

**DATA MINING: A CASE STUDY OF CARAVAN INSURANCE**

MERCER UNIVERSITY ATLANTA



December 16, 2022

KARIM [11033410] & MUSTAFA [11001954]

MSBA

Table of Contents

[INTRODUCTION 2](#_Toc122105457)

[DATA COMPOSITION 3](#_Toc122105458)

[STATISTICAL INFORMATION: 3](#_Toc122105459)

[MODELS 4](#_Toc122105460)

[RESULTS 4](#_Toc122105461)

[GOODNESS OF FIT 7](#_Toc122105462)

[RECOMMENDATIONS 7](#_Toc122105463)

[REFERENCES 7](#_Toc122105464)

[APPENDIX 8](#_Toc122105465)

[CHART 1 – Customer Main Type + Customer Sub Type 8](#_Toc122105466)

[CHART 2 – Income Distribution 8](#_Toc122105467)

[CHART 3 – Customer Sub Types 9](#_Toc122105468)

[CHART 4 - Distribution of customer subtypes 9](#_Toc122105469)

[CHART 5 – Age vs Insurance 10](#_Toc122105470)

[CHART 6 – Class Imbalance 10](#_Toc122105471)

[CHART 7 – Histogram 11](#_Toc122105472)

[CHART 8 – Residuals 11](#_Toc122105473)

[CHART 9 – ROC 12](#_Toc122105474)

[CHART 11 – DECILE-WISE-CHART 12](#_Toc122105475)

[CHART 12 – DECISION TREE 13](#_Toc122105476)

# INTRODUCTION

Sentient Machine Research, is a Dutch company which has provided with this data set. The data set is a collection of products that customer prefer, these products being types of insurances. The company coined the data set as a challenge to see to who could come with a better predictive models in order to answer a simple business question; who would be interested in buying CARAVAN Insurance. It should be also remembered that the dataset consists of 86 variables which collect various socio and economic conditions of the customers derived from area zip codes. It should also be noted that these variables by nature are categorical. Out of 86 variables, only 2 are numeric.

Apart from the data mining, we also considered many possible outcomes with respect to **Marketing Analytics**. The reason why we have done is because CARAVAN insurance is a product offered by the company. As a product, it is imperative that it reaches the correct audience. While data mining would tell us the groups/classes/segments that are important, marketing through proper targeting and advertisements would help us actually cater those particular segments. The idea behind that is to **boost sales** for CARAVAN Insurance by targeting the correct segments. An advantage of data mining is that it also tells us about the bad predictors and how marketing can help us avoid high costs associated with such groups. Another advantage of this research in the marketing domain would be that it could help us understand who are the most important to least important groups. This could easily be identified with the help of **BCG Matrix**, where in we could do product classification with respect to Cash Cows, Stars and Dogs.

Since, it was offered out as a challenge, the idea was to describe actual or potential customers that could or might actually buy the CARAVAN Insurance. Originally, the data set was divided into training and validation. The reason as to why it was done is because we use training data set to access the behaviour of the dependent variable and see how well it fits with other models and eventually, the entire data. This dataset had no missing values and included attributes related to the customers' socio-demographic data as well as information about their insurance product situation.

# DATA COMPOSITION

Before we move forward, we feel that it is imperative to be familiar with the data composition so that when we explain our models, they could be better understood. Following are the attributes of the data:

1. 86 Variables
   1. 84 Categorical Variables
   2. 2 Numeric Variables
2. Categorical data is a representation of a range of a group
   1. For example income marked by 2 means income range is between $30k - $45k
3. Socio-Economic Information for example:
   1. Household size
   2. Education
   3. Skilled
   4. Unskilled
4. No missing values

The composition of the data is meticulous and it presented many opportunities and challenges for us. However, we have tried our best to apply the techniques learnt in this course and to be able to see what we can come up with. We have tried to remain relevant and true to the techniques learnt as well.

# STATISTICAL INFORMATION

We analyze the raw variables by labeling and categorizing them in order to better understand the insurance product and its buyers.

The division of socio economic variables is primarily based on **customer main type [MOSHOOFD]** which consists of 10 categories and **customer sub-type [MOSTYPE]** consisting of 41 categories.

