The Insurance Company (TIC) Customer Policy Insight Report

by

Random Forest Predictive Modeling

Charles Adkins

MARCH 25, 2020

The Insurance Company (TIC) Data Set Summary

Organizational Background

The Insurance Company (TIC), is an insurance company that places significant value in providing their customers, both current and potential, with the most beneficial coverage and protection for their cherished assets. The company has been in business for the past 30 years and has over 10,000 customers, whom all have a wide range of personal policy dimensions. Products range from auto insurance, home insurance, life insurance and property. TIC wants to maintain the highest level of customer satisfaction by providing their customers with comprehensive protection at every level, for every asset, as well as market to potential customers with the least amount of expense and the most amount of benefit and value, for both the customer and the organization. Currently, TIC has focused its efforts in identifying customers that may have a need for caravan insurance, a subset asset of the auto insurance sector. Identifying these customers will not only mean more business for the organization, but will also mean more protection and peace of mind for the customer.

TIC has relied heavily on direct mailing marketing to remain in the forefront of the minds of their current customers and to gain the attention of new customers. However, these direct mailings are expensive in that, not every mailing results in a new business partnership. TIC must focus their direct mailing efforts towards the most receptive customers, in terms of prospective business partnerships, to not only gain new customers, but to also lower the amount of wasted effort they spend on advertising to populations who throw the literature away.

Data Set

The data was collected by a Dutch data mining company called, Sentient Machine Research, of which has been contracted by TIC, to assist them in identifying the customers who may be interested in caravan insurance policies. The research question is as follows:

- Can it be predicted who would be interested in buying a caravan insurance policy and give an explanation why?

By identifying and catering to the needs of such customers, TIC will not only provide such customers with the most comprehensive service and products available, but they will also lower marketing costs with a more focused customer acquisition strategy and reduce overall waste.

The dataset provided includes 5822 customer observations, or rows and 86 variables (of integer type)(**Appendix A**), of which cover both product data and socio-demographic data, based upon zip codes (P. van der Putten and M. van Someren, 2000). All customers living in the same zip code area have the same socio-demographic data, while variable 86, ‘CARAVAN’, specifically contains the number of current mobile home policies, of which is the target variable (P. van der Putten and M. van Someren, 2000). Variables 1-43 contain the socio-demographic data, while variables 44-86 contain the product ownership data (P. van der Putten and M. van Someren, 2000). The data types include several variations of factorial values, of which are explained in detail in the metadata. Access to specific metadata on each variable and its value is contained in reference material.

Introduction

Analytical Question

The Insurance Company (TIC) has placed considerable effort in their marketing strategy to acquire and maintain beneficial customer partnerships, but has ultimately found its efforts falling short of steady returns. Direct mailings have been known gain new customer business, but the costs of doing so, has proven to be less efficient than focusing such resources towards specific customers, in particular, their current customers, for the purpose of product and service promotion. In order to establish a more efficient strategy, TIC must begin by addressing the analytical question that this report intends to layout, analyze and answer. The question is:

- **Can it be predicted which current customers would be interested in buying a caravan insurance policy?**

Organizational Background

With respect to the TIC organizational mission, that which is to provide excellent customer service, it is imperative that the company understand their customers needs and how those needs can be met with TIC products and services. To address the analytical question, we must first understand the customers that currently have a caravan policy and the factors that led to such business. By identifying influential factors that have led those customers to purchase a caravan policy, we can then also identify those same factors in other TIC customer accounts. Such a strategy will be accomplished by implementing a predictive analytical plan, by which we will then further understand customer behavior and how those characteristics can be valued in business projections.

TIC will be more prepared and better equipped to assist their customers, in addition to addressing the needs and concerns of potential customers, with the use of predictive analytics. “Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future” (Faggella, 2019). Examples of such predictive models include decision trees, random forests and logistic regression. For this specific case, a random forest model will be utilized, as it offers a robust dimension of analysis and features functionality traits that limit bias, in addition to its ability to accommodate large datasets with several variables.

Forward-thinking, rooted in supportive data-derived evidence, will provide TIC customers an enhanced experience and greater satisfaction their choice to do business with TIC, while at the same time, providing TIC with value that optimizes their marketing efforts, produces organizational leanness, promotes excellent customer care and establishes an enterprise-wide strategy that will rest on the foundation of predictive business ideals. Maintaining a high-level understanding of customer care and knowledge regarding the desires and needs of such customers will grant TIC with an intuitive edge over its competitors. For example, “highly loyal customers with a high likelihood to purchase or continue business, are less likely to be affected by marketing on any of their usual channels” (Thys, n.d.), which can help TIC focus their promotional efforts towards customers who would benefit from new services and products, rather than washing over all possibilities of business, with little to no insight. Utilizing predictive analytics will ultimately enhance service and help personalize the products and relationship that each customer has with TIC, while also reducing inefficient resource allocation efforts, like that which TIC currently finds themselves within.

Problem Statement and Research

Overview

The Insurance Company (TIC) is an insurance company that has taken pride in providing excellent customer experience and satisfaction and to maintain such pinnacle performance, TIC must begin to optimize their organizational efforts. With optimization in mind, TIC must maximize the proportion of effort to returns, specifically in customer account proliferation. TIC grows, not only when they acquire business from new customers, but also when their current customers expand the dimension of their partnership.

When a customer adds a new policy or additional coverage options to their account, TIC revenue, as well as their trust and reputation, grows. To streamline this growth, TIC has the goal of identifying customers that may be interested in adding additional coverage options to their policy, specifically caravan coverage. The challenge with this task is beyond simply identifying such customers, but rather in communicating such needs to the customer and why such an addition of coverage would be in their best interest. The soft skills of business negotiation and product promotion will rest on the shoulders of TIC business relations specialists, while the identification of such potential customers will be the responsibility of myself and this report.

Analytical Approach

In order to grant this portion of the explanation a sense of conciseness, we will refrain from extrapolating the details of the process that will be undertaken for cleaning and organizing the data. However, these steps are crucial as they prepare the data to meet the specifications and requirements for the analytical technique and algorithm chosen. This process will include steps such as categorizing variable types, imputing NAs from the dataset, if applicable, removing irrelevant variables, such as variables 1-43, (as they are sociodemographic and represent statistics for all persons living in specific zip codes, not specific individuals), cross-tabulation analysis and more.

