STAT515 – 005 – Spring 2023 – Final Project

Team 2: Preethal Reddy Yellareddygari, Nivedita J, Grace Manasseh

1. Project Overview

The aim of this Project is to analyze a dataset related to Caravan Insurance Policies and try to answer the following main questions based on statistical and machine learning approaches learned during this course.

a. Main Questions:

1. When given a new customer profile, how likely are they to buy a caravan insurance policy?
2. Can we group caravan insurance policy holders into clusters based on their income, social class, family size and other demographic characteristics? And Can we classify the caravan insurance policy holders into the target variables?
3. Is there a relationship between buying a caravan insurance and having other insurance policy types (fire, property, life, car, etc…)?

b. Dataset

Source of the dataset: UCI Machine Learning Repository: Insurance Company Benchmark (COIL 2000) Data Set

This data set used in the CoIL 2000 Challenge contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data. The data was supplied by the Dutch data mining company Sentient Machine Research and is based on a real-world business problem. The training set contains over 5000 descriptions of customers, including the information of whether they have a caravan insurance policy. A test set contains 4000 customers of whom only the organizers know if they have a caravan insurance policy.

c. Attribute Description

Col	Name	Description Domain	Renamed Columns	
1	MOSTYPE	Customer Subtype see L0	CustomerSubtype	Socio-Demographic Data
2	MAANTHUI	Number of houses 1 – 10	NbHouses	
3	MGEMOMV	Avg size household 1 – 6	AvgSizeHousehold	
4	MGEMLEEF	Avg age see L1	AvgAge	
5	MOSHOOFD	Customer main type see L2	CustomerMainType	
6	MGODRK	Roman catholic see L3	RomanCatholic	
7	MGODPR	Protestant ...	Protestant	
8	MGODOV	Other religion	OtherReligion	
9	MGODGE	No religion	NoReligion	
10	MRELGE	Married	Married	
11	MRELSA	Living together	LivingTogether	
12	MRELOV	Other relation	OtherRelation	
13	MFALLEEN	Singles	Singles	
14	MFGEKIND	Household without children	HouseholdNochildren	

15	MFWEKIND	Household with children	HouseholdWithchildren	
16	MOPLHOOG	High level education	HighLevelEducation	
17	MOPLMIDD	Medium level education	MediumLevelEducation	
18	MOPLLAAG	Lower level education	LowerLevelEducation	
19	MBERHOOG	High status	HighStatus	
20	MBERZELF	Entrepreneur	Entrepreneur	
21	MBERBOER	Farmer	Farmer	
22	MBERMIDD	Middle management	MiddleManagement	
23	MBERARBG	Skilled labourers	SkilledLabourers	
24	MBERARBO	Unskilled labourers	UnskilledLabourers	
25	MSKA	Social class A	SocialclassA	
26	MSKB1	Social class B1	SocialclassB1	
27	MSKB2	Social class B2	SocialclassB2	
28	MSKC	Social class C	SocialclassC	
29	MSKD	Social class D	SocialclassD	
30	MHHUUR	Rented house	RentedHouse	
31	MHKOOP	Home owners	HomeOwners	
32	MAUT1	1 car	1car	
33	MAUT2	2 cars	2cars	
34	MAUT0	No car	Nocar	
35	MZFONDS	National Health Service	NationalHealthService	
36	MZPART	Private health insurance	PrivateHealthInsurance	
37	MINKM30	Income < 30	Income<30	
38	MINK3045	Income 30-45.000	Income30-45.000	
39	MINK4575	Income 45-75.000	Income45-75.000	
40	MINK7512	Income 75-122.000	Income75-122.000	
41	MINK123M	Income >123.000	Income>123.000	
42	MINKGEM	Average income	AverageIncome	
43	MKOOPKLA	Purchasing power class	PurchasingPowerClass	
44	PWAPART	Contribution private third party insurance seeL4	PrivateThirdPartyInsurance	Product Ownership
45	PWABEDR	Contribution third party insurance (firms) ...	ThirdPartyInsuranceFirms	
46	PWALAND	Contribution third party insurane (agriculture)	ThirdPartyInsuraneAgriculture	
47	PPERSAUT	Contribution car policies	CarPolicies	
48	PBESAUT	Contribution delivery van policies	DeliveryVanPolicies	
49	PMOTSCO	Contribution motorcycle/scooter policies	MotorcycleScooterPolicies	
50	PVRAAUT	Contribution lorry policies	LorryPolicies	
51	PAANHANG	Contribution trailer policies	TrailerPolicies	
52	PTRACTOR	Contribution tractor policies	TractorPolicies	
53	PWERKT	Contribution agricultural machines policies	AgriculturalMachinesPolicies	
54	PBROM	Contribution moped policies	MopedPolicies	

55	PLEVEN	Contribution life insurances	LifeInsurances	
56	PPERSONG	Contribution private accident insurance policies	PrivateAccidentPolicies	
57	PGEZONG	Contribution family accidents insurance policies	FamilyAccidentPolicies	
58	PWAOREG	Contribution disability insurance policies	DisabilityInsurancePolicies	
59	PBRAND	Contribution fire policies	FirePolicies	
60	PZEILPL	Contribution surfboard policies	SurfboardPolicies	
61	PPLEZIER	Contribution boat policies	BoatPolicies	
62	PFIETS	Contribution bicycle policies	BicyclePolicies	
63	PINBOED	Contribution property insurance policies	PropertyInsurancePolicies	
64	PBYSTAND	Contribution social security insurance policies	SocialSecurityInsurancePolicies	
65	AWAPART	Number of private third party insurance1- 12	NbPrivateThirdPartyInsurance	
66	AWABEDR	Number of third party insurance (firms)...	NbThirdPartyInsuranceFirms	
67	AWALAND	Number of third party insurane (agriculture)	NbThirdPartyInsuranceAgriculture	
68	APERSAUT	Number of car policies	NbCarPolicies	
69	ABESAUT	Number of delivery van policies	NbDeliveryVanPolicies	
70	AMOTSCO	Number of motorcycle/scooter policies	NbMotorcycleScooterPolicies	
71	AVRAAUT	Number of lorry policies	NbLorryPolicies	
72	AAANHANG	Number of trailer policies	NbTrailerPolicies	
73	ATRACTOR	Number of tractor policies	NbTractorPolicies	
74	AWERKT	Number of agricultural machines policies	NbAgriculturalMachinesPolicies	
75	ABROM	Number of moped policies	NbMopedPolicies	
76	ALEVEN	Number of life insurances	NbLifeInsurances	
77	APERSONG	Number of private accident insurance policies	NbPrivateAccidentPolicies	
78	AGEZONG	Number of family accidents insurance policies	NbFamilyAccidentsPolicies	
79	AWAOREG	Number of disability insurance policies	NbDisabilityInsurancePolicies	
80	ABRAND	Number of fire policies	NbFirePolicies	
81	AZEILPL	Number of surfboard policies	NbSurfboardPolicies	
82	APLEZIER	Number of boat policies	NbBoatPolicies	
83	AFIETS	Number of bicycle policies	NbBicyclePolicies	
84	AINBOED	Number of property insurance policies	NbPropertyInsurancePolicies	
85	ABYSTAND	Number of social security insurance policies	NbSocialSecurityInsurancePolicies	
86	CARAVAN	Number of mobile home policies 0-1	NbMobileHomePolicies	Target

