Group ONE – Final Report

Using predictive modelling techniques for product recommendation

Varad Satam
MS Industrial Engineering
Purdue University
West Lafayette, IN
vsatam@purdue.edu

Anas Patankar
MS Industrial Engineering
Purdue University
West Lafayette, IN
patankaa@purdue.edu

Gokulakrishnan Swaminathan
MS Industrial Engineering
Purdue University
West Lafayette, IN
gswamina@purdue.edu

ABSTRACT

The objective of our project is to develop a model which could accurately recommend a product to a customer. Our focus was primarily to correctly recommend a credit card to a potential customer. Over the course of this project, a unique problem of target imbalance was encountered. In dealing with this, it was found that the standard metrics and assumptions about the binary target variable were ineffective. This played a crucial role in driving the approach to model building and evaluation. Consequently, a variety of different predictive modeling algorithms were implemented and tuned to maximize the metrics that support the core business objective of minimizing risk while maximizing profit. This report intends to describe the activities that were taken up by the team to address the stated objectives.

1. INTRODUCTION

Traditionally, communities functioned in small sizes. As a trading merchant, the owner knew customers personally and could recommend products based on history and behavior. By mid 1980s this was the role of an agent in industries. The similarity in this approach of functionality is the personal relationship established between the customer and organization. The personal relationship allowed recommendations of production to customers based on knowledge of past purchases and customer behavior. It also ensured that customers received great customer service thereby allowing businesses to reap the benefit of brand loyalty.

Cut to 2019, products are now recommended by a recommender system based on customer transaction history, demographic and search history. This idea of narrowing down the pool of selection allows more meaningful choices of products. Recommender systems allow businesses to stay relevant to the product choices of the customers thereby improving customer retention. Hence, in this project keeping in mind the same goal, we trained a predictive model which could recommend a product to a customer. On point product recommendations increase the sales of the company. We are building a predictive model which will provide customers with potential products based on customer traits and features. The aim of this project is to develop an accurate model to predict the right products for recommendation. We are implementing different classification models and model ensemble techniques to find the best fit based on different metrics which will be discussed in detail later in the report.

2. BACKGROUND

Santander Bank offers several products to their customers through personalized product recommendations. Under the current system, a small number of Santander's customers receive many recommendations while many others rarely see any resulting in an uneven customer experience. With a more effective recommendation system in place, Santander can better meet the individual needs of all customers and improve customer retention. Hence, are objective will be to correctly recommend products to customers.

3. DATASET DESCRIPTION

3.1 Source

The dataset consists of 1.5 years of customer data of Santander Bank taken from Kaggle. Each row consists of different characteristics of an individual which constitute the independent variables. The details of each column can be viewed from the data description table below.

The original dataset consists of 17 products which could be recommended to a customer. Each product denotes a product which the bank sells. In our projects, we have taken the four most common products which could be recommended to narrow our focus thus increasing the model's precision

Later, we thought of implementing models on a dataset which comprised of only two products which made our project a binary classification problem. We strategically selected the two products which are completely different from each other.

3.2 Data Description

The description of different variables and the two products to recommend are given below:

Column Name	Type	Description
ID	Int	Customer code
Customer Index	Factor	Employee index: A active, B ex employed, F filial, N not employee, P passive
Sex	Factor	Customer's sex
Age	Factor	Age in years
Seniority	Int	Customer seniority (in months)
Indrel	Factor	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
Customer Type	Factor	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
Relation type	Factor	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
Residential Index	Factor	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
Foreigner	Factor	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
Spouse Index	Factor	Spouse index. 1 if the customer is spouse of an employee
Channel	Factor	channel used by the customer to join
Deceased	Factor	Deceased index. N/S
Address type	Factor	Address type. 1, primary address
Province Code	Factor	Province code (customer's address)
Activity index	Factor	Activity index (1, active customer; 0, inactive customer)
Gross Income	Numeric	Gross income of the household
Segmentation	Factor	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated
Label	Factor	0 – Current Account, 1 – Credit Card

Table 1. Data Description

4. EXPLORATORY DATA ANALYSIS

Data visualization revealed important insights. It is important to properly study the consumer based prior to building a recommendation system.

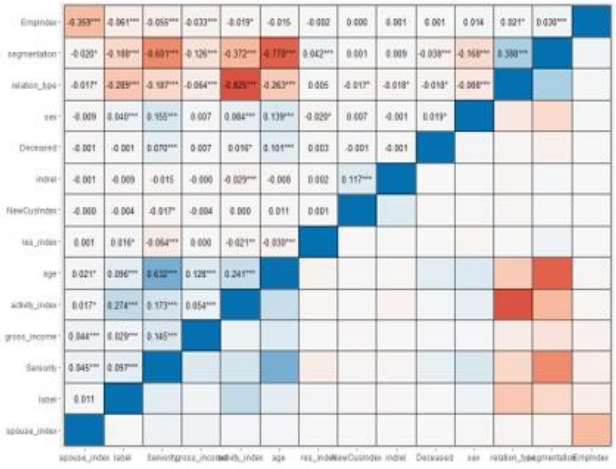


Figure 1: Correlation Plot

The correlation plot shows the relation between variables and their relationship with the response variable. The blue shades are positive correlations and the red shades show negative correlations. From the plot we can see that our response label is negatively correlated to age and type of relation and positively correlated to activity index.

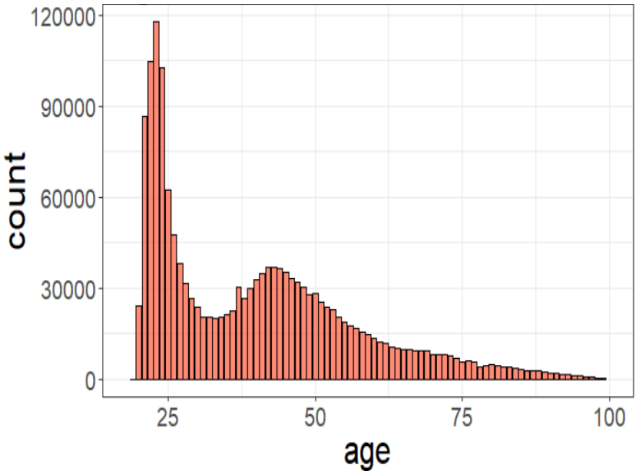


Figure 2: Variation of Age with customer base

It is evident that the most frequently, the customer age is either mid-20s or mid-40s.

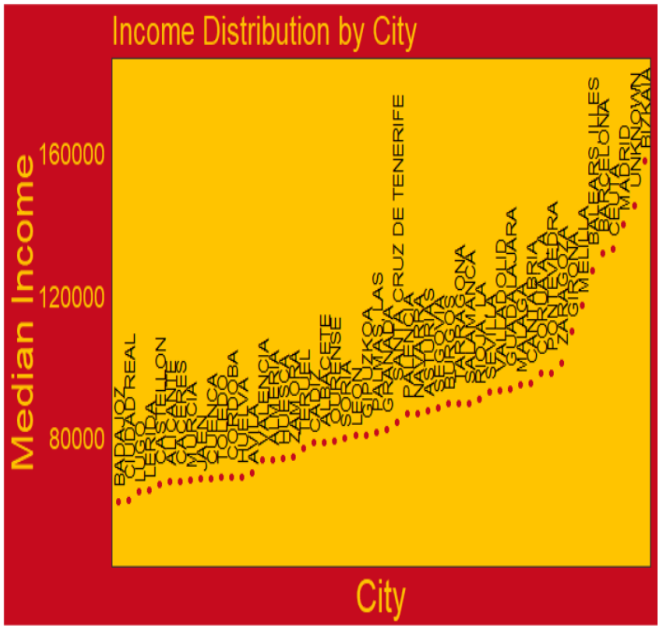


Figure 3: Variation of income based on a region

Income significantly governs the purchasing power of a customer and could explain which banking product a customer will use.

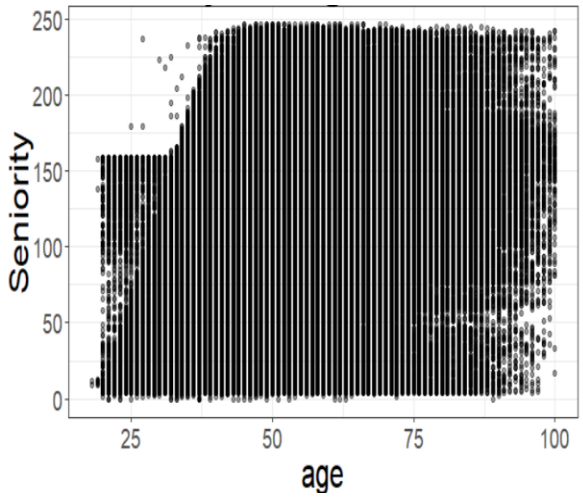


Figure 4: Variation of seniority with age

Seniority is given as the time of an individual being the bank's customer in months. Usually, more senior customers would have a better idea about the service provider. Hence, a part of our hypothesis is that seniority could be a variable of extreme importance. It is low for age 20-35 indicating customers are relatively newer compared to 35-55.

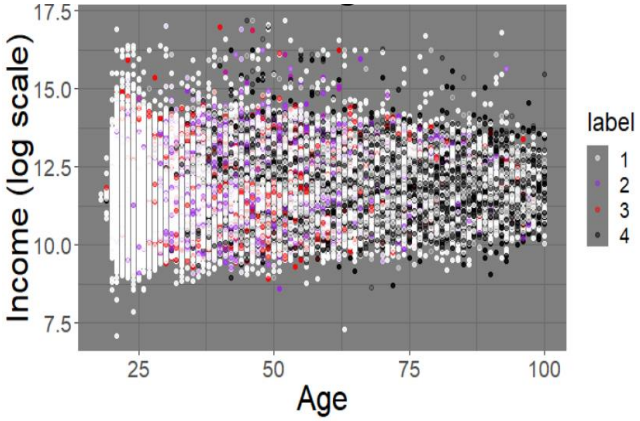


Figure 5: Distribution of products used based on income and age

1 denotes product as current account. 2 denotes "Debit Card". 3 denotes "Payroll Account" and 4 denotes "Particular Accounts". We can observe that the reading of 2 and 3 spike in the age range of 25-55. This is the general working age. Furthermore, the similar locations of points show that 2 and 3 might have similar explanatory features. A customer would only own a debit card if they have enough balance in their accounts to spend. This provides further explanation regarding the overlap of locations for 2 and 3.

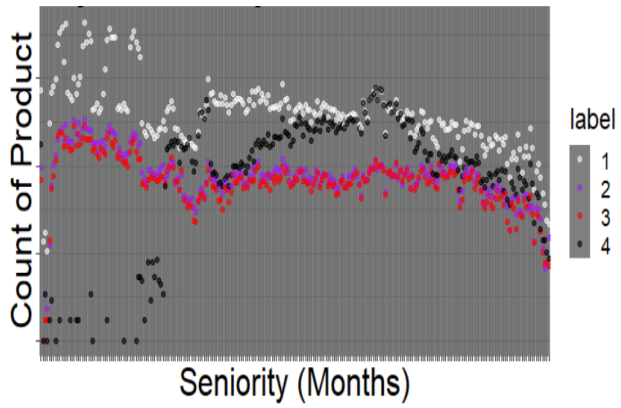


Figure 6: Variation of Count of product with Seniority

The initial hypothesis is further supported by the trend shown in increasing seniority. As the seniority value increases, the areas of count are getting denser. Furthermore, black dots which represent the "Particular Account" products are increasing in count as seniority increases. A

possible explanation for this is as a consumer uses more products he/she develops more trust with the service provider and tends to use different products. Hence, there are very few black dots at lower seniority values. However, the general trend shows that as seniority increases the number of products used decreases. This might be because as seniority and age increases simultaneously. Maybe aged customers are not using the products due to their age.

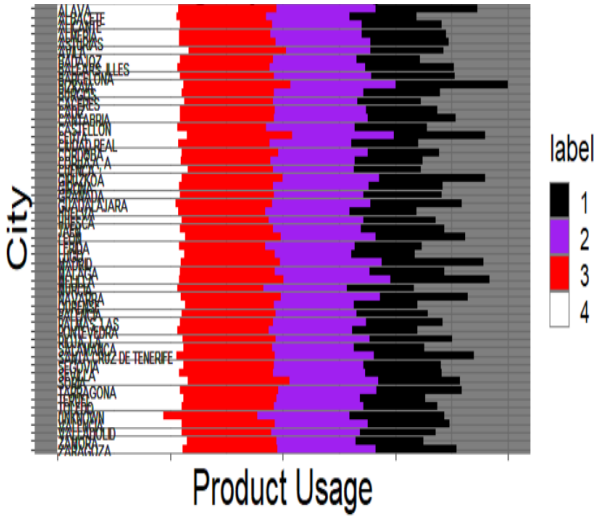


Figure 7

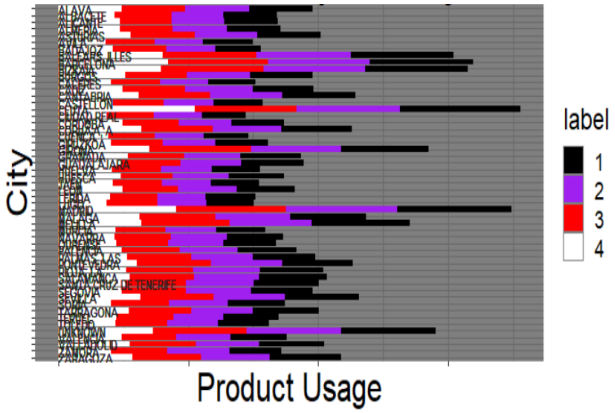


Figure 8

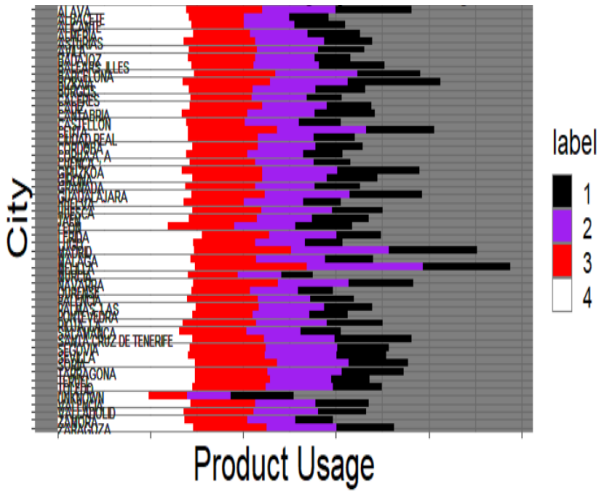


Figure 9

Figure 7 – 9 show the distribution of product count based on age, income and seniority based on provinces in Spain. Age isn't significantly differentiating between locations. However, income plot shows high variation. This can be also backed with the findings of figure 2. Hence, income is significantly explaining the classification.

