sensitivity: 0.8750

1 26 294 Specificity: 0.8647

**Random Forest**

1647 samples

6 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 1483, 1482, 1483, 1483, 1482, 1483, ...

Resampling results across tuning parameters:

mtry Accuracy Kappa

10 0.8991913 0.7862140

11 0.9060760 0.8006080

12 0.9060809 0.8006777

13 0.9101348 0.8088806

14 0.9119592 0.8124793

15 0.9129706 0.8149807

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 15.

Confusion Matrix

Reference Accuracy: 0.9197

Prediction 0 1 Kappa: 0.8302

0 188 24 Sensitivity: 0.9038

1 20 316 Specificity: 0.9294

**2.3 Model performance in comparison**

Models: C5, rf

Number of resamples: 30

Accuracy

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

C5 0.7317073 0.8401333 0.8936068 0.8773816 0.9257576 0.9575758 0

rf 0.8787879 0.8971249 0.9151515 0.9129706 0.9268293 0.9515152 0

Kappa

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

C5 0.4503352 0.6719457 0.7725685 0.7455498 0.8407122 0.9093051 0

rf 0.7478221 0.7799371 0.8174906 0.8149807 0.8436499 0.8966651 0

These results show that the “Random Forest” model is performing better on this data set

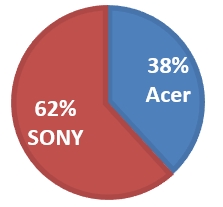
then the C5.0 model in regard to “Accuracy” and “Kappa”.

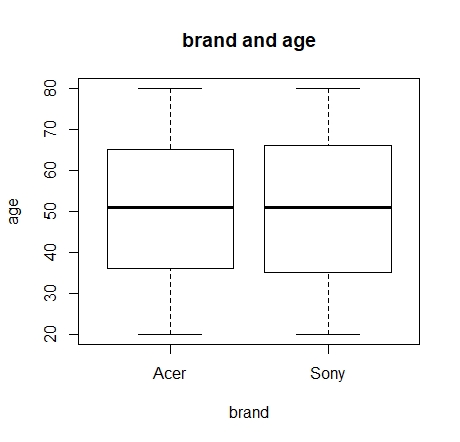
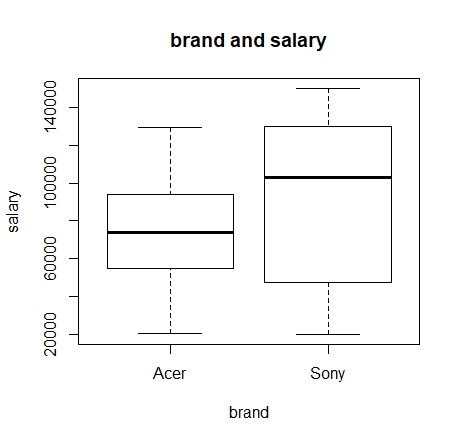
That’s why I have chosen “Random Forest” for making the predictions.

**4. Results**

The predictions calculated by the model show that 743 observations of the incomplete survey data set prefer “Sony” as a brand compared to “Acer”. Adding to the complete survey there are 2.014 people in total that prefer “Sony”. This is a majority with 61,8%.

410 observations are predicted to prefer “Acer”. In total these are 1.244 observations and therefore 38,2% of all survey observations.

****



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **brand** | **mean salary** | | **median salary** | | **mean age** | **median age** |
| **Acer** | | 74501.24 | | 73985.98 | 50.2 | 51 |
| **Sony** | | 91436.89 | | 102840.8 | 50.3 | 51 |

|  |
| --- |
|  |
|  |

Analysis of the model have shown that “salary” and “age” are the most decisive attributes of customers in regard to brand preference.

While a closer look on customers average age can’t confirm that there is a preference for one of the brands in dependence of age, there is a clear difference in regard to customers salary.

The median salary of customers that prefer “Sony” as a brand is with 102.840 $ per year much higher than the salary of customers that prefer “Acer” with 73.985 $ per year.

**4. Recommendation**

The obvious lack of quality of the data set makes data analysis and predictive analytics very unreliable. Therefore it will be necessary to get into touch with the sales team and the market research company to investigate and discuss the reasons for this[irregularities](https://context.reverso.net/übersetzung/englisch-deutsch/irregularities)*.*

That’s why the results of the analytics process should be taken with caution. They show a clear preference of our customers for “Sony” as a brand with a majority of 62 % compared to “Acer”.

Especially customers with a relative higher income prefer “Sony”. This position of our key customers just strengthens a clear plea for “Sony” as future cooperation partner.

However, as long as the quality of data hasn’t been investigated, a clear decision should not be taken.