International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Bank Customer Segmentation & Insurance Claim Prediction

Yashi Rajput[1], Prof. Vikash Singhal[2], Manish Saraswat[3], Vanshika Chitranshi [4], Mohd. Talib Khan [5], Shipra Shrivastava [6]

[1,3,4,5]B-Tech Student, Greater Noida Institute of Technology, Greater Noida, India
[2]Professor, Greater Noida Institute of Technology, Greater Noida, India
[6]Ass.Professor, Greater Noida Institute of Technology, Greater Noida, India

---

Abstract - *This project will help a bank in segmenting their client and an Insurance company to study the claim Prediction pattern. This design is majorly grounded on data mining and its ways. It majorly focuses on clustering and Prediction using machine literacy and python libraries similar as NumPy, pandas, seaborn, and matplotlib. It also contributes to the enhancement of the pricing models. This helps the insurance company to be one step ahead of its contender. Carrying and acting on client data through the lens of segmentation can have a massive impact on marketing and deals, retention sweats, client service, and more.*

Key Words: **NumPy, Pandas, Seaborn, Matplotlib, Machine learning, Decision Tree, Random Forest, Binary logistic Regression.**

1. INTRODUCTION

Client segmentation is the approach of dividing a large and different client base into lower groups of affiliated guests that are analogous in certain ways and applicable to the marketing of a bank's products and services. Some introductory segmentation criteria include terrain, income, and spending habits. Through client segmentation, banks can get to know their guests on a core particular position and offer them more customized products and services.

Insurance companies are extensively interested in the Prediction of the future. Accurate Prediction gives a probability to drop fiscal loss for the company. The insurers use rather complex methodologies for this purpose. The major models are a decision tree, a arbitrary timber, a double logistic retrogression, and a support vector machine. A great number of different variables are under analysis in this case.

A bank's client segmentation approach can vary extensively and must be grounded on the association's business model and precedences. Parts can be quantitative, similar as by age and gender, or they can be qualitative, similar as separation by values and interests.

The maximum value is attained when banks combine both types of data to more understand the wants and requirements of their client parts, allowing them to offer the right product or service at the right time.

1.1 CLUSTERING

A bank's client segmentation (Clustering) approach can vary extensively and must be grounded on the association's business model and precedences. Parts can be quantitative, similar as by age and gender, or they can be qualitative, similar as separation by values and interests.

The maximum value is attained when banks combine both types of data to more understand the wants and requirements of their client parts, allowing them to offer the right product or service at the right time.

A leading bank wants to develop a client segmentation to give promotional offers to its guests. They collected a sample that summarizes the conditioning of druggies during the once many months. We've to identify the parts grounded on credit card operation.

1.2 CART-NF-ANN

An Insurance Claim Prediction (Cart-NF) firm providing tour insurance was facing higher claim frequency. The management decided to collect data from the past few years. We will make a model that predicts the claim status and provides recommendations to management. Using CART, RF & ANN or comparing the models' performances in train and test sets.

2. WORKING

It is divided into two sub problems:

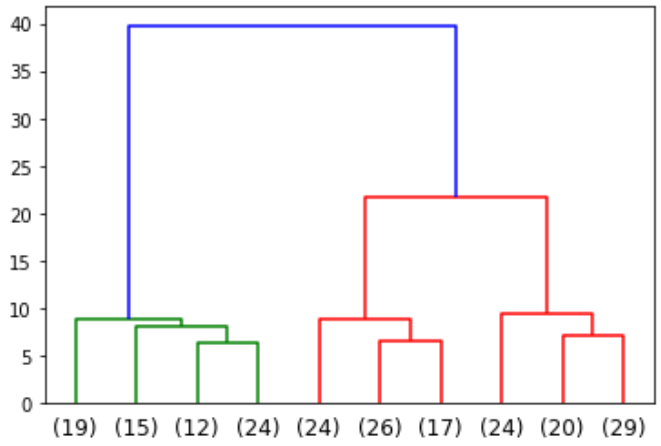
2.1 Bank Customer Segmentation

Step 1: Doing all the initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

Step 2: Justifying that scaling is necessary for clustering in this case.

Step 3: Applying hierarchical clustering to scaled data. Identifying the number of optimum clusters using Dendrogram and briefly describing them.