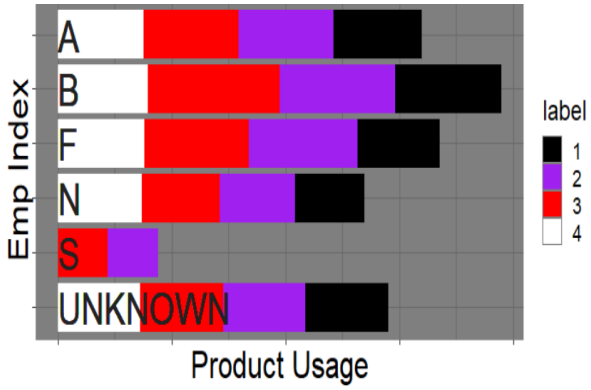


Figure 10

Figure 10 explains the variation of product based on employee index value. These indexes explain whether the customer is an active, inactive, passive foreign, retired or not an employee itself. Higher variation is explained by income and seniority.

5. DATA PREPROCESSING

Data Preprocessing involved cleaning the data set and removal of unnecessary variables. We first ran codes to check for NA values and data types of different variables. Unnecessary characters were removed from the dataset. The NA values were replaced with mean, median and a factor value based on the complexity, importance and data type of the variable. Variables with string data type and lower counts factors were label encoded. However, variables like "name of province" had a larger count. So, these were converted as factors but weren't label encoded. This allowed easier interpretation of graphs during data visualization. Although the entire dataset was processed, due to computational limitations the data used for modelling was limited as a subset of first 3 months. This was backed by data visualization as any significant/alarming trends were not spotted with increasing time. Similarly, we dropped variables like pin code, province code etc as they did not explain anything different from province name.

6. METHODOLOGY

As explained in the description of data, we had data stretched across a period of eighteen months, starting from 2015 and ending in min 2016. This was our initial dataset which we had to work upon to get our final model. Below we'll discuss our plan of action and our approach towards making out training dataset.

6.1 Subset of Data

Our approach was to narrowly study the behavior of people using the products of the bank. The huge chunk of data did not allow us to focus on the behavioral pattern of customers. Hence, we decided to focus on only a subset of months.

We saw that the trend of customers buying a particular product was similar in all the months. We checked that using bar plots for each month to see the variation. The distribution was similar for each month. Hence, we randomly chose three months for our analysis.

We chose three months keeping in mind the computational constrains of our laptop. Initially, or dataset had more than 10 million rows which comprised of data for 18 months. It was impossible to manipulate and analyze this large chunk of data. Thus, for simplicity, we took 3 months as mentioned in the paragraph, reducing the row count to less than a million.

6.2 Type of products selection

As mentioned in the description of the dataset, we had sixteen different types of products offered by the bank. The types of products are described in the data description subpart.

Initially, we analyzed each product for the three months considered just to get an idea about the distribution of each product with respect to the customers of the bank. We saw that there were only four products of significance in the dataset. By significance, what we mean is most of the customers were buying only those four types of products. Since the products of the bank were in the form of different type of accounts, the four important products were namely, current account, debit card, particular account and payroll account.

As our mindset was to focus on the significant products, we took a subset of data which two most important products, current account and credit card. This further reduced the size of the dataset, thus having a positive effect on the computational complexity of the data. In summary, we had two different labels in our final dataset. Our main goal was to predict whether a credit card should be recommended to a person or not.

6.3 Dataset Imbalance

Imbalance data means that the number of data points belonging to a particular class is different. Ideally, the distribution of classes should be equal. This means that in our dataset, each product should have comprised of 50 percent of the values. However, the product current account was taken by almost 90 percent of the people and only 10 percent of the people were recommended credit cards. Hence, the data set that we had was imbalanced.

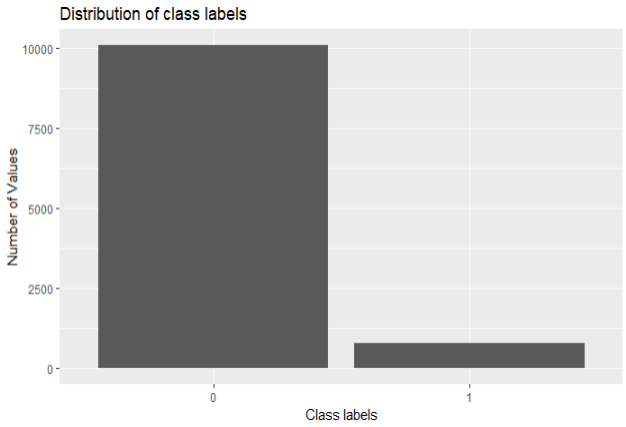


Figure 11: Imbalance in response

6.3.1 Dealing with imbalanced data

The best way for us to solve the data imbalance problem was to oversample the points belonging to the minority class and under-sample the points belonging to the majority class. This aids in creating a balanced dataset. However, in practice, these simple sampling approaches have flaws. Oversampling the minority can lead to model overfitting, since it will introduce duplicate instances, drawing from a pool of instances that is already small. Similarly, under-sampling the majority can end up leaving out important instances that provide important differences between the classes. We addressed the problem using Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE creates new instances of the minority class by forming convex combinations of neighboring instances. As the graphic below shows, it effectively draws lines between minority points in the feature space, and samples along these lines.

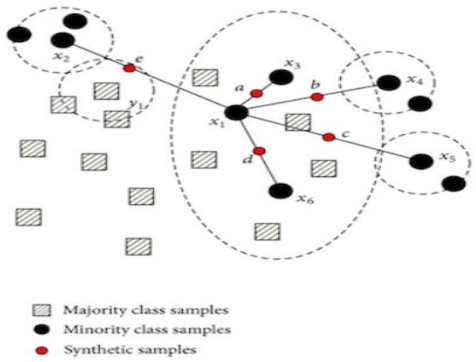


Figure 12: Mechanism of SMOTE

This allows us to balance our data-set without as much overfitting, as we create new synthetic examples rather than using duplicates. This however does not prevent all overfitting, as these are still created from existing data points.

6.3.2 Final Balanced Data

The parameters of SMOTE are the percentage of different labels which are to be taken in the final dataset and the imbalanced data.

Low percentage of the majority samples and high percentage of the minority class were chosen. The actual value of the different percentages was done by trial and error. Multiple iterations were computed based on changes in the percentage of values of different labels until we got our final balanced data set.

6.3.3 Final Data before modeling

As discussed, we had the data for three months as our processed dataset. The original data was arranged monthly. Hence, we randomly sampled the data.

The dataset had about five hundred thousand rows. Due to the computational complexity, keeping in mind the constraints on the computational power of our laptops, we considered only 10 percent of the data for modeling.

6.3.4 Metrics

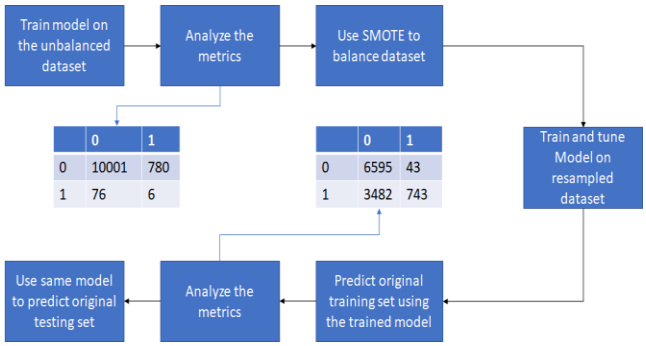
As mentioned throughout the report, we are dealing with an imbalanced class problem related to the financial sector. Our focus is to correctly predict which customer is likely to get a credit card. We are not focused on the number of people who get a current account. Since this is an imbalanced problem, generic modeling metrics such as accuracy cannot be used for model comparison and selection. This is because predicting all customers to get a current account will give us an accuracy of 90 percent which is not our goal. Selection based on accuracy will lead to inaccurate results.

For our model, the metrics used is Recall and Precision. The former is the fraction of positive instances that are predicted correctly, and the latter is the fraction of positive predictions that are correct. Hence our focus, is to increase the recall and precision, thus increasing the F-1 score of the

model which is the harmonic mean of precision and recall.

6.3.5 Modeling Methodology

The figure below shows the pattern we followed for all classification algorithms. We trained the model using the sampled dataset and the trained model was later used to predict on the original dataset. We got convincing results using this strategy. All the models were trained and tuned using Cross – Validation. Initially, we tried using the original dataset for training, but we got a low precision and recall in all the models. Confusion matrix was used for evaluation. The chart below gives you a visual representation of the effect of sampling on the confusion matrix.



We can see the difference in the confusion matrix before and after sampling.

7. Classification Models and Results

We implemented existing predictive modeling algorithms to learn the model. Since the target variable are in the form of labels, we applied classification modeling techniques. This would allow us to evaluate the model performance under different set ups.

From the perspective of learning the model, the goal is to tune the parameters of the model to get high recall and precision to get a high F-1 score.

7.1 Decision Tree

This is our first classification model is used to recommend a product. We used Cross Validation to select the optimal value for cp based on the correct depth. We repeated cross validation thrice for different samples of data to get a model which does not overfit and finally got a model which has no bias due to the sampling procedure used. Also, since we use SMOTE to get our balanced dataset, our model was likely to overfit. Hence, for all models, we have used Cross validation to tune the parameters.

From the picture below, we can see that as complexity parameter increases, cross-validated accuracy decreases due to overfitting. Hence cp with the highest accuracy for a depth is selected for modeling.

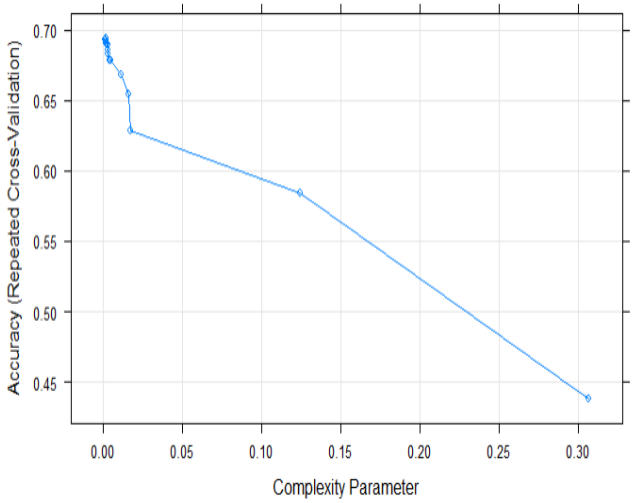


Figure 13: Distribution of cp parameter

7.1.1 Decision Tree Results

As mentioned earlier, we initially used the original dataset to train the model. The F-1 score, which considers both precision and recall was 0. After training using the resamples data, the training and testing F-1 score came out to be 0.3069 and 0.2895 respectively. Hence, we see an improvement in the model when resampling is done. The detailed values of precision and recall can be viewed from the table which is displayed after discussing all the models. The advantage of model is that it has good accuracy and interpretability power.

7.2 Random Forest

Two main parameters which were modified for the model were the number of variables chosen for each tree in the random forest and the total number of trees. Used Cross Validation to select the ideal number of variables for each tree. From the plot below, we can see that at n = 8, the accuracy value is the highest. The optimal number of variables selected, minimizes the correlation of trees, thus increasing the accuracy.

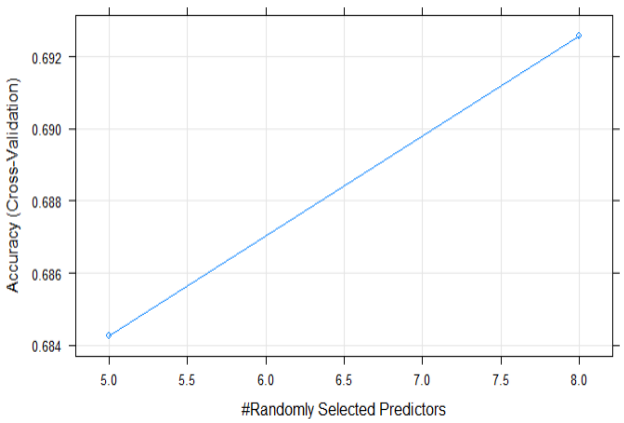


Figure 14

Also, after number of trees = 500, the error rate becomes constant. Hence, in our final model, we took the number of trees as 500 as adding more trees does not increase the accuracy.

7.2.1 Random Forest Result

As mentioned earlier, we used cross validation of the training data to tune the hyperparameters. Based on this, the F-1 score when the model was trained using the entire dataset was 0.012. After training using the resampled data, the training and testing F-1 score came out to be 0.3071 and 0.2904 respectively. The detailed values of precision and recall can be viewed from the table which is displayed after discussing all the models.