The random forest model will develop a series of iterative decision trees, with each iteration improving the error rate and drawing a more conclusive result, rooted in the established target variable. The model will display the variables that have the highest predictive power towards a customer having a caravan policy and the customers who do not. We will then take note of these predictor variables and apply such knowledge to our analysis of the customers, of whom do not currently have a caravan policy.

The most significant variables that relate to the existence of a caravan policy will be focused upon, as we place the caravan policy variable, once again, as our target variable, in our random forest model. By doing so, we will identify the predictive variables that are most significantly linked to a caravan policy and consolidate the customers who have significantly correlated values for such variables, into the groups of customers to whom marketing and promotion efforts should be focused.

With these customers, who have the most likeliest of interest in a purchasing a caravan policy, in mind, TIC can focus their marketing efforts towards these individuals, understanding that based upon the data, they may also be in need of a caravan policy.

Pilot Plan

We have established that TIC organizational goals include optimizing their marketing efforts and understanding their customers' needs and potential areas of new product acquisition and business. TIC has expressed their desire to market caravan policies towards customers that may be in a position or in need of such a policy, by which they can narrow their promotional and advertising efforts. To better understand who these customers are, we must analyze the data and determine specific variables that predict when customers already have such a policy and which customers, who do not currently have said policy, share in such predictive variables.

To uncover such relationships within the data, we will draft our model by utilizing a random forest algorithm. Random forests are collections of decision tree that manage error and data bias well, giving us the most accurate and comprehensive result. The Random Forest model algorithm utilizes two concepts by which details the origin of its name 'random'. Such concepts include the fact that during the training set model construction, random sampling develops the points within the trees and random subsets of features are considered when splitting nodes within those trees (Koehrsen, 2018). The CARAVAN variable is the target variable and the first forty-three variables have been removed from the dataset because they represent socioeconomic data for entire zip codes, not specific individuals or customers (**Appendix B**)(App A in 4). The data frame is organized by way of column headers and adjusted for correct variable types (factors) to assist in the random forest creation. Potentially useful packages are installed and loaded prior to establishing the model as well (**Appendix B**)(App A in 4).

The following steps should be implemented to generate the random forest model. Firstly, after initializing the Rattle package, the dataset will be imported by selecting the 'File' option within the 'Data' tab of the Rattle interface window. After viewing the characterization of the included variables, with 'CARAVAN' being the TARGET variable, select 'Execute' to establish the data within Rattle. Then, select the 'Model' tab to create the Random Forest, by selecting the 'Forest' radio button. Ensure that the 'Trees' option states "500" and the 'Variables' option states "6", as well as the 'Impute' option checked, to account to any N/A values. (In this case, there are no N/As). Finally select 'Execute' to generate the model. The view window will display the results.

Appendix C & D (App B & C in 4), show the results of the Rattle random forest model. We have utilized 500 trees, iterating through six variables at each branch, which is appropriate, given it is typically represented by the square root of the variables used (43 in this case). I've also ensured that the values are imputed as well.

The Out-of-Bag (OOB) estimate shows us an error rate of 5.92%, meaning that the model predicted the correct value for the target variable (CARAVAN) 94.08% of the time. The confusion matrix shows the breakdown of such classification, where "0" is no CARAVAN policy and "1" is a yes CARAVAN policy.

Appendix D (App C in 4) shows the variable importance as it relates to the binary values of the CARAVAN variable.

We can see that:

‘APLEZIER’ (# of boat policies),
‘PPLEZIER’ (contribution boar policies),
‘PPERSAUT’ (contribution car policies),
‘APERSAUT’ (# of car policies)

have the highest predictive values for observations have “1” for the CARAVAN variable value. This means that observations with these highly-predictive variables, lead us to the notion that they may already have a CARAVAN policy or if they do not, they may be in need or interested in obtaining such a policy. On the other side of things, we can see which variables are most predictive in observations not having a CARAVAN policy, value = “0”, such as:

‘AWAPART’ (# of private 3rd party insurance),
‘PINBOED’ (contribution property insurance policies),
‘AINBOED’ (# of property insurance policies).

Given the quick output of the model, we can temporarily focus our efforts towards customers who have either a boat policy and a car policy, or both. These customers may be interested in or in need of a CARAVAN policy as well. In addition to this conclusion, we may add that customers who have private 3rd party insurance policies may not be interested in a caravan policy, as it may be covered in that policy, as well as customers who have property insurance policies, as it too may cover CARAVAN needs. However, as seen in Appendix C, some values related to CARAVAN value “0”, are close in value to those representing CARAVAN “1”, as these values may warrant additional focus, as they may have some contributing predictive power that can be exploited for CARAVAN policy marketing.

Model Optimization

Model optimization involves tuning model parameters to meet the specifications for the scope of the project and the anticipated and required accuracy. Harnessing the power and accuracy of a model, prior to deployment, involves partitioning the data into subsets, of which includes the training, validation and testing sets. In doing so, the model can be developed to the internal patterns or limitations of the training set, of which the results will then be used to tune the model for the validation set. Once the parameters have been appropriately finalized, the testing set will be incorporated to test the effectiveness and accuracy of the model, which will then be transposed towards external datasets/projects in the future. These results will be considered the scoring of the model. “The process of applying a predictive model to a set of data is referred to as scoring the data” (IBM, n.d.).

With respect to scoring, we can understand the usability of the model by how well it incorporates the validation and testing datasets, as it relates to the real-world application of said results. This method proves to be the most comprehensive method for scoring purposes as it promotes a deep understanding of the data and it evaluates the internal patterns with little to no bias. The results of said model, while incorporating this method, provides the most accurate results, of which supports its own reproducibility.