2. Descriptive Statistics

a. Summary

Dependent Variable: NbMobileHomePolicies (which is 1 when customer has purchased Caravan Insurance, and 0 otherwise)

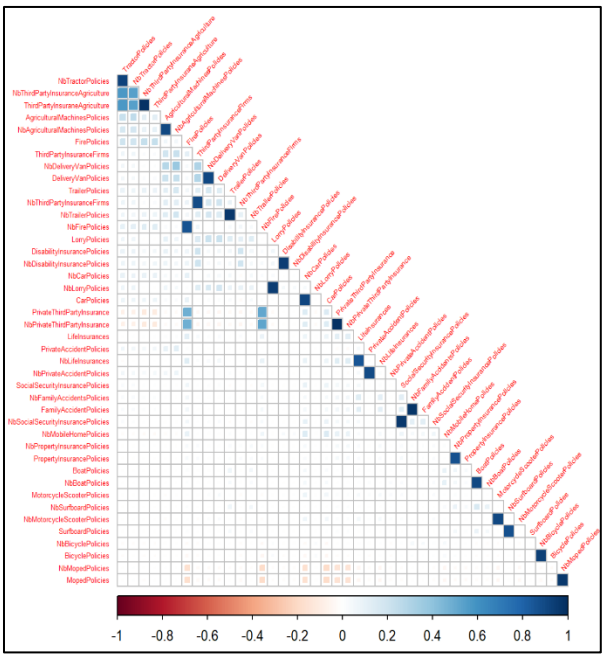
Numeric Variables (85): Socio-Demographic Attributes (1-43) and Product Purchase Attributes (44-85)

Training Dataset = 5822 Observations having 348 classified with Target = 1 and 5474 with Target = 0. (~ 6.3% having 1)

Test Dataset = 4000 Observations having 238 classified with Target = 1 and 3762 with Target = 0. (~6.3% having 1)

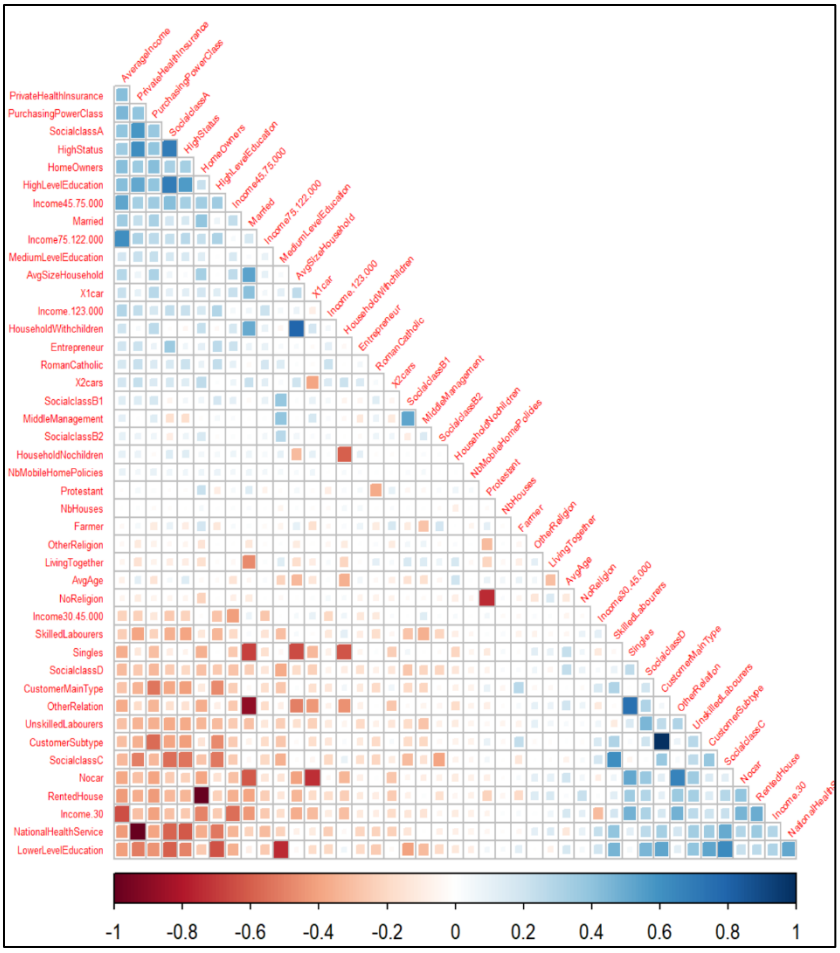
b. Correlations

Check the Correlation Matrix among all attributes. This was done in chunks, by looking first at the insurance product purchase attributes 44-85, the following matrix was obtained:



One can notice that the 2 values, whether the customer has insurance and the number of policies of each insurance product are very highly correlated. This is occurring for each type of insurance attribute, so we kept only one of each, the one related if the customer had this type of insurance.