We have taken the approach of classification and grouping because this is what we saw best drives our model best in the light of our business question that is, identifying the attributes of those customers that could buy the CARAVAN Insurance. This will help us know which group of people are more likely to buy insurance as compared to others. Variables 44 to 85 and 35,36 represent the interest of customers in various types of insurance policies, ranging from essential policies such as life, health, disability, and family/private accident insurance, to more optimal policies such as property, small automobile, and delivery vehicle insurance. Additionally, more sophisticated policies such as private third party insurance, car, fire, and social security insurance may provide greater safety and luxury. Similarly, variables related to income **[MINKM30, MINK3045, MINK4575, MINK7S12, MINK123M]** we have computed average income for each record based on the highest limit for that group. We did that to convert the categorical income variables to numeric. We also grouped number of houses between having 1 or more.

# MODELS

In order to solve the prediction tasks of the challenge, we used a variety of data mining techniques. Utilizing R programming language, we explored the data to make logical inferences through visual and mathematical aids. Through this process, we were able to reduce the number of predictors or variables used in the prediction, which is beneficial as an increase in noise is likely to occur when more variables are included.

Our first and foremost challenge was to identify the variables that impact our dependent variable **[CARAVAN]** along with imbalanced data. In order to resolve these issues, we created 4 models. 3 models based on Logistic Regression [GLM] and 1 for decision tree. We also utilized ROSE library to resolve the imbalanced data when doing data partitioning. We identified our variables of interest based on the following:

1. Visualization
2. Correlation
3. Domain Knowledge
4. Decision Tree

# RESULTS

When making predictive models, accuracy is at the heart of every model but then it is dependent on the business question. We utilized confusion matrix to access accuracy, error rate, specificity and sensitivity for our models. Following are the results of these models:

|  |  |  |  |
| --- | --- | --- | --- |
| **MODELS** | **ACCURACY** | **SENSITIVITY** | **SPECIFICITY** |
| Mod 1 | 11% | 95% | 5.4% |
| Mod 2 | 54% | 59% | 54% |
| Mod 3 | 60% | 65% | 59% |
| Mod 4 | 65% | 55% | 65% |

**For Mod 1**

***What are the variables?***

We had to go through all data dictionary to understand what type of variables are there. Based on our intuitional inference, we picked variables that could potentially be a good predictor for customers who might buy CARAVAN insurance. These variables are Customer Main Type, Average Household Size, One House [Dummy Variable], Average Income [Numeric] and Average Age [Numeric]. It should be remembered that all of them are categorical but we converted income and age to numeric form.

***Which method do we use?***

We used a simple GLM method on the above mentioned variables, on which we were able to draw a confusion matrix.

***What is/are the results?***

Our results indicate that Driven Growers, Living Well, Seniors, Family with grown ups are significant along with income as well. 1 house is least significant.

**For Mod 2**

***What are the variables?***

In this model we found that here are 34 variables which are correlated with each other, before dropping them we need to see how they are correlated with response variable.

***Which method do we use?***

We have used a deep correlation analysis. Through this analysis we found that the max positive relation is 1 and the negative minimum is -.99. Then we tried to figure out how many predictors are correlated at cut-off = .75.

As we had 27 variables and we wanted to do dimension reduction, we then proceeded to do correlation of these variables with our dependent variable.

***What is/are the results?***

As a result of the above dimension reduction, we took these and the above 4 variables and thus, reduced our variables from 86 to 59. After this we applied we step-wise forward regression and we got all our predictors from least to most. We included, first 12 significant variables from the 59 variables mentioned above. Our results indicate that all levels of medium education are significant among households with children.

**For Mod 3**

***What are the variables?***

Based on our domain knowledge we made model 3 and the variables are customer sub-type, contribution to fire policies, contribution to car policies, purchasing power class and homeowners. We went through some articles to be able to pin point relevant variables with respect to our data.

***Which method do we use?***

Here we again applied a simple GLM model to make a prediction

***What is/are the results?***

The result of this model is promising as the value of RMSE is .8. Indicating that the model is predicting well. Also the ROC value is .67 ~ 1. Based on this we see that fire policies, homeowners and power class positively respond to CARAVAN insurance.