By utilizing the ‘Score’ functionality of the ‘Evaluate’ tab in Rattle, we are able to evaluate the accuracy of model prediction as it relates to the probability of observations resulting in the positive value of the binary target variable, in this case, the presence of a CARAVAN policy (**Appendix E**)(App A in 5). The results will be saved into a csv-format file that can be reviewed outside the Rstudio platform. The display of said results can be toggled from probability values to either class. In this case, I’ve used the probability standard. These results present how the model prediction values correspond to the actual values of the target variable values. If the model prediction values are at or above an acceptable rate of accuracy, the model can be deemed useful for this particular dataset, as well as future datasets and new observations. In this particular case, based upon the scoring results, for the value of “1” for the case of CARAVAN policy, meaning the policy exists, the range of accuracy generated is 0.786 (78.6%) to 0.964 (96.4%). In conjunction with this finding, we can observe from the testing results, the error rate has been reduced to 0%, properly identifying and partitioning the “0” and “1” values for the CARAVAN target variable in its entirety (**Appendix F**)(App B in 5).

If the results, both the scoring probability values and/or the error rate is not within the acceptable range of accuracy, the parameters of the model can be tuned to better form the results to achieve such values. For example, the number of trees and number of variables tried at each split can be changed, in which case, should be referenced by the lift value, previously established during the model development.

Conclusion

Based upon the aforementioned details regarding the supplied data and the analytic model developed, it is recommended that TIC direct their marketing and promotional efforts towards a select group of customers to achieve the a beneficial ROI. Those individuals would be customers that have one or more boat policies and or one or more car policies (’APLEZIER’ (# of boat policies), ‘PPLEZIER’ (contribution boar policies), ‘PPERSAUT’ (contribution car policies), ‘APERSAUT’ (# of car policies)). These customers, based upon the results, show a direct correlation to also having a caravan policy.

TIC can focus their efforts towards identifying customers who have one or both of these policies, or may be interested in such, to also suggest or investigate their interest in investing in a caravan policy as well. The probability of success, when utilizing this approach, will prove to be more efficient in terms of both time spent and financial investment and return. This increase in marketing efficacy, by way of business intelligence, will provide more financial freedom for TIC and in turn, will provide more exceptional care and service to the customers of TIC.

References

1. Davis, B. (2017). Predictive Analytics: Four prerequisites of an effective strategy. Retrieved February 8, 2020 from https://econsultancy.com/predictive-analytics-four-prerequisites-of-an-effective-strategy/

2. Faggella, D. (2019). Predictive Analytics for Marketing - What's Possible and How it Works. Retrieved February 7, 2020 from https://emerj.com/ai-sector-overviews/predictive-analytics-for-marketing-whats-possible-and-how-it-works/

3. IBM. (n.d.). Scoring data with predictive models. Retrieved March 15, 2020 from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/base/scoring_wizard_intro.html

4. Koehrsen, W. (2018). Towards Data Science. An Implementation and Exploration of the Random Forest in Python. Retrieved March 19, 2020 from https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

5. Lewinson, E. (2019). Explaining Feature Importance by example of a Random Forest. Retrieved February 26, 2020 from https://towardsdatascience.com/explaining-feature-importance-by- example-of-a-random-forest-d9166011959e

6. P. van der Putten and M. van Someren (eds). CoIL Challenge 2000: The Insurance Company (TIC) Benchmark. Retrieved January 31, 2020 from http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/

7. P. van der Putten and M. van Someren (eds). CoIL Challenge 2000: The Insurance Company (TIC) Benchmark. **Detailed Data Description.** Retrieved January 31, 2020 from http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/data.html. **(includes metadata)**

8. P. van der Putten and M. van Someren (eds). CoIL Challenge 2000: The Insurance Company (TIC) Benchmark. **Original Problem Task Description.** Retrieved January 31, 2020 from http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/problem.html

9. RDocumentation. (n.d.). rfImpute. Retrieved February 26, 2020 from https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/rfImpute

10. StatQuest with Josh Stramer. (2018). StatQuest: Random Forests in R. Retrieved February 23, 2020 from https://www.youtube.com/watch?v=6EXPYzbfLCE

11. StackExchange. (2017). What is the proper way to use rfImpute? (Imputation by Random Forest in R). Retrieved February 27, 2020 from https://stats.stackexchange.com/questions/226803/what- is-the-proper-way-to-use-rfimpute-imputation-by-random-forest-in-r