7.3 Logistic Regression

The caret function was used to train the logistic regression model. Like the previous models, it was first trained on the unsampled dataset, then on the resampled balanced dataset. The regularization parameter L2 norm was tuned using cross validation to get the best F-1 score.

7.3.1 Logistic Regression Accuracy

On running the model, the training accuracy came out to be 65.21 percent and the testing accuracy is equal to 64.3 percent. The F1 scores for training and testing are 0.2822 and 0.2688 respectively. The models are tuned using caret package along with the inbuilt cross validation function.

7.4 Gradient Boosting

The factors considered for making the model were the number of trees, maximum depth of each tree, minimum number of observations in the terminal node of tree and the shrinkage or the step size. The optimal values of the factors were calculated using cross-validation.

7.4.1 Gradient Boosting Accuracy

The F-1 score when the model was trained using the entire dataset was 0.2260. After training using the resampled data, the training and testing F-1 score came out to be 0.3754 and 0.2792 respectively. The detailed values of precision and recall can be viewed from the table which is displayed after discussing all the models.

7.5 Support Vector Machines

The model was initially built on the training set. Once, the training set was resampled, the model was fitted using the resampled dataset. The hyper parameters that were tuned for SVM were the kernel type and C value using 10-fold cross validation. Kernel type governs the shape of the decision boundary. C values decides how much the model should avoid misclassification. It controls the tradeoff between achieving low error on training data and minimizing the norm of the weights. Once the model was tuned with the best parameter values, it was used to calculate the metric scores of training and testing data.

7.5.1 SVM Accuracy

The F1 scores for training and testing are 0.2670 and 0.2549 respectively. The models are tuned using caret package along with the inbuilt cross validation function. The detailed comparison of different models can be seen in the model comparison section below

7.6 BART

Out of all models, BART gave the best F1 score, precision and recall values. BART is a Bayesian approach to nonparametric function estimation using regression trees. Regression trees rely on recursive binary partitioning of predictor space into a set of hyper-rectangles. The prior for the BART model has three components: (1) the tree structure itself, (2) the leaf parameters given the tree structure, and (3) the error variance σ^2 which is independent of the tree structure and leaf parameters.

There are 6 hyper parameters which can be tuned while building a BART model. They are α, β, k, q, v and m. The hyper- parameters α and β govern the

tree prior probabilities, k and q control the model regularization i.e. σ^2, v is the associated with leaf priors and m governs the number of trees. The whole model is based on MCMC samplings. The importance of every parameter is explained in detail in the appendix. This component of the tree structure prior can enforce shallow tree structures, thereby limiting complexity of any single tree and resulting in more model regularization.

7.6.1 BART Accuracy

Post tuning, the final model of BART gave a F1 score of 0.3049 while training and 0.2926 with the testing dataset. The model also gave strong recall values of 0.9465 and 0.9506 for training and testing sets respectively.

8. MODEL SELECTION

On comparing different models, we get:

Model Name	Before Sampling (CV train)				After sampling(train)				After sampling (test)			
	Acc	Recall	Prec	F1	Acc	Recall	Prec	F1	Acc	Recall	Prec	F1
Decision Tree	0.9276	0.0000	0.0000	0.0000	0.707	0.8969	0.1859	0.3069	0.7203	0.8734	0.1735	0.2895
Random Forest	0.6755	0.006	1.0000	0.012	0.6918	0.9440	0.1811	0.3071	0.6867	0.9228	0.1722	0.2904
Logistic Regression	0.9279	0.003	1.0000	0.007	0.6441	0.9656	0.1650	0.281	0.632	0.9722	0.1560	0.2688
Gradient Boosting	0.9341	0.1335	0.7500	0.226	0.763	0.9847	0.2319	0.3754	0.7282	0.7561	0.1712	0.2792
BART	0.9277	0.0025	1.0000	0.0050	0.6977	0.9363	0.1853	0.3095	**0.6887**	**0.9321**	**0.1746**	**0.2942**
Support Vector Machine	0.9277	0.0001	1.0000	0.0002	0.6109	0.9796	0.1545	0.2670	0.6006	0.9814	0.1464	0.2549

Table 2: Metric Comparison of different Models

The table above depicts the precision, recall and F-1 scores of different models before sampling, after sampling training and testing sets. It is important for our model to get a high precision and recall score to recommend the correct product to a customer. As mentioned earlier in the report, we have trained the model to get a confusion matrix with low False Positives and False Negatives to get a high recall and precision. Thus, keeping these two metrics in mind, we have the third metric F-1 score which is the harmonic mean of Precision and Recall. Hence, for selecting the best model, we strictly looked at the model with the highest testing set F-1 score.

From the table we can clearly see that BART gives us the best F-1 score. It even has good interpretability power in terms of variable importance and dependence plots. Hence, we select BART as the best model. A bigger table is present in the Appendix of the report.

9. INSIGHTS

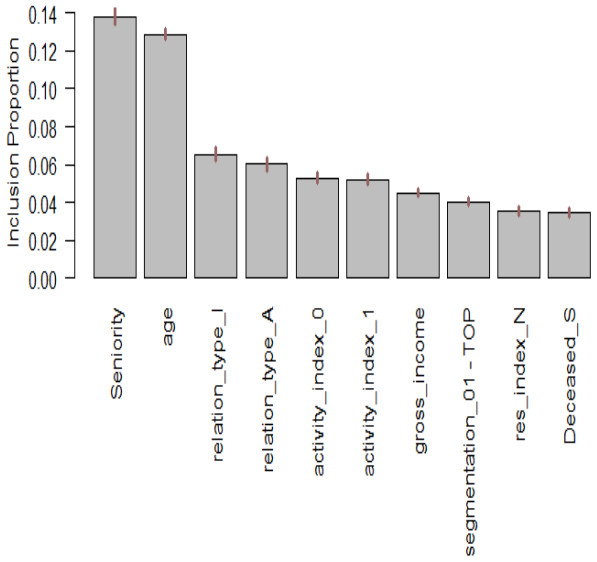


Figure 15: BART Variable Importance Plot

Based on the variable importance plot obtained from the BART model, we can see that Seniority and age of customer are the most significant factors. This makes sense as banks would want to recommend a product like credit card to customers who can make the payments on time. Seniority and Age depict the capacity and capability of an individual to make timely payments. Seniority shows since how long the applicant has been a customer of the bank. The next variables in line are 'relation_type' which explain whether the customers have any personal relationship with an employee of the bank. 'Activity_index' depicts how frequently the customer used the current banking products of the bank. 'Income' also explains the individual's capacity to pay. Based on the variable importance plot, we can infer that the significant features selected by the model strongly explain capacity of an individual to make timely payments which is of utmost importance in products like credit cards.

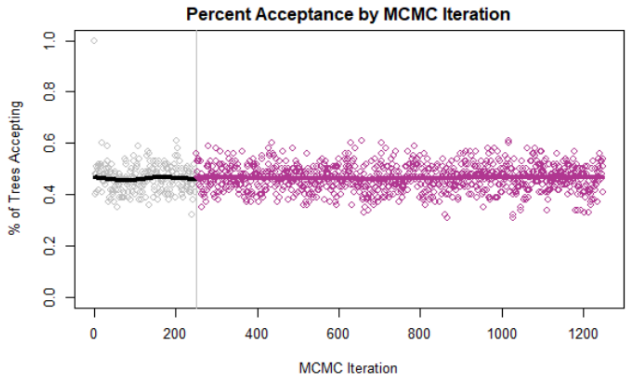


Figure 16: Percent acceptance of Sampling

The above figure shows the percent acceptance of the samplings by the trees. Every point depicts a single iteration. It denotes the number of Metropolis-Hastings proposals accepted across every tree. The values to the left of the vertical line in the gray area denote the burn-in sample. The acceptance is between 30%-60% and is mostly concentrated at 50%.

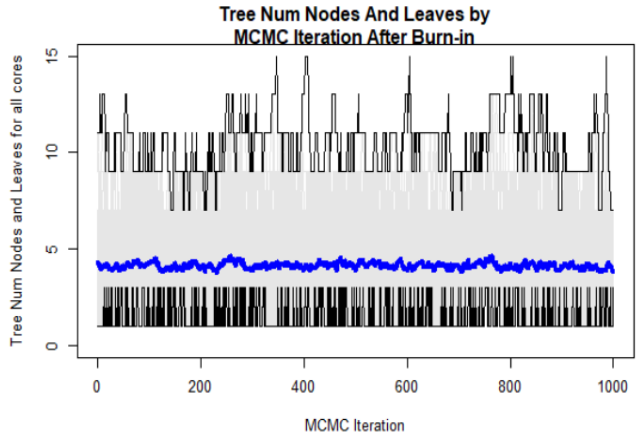


Figure 17: Diagnostic plot for BART leaf

The plot above shows the average number of leaves across the m trees by iteration. Similarly, figure 18 explains the varying depth of the trees in the model. These diagnostic plots only consider the post burn-in samples. The number of leaves and depth explain the complexity of every tree out of the m trees used. The number of leaf nodes vary from 8 to 15 and the depth varies from 3 to 5 based on the random samplings generated by the MCMC iterations.

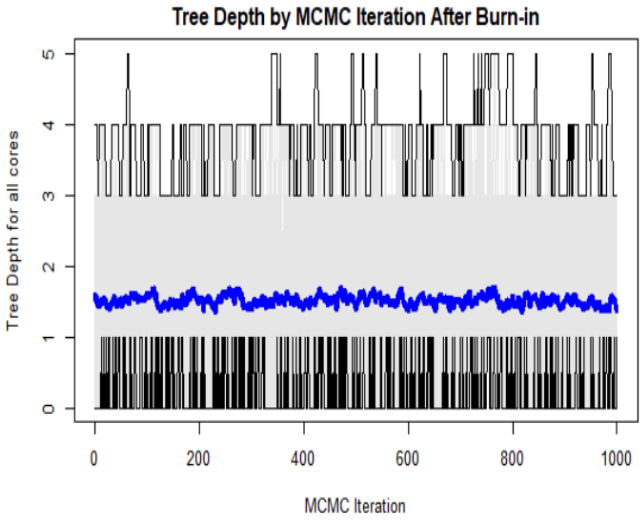


Figure 18: Diagnostic plot for depth of tree

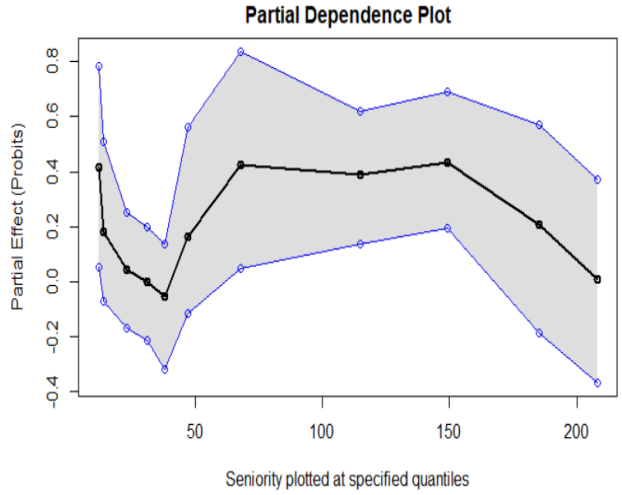


Figure 19: Partial Dependence plot (Seniority)

Figure 19 explain the probability values of the response label considering only changes in Seniority. The values decrease in the start, however, once seniority increases beyond 50, the values shoot up. The function maintains that that value till seniority reached 150 and declines beyond that. Seniority is the time in months for which the individual has been a customer of the bank. Financial organizations study the behavior and backgrounds of customers. As seniority increases, banks study the financial condition and capacity of an individual. Thus, seniority could explain why a particular product would be recommended to a particular customer.

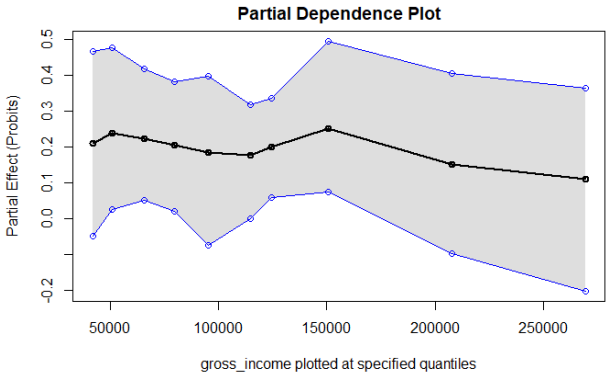


Figure 20: Partial Dependence Plot (Gross income)

Similarly, the pd plot of income also explains how probit values change marginally with an increase or decrease in income. The value increases slightly post the 100000. Income also explains the capacity of an individual to pay and hence, has a strong relationship with the response. Refer the appendix for more partial dependence plots.

10. PROJECT OUTCOME EVALUATION

The project gave the team valuable experience in terms of working with a real-world dataset which involved significant data cleaning, processing, exploration and implementation.

The major issue was dealing with the imbalance of the dataset. Working with this dataset didn't only introduce us to solving imbalance; an omnipresent issue in business datasets, but also towards selecting which solution out of the possible ones is the more appropriate solution.

The project involved training, tuning, testing and interpreting various models using powerful packages in R.

11. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Prof. Roshanak Nateghi and Min Soo Choi for their invaluable guidance and insights. We would also like to thank our classmates for their constructive criticism and suggestions.

12. REFERENCES

[1] Galit Shmueli (2010), To Explain or to Predict? Statistical Science (2010), Vol. 25, No. 3 289-310. DOI= 10.1214/10-STS330.