By looking next at the second chunk of attributes (1:43) related to socio-demographics, the following matrix is observed:

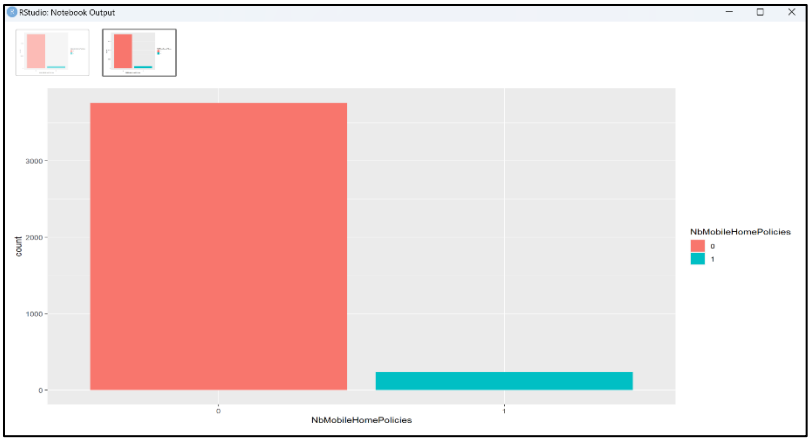


The following attributes seem to be very highly correlated, and we can keep only one of them:

OtherRelation v.s. Married
Singles v.s. Married
RentedHouse v.s HomeOwners
HouseholdNoChildren v.s. HouseholdWithChildren
CustomerMainType v.s. CustomerSubType

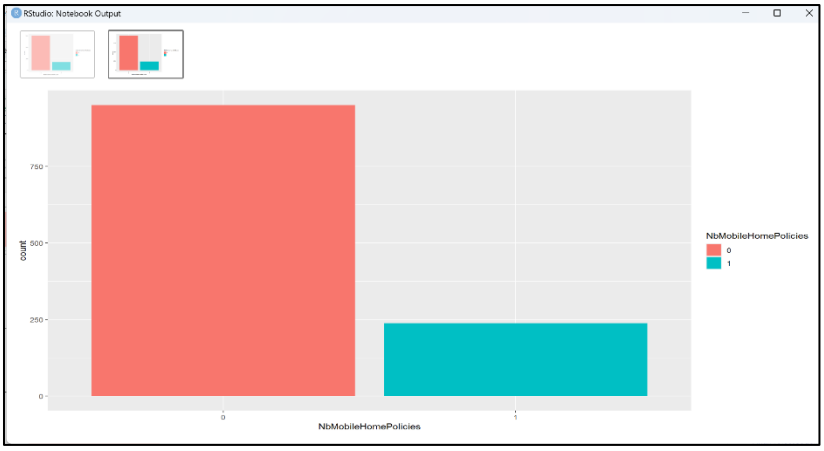
c. Imbalanced Data

The data seems imbalanced in both Training and Test samples, both with the same % of target observations classified as 1's being 6.3%.



We have created another dataset by removing some rows that have 0 as Target in an attempt to create a somewhat balanced dataset. However, during the modeling process we were using both the original and the dataset that treated the imbalanced data to compare results.

Below is a chart showing the Treated dataset after removing some of the 0's from both the training and test datasets.



3. Question 1: Predicting Customers Purchasing Mobile Home Insurance

To tackle this question, we applied 2 methods: Random Forest and Logistic Regression and compared the results at the end by comparing the ROC curve for the True Positives.

a. Random Forest

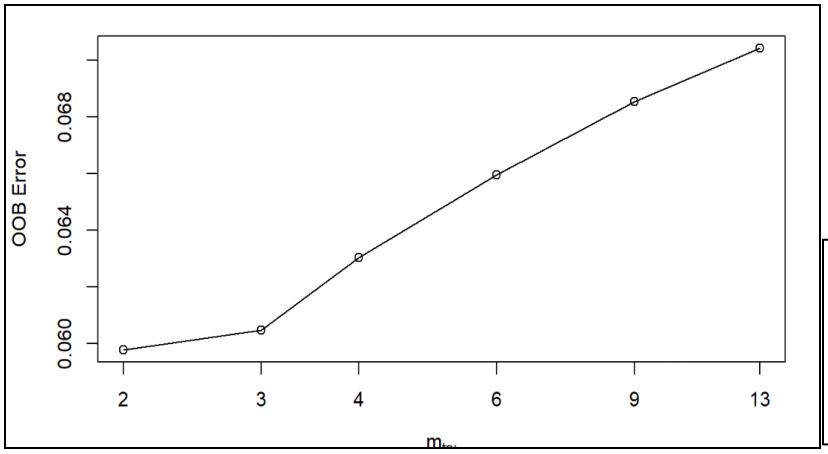
By applying random forest classification on the **Full Dataset as a first model** we got the below confusion matrix results:

	Actual 0 = No	**Actual 1 =Yes**
Predicted 0 = No	5415	59
Predicted 1 = Yes	337	11