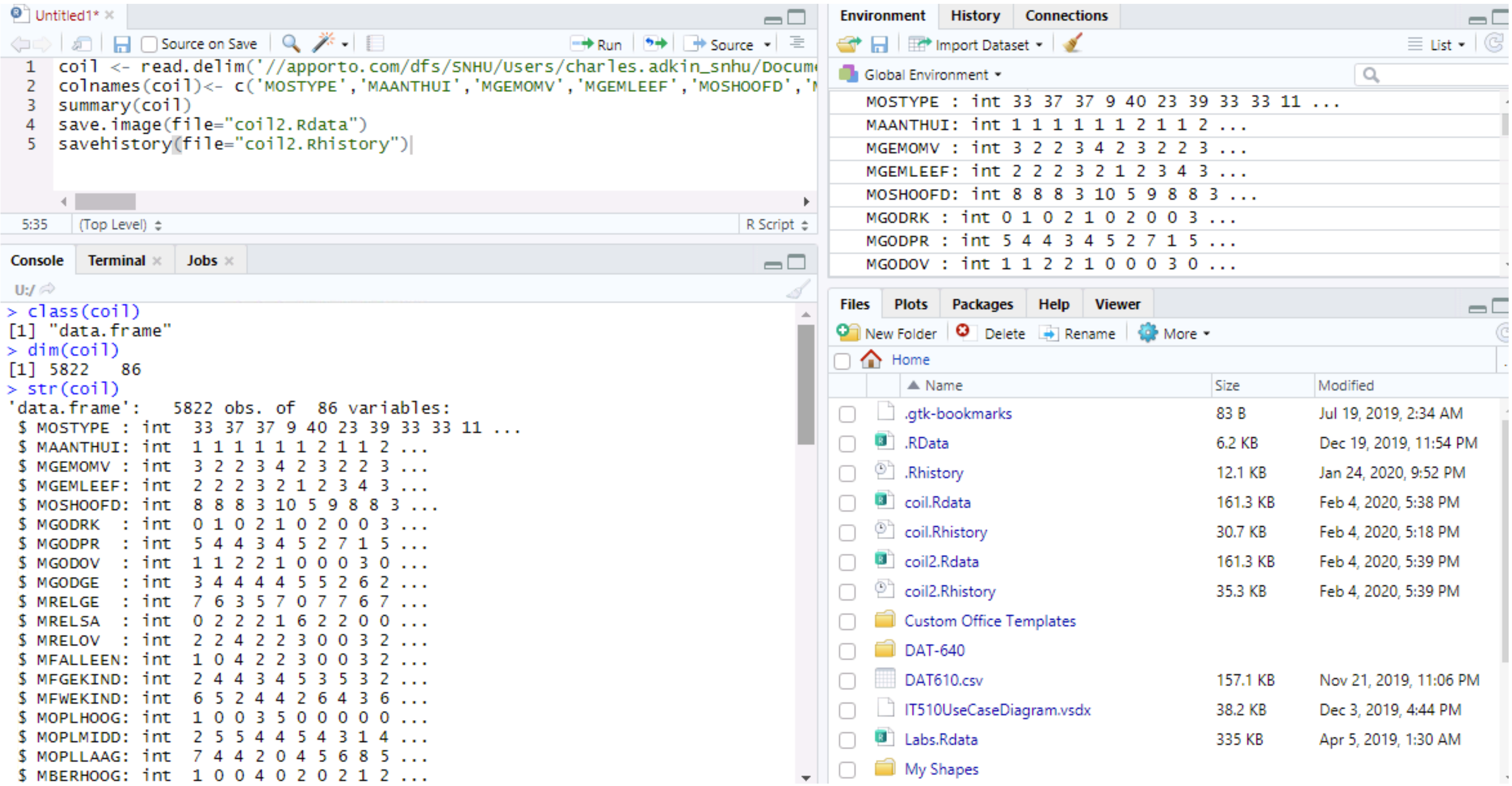
12. Thys, F. (n.d.) Marketing attribution: Web analytics won't work, but predictive analytics will. Retrieved February 8, 2020 from https://www.sas.com/en_us/insights/articles/marketing/marketing-attribution-predictive-analytics.html

13. Williams, G. (2010). DATA MINING. Desktop Survival Guide. Scoring. Retrieved March 17, 2020 from https://www.togaware.com/datamining/survivor/Scoring.html

14. Williams, G. (2011). Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discovery. Ch. 12. Retrieved February 20, 2020.

Appendices

Appendix A: Data set dimension and structure summary



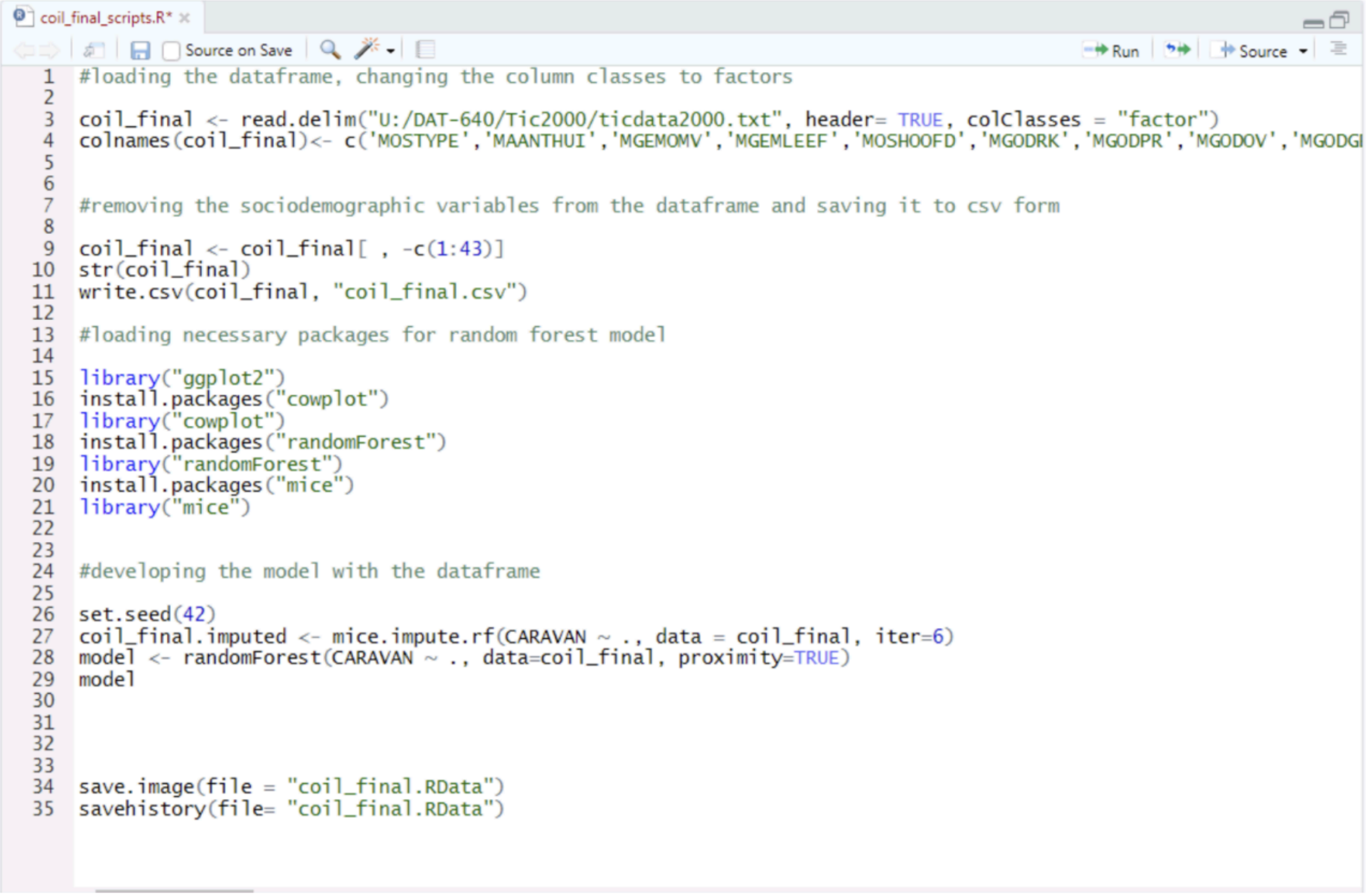
```r
coil <- read.delim('//apporto.com/dfs/SNHU/Users/charles.adkin_snhu/Docume
colnames(coil)<- c('MOSTYPE','MAANTHUI','MGEMOMV','MGEMLEEF','MOSHOOFD','M
summary(coil)
save.image(file="coil2.Rdata")
savehistory(file="coil2.Rhistory")
```

```
> class(coil)
[1] "data.frame"
> dim(coil)
[1] 5822   86
> str(coil)
'data.frame':	5822 obs. of  86 variables:
 $ MOSTYPE : int  33 37 37 9 40 23 39 33 33 11 ...
 $ MAANTHUI: int  1 1 1 1 1 1 2 1 1 2 ...
 $ MGEMOMV : int  3 2 2 3 4 2 3 2 2 3 ...
 $ MGEMLEEF: int  2 2 2 3 2 1 2 3 4 3 ...
 $ MOSHOOFD: int  8 8 8 3 10 5 9 8 8 3 ...
 $ MGODRK  : int  0 1 0 2 1 0 2 0 0 3 ...
 $ MGODPR  : int  5 4 4 3 4 5 2 7 1 5 ...
 $ MGODOV  : int  1 1 2 2 1 0 0 0 3 0 ...
 $ MGODGE  : int  3 4 4 4 4 5 5 2 6 2 ...
 $ MRELGE  : int  7 6 3 5 7 0 7 7 6 7 ...
 $ MRELSA  : int  0 2 2 2 1 6 2 2 0 0 ...
 $ MRELOV  : int  2 2 4 2 2 3 0 0 3 2 ...
 $ MFALLEEN: int  1 0 4 2 2 3 0 0 3 2 ...
 $ MFGEKIND: int  2 4 4 3 4 5 3 5 3 2 ...
 $ MFWEKIND: int  6 5 2 4 4 2 6 4 3 6 ...
 $ MOPLHOOG: int  1 0 0 3 5 0 0 0 0 0 ...
 $ MOPLMIDD: int  2 5 5 4 4 5 4 3 1 4 ...
 $ MOPLLAAG: int  7 4 4 2 0 4 5 6 8 5 ...
 $ MBERHOOG: int  1 0 0 4 0 2 0 2 1 2 ...
```

```
MOSTYPE : int 33 37 37 9 40 23 39 33 33 11 ...
MAANTHUI: int 1 1 1 1 1 1 2 1 1 2 ...
MGEMOMV : int 3 2 2 3 4 2 3 2 2 3 ...
MGEMLEEF: int 2 2 2 3 2 1 2 3 4 3 ...
MOSHOOFD: int 8 8 8 3 10 5 9 8 8 3 ...
MGODRK : int 0 1 0 2 1 0 2 0 0 3 ...
MGODPR : int 5 4 4 3 4 5 2 7 1 5 ...
MGODOV : int 1 1 2 2 1 0 0 0 3 0 ...
```