[2] Adam Kapelner and Justin Bleich, (2013), Machine Learning with Bayesian Additive Regression Trees, Journal of Statistical Software (2016), Vol. 70, No. 1, arXiv: 1312.2171

[3] Andrea Dal Pozzolo, Oliver Caelen and Gianluca Bontempi, (2015), When is undersampling effective in unbalanced classification tasks? Machine Learning and Knowledge Discovery in Database (2015), Springer, 200-215

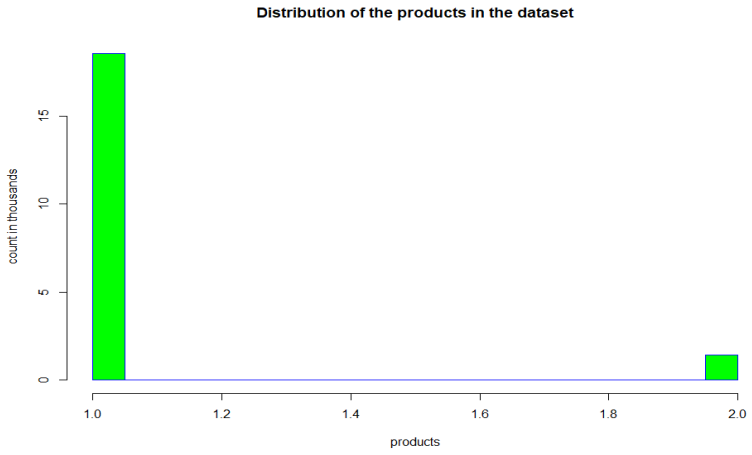
[4] Paula Branco, Luis Torgo and Rita P. Ribeiro (2016), A Survey of Predictive Modeling on Imbalanced Domains, ACM Computing Surveys (CSUR) (Nov 2016), Vol. 49 Issue 2, No. 31, DOI: 10.1145/2907070

Appendix

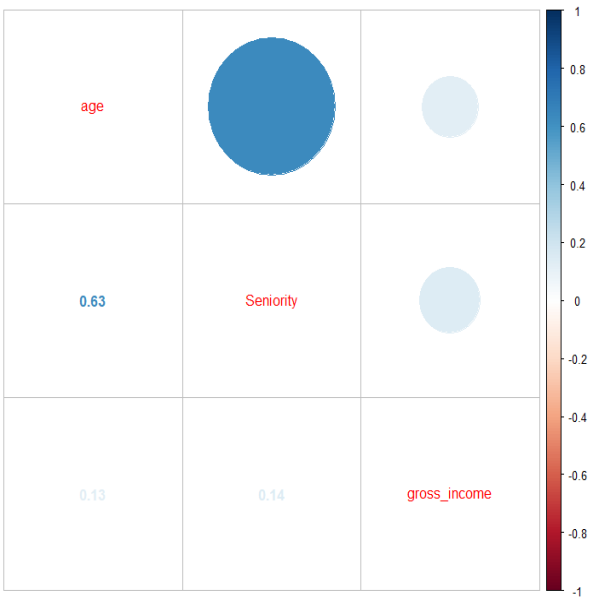
METRIC COMPARISION OF MODELS

Model Name	Before Sampling (CV train)				After sampling(train)				After sampling (test)			
	Acc	Recall	Prec	F1	Acc	Recall	Prec	F1	Acc	Recall	Prec	F1
Decision Tree	0.9276	0.0000	0.0000	0.0000	0.707	0.8969	0.1859	0.3069	0.7203	0.8734	0.1735	0.2895
Random Forest	0.6755	0.006	1.0000	0.012	0.6918	0.9440	0.1811	0.3071	0.6867	0.9228	0.1722	0.2904
Logistic Regression	0.9279	0.003	1.0000	0.007	0.6441	0.9656	0.1650	0.281	0.632	0.9722	0.1560	0.2688
Gradient Boosting	0.9341	0.1335	0.7500	0.226	0.763	0.9847	0.2319	0.3754	0.7282	0.7561	0.1712	0.2792
BART	0.9277	0.0025	1.0000	0.0050	0.6977	0.9363	0.1853	0.3095	**0.6887**	**0.9321**	**0.1746**	**0.2942**
Support Vector Machine	0.9277	0.0001	1.0000	0.0002	0.6109	0.9796	0.1545	0.2670	0.6006	0.9814	0.1464	0.2549

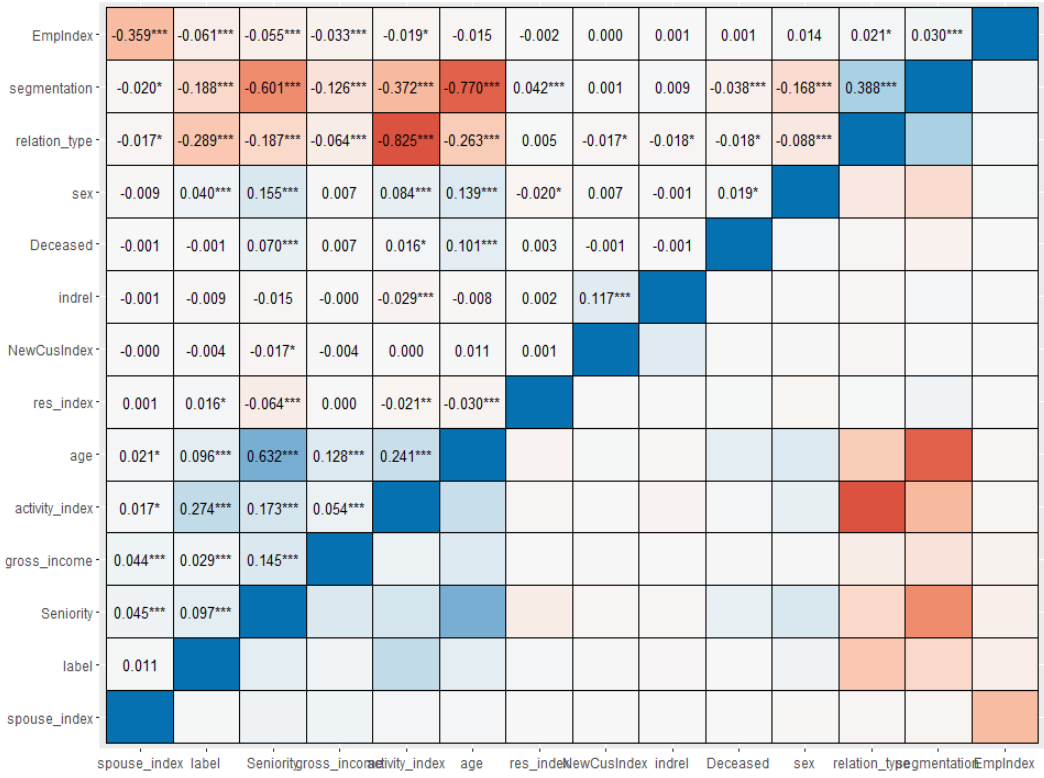
EDA



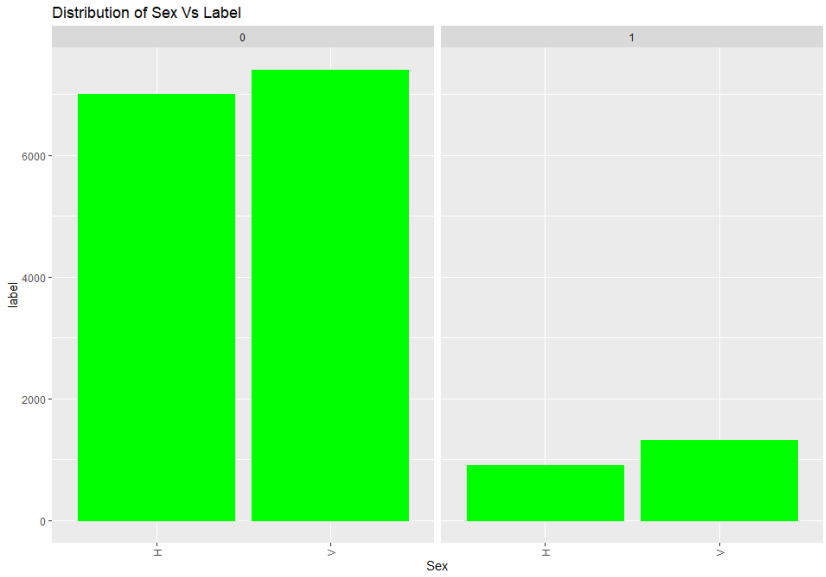
From the above graph we could see that labels present in the dataset are distributed unevenly. This depicts the imbalance of classes present in the dataset.



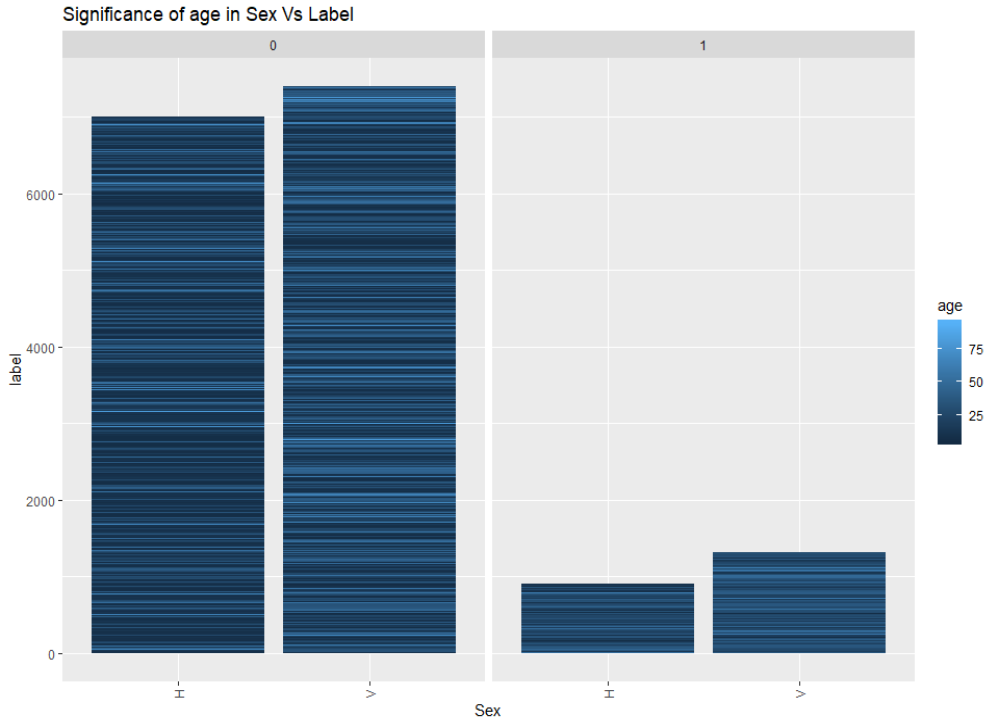
From the correlation plot we could see that there is high positive correlation between age and seniority.



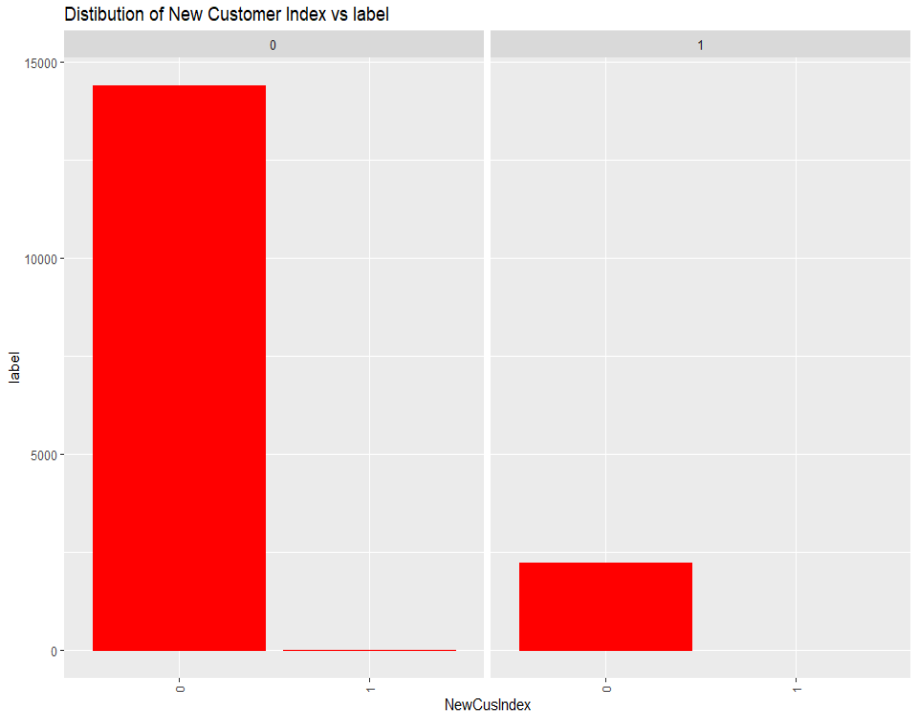
From the plot above, we are able to observe the relation between variables present in the dataset. We could observe there is a high negative correlation between the activity index and the relation type; age and segmentation; seniority vs segmentation.