**For Mod 4**

***What are the variables?***

In this model, we have 5 variables of interests, namely, contribution car policies, customer sub type, contribution fire policies, high status, middle management.

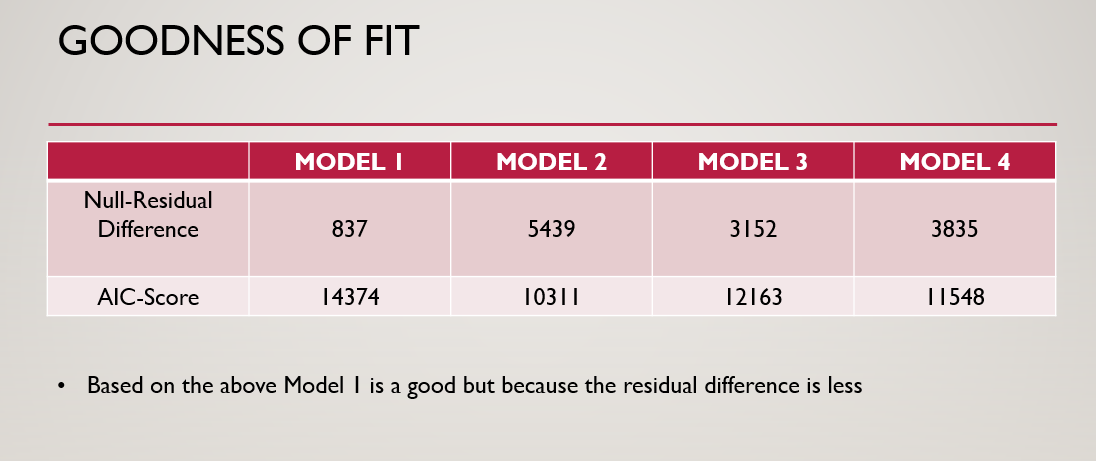
***Which method do we use?***

We performed the decision tree method

***What is/are the results?***

The above mentioned variables are the results of the decision tree which it has factored out as to be the most significant. We verified this using the confusion matrix and achieved an accuracy rate of 70% at cutoff .5 but should the cutoff be changed, accuracy will change as well.

# GOODNESS OF FIT



From the above table for goodness of fit, we know that the lesser the Null-Residual Difference and higher the AIC-Score, the better the model is. However, if we want to target the AIC-Score, Model 2 is most suitable but it is prone to over-fitting.

However, at the end, it depends on the user what the user is trying to do. Our goal has been simple, to provide the user with various models, each a best fit for a particular situation/group/class.

# RECOMMENDATIONS

* Model I is recommended because of high sensitivity
* Model II is recommended over Model 1 because of high accuracy
* Model III is recommended over I & II because of high accuracy and sensitivity from II
* Model IV is recommend overall because of high accuracy

# REFERENCES

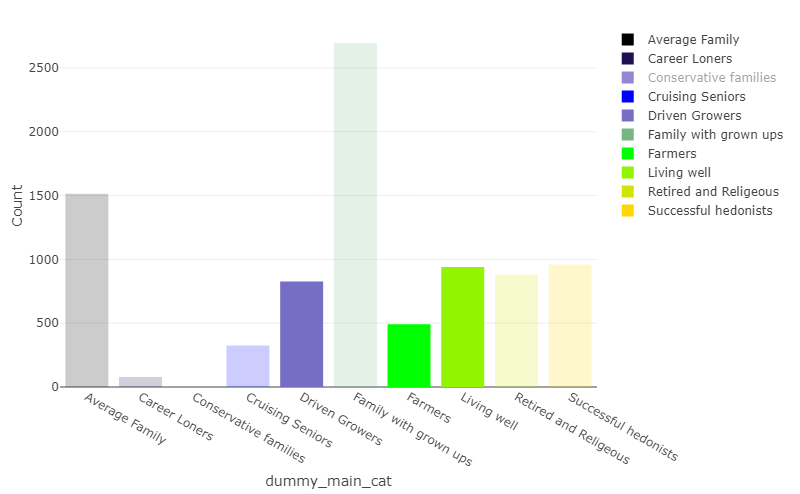
<https://www.kaggle.com/datasets/uciml/caravan-insurance-challenge>