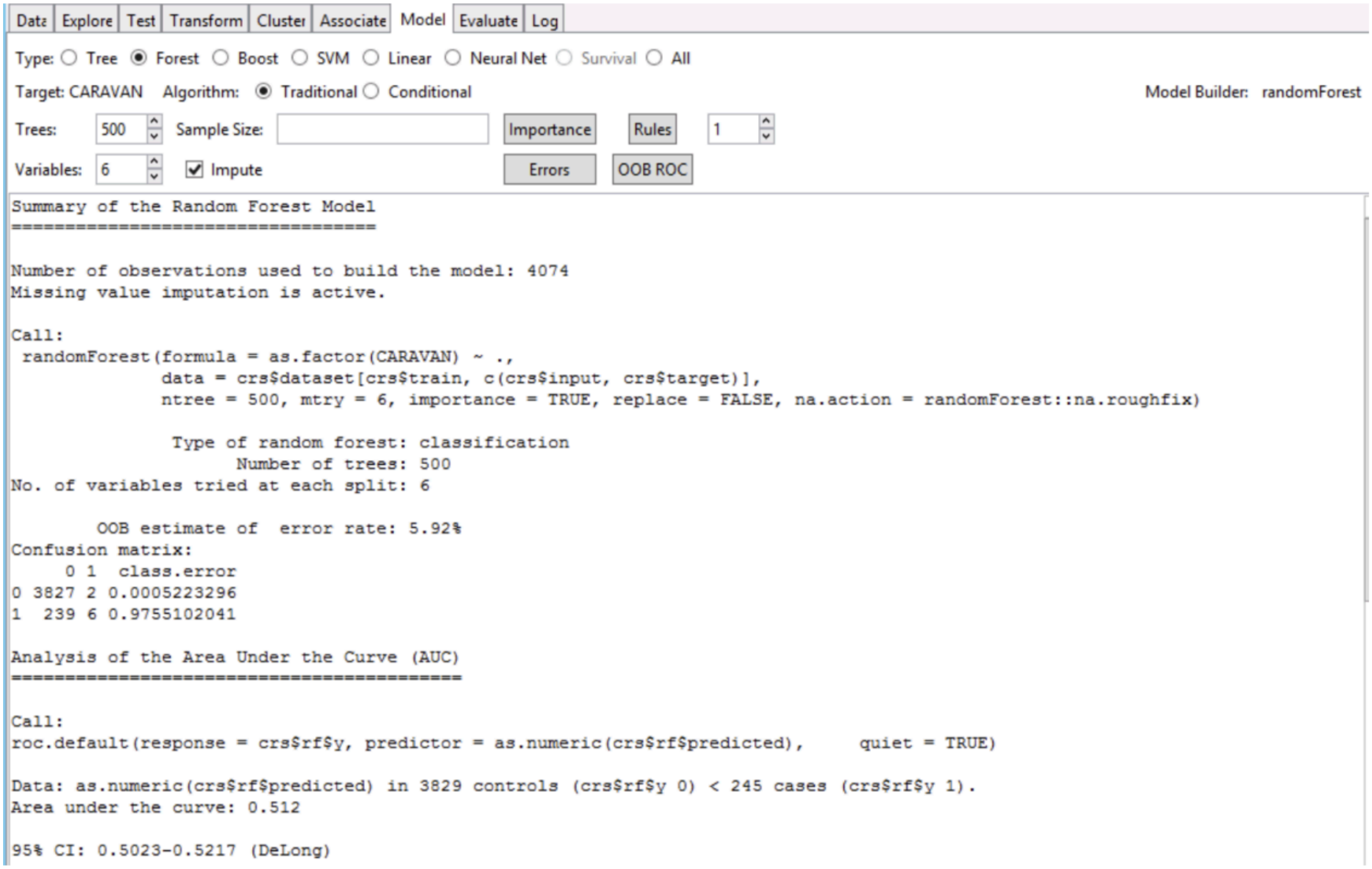
	Name	Size	Modified
□	.gtk-bookmarks	83 B	Jul 19, 2019, 2:34 AM
□	.RData	6.2 KB	Dec 19, 2019, 11:54 PM
□	.Rhistory	12.1 KB	Jan 24, 2020, 9:52 PM
□	coil.Rdata	161.3 KB	Feb 4, 2020, 5:38 PM
□	coil.Rhistory	30.7 KB	Feb 4, 2020, 5:18 PM
□	coil2.Rdata	161.3 KB	Feb 4, 2020, 5:39 PM
□	coil2.Rhistory	35.3 KB	Feb 4, 2020, 5:39 PM
□	Custom Office Templates		
□	DAT-640		
□	DAT610.csv	157.1 KB	Nov 21, 2019, 11:06 PM
□	IT510UseCaseDiagram.vsdx	38.2 KB	Dec 3, 2019, 4:44 PM
□	Labs.Rdata	335 KB	Apr 5, 2019, 1:30 AM
□	My Shapes		