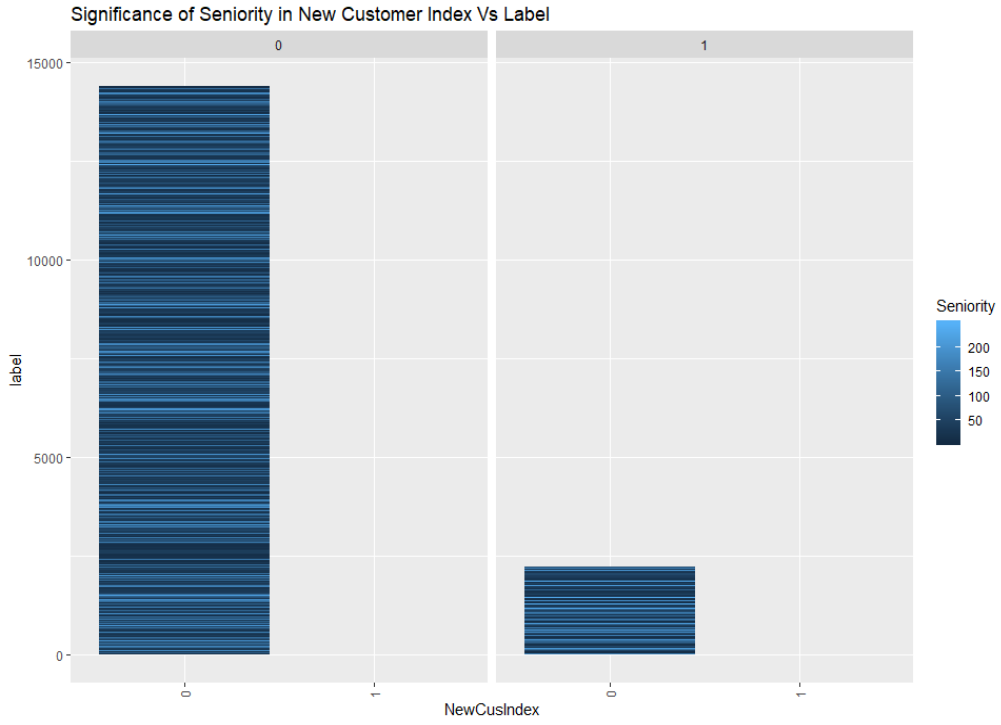
The plot shows the number of “H” and “V” factors present in the label classes in the dataset. It could be seen that there more number “v” factor in both classes.



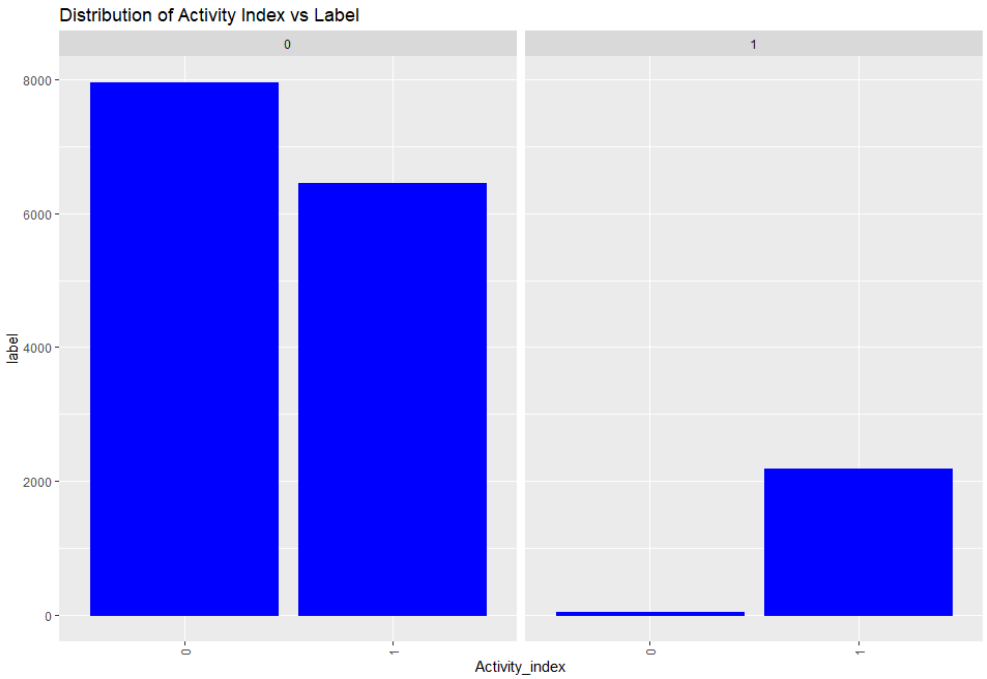
The plot constructed shows the significance of age in sex vs label. We could see that age greater than 75 are vaguely present in the dataset.



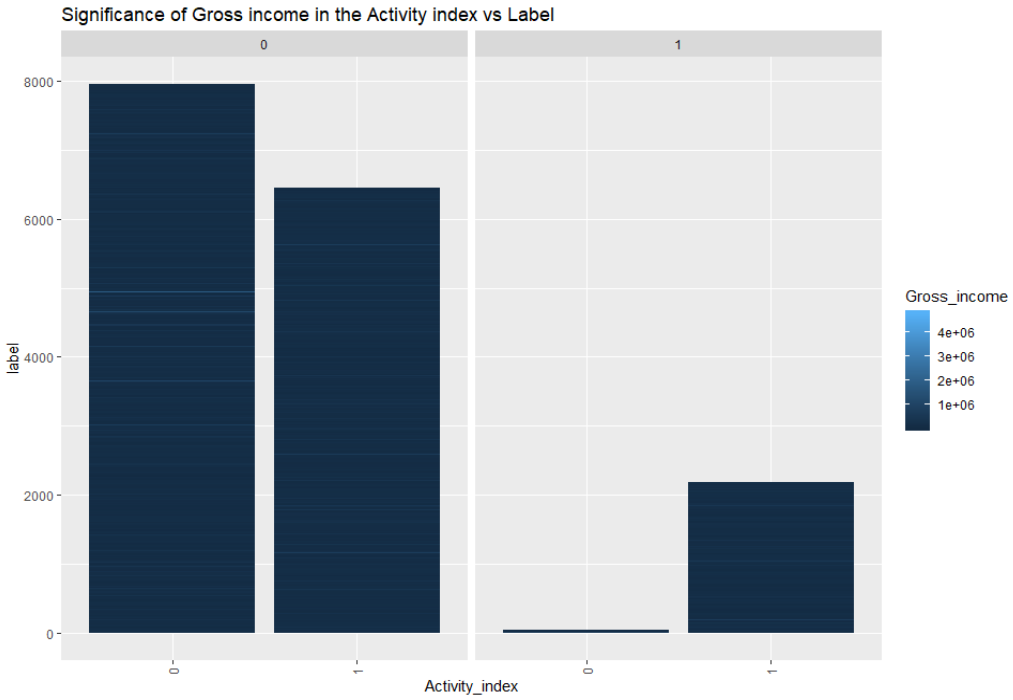
The plot shows distribution of factors of new customer index “0” and “1” in the dataset. The majority of the new customer belongs to the “0” class.



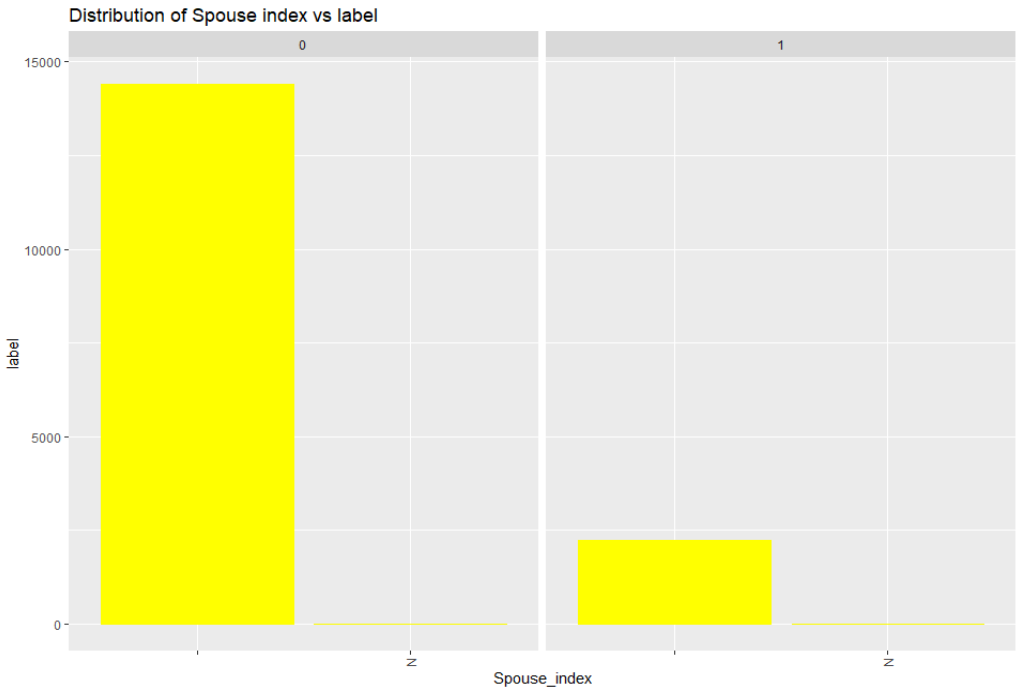
This plot is generated to understand the effect of seniority in new customer index vs label. Seniority greater than 200 months is scarce in the dataset.



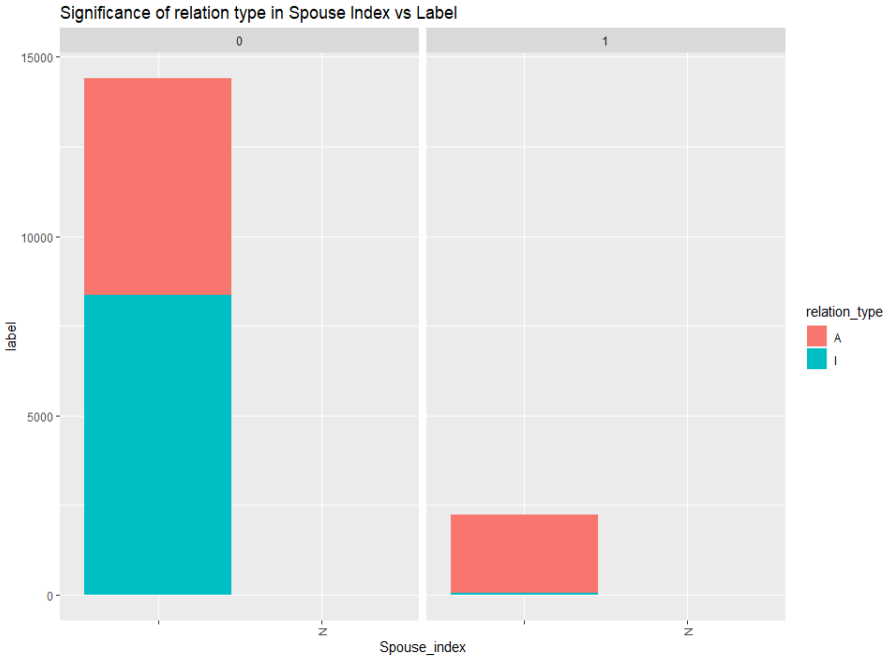
The plot shows the activity index distribution with respect to the labels in the dataset. We could see that in class “0” the factors of activity index are high compared to the class “1”.



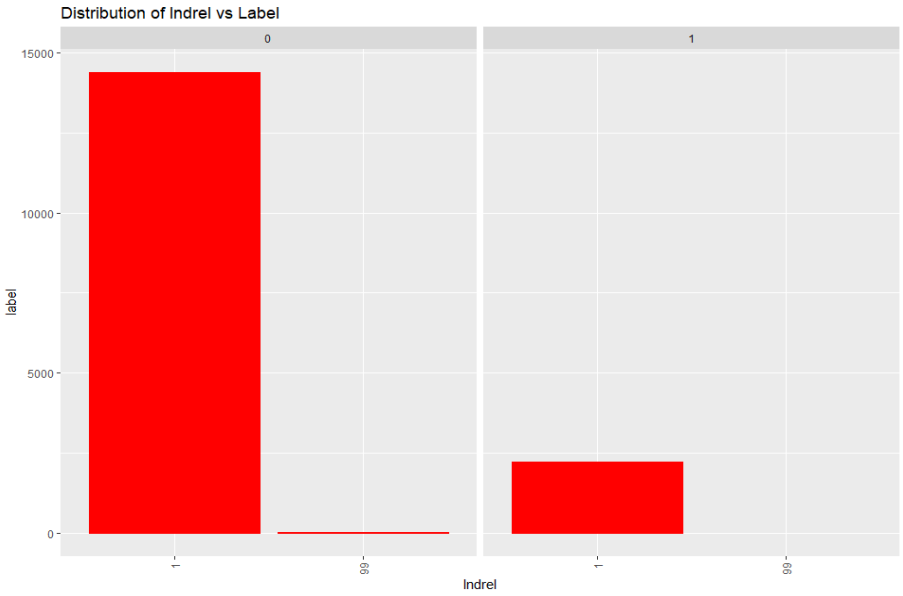
The plot is generated to understand the effect of gross income in the activity index vs label. This is done because from the previous plot we were able to understand that the activity index factor distribution was high in class “0” of the label.