[Feature Selection to Kaggle Caravan Insurance Challenge on R | by Kieran Tan Kah Wang | The Startup | Medium](https://medium.com/swlh/feature-selection-to-kaggle-caravan-insurance-challenge-on-r-bede801d3a66)

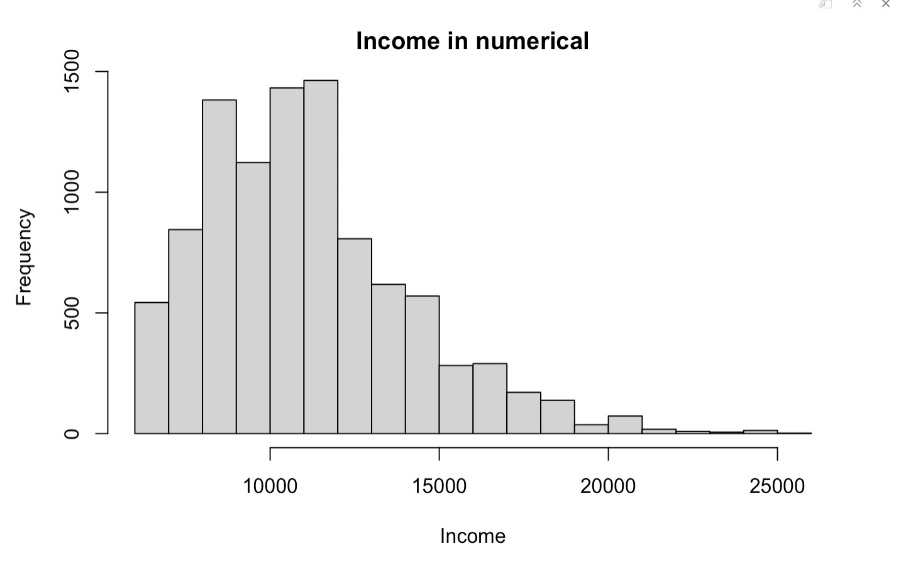
[UCI Machine Learning Repository: Insurance Company Benchmark (COIL 2000) Data Set](https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+(COIL+2000))

# APPENDIX

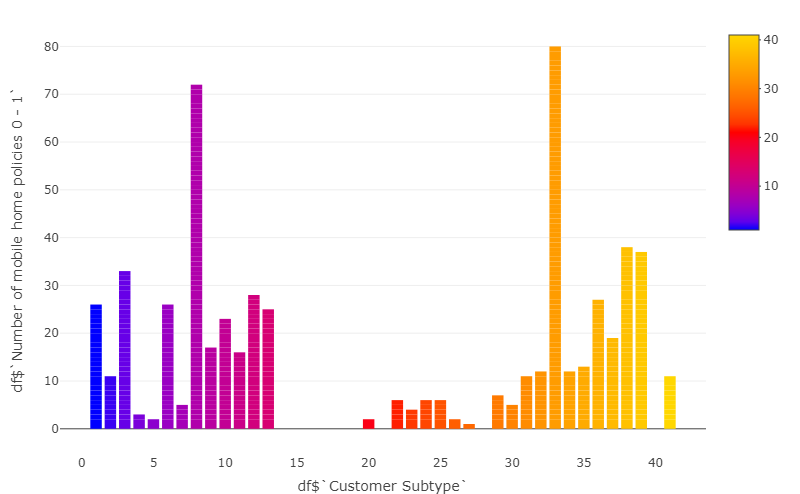
## CHART 1 – Customer Main Type + Customer Sub Type