Appendix B. Loading the dataset, changing data types, packages and model structure.



```r
#loading the dataframe, changing the column classes to factors

coil_final <- read.delim("U:/DAT-640/Tic2000/ticdata2000.txt", header= TRUE, colClasses = "factor")
colnames(coil_final)<- c('MOSTYPE','MAANTHUI','MGEMOMV','MGEMLEEF','MOSHOOFD','MGODRK','MGODPR','MGODOV','MGODG


#removing the sociodemographic variables from the dataframe and saving it to csv form

coil_final <- coil_final[ , -c(1:43)]
str(coil_final)
write.csv(coil_final, "coil_final.csv")

#loading necessary packages for random forest model

library("ggplot2")
install.packages("cowplot")
library("cowplot")
install.packages("randomForest")
library("randomForest")
install.packages("mice")
library("mice")


#developing the model with the dataframe

set.seed(42)
coil_final.imputed <- mice.impute.rf(CARAVAN ~ ., data = coil_final, iter=6)
model <- randomForest(CARAVAN ~ ., data=coil_final, proximity=TRUE)
model




save.image(file = "coil_final.RData")
savehistory(file= "coil_final.RData")
```

Appendix C. Rattle Random Forest output



Type: ○ Tree ● Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

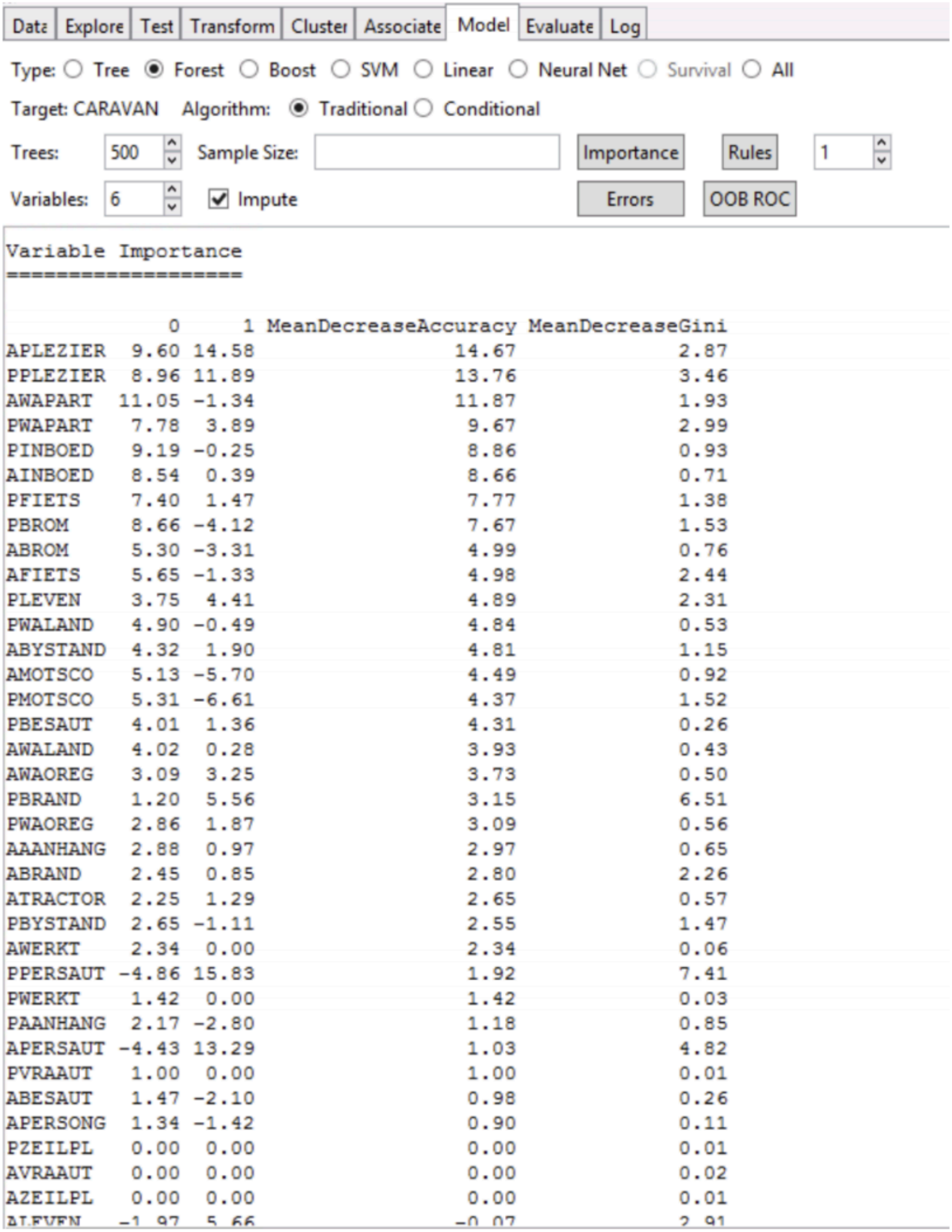
Target: CARAVAN Algorithm: ● Traditional ○ Conditional Model Builder: randomForest

Trees: 500 Sample Size: 1

Variables: 6 ☑ Impute

```
Summary of the Random Forest Model
==================================

Number of observations used to build the model: 4074
Missing value imputation is active.

Call:
 randomForest(formula = as.factor(CARAVAN) ~ .,
              data = crs$dataset[crs$train, c(crs$input, crs$target)],
              ntree = 500, mtry = 6, importance = TRUE, replace = FALSE, na.action = randomForest::na.roughfix)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 5.92%
Confusion matrix:
     0 1  class.error
0 3827 2 0.0005223296
1  239 6 0.9755102041

Analysis of the Area Under the Curve (AUC)
==========================================

Call:
roc.default(response = crs$rf$y, predictor = as.numeric(crs$rf$predicted),     quiet = TRUE)

Data: as.numeric(crs$rf$predicted) in 3829 controls (crs$rf$y 0) < 245 cases (crs$rf$y 1).
Area under the curve: 0.512

95% CI: 0.5023-0.5217 (DeLong)
```

Appendix D. Random Forest Variable Importance output



Type: ○ Tree ● Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: CARAVAN Algorithm: ● Traditional ○ Conditional

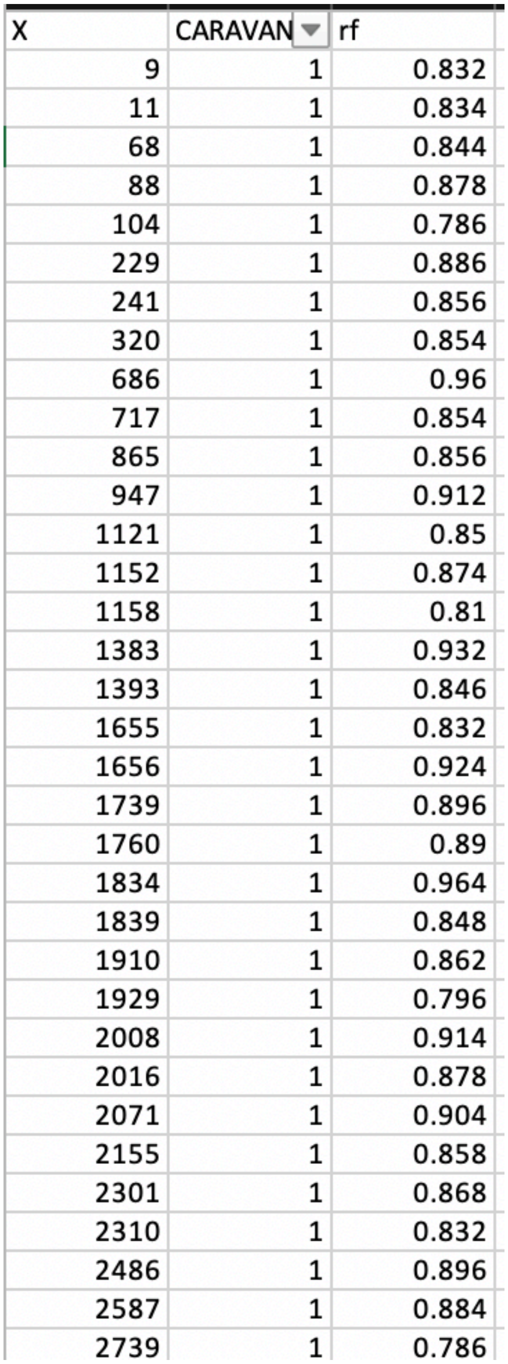
Trees: 500 Sample Size: 1

Variables: 6 ☑ Impute

```
Variable Importance
===================

                0     1 MeanDecreaseAccuracy MeanDecreaseGini
APLEZIER   9.60 14.58                14.67             2.87
PPLEZIER   8.96 11.89                13.76             3.46
AWAPART   11.05 -1.34                11.87             1.93
PWAPART    7.78  3.89                 9.67             2.99
PINBOED    9.19 -0.25                 8.86             0.93
AINBOED    8.54  0.39                 8.66             0.71
PFIETS     7.40  1.47                 7.77             1.38
PBROM      8.66 -4.12                 7.67             1.53
ABROM      5.30 -3.31                 4.99             0.76
AFIETS     5.65 -1.33                 4.98             2.44
PLEVEN     3.75  4.41                 4.89             2.31
PWALAND    4.90 -0.49                 4.84             0.53
ABYSTAND   4.32  1.90                 4.81             1.15
AMOTSCO    5.13 -5.70                 4.49             0.92
PMOTSCO    5.31 -6.61                 4.37             1.52
PBESAUT    4.01  1.36                 4.31             0.26
AWALAND    4.02  0.28                 3.93             0.43
AWAOREG    3.09  3.25                 3.73             0.50
PBRAND     1.20  5.56                 3.15             6.51
PWAOREG    2.86  1.87                 3.09             0.56
AAANHANG   2.88  0.97                 2.97             0.65
ABRAND     2.45  0.85                 2.80             2.26
ATRACTOR   2.25  1.29                 2.65             0.57
PBYSTAND   2.65 -1.11                 2.55             1.47
AWERKT     2.34  0.00                 2.34             0.06
PPERSAUT  -4.86 15.83                 1.92             7.41
PWERKT     1.42  0.00                 1.42             0.03
PAANHANG   2.17 -2.80                 1.18             0.85
APERSAUT  -4.43 13.29                 1.03             4.82
PVRAAUT    1.00  0.00                 1.00             0.01
ABESAUT    1.47 -2.10                 0.98             0.26
APERSONG   1.34 -1.42                 0.90             0.11
PZEILPL    0.00  0.00                 0.00             0.01
AVRAAUT    0.00  0.00                 0.00             0.02
AZEILPL    0.00  0.00                 0.00             0.01
ALEVEN    -1.97  5.66                -0.07             2.91
```

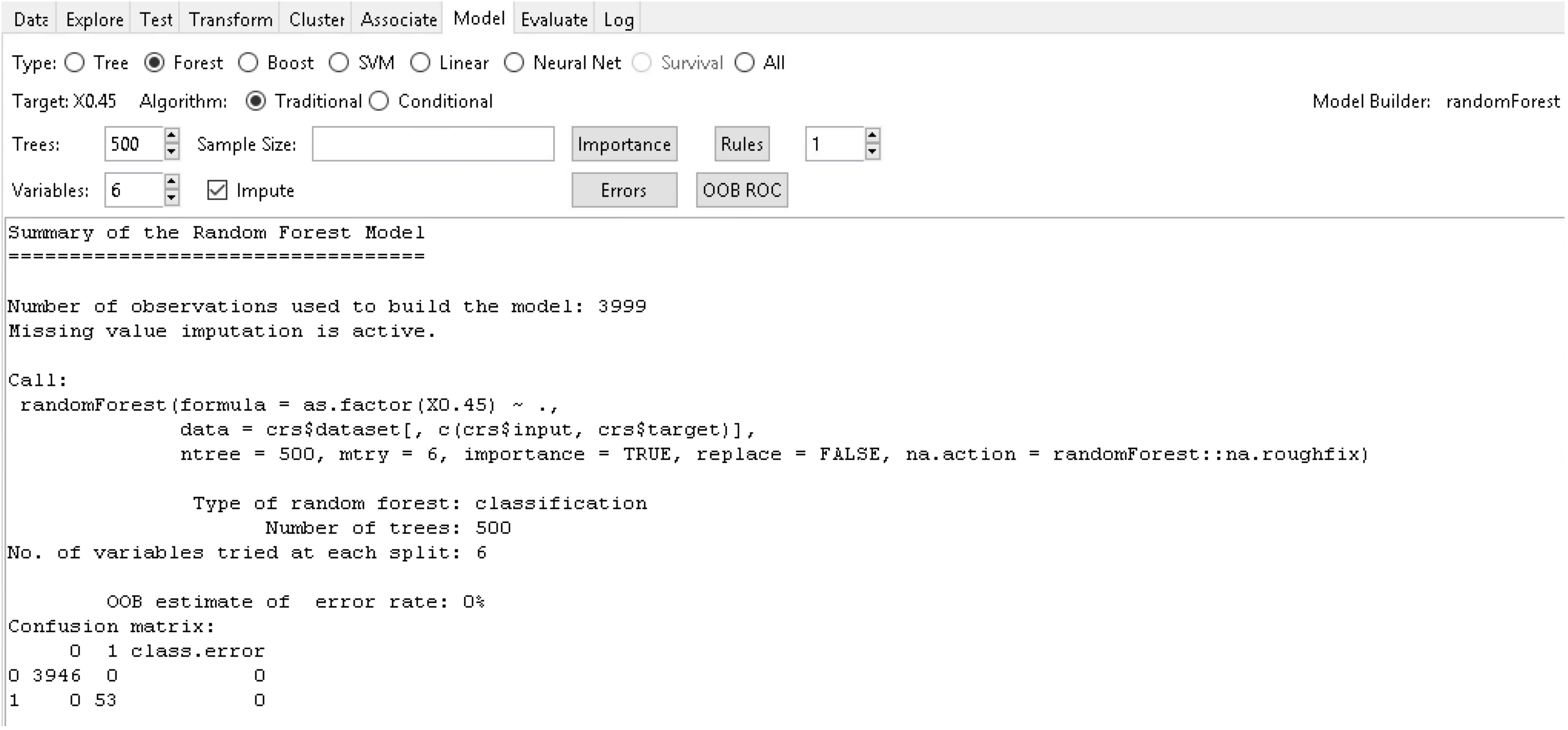
Appendix E. Model Scoring probability results.

(The results continued further, of which included all 3999 observations)



X	CARAVAN	rf
9	1	0.832
11	1	0.834
68	1	0.844
88	1	0.878
104	1	0.786
229	1	0.886
241	1	0.856
320	1	0.854
686	1	0.96
717	1	0.854
865	1	0.856
947	1	0.912
1121	1	0.85
1152	1	0.874
1158	1	0.81
1383	1	0.932
1393	1	0.846
1655	1	0.832
1656	1	0.924
1739	1	0.896
1760	1	0.89
1834	1	0.964
1839	1	0.848
1910	1	0.862
1929	1	0.796
2008	1	0.914
2016	1	0.878
2071	1	0.904
2155	1	0.858
2301	1	0.868
2310	1	0.832
2486	1	0.896
2587	1	0.884
2739	1	0.786

Appendix F. Model Evaluation results



Type: ○ Tree ● Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

Target: X0.45 Algorithm: ● Traditional ○ Conditional Model Builder: randomForest

Trees: 500 Sample Size:

Variables: 6 ☑ Impute

```
Summary of the Random Forest Model
==================================

Number of observations used to build the model: 3999
Missing value imputation is active.

Call:
 randomForest(formula = as.factor(X0.45) ~ .,
              data = crs$dataset[, c(crs$input, crs$target)],
              ntree = 500, mtry = 6, importance = TRUE, replace = FALSE, na.action = randomForest::na.roughfix)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 0%
Confusion matrix:
     0  1 class.error
0 3946  0           0
1    0 53           0
```