OOB estimate of error rate: 6.8%

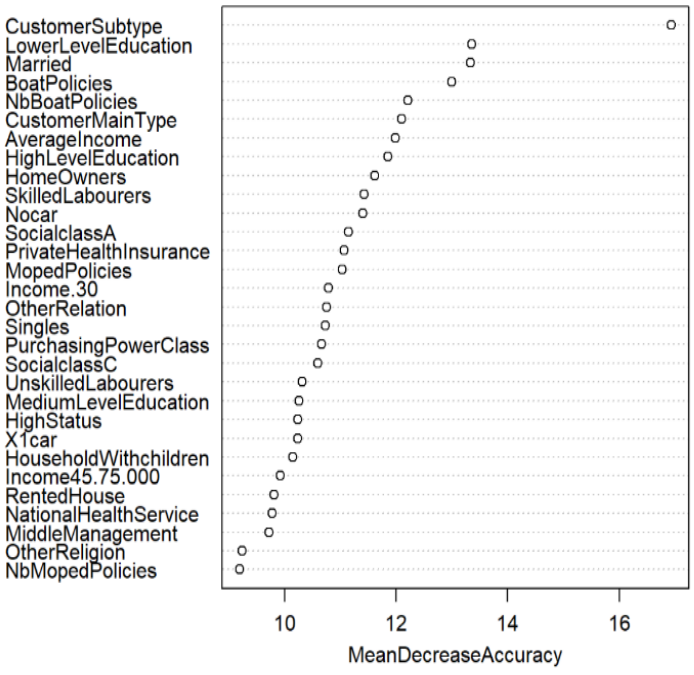
True Positives = 11/70 = 16%

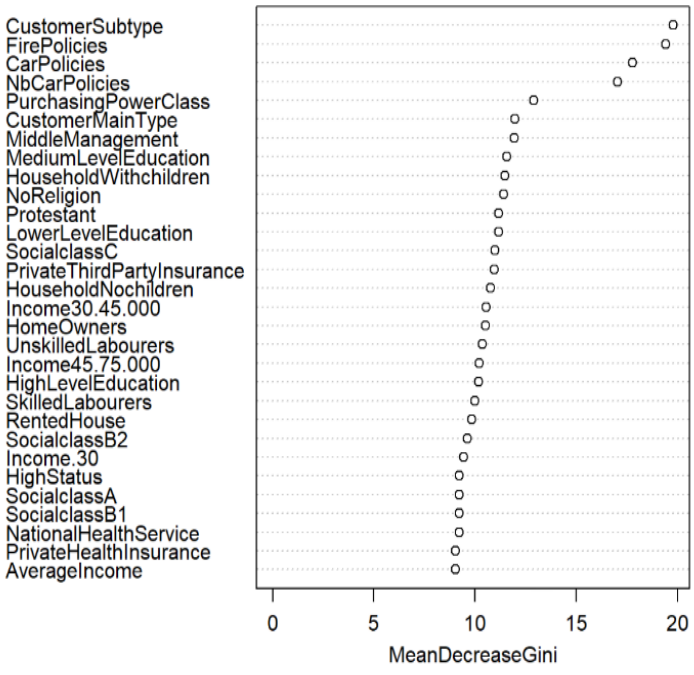
After searching for **the best Number of random variables used in each tree** (**mtry**), the algorithm was run based on the best mtry = 2. However, when applying the mtry = 2, the number of true positives was 0, so we kept the **default mtry = 9** with the results indicated above.



	mtry	OOBError
2.OOB	2	0.05977327
3.OOB	3	0.06046032
4.OOB	4	0.06303676
6.OOB	6	0.06595672
9.OOB	9	0.06853315
13.OOB	13	0.07042254

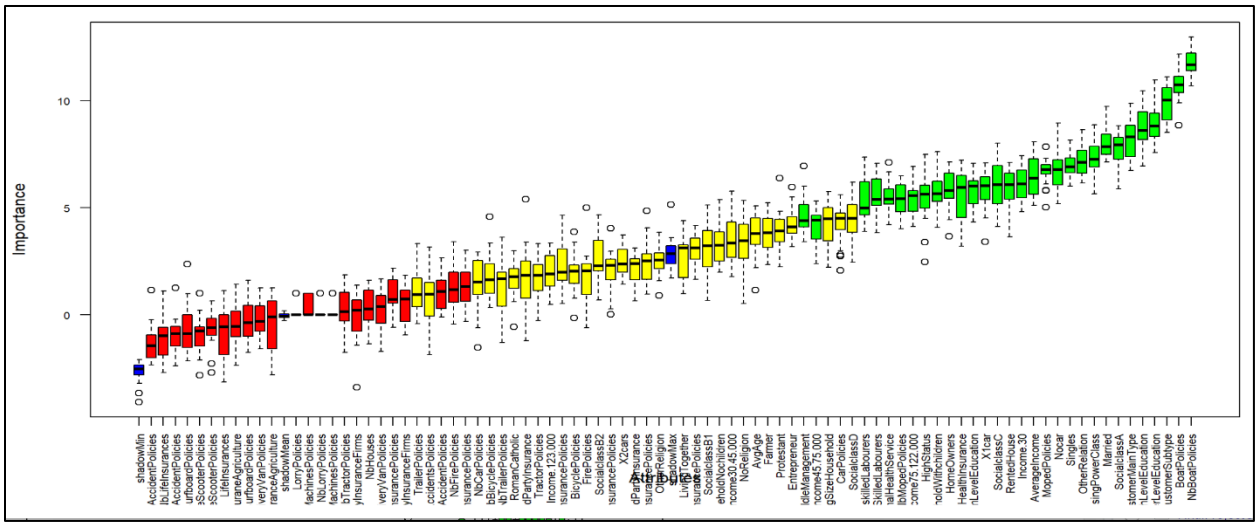
By examining the importance of the attributes in the classification random forest, the below chart displays the **Mean Decrease Accuracy** (*How much the model accuracy decreases if we drop that variable*) and the **Mean Decrease Gini** (*Measure of variable importance based on the Gini impurity index used for the calculation of splits in trees*):



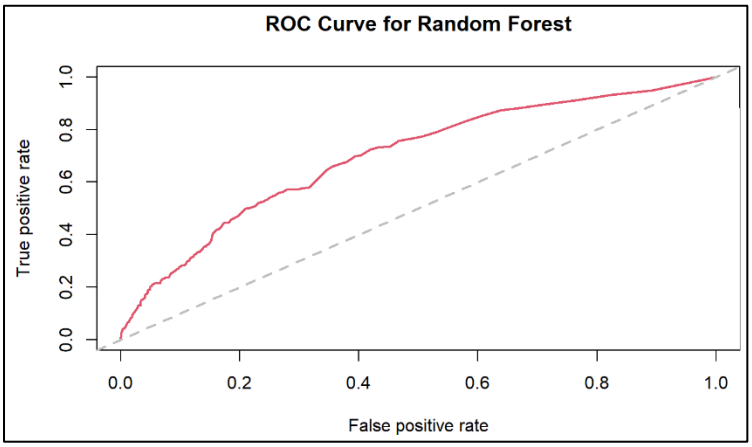


In the plot above, we notice that the **CustomerSubType** is the most important variable.
Other Important Variables are: **FirePolicies, CarPolicies, NbCarPolicies, PurchasingPowerClass**
And also: **NbBoatPolicies, BoatPolicies, Married, SocialClassC, LowerLevelEducation**

By applying the **Boruta library**, the above information is displayed in a more attractive chart:



The **ROC Curve** for the first random forest model is displayed below, showing the True Positive rate being higher than the False Positive rate:



A **second model was built using Random Forest while using the dataset having the treated imbalanced data**. The best mtry was set to 12 in this case. Below confusion matrix results were obtained:

	Actual 0 = No	**Actual 1 =Yes**
Predicted 0 = No	1416	87
Predicted 1 = Yes	285	63

OOB estimate of error rate: 20.1%

We noticed the OOB error rate increased to 20.1% when compared with the the first model. However, if we look at the True Positive matches, we notice a much higher number 63 out of 150 total 1 were true positives (42%). And the false positive rate is 285/1701 (17%).

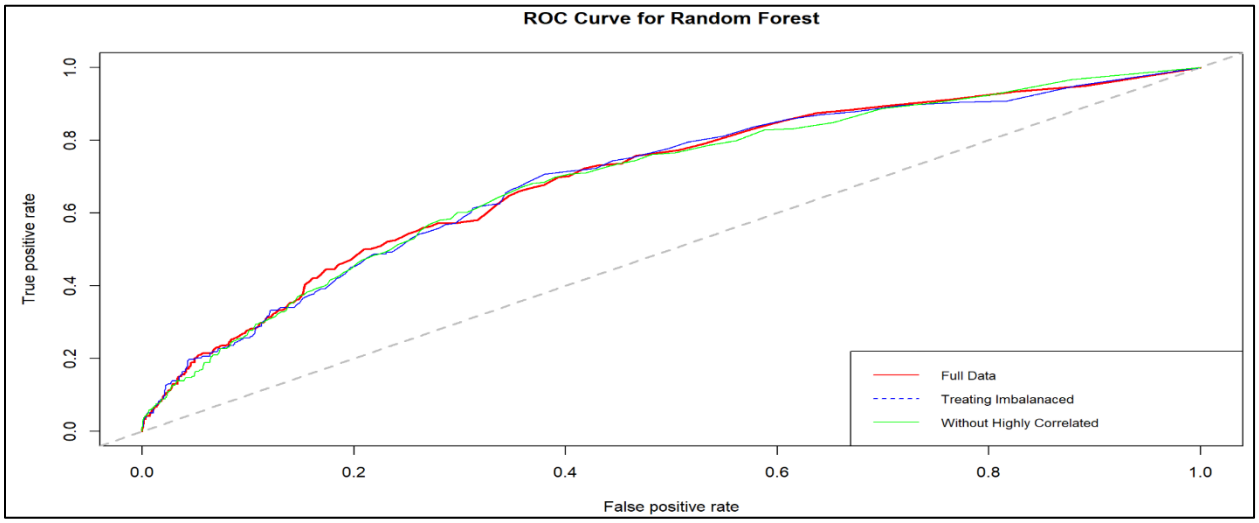
The first model had only 11% matching true positives.

A **third model was built using Random Forest while the highly correlated attributed** were removed. The best mtry variable in this case = 3. Below confusion matrix results were obtained:

	Actual 0 = No	**Actual 1 =Yes**
Predicted 0 = No	5429	45
Predicted 1 = Yes	338	10

OOB estimate of error rate: 6.58%

The OOB error rate is less than model 1 as well as model 2. The below chart displays the comparison of the predictions of the 3 models. They are very close to one another.



b. Logistic Regression

A **fourth model was built to do prediction using Logistic Regression using the Treated Imbalanced as well as removing the highly correlated attributes** and was compared with Random Forest.

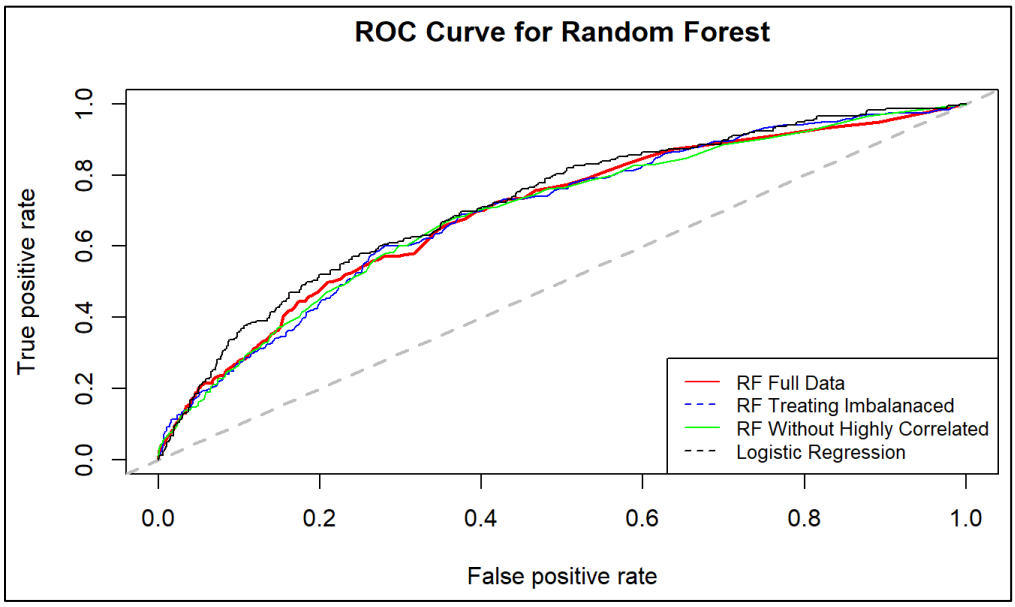
When using 50% as classifying probability, only 25 of the test observations are predicted to purchase insurance. The true positive rate is ~11% (25/238).

	Actual 0 = No	**Actual 1 =Yes**
Predicted 0 = No	928	213
Predicted 1 = Yes	22	25

When using 25% as classifier, we get better results, we have 124 predicted to purchase insurance with a true positive rate of ~52% (124 / 238).

	Actual 0 = No	**Actual 1 =Yes**
Predicted 0 = No	757	114
Predicted 1 = Yes	193	124

c. Comparing Prediction Models

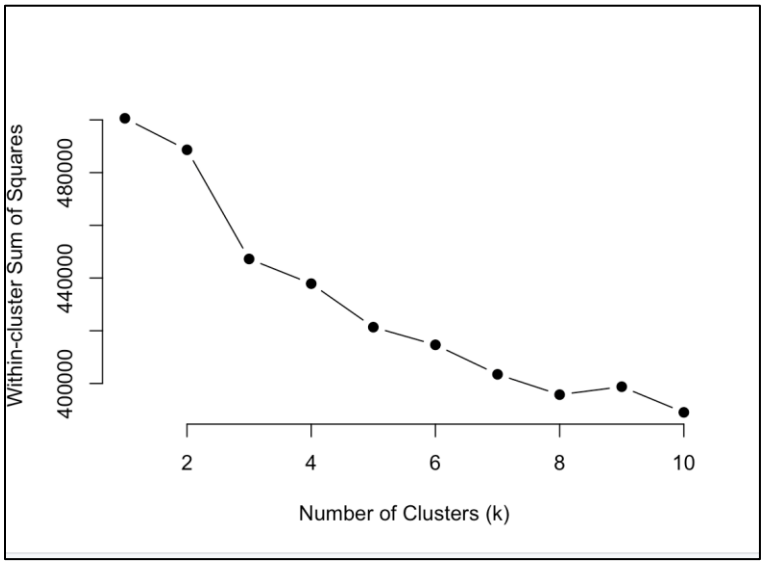


We can see the Logistic Regression which is represented by the black line on the ROC graph seems to perform a bit better than the 3 Random Forest models.