DENDROGRAM :

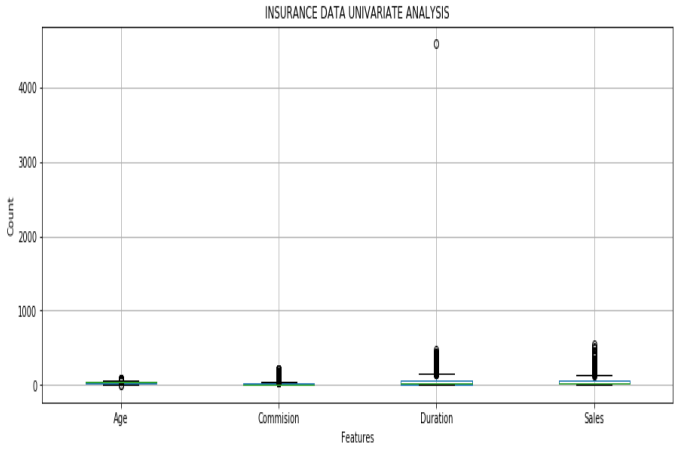


Step 4: Applying K-Means clustering on scaled data and determining optimum clusters. Applying elbow curve and silhouette score. Explaining the results properly. Interpreting and writing inferences on the finalized clusters.

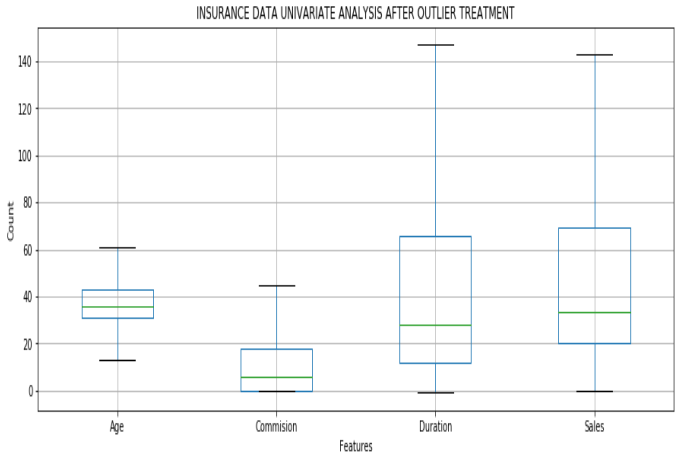
Step 5: Describing cluster profiles for the clusters defined. Recommending different promotional strategies for different clusters.

2.2 Insurance Claim Prediction

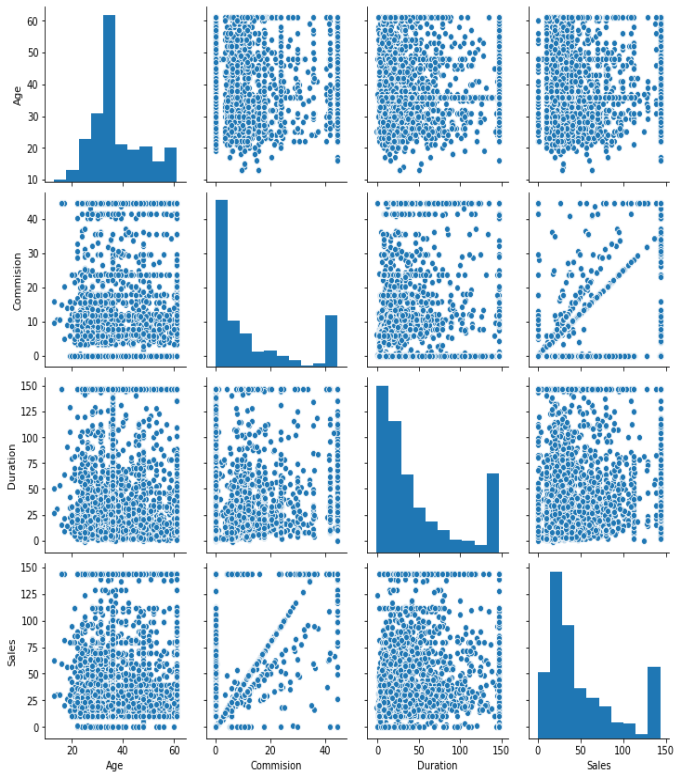
Step 1: Doing all the initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).



We see that all 4 continuous variables : Age , Commission , Duration and sales have outliers present.



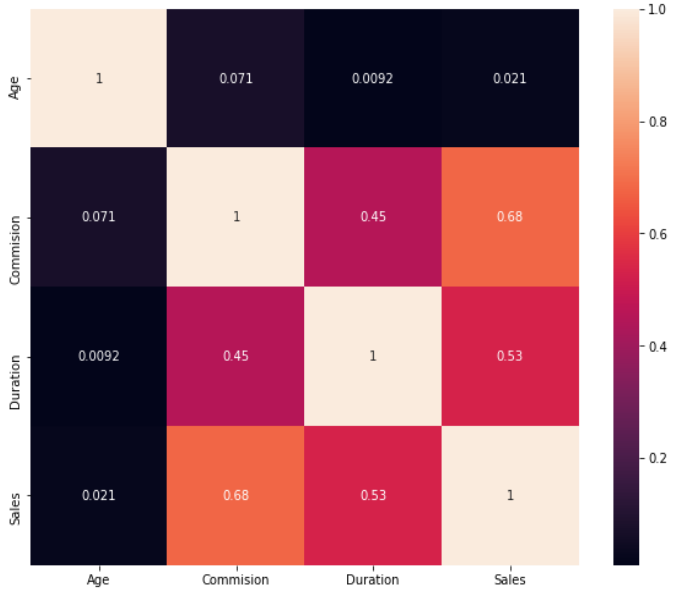
Now we have treated the outliers although in this problem we shall proceed with complete dataset as the data is mostly relevant for analysis and we shall get a more generalized result.



#We can see from above BIVARIATE ANALYSIS that Sales and Duration have linear relation only and

#None of the other features have linearity.

#This implies that with higher commission there is more sales in the insurance marketplace.



#Again we can conclude from heatmap that Sales and commision have a high correlation .

Step 2: Data Splitting: Splitting the data into test and train, building classification model CART, Random Forest, Artificial Neural Network.

Step 3: Performance Metrics: Checking the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model.

Step 4: Final Model: Comparing all the models and writing an inference about which model is best/optimized.

Step 5: Inference: Based on the whole Analysis, what are the business insights and recommendation.

3. ARTIFICIAL NEURAL NETWORK MODEL:

ANN on train dataset:

Confusion matrix:

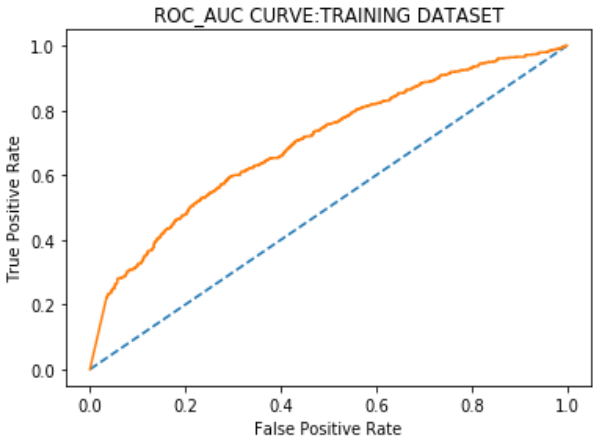
[[1258(TN), 101(FP)],

[398(FN), 245(TP)]],

Classification report:

	precision	recall	f1-score	support
0	0.76	0.93	0.83	1359
1	0.71	0.38	0.50	643
accuracy			0.75	2002
macro avg	0.73	0.65	0.66	2002
weighted avg	0.74	0.75	0.73	2002

Roc Auc Curve:



AUC: 0.700

Accuracy: 55.69%

ANN on test dataset:

Confusion matrix:

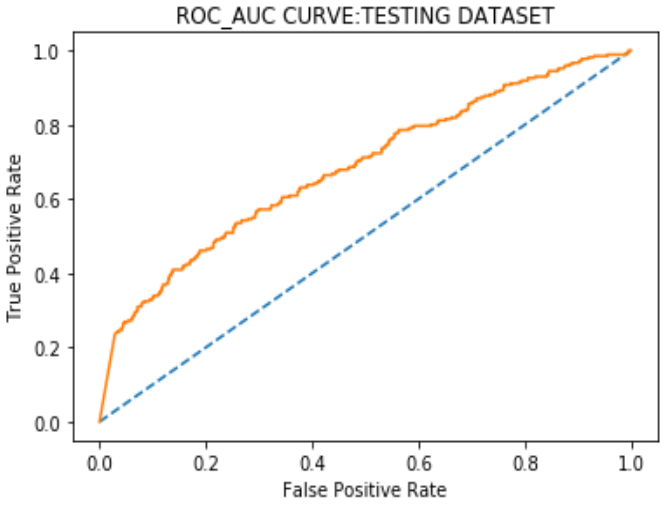
[[547(TN), 41(FP)],

[170(FN), 101(TP)]],

Classification Report:

	precision	recall	f1-score	support
0	0.76	0.93	0.84	588
1	0.71	0.37	0.49	271
accuracy			0.75	859
macro avg	0.74	0.65	0.66	859
weighted avg	0.75	0.75	0.73	859

Roc AUC Curve:



AUC: 0.683

Accuracy: 54.71%

From the above Evaluations we can again conclude that the evaluation

values are close to the same for both the train and test datasets therefore we can accept this model as well since it has an acceptable range of accuracy and is neither over fitted nor under fitted.

4. Objectives

Problem 1 Objectives:

Making clusters according to the dataset or segmenting customers according to their credit card usage. After making clusters we will study whom to give extended benefits and whom to educate about digital transaction benefits and whom to reduce into single monthly EMIS and various other important aspects which should be taken care of.

Problem 2 Objectives:

The main objective here is to reduce Claim frequency and improve sales. Accordingly we will proceed in data set and do the further analysis resulting into our objective.

5. Business Insights:

Based on the whole Analysis, what are the business insights and recommendation, The following inferences and recommendations can be put forward to the insurance company's management:

The Business objective here is to reduce Claim frequency and improve sales:

1. We have initially analyzed that Agency code i.e. firms providing tour packages have a high importance in claim status. A detailed analysis of claim per Agency can be worked out and products can be limited or modified to agencies rendering high claims in order to cover the settlement cost.

2. There is a close to linear relation between commission and sales which is a clear indicator that insurance lead generators who make a good amount of sales earn high commission , management can work out some schemes or added benefits for sales people to motivate them and induce a sense of competition among them to achieve higher and higher sales.

3. The cases where prediction is False Positive can be looked at closely and the reason for decline of claim settlement could be studied and improved to provide customer satisfaction and retention.

4. Other features like Age and Destination can be grouped into Low, Medium and High Risk categories and subsequent pricing could be done in order to limit the risk and reduce claims.

5. In order to reduce claims, the False Negative cases have to be minimized thus improving model accuracy.

6. CONCLUSIONS

Conclusion of problem 1:

The cluster 1 customers could be provided with an add-on card in order to cover their high spending or credit card limit could be enhanced resulting in added revenue to the Bank. The cluster 1 customers are already generating high revenue for the bank therefore various discounts could be offered for Ecom based on spending patterns. This will increase customer retention and decrease the churn rate. Also, the annual fee could be waived off based on a certain amount of

spending. Cluster 0 the middle-income group is the one who has a high spending potential but may be dealing mostly in cash transactions and therefore Bank should chart out strategies to educate these customers about the advantages of making Digital/Card Transactions and onboard them on the Digital Platform. Once Cluster 0 customers switch to the Digital/Card mode of payment they could be laddered up to Cluster 1 by giving them the right offer at the right time increasing their spending and thus enhancing the bank's wallet share in this cluster as well. Cluster 2 is a comparatively low-income group, here Bank may provide offers to break down big single spent into EMIs to ease the monthly bill and ultimately reduce the possibility of default.

Conclusion of problem 2:

We have initially analyzed that Agency code i.e. firms providing tour packaged have a high importance in claim status. A detailed analysis of claim per Agency can be worked out and products can be limited or modified to agencies rendering high claim in order to cover the settlement cost. There is a close to linear relation between commission and sales which is a clear indicator that insurance lead generators who make a good amount of sales earn high commission , management can work out some schemes or added benefits for sales people to motivate them and induce a sense of competition among them to achieve higher and higher sales. The cases where prediction is False Positive can be looked closely and the reason for decline of claim settlement could be studied and improved to provide customer satisfaction and retention. Other features like Age and Destination can be grouped into Low , Medium and High Risk category and subsequent pricing could be done in order to limit the risk and reduce claims. In order to reduce claims, the False Negative cases have to be minimized thus improving model accuracy.

REFERENCES

[1] ACRL Research Planning and Review Committee. (2014). Top trends in academic libraries: A review of the trends and issues affecting academic libraries in education . College & Research Libraries News, 75(6),294-30

[2] ZTM (Andrei Neagoei: Data science Instructor)

[3] Tayyab (Analytics Consultant at EXL)

[4] Great Learning

[5] Stack overflow

[6] Google Professional data analytics course

[7] Kaggle