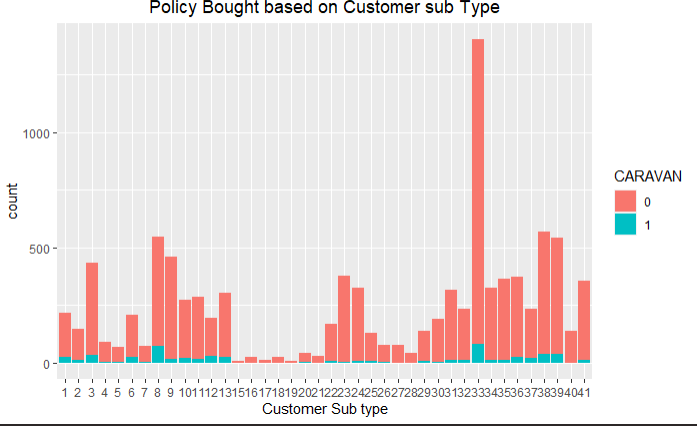
## CHART 2 – Income Distribution



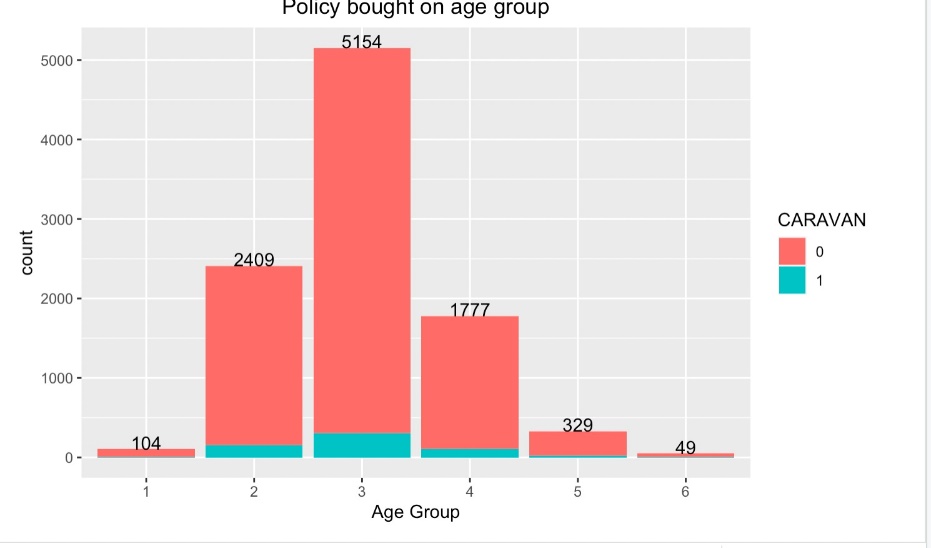
## CHART 3 – Customer Sub Types



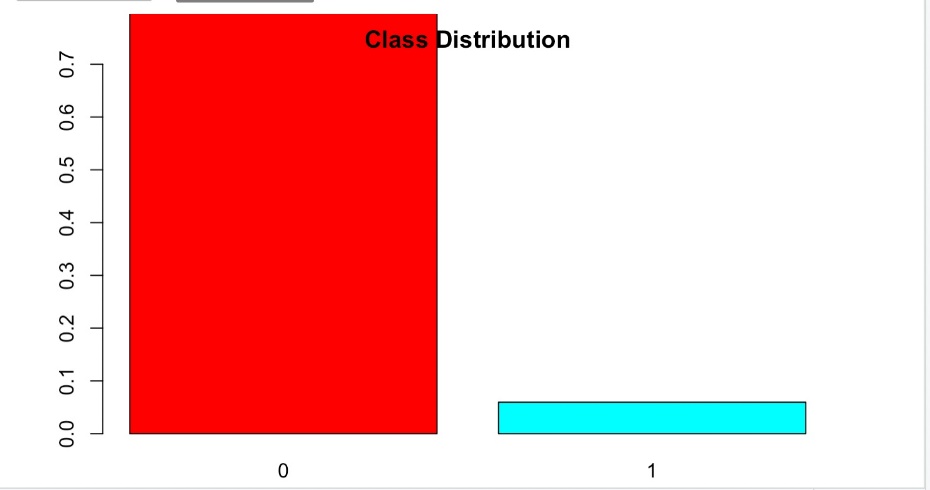
## CHART 4 - Distribution of customer subtypes

******

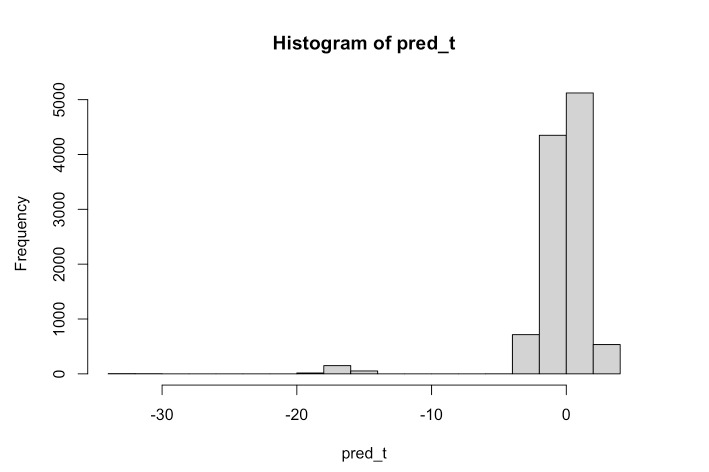
## CHART 5 – Age vs Insurance



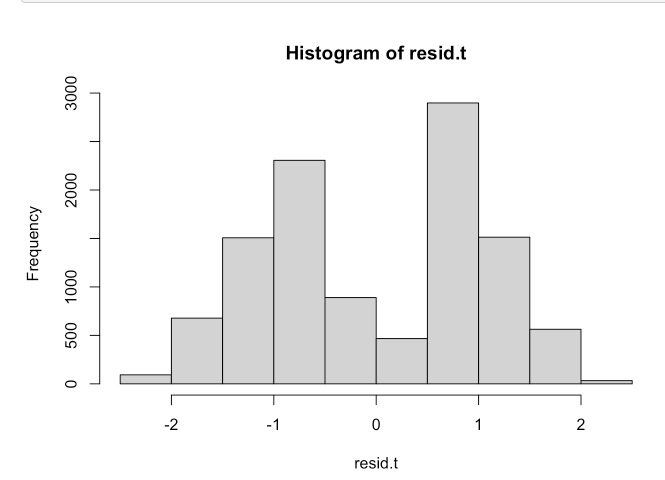
## CHART 6 – Class Imbalance



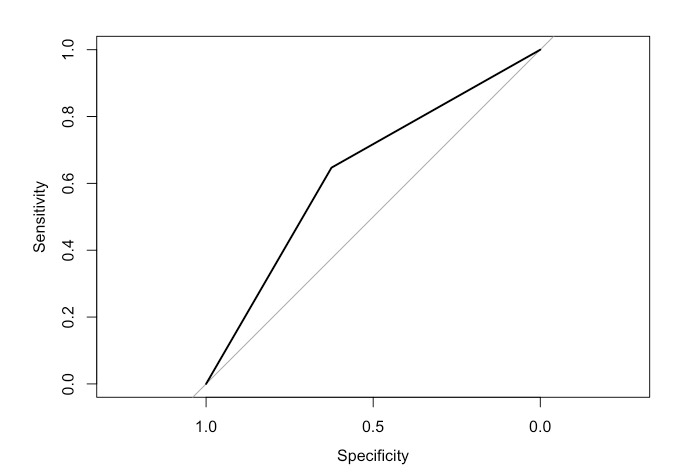
## CHART 7 – Histogram



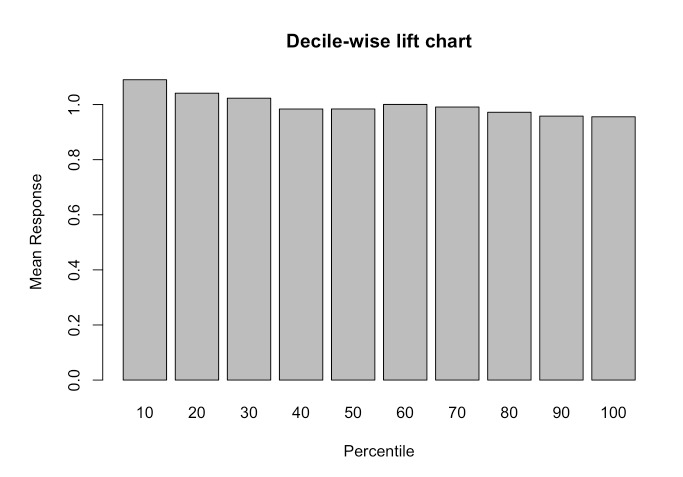
## CHART 8 – Residuals



## CHART 9 – ROC

****

## CHART 11 – DECILE-WISE-CHART



## CHART 12 – DECISION TREE