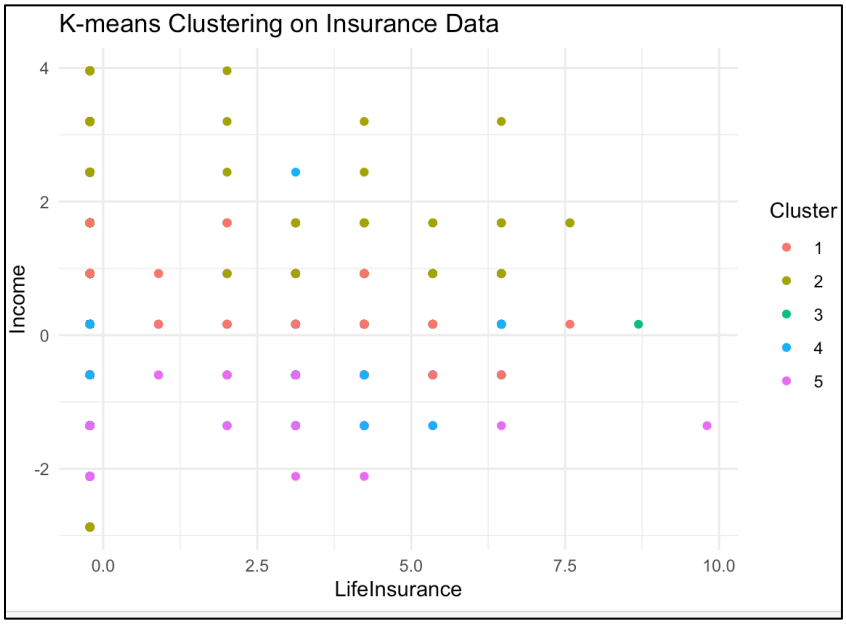
4. Question 2: Grouping caravan insurance policy holders into clusters based on their income, social class, family size and other demographic characteristics?

a. K-Means Clustering

Kmeans clustering model on Insurance data to find customer segmentation. Trained a clustering model with different K ranging from [1,10] and plotted the elbow curve. This plot is an objective function(Sum of Squared errors) on the y-axis and the number of clusters k on the x-axis. Considering this, we have chosen an optimal k value of 5.



To understand the different segmentation, we have plotted the Income vs Number of Life Insurance Policies different customers have brought with their income range.



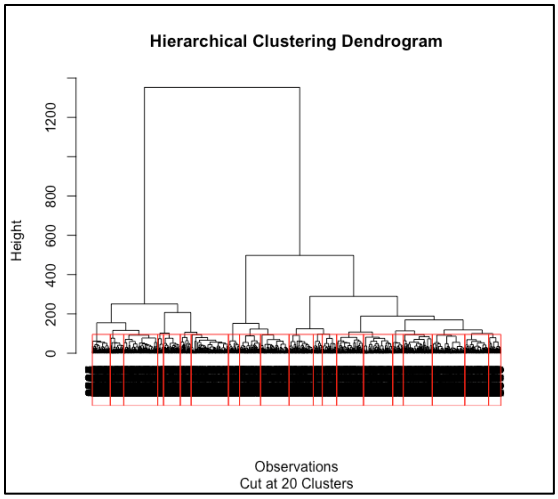
The potential reasons why this could have happened, even after removing outliers from the data using Inter Quartile Range Method and normalizing values.

1. While k means using Euclidean distance to calculate similarities between different instances in the dataset, making the clusters look spherical and equal in size. While the data has distinct scales, as some values are categorical and some are numerical and Euclidean distance does not work well with this kind of data type, or the clusters probably have different shapes and densities.
2. The centroids in the k-means clustering algorithm are initialized by random selection, which could have produced less than ideal-outcomes.
3. This dataset could have overlapping clusters or lack a distinct separation, and k-means clustering would not work well for this data.

With this limitation in mind, we have considered using a hierarchical clustering algorithm to perform clustering.

b. Hierarchical Clustering

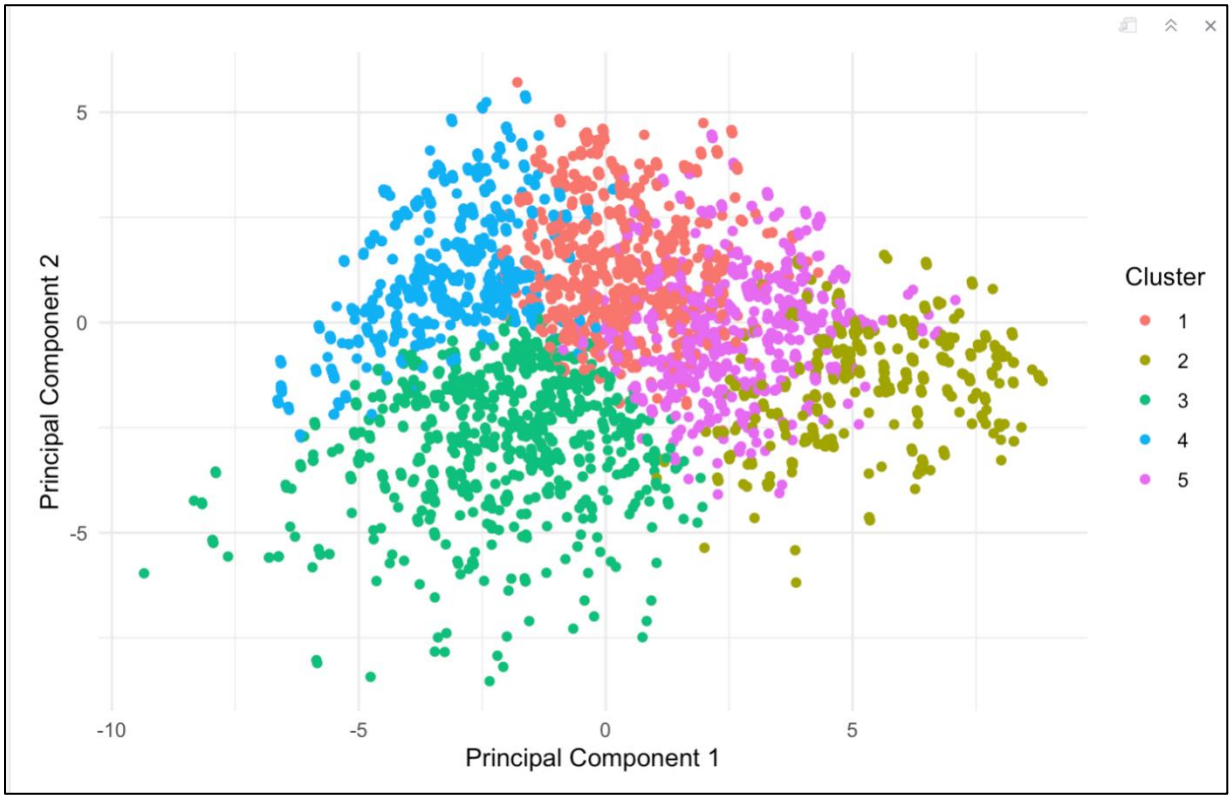
Ward's linkage, a method to minimize the rise in within-cluster variance, has been used as the distance measure. With this technique, we can capture intricate relationships in the data. Dendrogram visualization, below is a dendrogram with 20 clusters cut in observations,



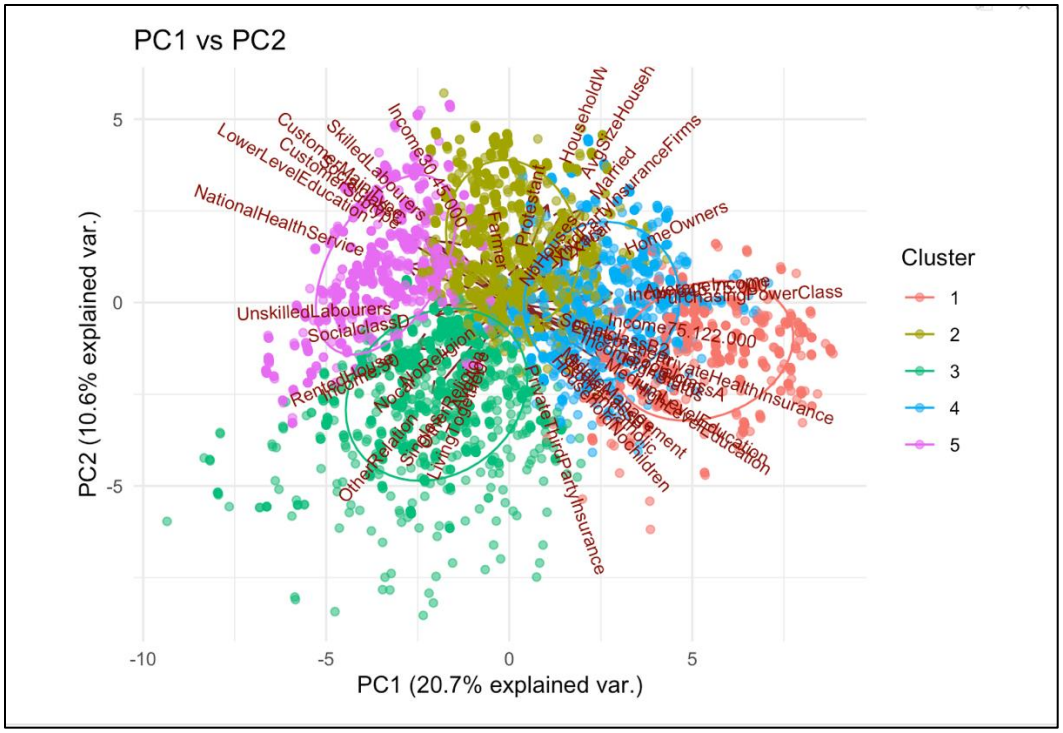
The drawback of this model for our dataset is that we have many observations and variables; it is tough to interpret this large dendrogram visualization, making it difficult to identify meaningful clusters and patterns.

c. PCA + K-Means Clustering

Tested on Dimensionality reduction techniques to have meaningful clusters. Used PCA to reduce the dimensionality of the data as it works by capturing linear combinations of variables.



There is some reasonable separation between the clusters, and some clusters overlap. Understanding clusters and their demographic information,



Here we can clearly see which Demographic Variables have influenced formation of a cluster. For example, in Cluster number 3, Customers who have rented a house, have no car and have income less than or equal to 30 thousand dollars a month with no religion differentiation have been grouped. By identifying these groups, we can custom our marketing strategies and provide better products to meet the needs and preferences of each segment.

5. Question 3: Relationship between Caravan Insurance and Other Insurance Policies

To thoroughly investigate the relationship between buying caravan insurance and other insurance policies, we employed a statistical analysis technique known as the Chi-squared test. This test allowed us to determine if the presence of certain insurance policies influenced the decision to purchase caravan insurance. We also visualized the results using line plots to gain further insights and facilitate interpretation.

a. Chi Squared Test

The analysis showed that purchasing caravan insurance was significantly associated with certain insurance policy types. These included private third-party insurance, car policies, motorcycle/scooter policies, bicycle policies, fire policies, and social security insurance policies. The low p-values obtained for these variables indicated a strong relationship between buying caravan insurance and the presence of these specific insurance policies.

Conversely, there was no significant relationship found between buying caravan insurance and other insurance policy types. These non-significant associations encompassed insurance from third-party insurance firms, agriculture-related third-party insurance, delivery van policies, lorry policies, trailer policies, tractor policies, agricultural machines policies, life insurances, family accident policies,

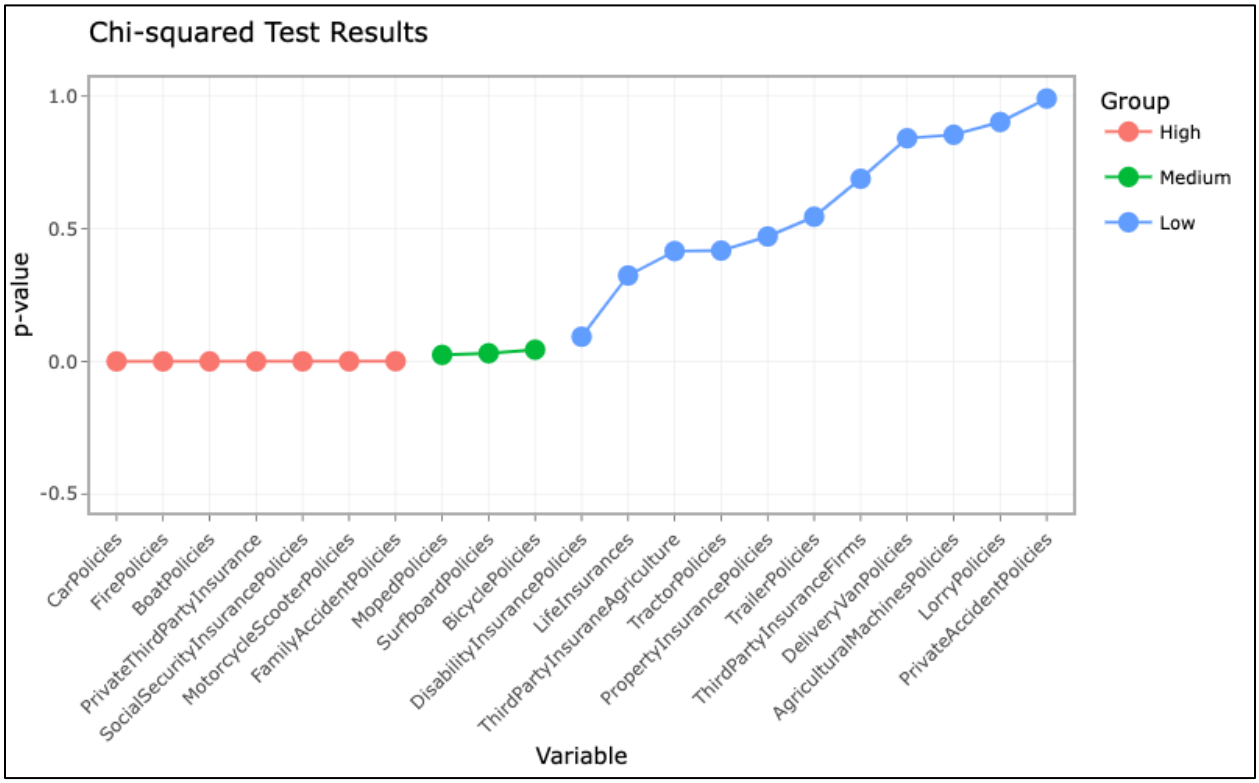
disability insurance policies, fire policies, surfboard policies, boat policies, and property insurance policies. The higher p-values for these variables suggested that the presence or absence of these insurance policies did not significantly impact the likelihood of purchasing caravan insurance.

Variable	P_Value	Numeric Value	Group
CarPolicies	3.89E-40	0.0000000000	High
FirePolicies	1.97E-26	0.0000000000	High
BoatPolicies	2.28E-15	0.0000000000	High
PrivateThirdPartyInsurance	2.03E-12	0.0000000000	High
SocialSecurityInsurancePolicies	6.60E-06	0.0000065984	High
MotorcycleScooterPolicies	2.50E-04	0.0002495952	High
FamilyAccidentPolicies	7.31E-04	0.0007309731	High
MopedPolicies	2.46E-02	0.0246277500	Medium
SurfboardPolicies	3.08E-02	0.0307710500	Medium
BicyclePolicies	4.41E-02	0.0440804400	Medium
DisabilityInsurancePolicies	9.31E-02	0.0931024600	Low
LifeInsurances	3.24E-01	0.3239193000	Low
ThirdPartyInsuraneAgriculture	4.16E-01	0.4158361000	Low
TractorPolicies	4.18E-01	0.4176891000	Low
PropertyInsurancePolicies	4.71E-01	0.4710528000	Low
TrailerPolicies	5.46E-01	0.5456894000	Low
ThirdPartyInsuranceFirms	6.89E-01	0.6885267000	Low
DeliveryVanPolicies	8.42E-01	0.8415373000	Low
AgriculturalMachinesPolicies	8.55E-01	0.8545754000	Low
LorryPolicies	9.03E-01	0.9025744000	Low
PrivateAccidentPolicies	9.91E-01	0.9911198000	Low

b. Visualization of Results

To effectively communicate the findings, we created line plots using the ggplot and plotly packages. The line plots offered a clear visual representation of the relationship between the different insurance policy types and their corresponding p-values.

The line plots provided a comprehensive overview of the relationships between buying caravan insurance and other insurance policy types. By visually examining the plot, we could easily identify the insurance policies that exhibited strong associations with caravan insurance, as indicated by their low p-values and position in the high significance category. Furthermore, the plot allowed us to observe any potential trends or patterns in the data, enhancing our understanding of the relationship between caravan insurance and other insurance policies.



c. Future Advancements:

To continue advancing our knowledge about the purchase of caravan insurance there are several avenues worth pursuing. One possibility is qualitative research, which would involve conducting interviews or surveys to uncover subjective insights into why people choose to buy this type of policy. This could shed light on important aspects that might be overlooked by quantitative analysis.

Another possibility is longitudinal analysis which helps in examining the changes in insurance coverage patterns over time to identify any evolving trends or shifts in the relationship between caravan insurance and other policy types so the insurance providers can change their strategies based on customer preference.

d. Limitations or Problems Faced:

Chi-squared test reveals associations between variables but does not establish causality. To gain a more profound understanding of the correlations at hand researchers may benefit from pairing the Chi squared test with additional research methods, such as experimental studies or causal modeling. Ultimately discerning meaningful insights requires evaluating all results within the appropriate context of the variables involved in a study.

The size of the dataset used for this analysis may have implications for the statistical power of the Chi-squared test. A larger sample size with better variant inputs can provide more reliable results and detect smaller effect sizes.

Another limitation we faced is that the results should have been interpreted within the context of the variables available. There might be several other insurance policies that might be a factor in customers considering caravan insurance policies.

6. References

- Insurance Company Benchmark (COIL 2000) Data Set, last accessed May 15, 2023, https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29
- A Complete Guide to Random Forest in R, last accessed May 15, 2023, https://www.listendata.com/2014/11/random-forest-with-r.html
- Random Forest Feature Selection, last accessed May 15, 2023, https://www.r-bloggers.com/2021/05/random-forest-feature-selection/
- Feature Selection in R with the Boruta R Package, last accessed May 15, 2023, https://www.datacamp.com/tutorial/feature-selection-R-boruta
- An Introduction to corrplot Package, last accessed May 15, 2023, https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
- Multiple ROC curves in one plot ROCR, last accessed May 15, 2023, https://stackoverflow.com/questions/14085281/multiple-roc-curves-in-one-plot-rocr
- Testing Independence: Chi-Squared vs Fisher's Exact Test, last accessed May 15, 2023, https://www.datascienceblog.net/post/statistical_test/contingency_table_tests/
- Chi-Square Test for Feature Selection – Mathematical Explanation, last accessed May 15, 2023, https://www.geeksforgeeks.org/chi-square-test-for-feature-selection-mathematical-explanation/amp/
- STAT515-005 Spring 2023 Lecture Slides – Week 13, Unsupervised Learning, Unsupervised Learning, Clustering, K-means clustering, Hierarchical Clustering by Dr. Isuru Dassanayake (2023-04-25)