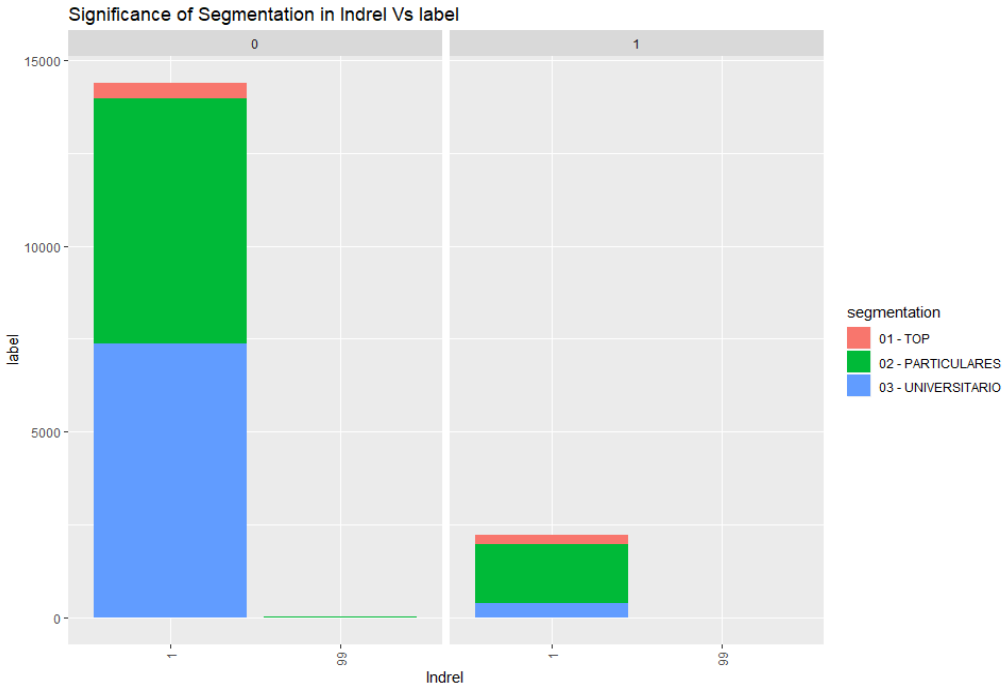
This plot is generated to understand the effect of spouse index with respect to the labels in the dataset.



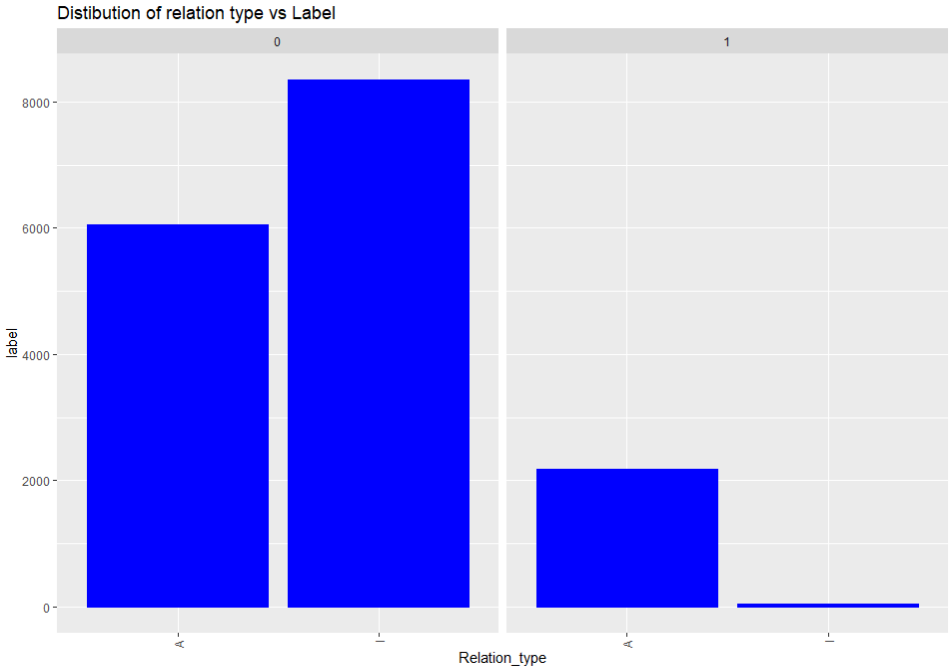
From this plot we are able to observe that in class “0” the relation type is distributed evenly in the spouse index and in class “1” we are able to observe that facotr “i” in relation type is very low in the spouse index.



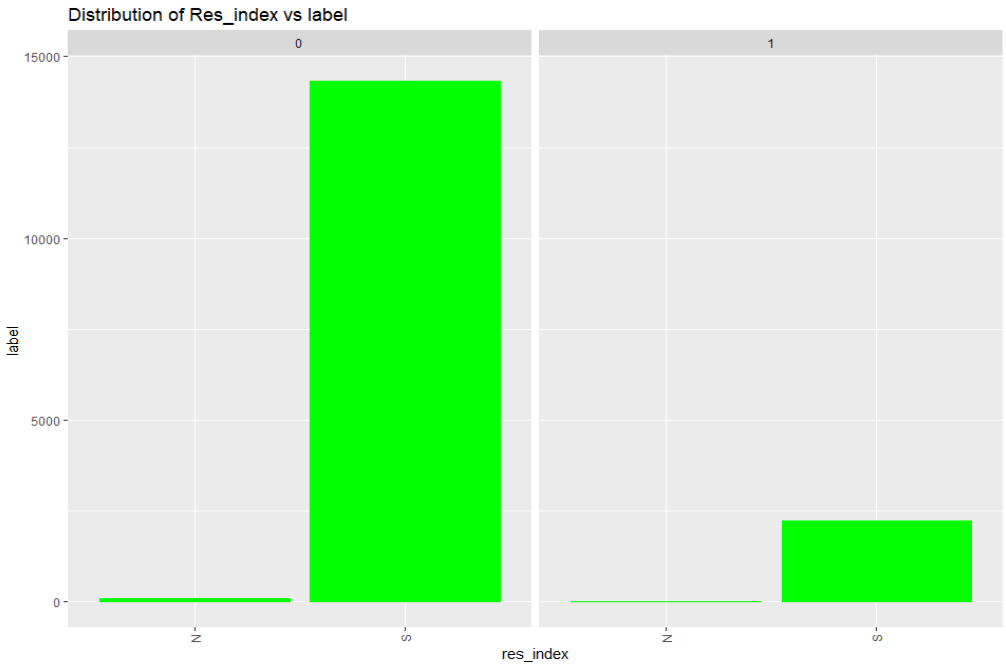
This plot is generated to understand the effect of Indrel with respect to the labels in the dataset.



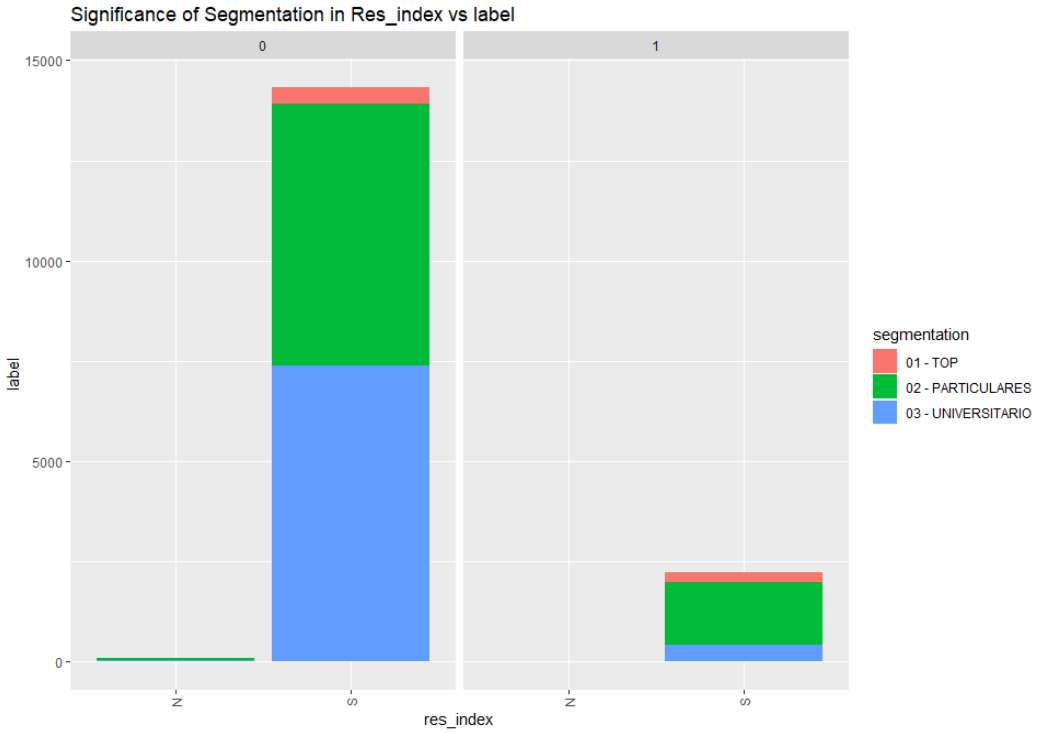
This plot is generated to understand the effect of segmentation in Indrel. We could see that in Class “0” of the label, the factor “1” of indrel the segmentation “02-Particulares” and “03-Universitario” are distributed almost evenly. But in Class “1”, majority is in “02- Particulares”



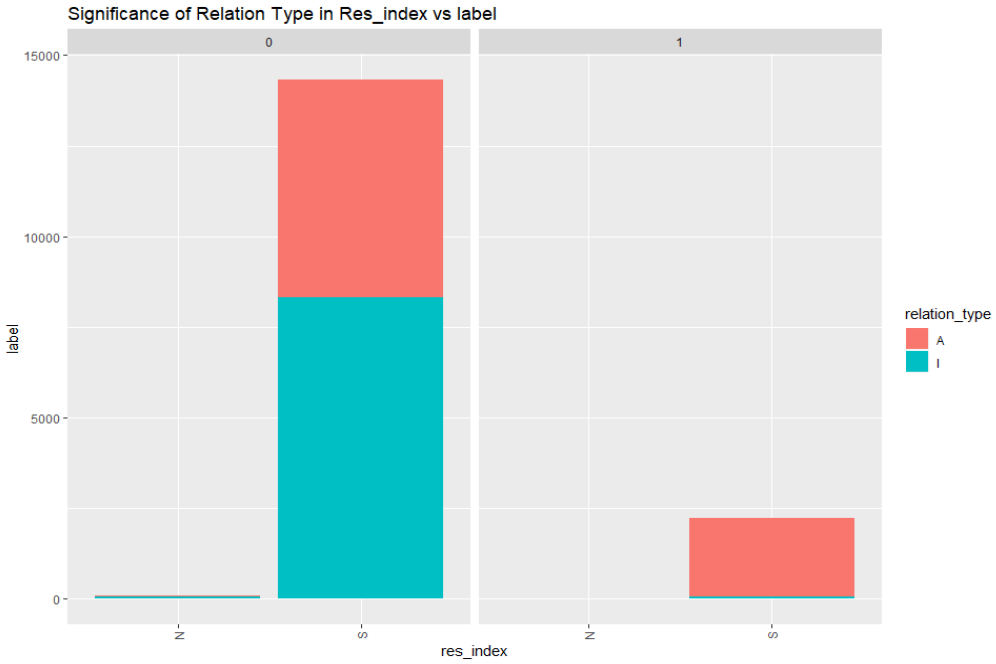
This plot is generated to understand the effect of Relation type with respect to the labels in the dataset.



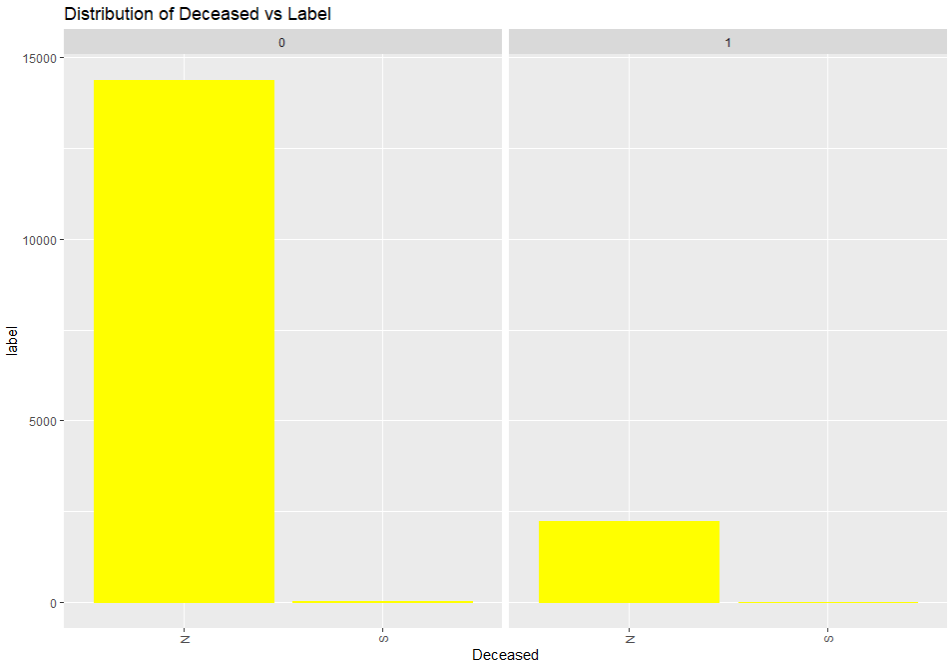
This plot is generated to understand the effect of Res_index with respect to the labels in the dataset. From the plot most of the values of the res_index belong to the factor“s”



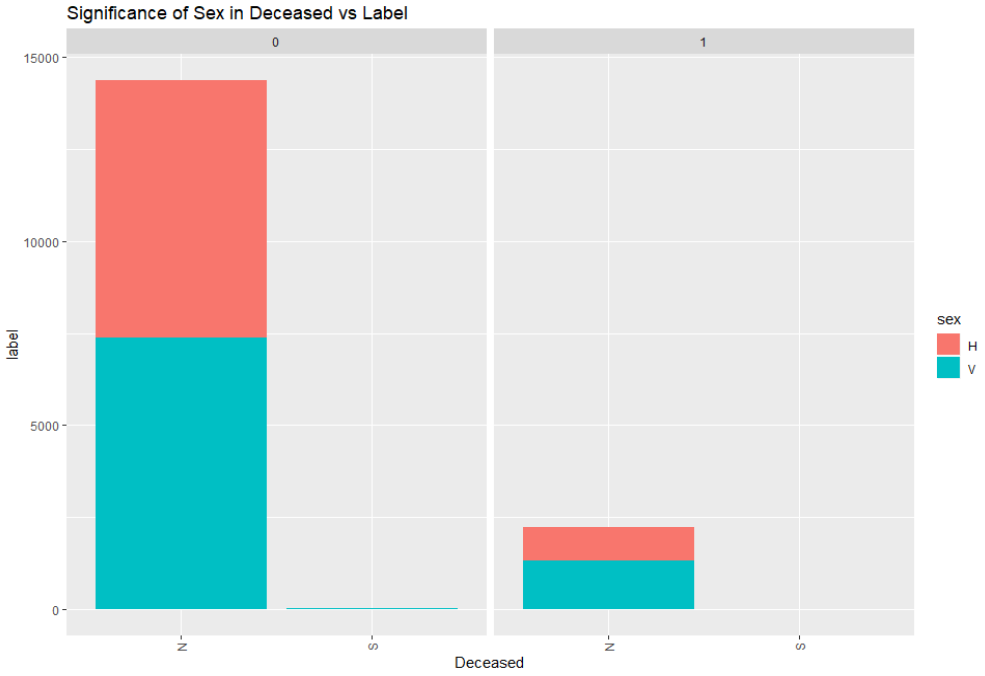
This plot is generated to understand the effect of segmentation in the res_index. From the observations of the previous plot it is seen that “02-Particulares” and “03-Universitario” are the distributed widely.



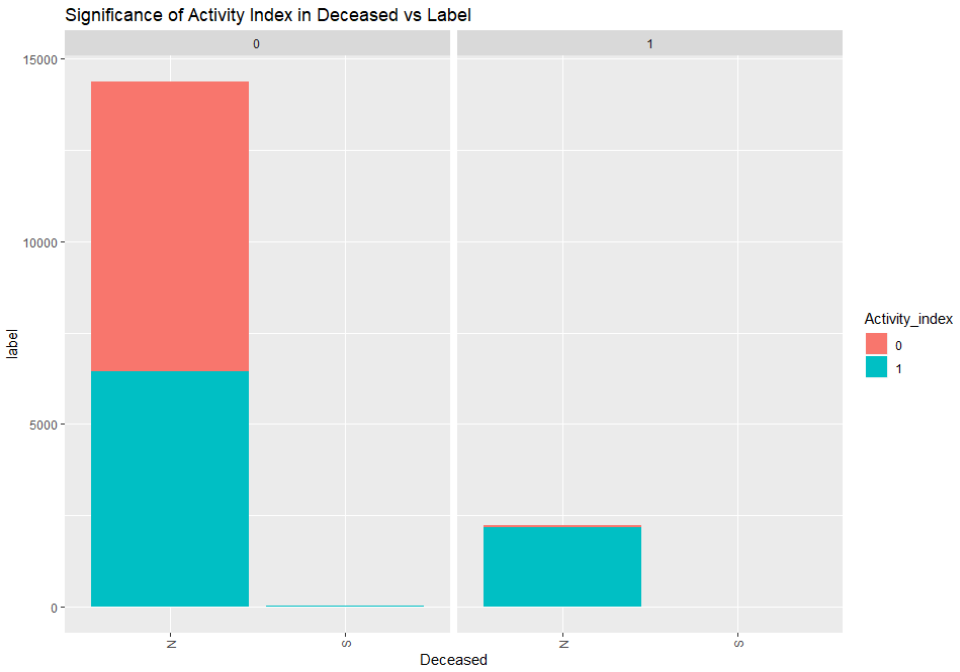
This plot is generated to understand the effect of Relation_type in the res_index. We can observe that relation types “A” and “I” are distributed evenly in the dataset. Whereas in class “1” the predominant relation type is “A”.



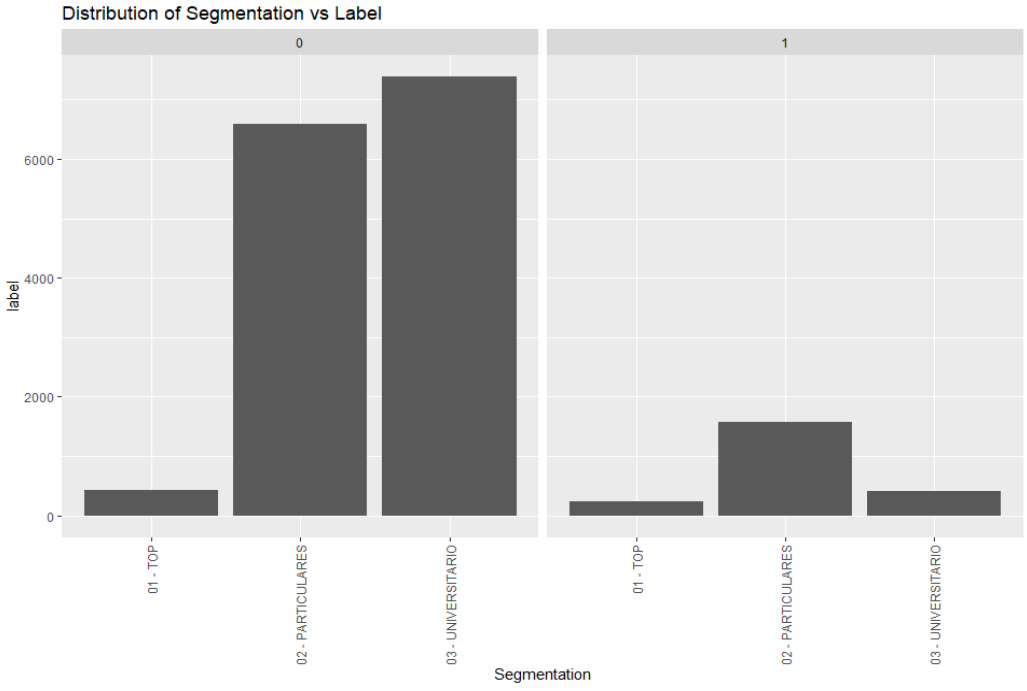
This plot is generated to understand the effect of Deceased with respect to the labels in the dataset.



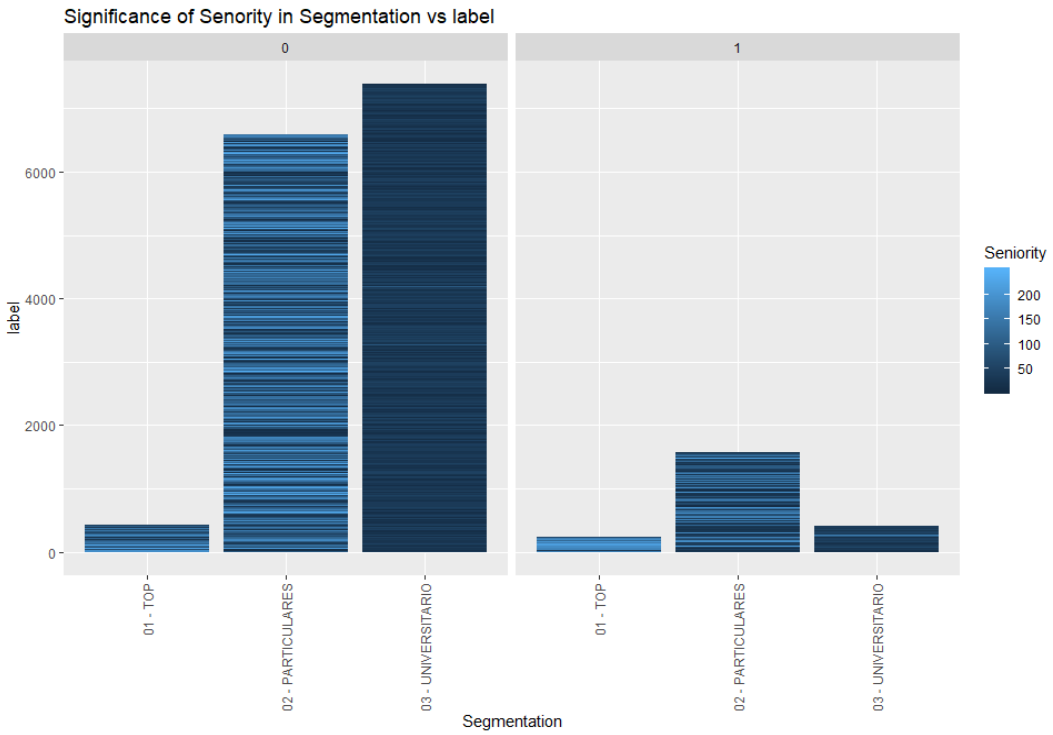
This Plot is generated to understand the influence of Sex in the Deceased vs label. We could observe in an even distribtuion in sex for factor “N” in both classes.



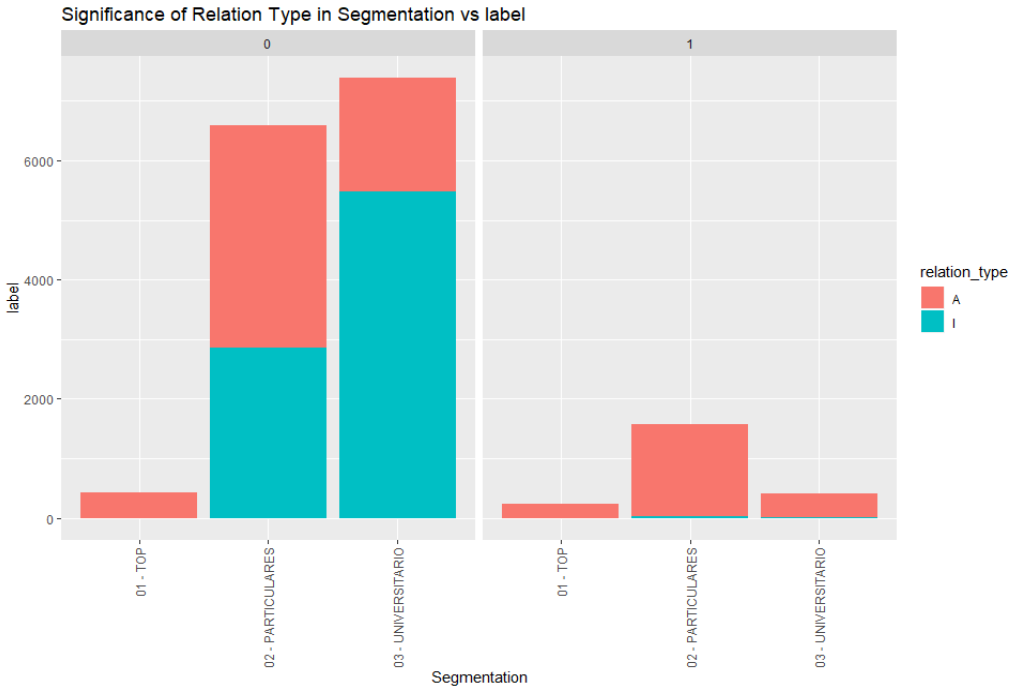
From the plot generated we could observe the factor “N” for class“0” and Class “1”. it is eveny distributed in the former and “1” is predominant in the latter.



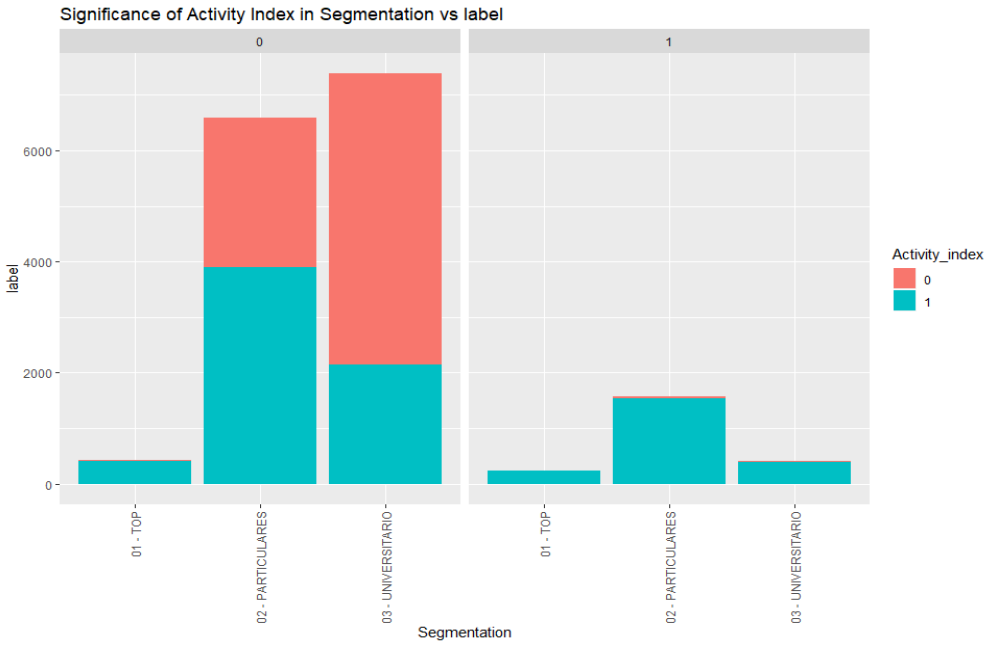
This plot is generated to understand the effect of Segmentation with respect to the labels in the dataset. “02-Particulares” is found to be majorrly distributed for both classes.



This Plot is generated to understand the influence of Senority in the Segmentation vs label. We could observe the majority of values less than 50 in the “03-Universitario” and 50 to 200 in “02-Particulares” and above 200 in the “01-Top”



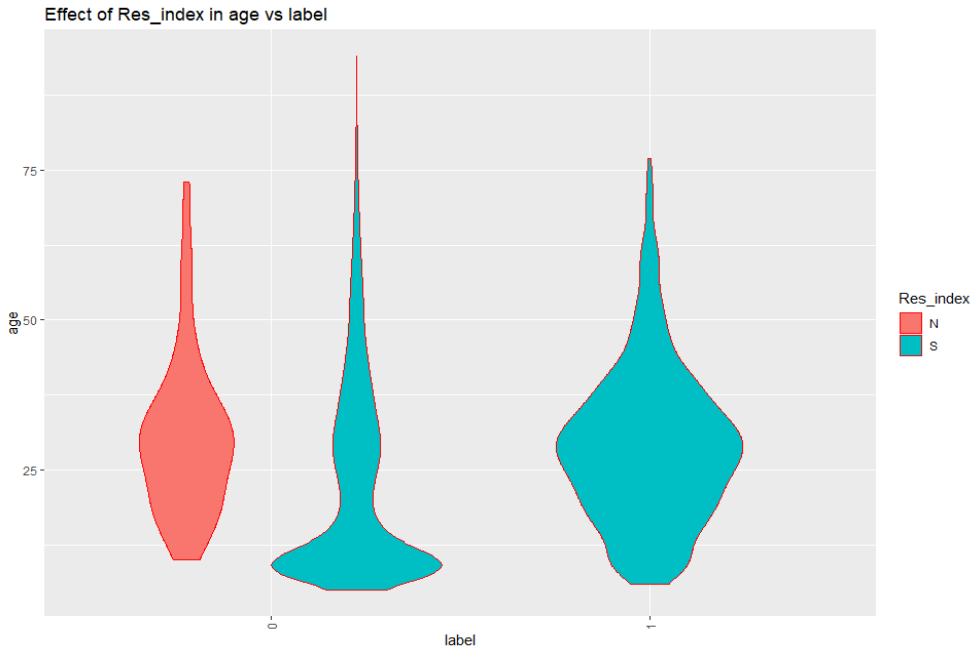
From this plot we can observe that in Class “1” we could see that relation type “I” is not present. In Class“0” the relation type“I” is present evently in “02-Particulares” and majorly present in “03-Universitario”

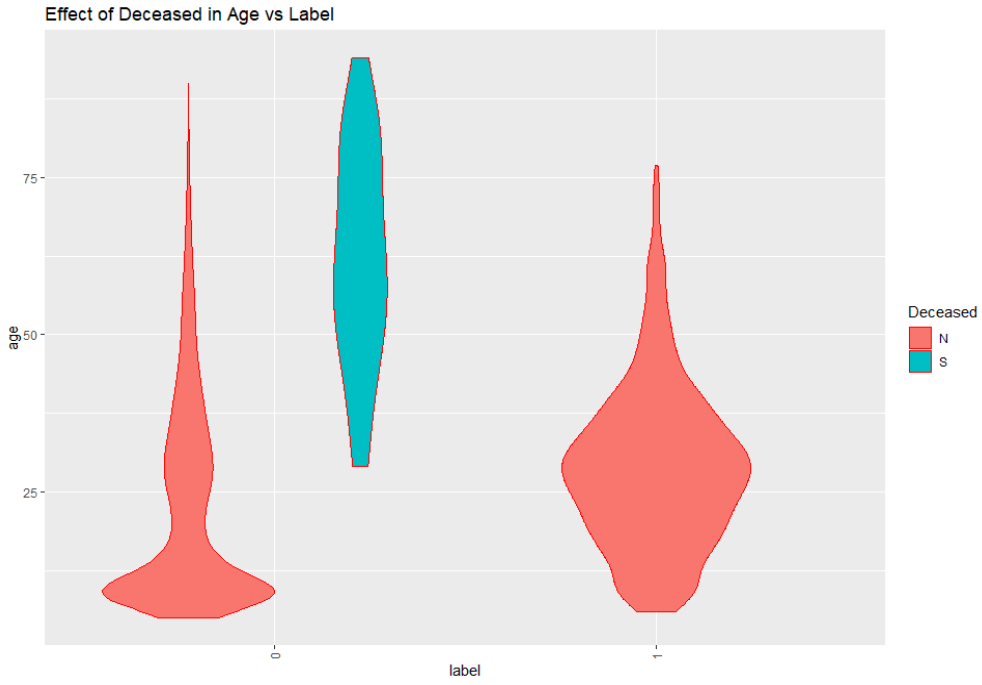


From this plot we can observe that in Class “1” we could see that Activity Index “0” is not present. In Class“0” the Activity Index “0” is present evently in “02-Particulares” and majorly present in “03-Universitario”

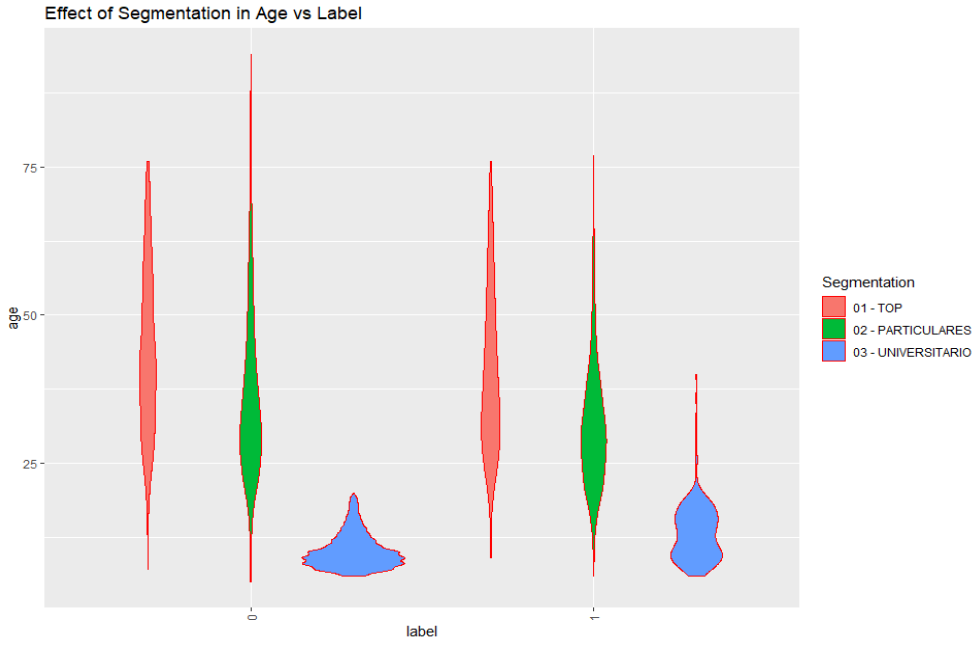
The violin plot sugests that the in class“0” the large number of values for the age are found below 25 and in class“1” the age is nornally distributed with the peak at 30.


Effect of Spouse index in age vs label
75
50
25
age
0
1
label
Spouse_index
N

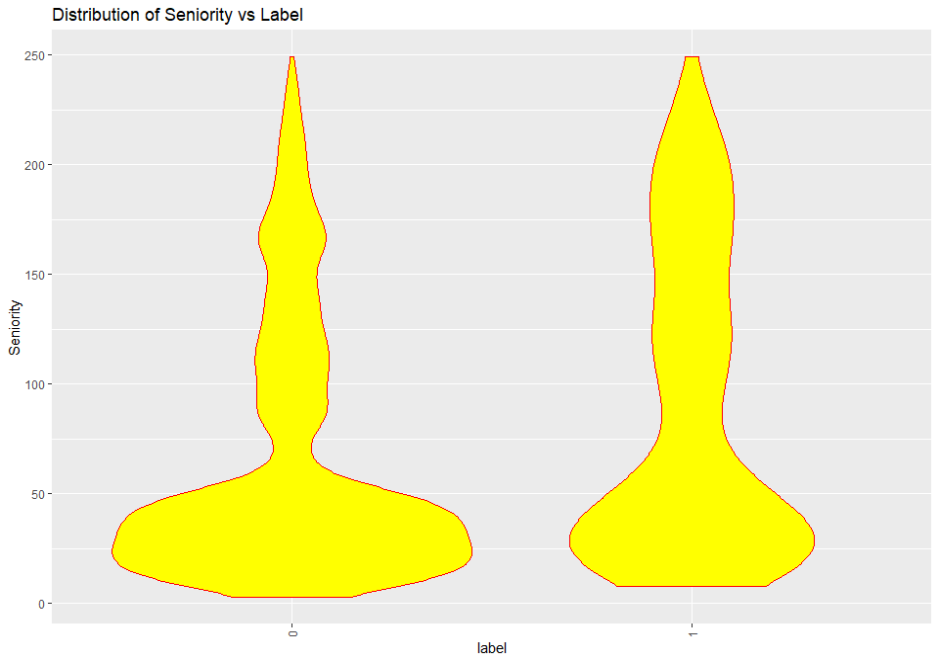

Effect of Res_index in age vs label
75
50
25
age
0
1
label
Res_index
N
S



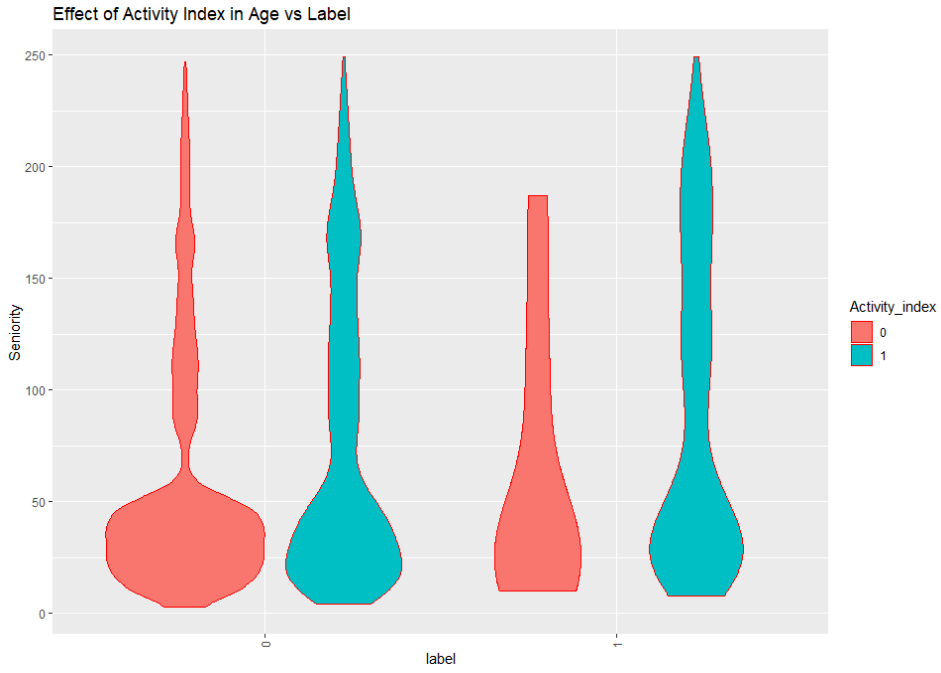
From the plot we could see that the factor “N” of deceased is spread in class “1” for ages 25 to 50 and for class“0” in less than 25.



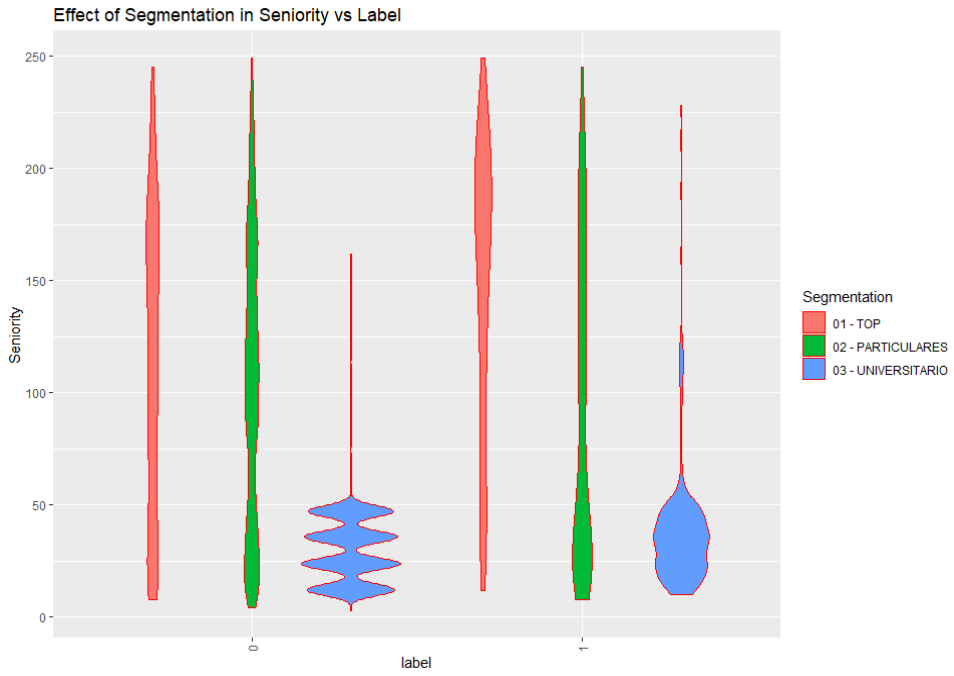
From the plot for ages less than 25 there’s a wide spread in the “03-Universitario”.



From this plot for class“1” we could see that seniority remains constant after 50 months whereas for class“0” the seniority varies periodically.



For both classes the activity index“1”" is similarly distributed whereas for activity index“0” in clas “0” there is a huge drop after 50 months and there’s a slight increase and decrease after 50 months. In class“1”there’s slow decay in the value.



From this plot we can observe huge vibrations present for "03-Universitario" in class"0" till 50 and then there is very minimal value till the 150 and it goes to zero. Similar pattern excluding the vibration present below 50 is observed in class"1".

PD plots for BART:

Age: