INSURANCE CROSS-SELL

Project Report

Sachin Sharma (IIIT Delhi)

TABLE OF CONTENTS

LIST OF FIGURES/TABLES

EXECUTIVE SUMMARY

This report introduces an analytics-based strategy for a consumer bank to effectively promote personal protection insurance (PPI) to its existing customer base. The primary goal is to identify customers who currently do not possess PPI but hold either secured or unsecured loans. Additionally, the report aims to determine the specific type of PPI product that should be targeted towards these customers.

To achieve this objective, a comprehensive analysis was conducted using provided sample data from the bank. The analysis involved a thorough examination of customer demographics, loan types, and other pertinent factors in order to identify potential candidates for PPI cross-selling.

Initially, a segmentation analysis was performed to categorize customers based on their loan types and PPI ownership status. This segmentation allowed for the identification of a specific group of customers who are eligible for PPI cross-selling.

Subsequently, predictive modeling techniques were employed to identify the customers most likely to respond positively to PPI offers. By leveraging historical data and customer attributes, the model identified key indicators that significantly influence the probability of PPI adoption. This information was then utilized to develop a prioritized list of potential targets for the bank's marketing efforts.

Furthermore, the analysis explored the various types of PPI products available and assessed their suitability for different customer segments. Factors such as customer profiles, loan types, and risk levels were taken into consideration, enabling the report to provide recommendations on the specific types of PPI products that should be targeted towards each customer group.

The findings of this analysis underscore the importance of utilizing data-driven insights to optimize cross-selling endeavors. By effectively targeting the appropriate customers with the most relevant PPI products, the bank can maximize the likelihood of success and enhance overall customer satisfaction.

The recommendations and conclusions outlined in this report offer a strong foundation for the bank's decision-making process. It is advised that the bank further refines its marketing strategies and communication approaches based on the insights provided, while continuously monitoring and evaluating the effectiveness of the cross-selling initiatives.

By adopting an analytics-driven approach, the bank can bolster its cross-selling capabilities, stimulate revenue growth, and cultivate stronger relationships with its customer base.

PROBLEM STATEMENT

A consumer bank with a range of products would like to cross-sell insurance to its consumer base (that is, cross-sell the personal protection insurance (PPI) product to those customers who have a secured or unsecured type of loan, but no PPI product as yet.

The bank would like to adopt analytics driven approach applied on this sample data for deciding:

1. Who should they target from the pool of customers that currently do not have a PPI, and
2. What type of PPI product they should be targeting them with.

SCOPE

The scope of the study is to perform an exploratory data analysis on the data to gain insights into the characteristics and trends exhibited by the variables in the dataset. The aim is to determine the target customers who are likely to be interested in purchasing PPI (Payment Protection Insurance).

1. The study will utilize suitable analytical methods including Exploratory Data Analysis (EDA), Predictive Modelling, and Machine Learning algorithms to identify potential customers for cross-selling PPI products.
2. Additionally, the study will analyse the combination of PPI products purchased by customers using EDA and ML Modelling to determine the optimal mix.

INTRODUCTION

What is Cross-Selling?

Although cross-selling is a widely adopted approach in various sectors, its significance and effectiveness are particularly pronounced in the banking industry. Given the highly competitive nature of the banking market, maximizing the value derived from each customer is vital for sustained prosperity. As Forbes reports, selling to existing customers is nearly 50% less challenging than acquiring new leads. Cross-selling enables bank personnel to capitalize on this advantage.

Advantages of Cross-Sell:

The effectiveness of cross-selling in the banking sector stems from the contrast between customer retention and customer acquisition. Products achieve a success rate of 60-70% when persuading current clients to make a purchase. Conversely, the success rate of selling to new customers ranges from only 5% to 20%.

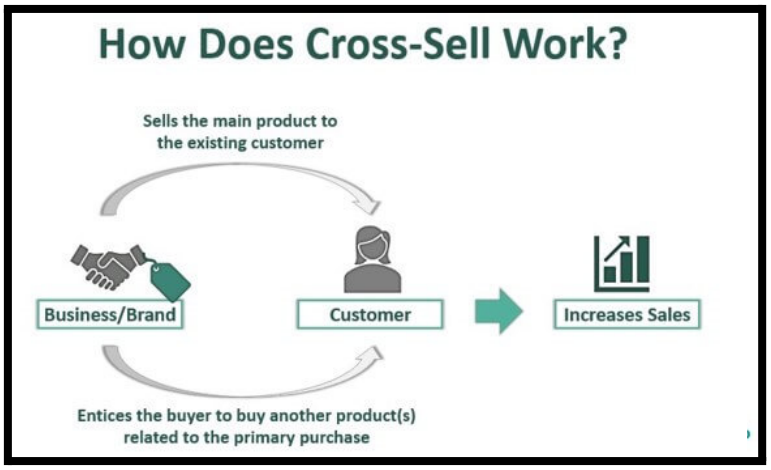


Figure 1: Cross Sell Concept

Cross-selling offers several additional benefits, including:

1. **Elimination of acquisition costs**: Huify (Inbound Marketing and Sales Agency) discovered that acquiring a new customer can cost five times more than serving an existing one. Cross-selling eliminates the need for additional expenses associated with customer acquisition for new sales.

2. **Cultivation of brand loyalty**: Particularly in banking, providing a range of products to consumers helps position financial services as a comprehensive solution, catering to all their needs in one place.

3. **Increased revenue:** By offering additional products at the right time when customers genuinely require them, banks can generate more revenue while simultaneously meeting the needs of their clients.

In this case, the focus is on cross selling the PPI (Personal Protection Insurance) to the existing customer base of the consumer bank.

Personal Protection Insurance:

Personal protection insurance refers to a type of insurance coverage designed to protect individuals and their loved ones from various risks and uncertainties. It encompasses a range of policies that provide financial security in case of unfortunate events such as disability, critical illness, or death.

Personal protection insurance typically includes life insurance, which pays out a lump sum to beneficiaries upon the insured person's death. It can help cover funeral expenses, replace lost income, or provide for the financial needs of dependents. Disability insurance offers income replacement if the insured individual becomes unable to work due to injury or illness. Critical illness insurance provides a lump sum if the policyholder is diagnosed with a specified critical illness, assisting with medical expenses and other financial burdens.

By offering financial support during challenging times, personal protection insurance offers peace of mind and safeguards against unforeseen circumstances, allowing individuals to focus on recovery and providing for their loved ones.

DATA DISTRIBUTION

The dataset consists of three main sections and can be divided into three categories: Demographic Data, Consumer Bank Data, and PPI Product Based Data. Each of these categories contains significant variables. In total, the dataset comprises 59 variables.

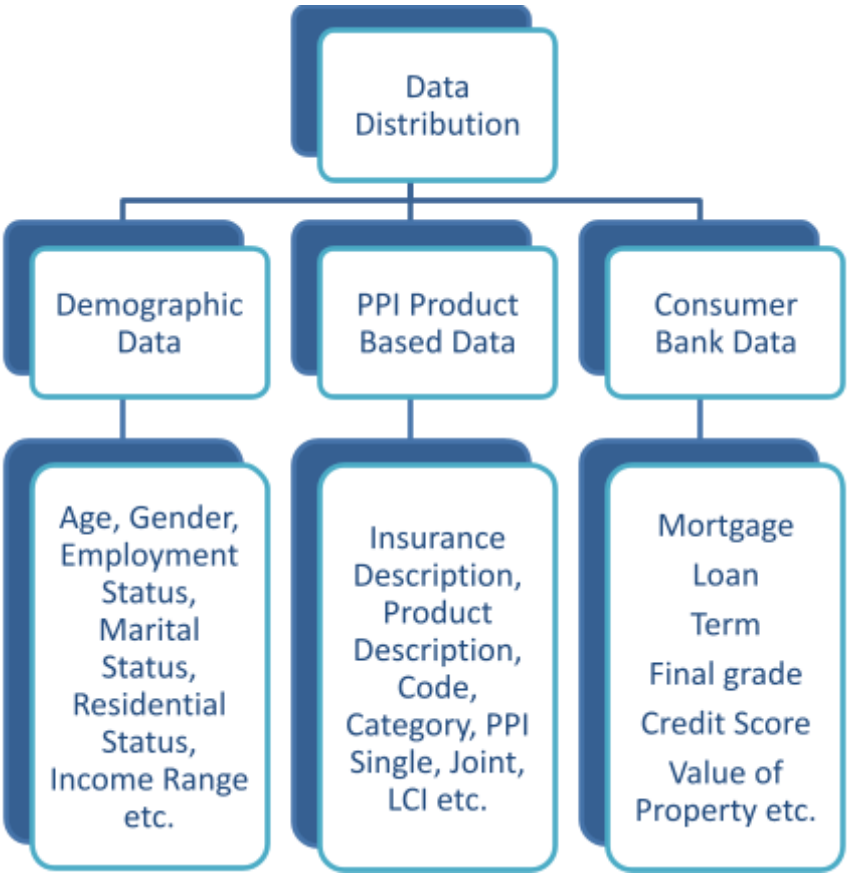


Figure 2: Data Segmentation

DATA PRE-PROCESSING

Data Description:

There are 16383 observations with 59 variables out of which 37 are numerical and 22 are categorical.

The case further required a comprehensive description of all the variables. As a result, the subsequent table presents a detailed description of each variable contained within -

Columns	Description
Ref	Reference No.
Credit_Score	Depicts an individual's creditworthiness, reflecting their financial history and indicating the likelihood of repaying debts responsibly.
Final_Grade	A classification system that involves assigning a quality score to a loan based on a borrower's credit history, quality of the collateral, and the likelihood of repayment of the principal and interest
Term	term refers to the duration or length of time over which a loan or financial instrument is agreed to be repaid.
Net_Advance	Net Advances means the principal amount of the outstanding Advances minus the amounts then on deposit in the Accounts representing Principal Proceeds.
APR	APR stands for Annual Percentage Rate. Represent the overall cost of borrowing, including both the interest rate and any additional fees or charges associated with the loan.
Loan_Type	Type of loan -- Secured or unsecured
Mosaic	Experian's Mosaic is a household-based consumer lifestyle segmentation system that classifies all households, providing a 360-degree view of consumers' choices, preferences and habits. Mosaic Global is a consistent segmentation system that covers over 284 million of the world's households. It is based on a simple proposition that the world's cities share common patterns of residential segregation
Mosaic_Class	Mosaic class has 11 categories of Mosaic
Time_at_Address	the time that the customer have lived at thier current address(in months)
Residential_Status	Residential status of customer. H- Own House; R -Rent; L-Lease; TTenant
Telephone_Indicator	A customer has Telephone or not. Y - ACTIVE; N - Inactive
Number_of_Dependants	No. of dependents of customer
Marital_Status	Marital Status of the customer
Gender	Gender of the customer
Time_in_Employment	Number of Months in employment
Employment_Status	Employment status of customer

Full_Part_Time_Empl _Ind	Status of employment: Fulltime or Part-Time
Perm_Temp_Empl_Ind	Status of employment: Permanent or Temporary
Income_Range	Income range of Customer, 0-6 Levels
Current_Account	Current account status of customer
ACCESS_Card	Does the customer hold an Access Card - True/False
VISA_Card	Does the customer hold a VISA Card - True/False
American_Express	Does the customer hold a American Express Card - True/False
Diners_Card	Does the customer hold a Diners Card - True/False
Cheque_Guarantee	When a customer presents a cheque to a merchant or another party, the cheque guarantee service provides a guarantee to the recipient that the cheque will be honoured up to the specified amount, even if the customer's account does not have sufficient funds to cover the payment. Cheque guarantee is typically offered to customers who have a good credit history and have been with the bank for a certain period of time
Other_Credit_Store_Card	Information on other Credit card customers. It refers to a type of credit card that is issued by a retail store or a retail chain, rather than a bank. These cards can only be used to make purchases at the store or chain that issued the card.
Time_with_Bank	Time with bank-how much time completed after account opening
Value_of_Property	The value of a property at any given time is determined by what the market will bear.
Outstanding_Mortgage_Bal	The amount of money it would take to pay off the loan in full in case of secured loan or when the customer has mortgaged something
Total_Outstanding_Balances	Total Outstanding Balance in relation to any Account Statement means the total of all the Outstanding Balances of all the Card Account(s) stated in the Account Statement.
Bureau_Data___Monthly_ Other_Co_R	Other Company's Rating
Worst_History_CT	History of Current Term
Payment_Method	The number of ways in which merchants can collect payments from their customers. 'C' - Cash , 'S' - Swipe(Card) , 'D' - Digital Wallet
Age	Age of the Customer
Total_outstanding_balance __mortg	any interest due and payable on any Top-Up Loans in respect of such Mortgage Loans which have been or will be added to the Mortgage Portfolio.
Total___Public_Info___CCJ ____ban	If your creditor has taken court action against you for a debt, they may have got a county court judgment (CCJ) or other court order against you. A court order means you have to pay the money back, either in instalments or in full by a certain date. (No of total cases)
Total_value__Public_Info ___CCJ	CCJ Value

Time_since_most_recent _Public_In	Time since last CCJ case (Assuming 99 to be null)
Total_value__CAIS_8_9s	CAIS stands for Consumer Credit Account Information Sharing (0 means the consumer has not shared the data to CAIS ; Rest are number of members to which the information has been shared under CAIS).
Worst_status_L6m	Last six months worst status
Worst_CUrrent_Status	Current Worst Status
__of_status_3_s_L6m	Last 6 months status
Searches___Total___L6m	last six months status
Bankruptcy_Detected__SP_	bankruptcy detected
Total___outstanding_CCJ_s	total outstanding of CCJ (# cases)
Total_outstanding_balance ___excl	any interest due and payable on any Top-Up Loans in respect of total outstanding loans except Mortgage Loans .
Total___of_accounts	total number of accounts
CIFAS_detected	Credit Industry Fraud Avoidance System detected
Time_since_most_recent _outstandi	Time since last payment
Insurance_Description	describe insurance and its type
PPI	Status of PPI sold or not
code	Code for the product of PPI
prdt_desc	Description of PPI Products
category	Category for a product of PPI
PPI_SINGLE	Its type of PPI Product is Single
PPI_JOINT	It's a type of PPI Product Joint
PPI_LCI	It's a type of PPI Product LCI

Table 1 – Description of all the columns

Missing Values Treatment: The dataset contains 6978 instances of missing values for the variable "code." These values correspond to the PPI product code, which is deemed irrelevant for the analysis. In regard to the variables "Insurance," "Product Description," and "Category," where the PPI value is '0,' the missing values have been substituted with the term "No Response."

Negative values: There were some negative values in some rows which has been removed.

Typing Mistakes: The typo mistakes have been corrected. For example – "**FALS"** was changed to "**FALSE"**

Outliers Treatment: Outliers in all the continuous variables are detected and have been capped & floored at a certain value using Percentile method.

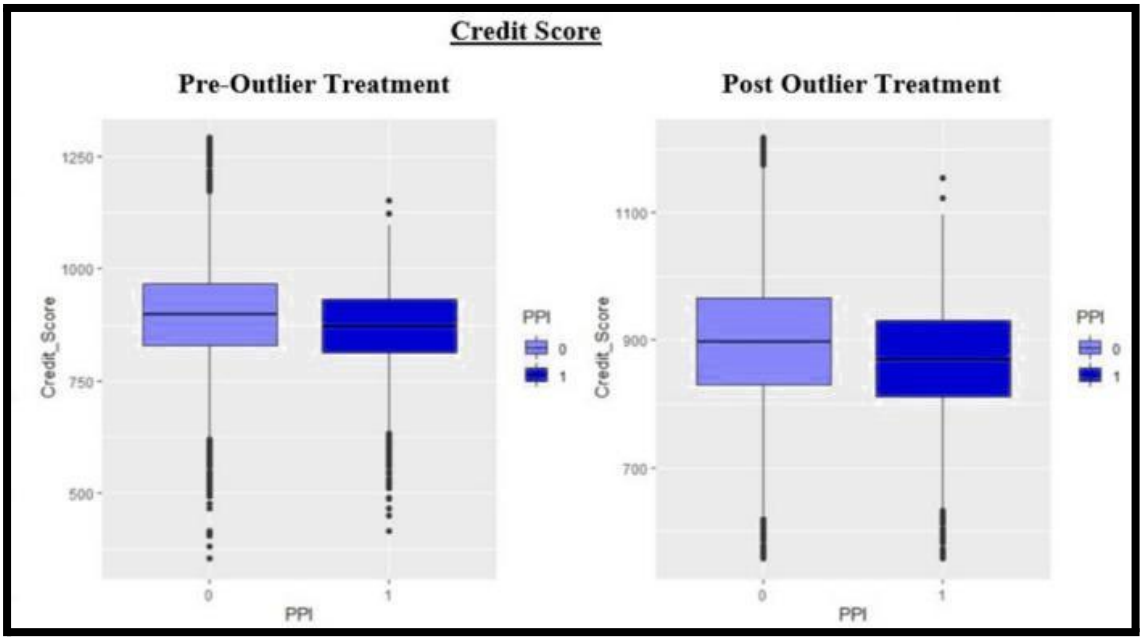


Figure 3.1 – Outliers Treatment of Credit Score

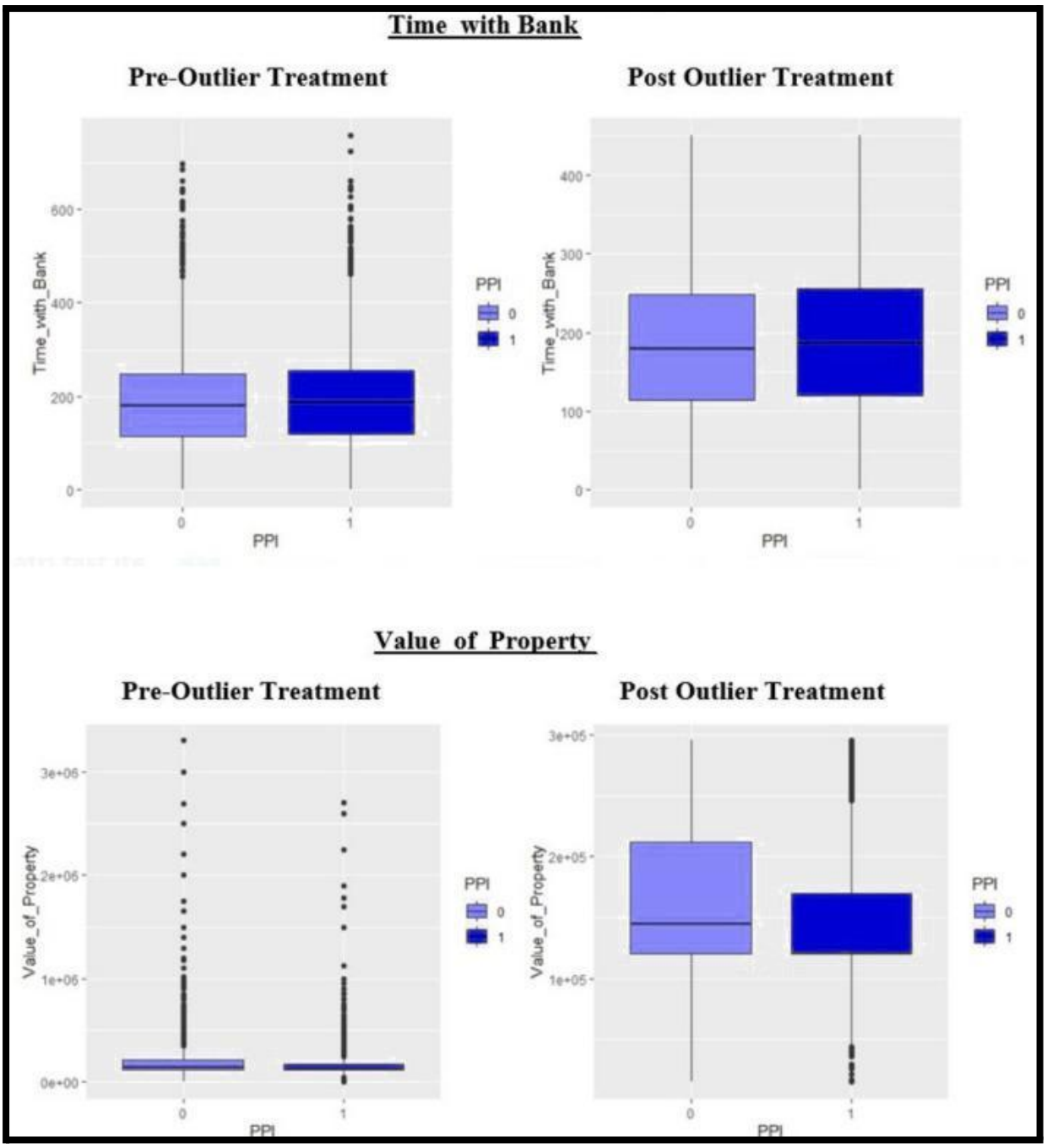


Figure 3.2 – Outliers Treatment of Time with Bank & Value of Property

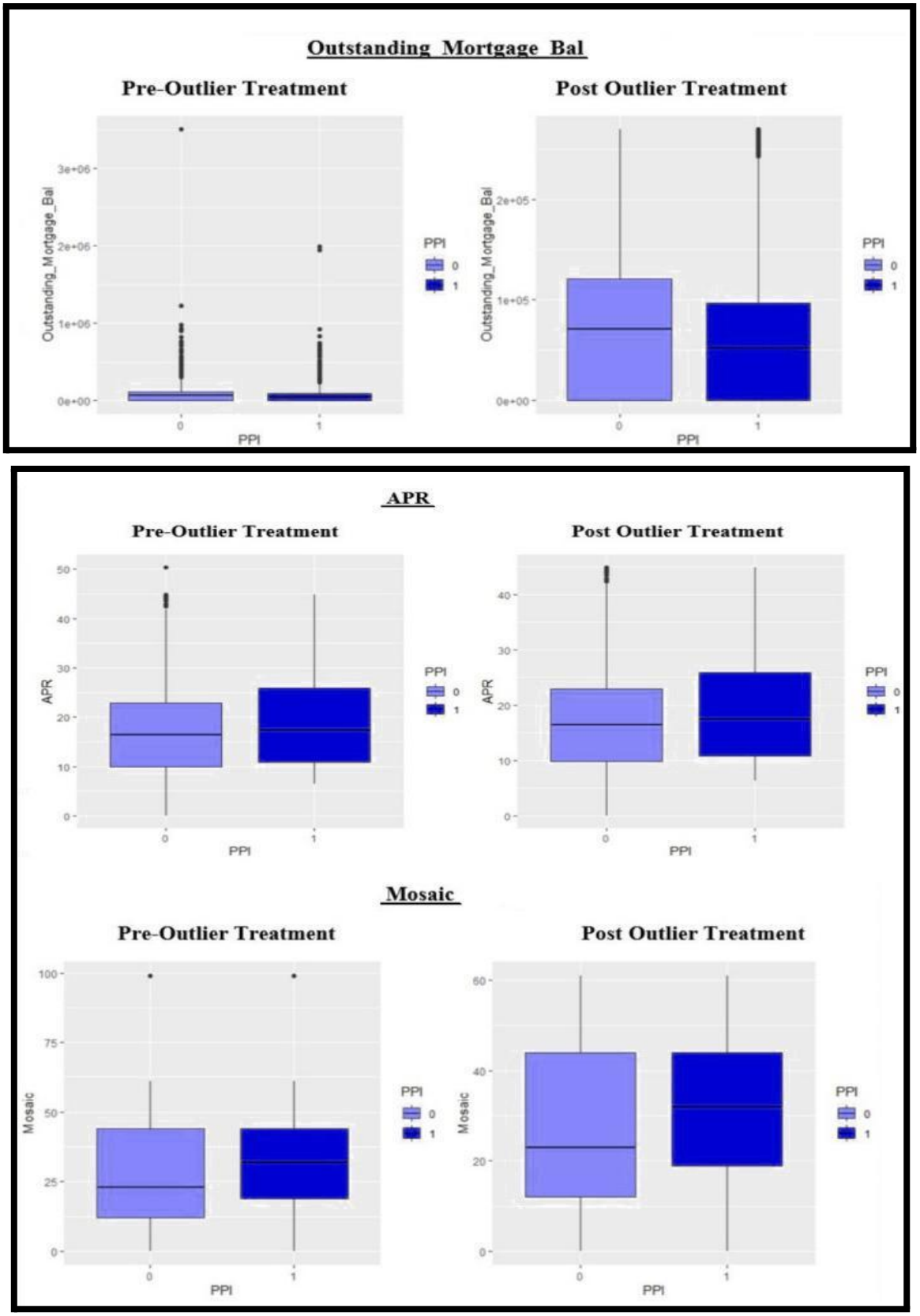


Figure 3.3 – Outliers Treatment of Mosaic, APR & Outstanding balance

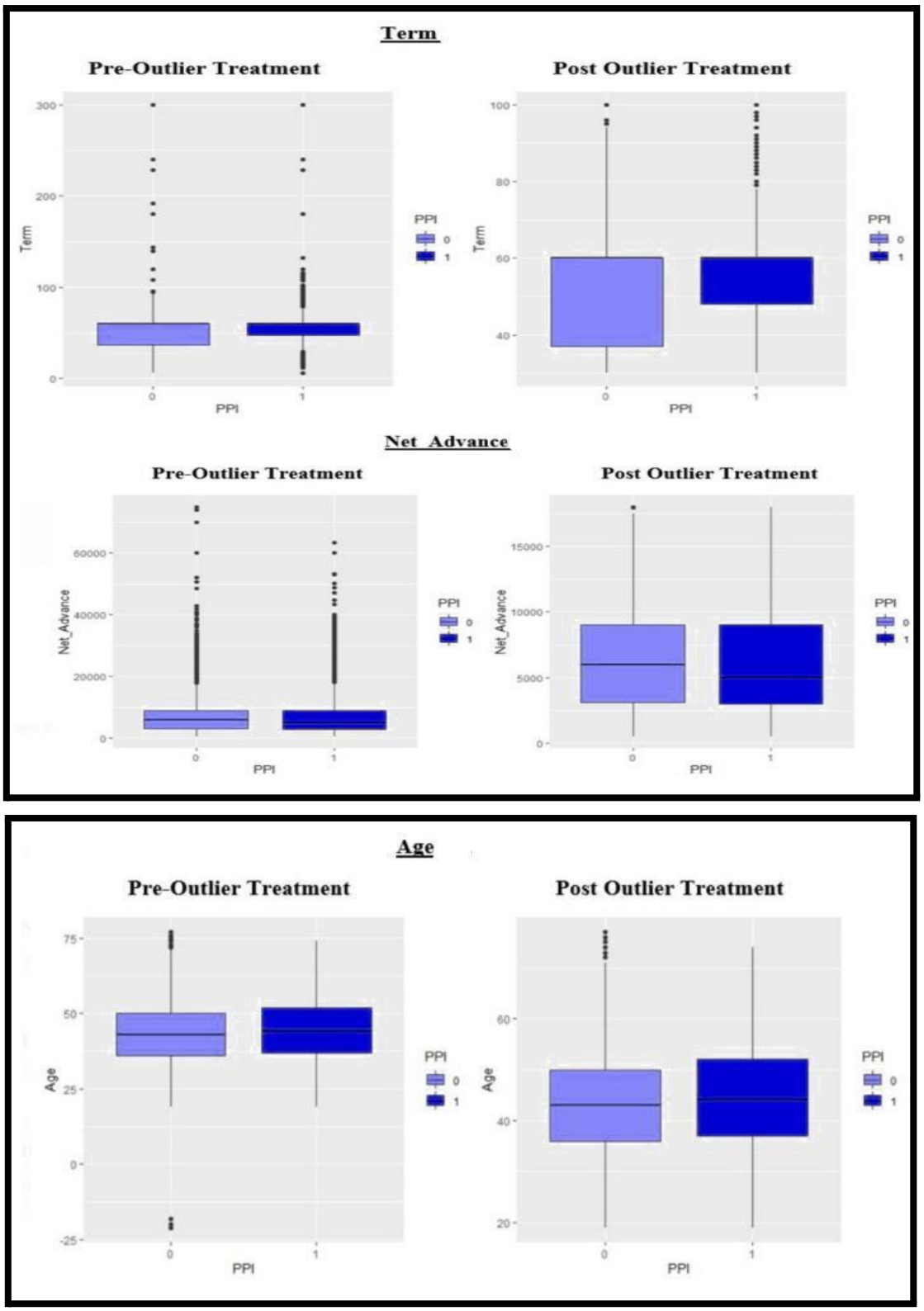


Figure 3.4 – Outliers Treatment of Age, Term & Net Advance

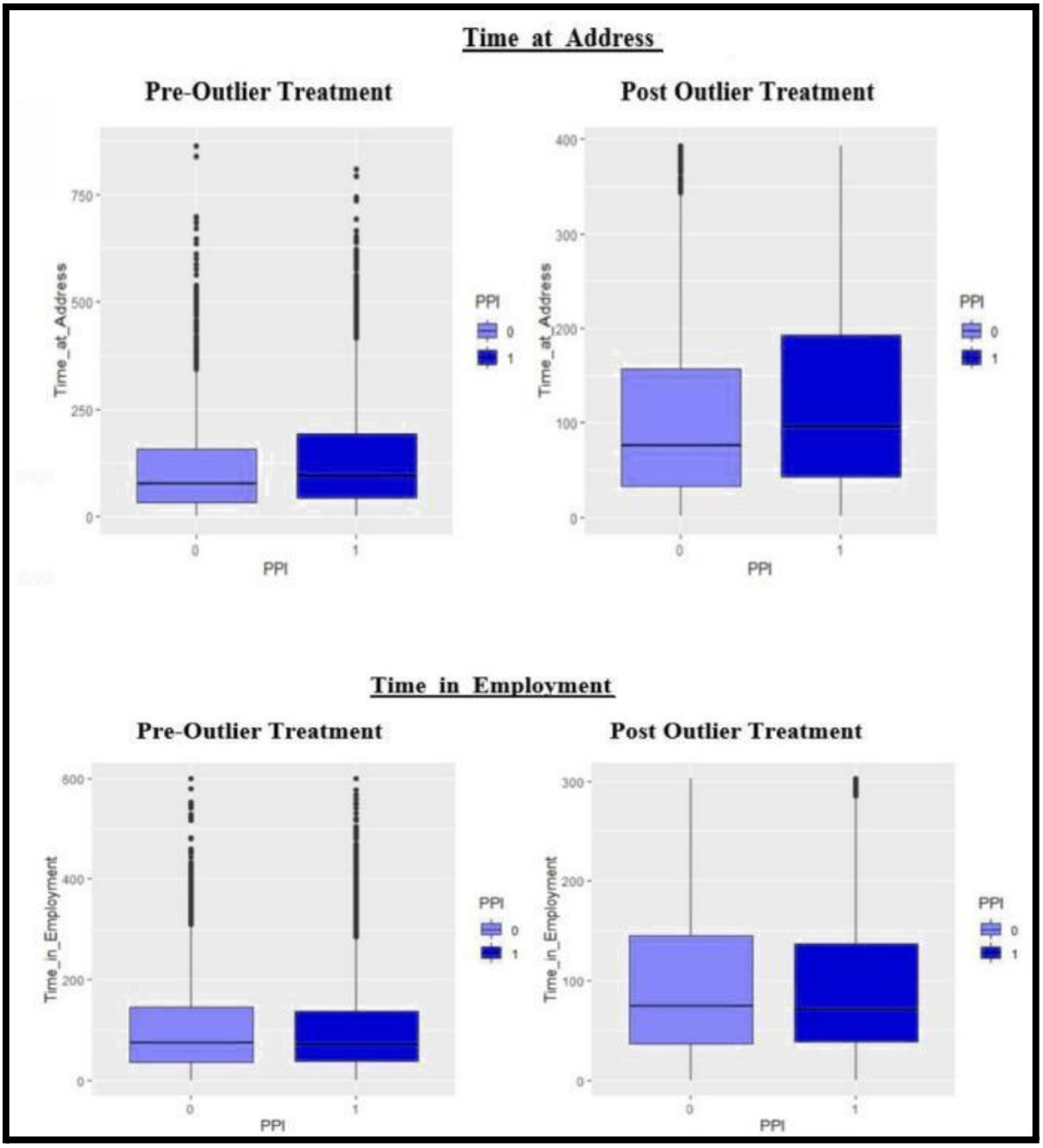


Figure 3.5 – Outliers Treatment of Time in Employment or in Address

UNDERSTANDING THE DATA

Out of all the customers, 57.8% are currently PPI Buyers, while the remaining 42.2% do not have any PPI holdings at present.

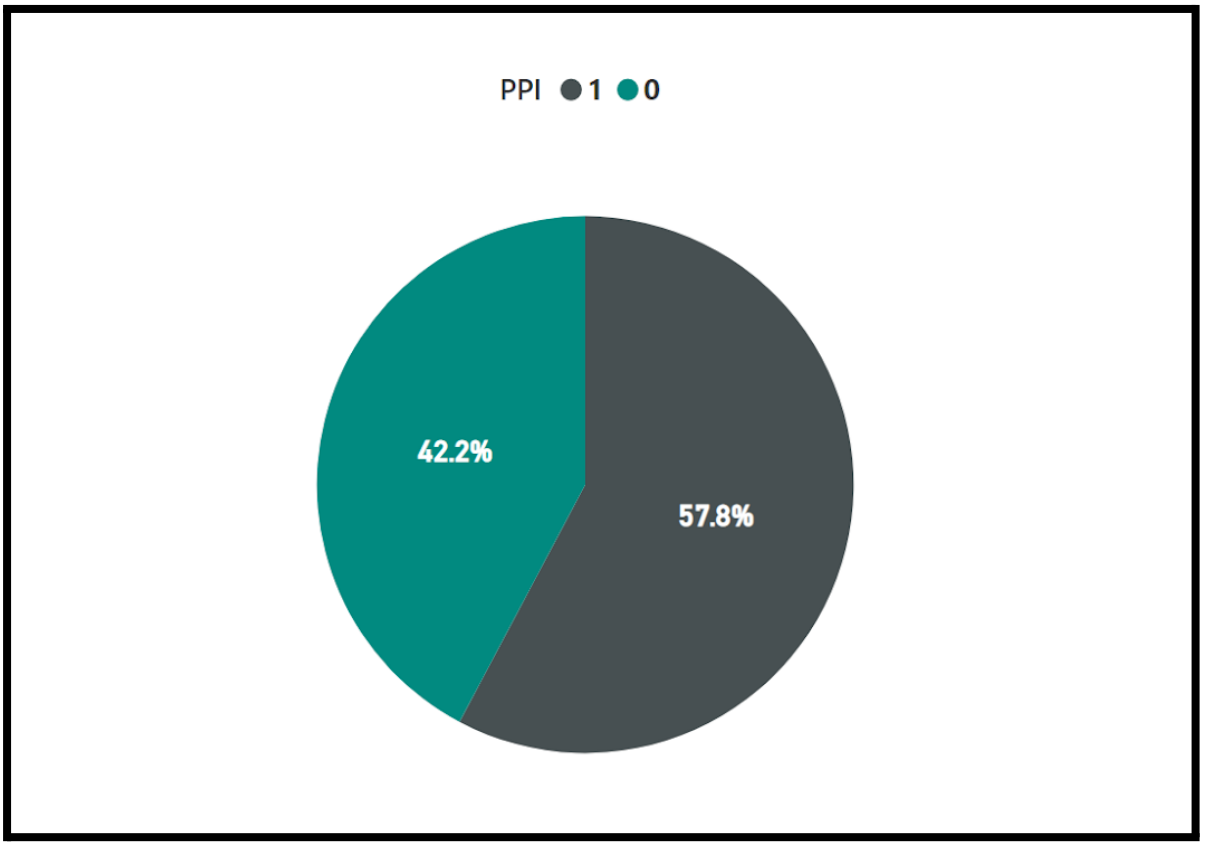


Figure 4: PPI-wise Customer Distribution

If we consider the customers having PPI, LASU is the top selling Insurance followed by Life & Critical Illness and Single is the best-selling Category followed by LCI.

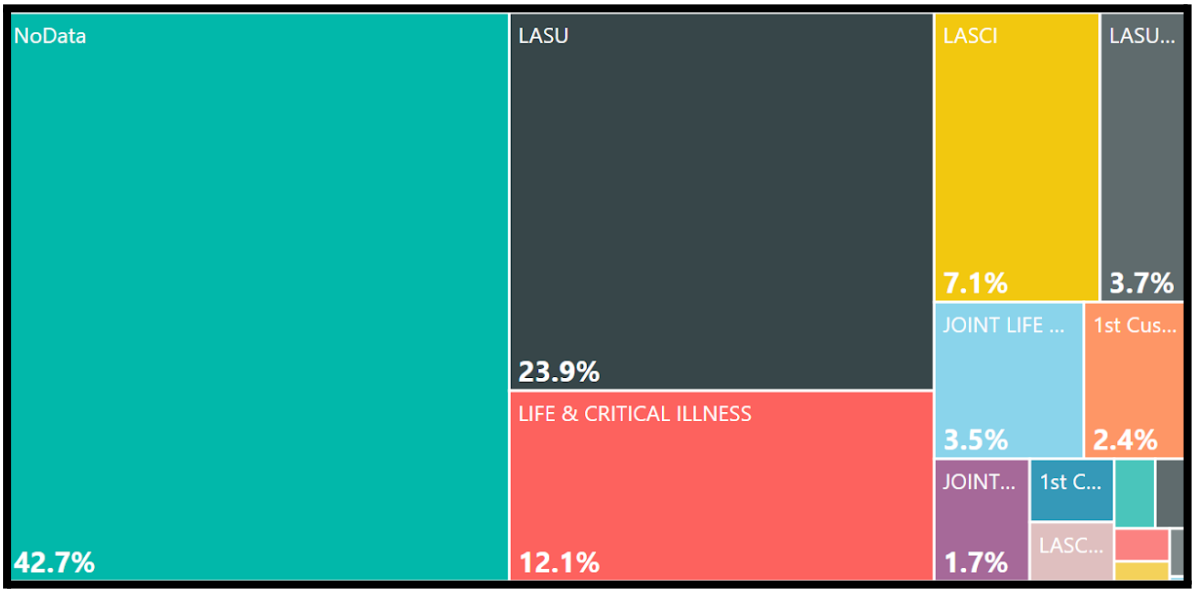


Figure 5: Product Description Distribution among Customers

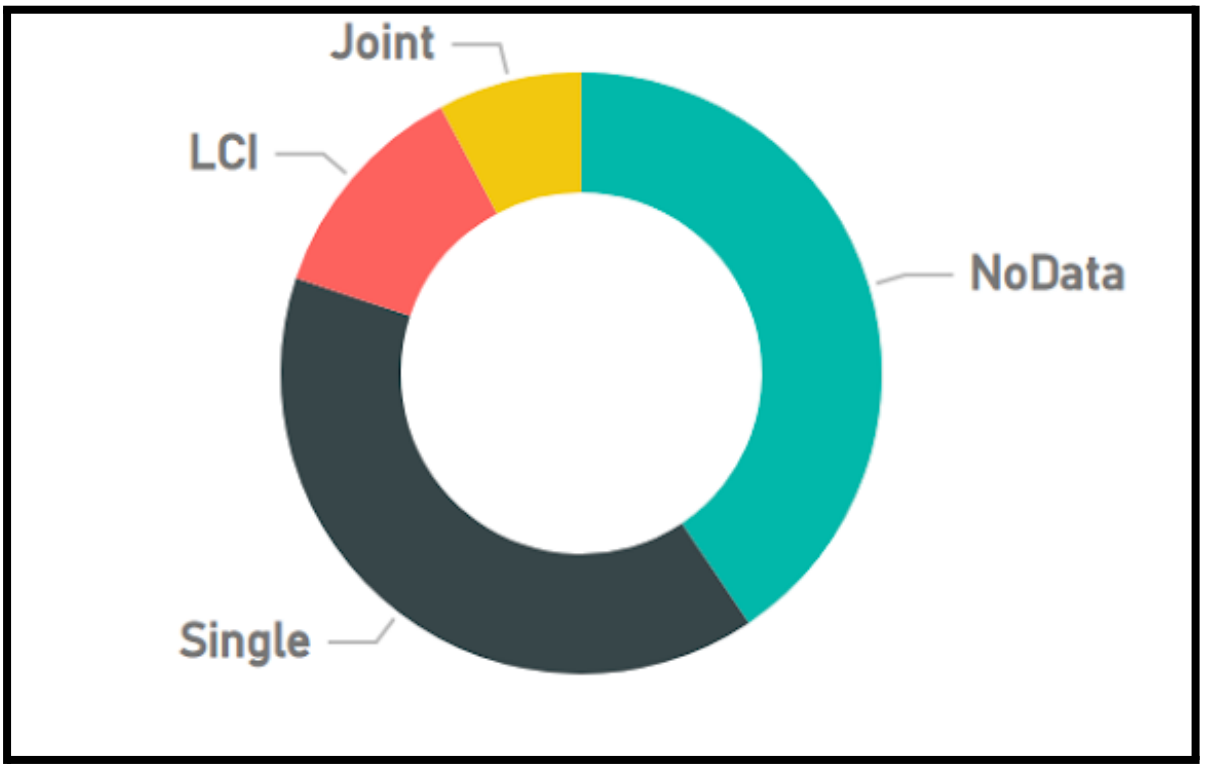


Figure 6: Category Distribution among Customers

FEATURE EXPLORATION

Using Correlation Matrix:

A thorough examination of the variables' correlation was conducted, and any variables exhibiting significant co-linearity were eliminated from the dataset. The correlation matrix, both before and after the removal of highly correlated variables, is presented below. This process was crucial to ensure the integrity and accuracy of the data analysis. Identifying and addressing high co-linearity is important because it helps avoid multicollinearity issues, which can distort statistical models and hinder the interpretation of results. By eliminating variables with high correlation, the dataset is refined, resulting in a more reliable and robust foundation for subsequent analyses. This step contributes to enhancing the overall quality and validity of the findings, ultimately facilitating more accurate and meaningful insights.

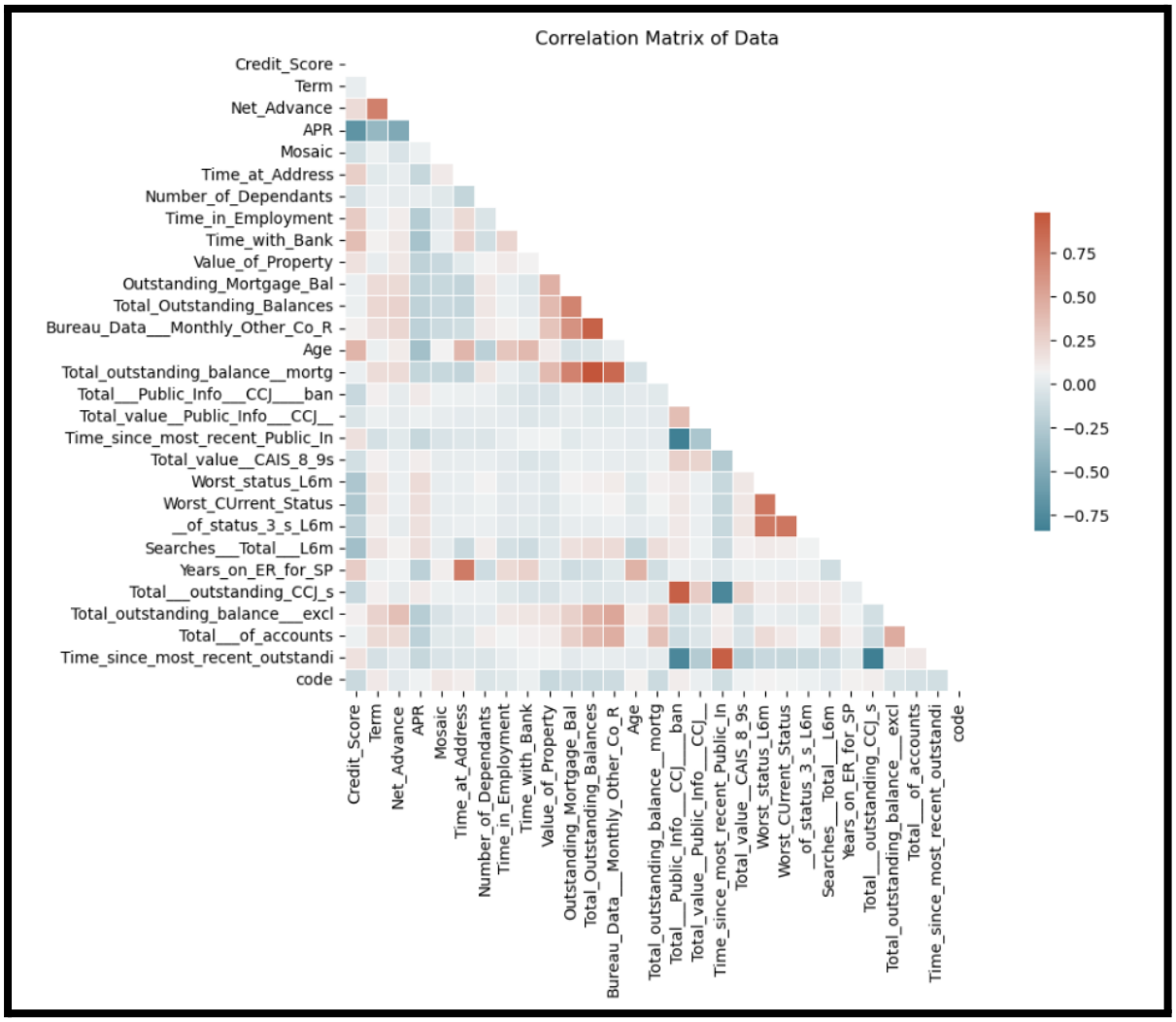


Figure 7: Correlation Matrix before removing the high-co-related variables.

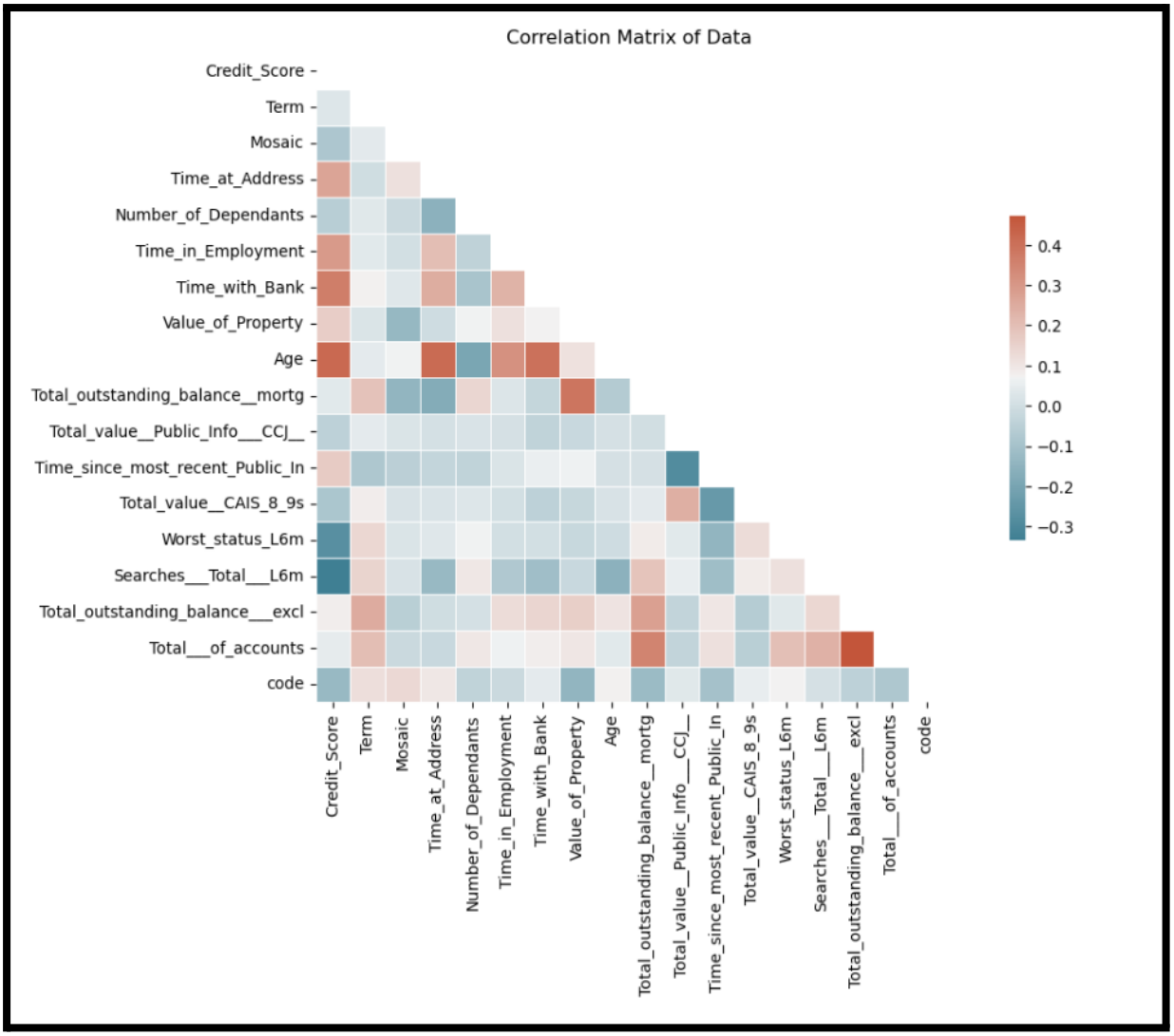


Figure 8: Correlation Matrix after removing the high-co-related variables.

Using EDA:

After conducting an exploratory data analysis (EDA), certain variables were deemed irrelevant in determining the individual's likelihood of purchasing the PPI (the target variable in this case). Consequently, these variables were excluded from any subsequent analyses. The following is a list of all the variables that were removed due to their lack of significance in the analysis –

- *Full time part time*
- *Telephone Indicator*
- *Permanent Temporary*
- *Payment Mode*
- *Current Account*

- *American Express*
- *Diners Card*
- *Cheque Guarantee, Other Credit Card*
- *Total_value__Public_Info___CCJ*
- *Time_since_most_recent_Public_In*
- *Total_value__CAIS_8_9s*

This careful variable selection process helps streamline the analysis by focusing only on the most influential and informative variables. By eliminating irrelevant variables, the subsequent analysis becomes more accurate and efficient, ensuring that only the most relevant factors are considered in predicting PPI purchases.

The graphical representation depicting the insignificance of some of these variables is provided below –

1. Full-time part-time employee: The stacked bar chart clearly illustrates that nearly all customers in the dataset are employed on a full-time basis. Consequently, this specific variable holds minimal importance in determining whether a customer will purchase a PPI or not.

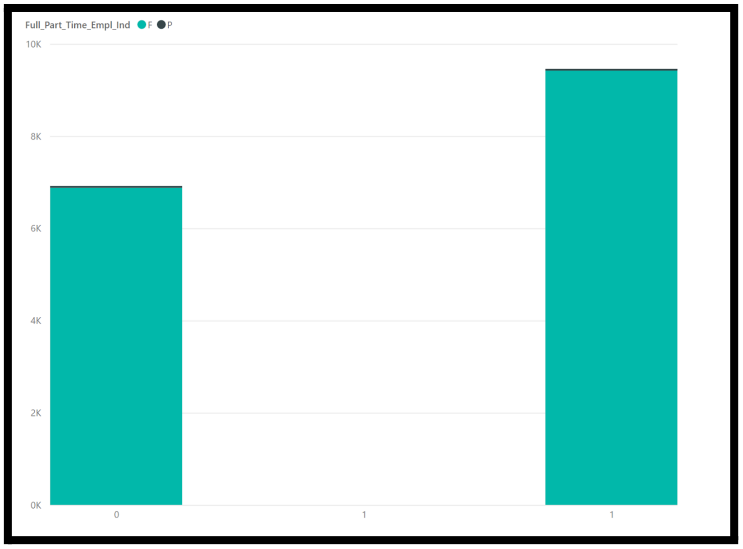


Figure 9: Customer Distribution as per their Employment type (Full-Time or Part-Time)

2. Total_value__Public_Info___CCJ: In a similar vein, it can be observed that a majority of the customers, regardless of whether they possess PPI or not, exhibit a "Total_value__Public_Info___CCJ" value of 0. Consequently, this variable also holds diminished significance when it comes to determining the customers who are likely to purchase the PPI.

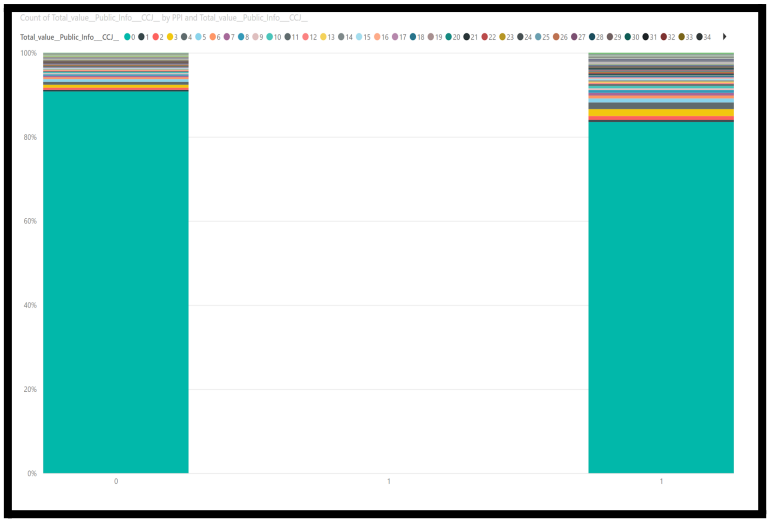


Figure 10: Customer Distribution as per Total_value__Public_Info___CCJ

Payment Method: The payment mode variable also shows a dominant field 'D' with nearly all customers utilizing this payment method. Consequently, this variable holds limited relevance in determining the customers who are likely to purchase PPI.

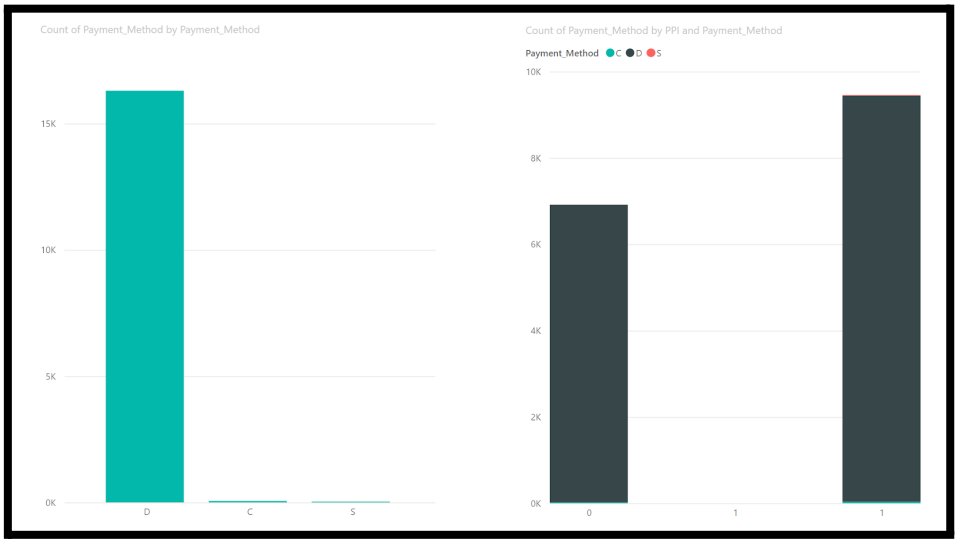


Figure 11: Customer Distribution as per the Payment Method they use.

FEATURE SELECTION USING INFORMATION GAIN

By utilizing the information gain technique, a comprehensive analysis was conducted to determine the most crucial variables. Information gain is a measure used in decision trees to quantify the reduction in entropy or uncertainty when a particular variable is included in the model. The resulting list of the most important variables, identified through information gain analysis, is presented below. These variables were found to have the highest impact on the outcome or target variable (i.e., PPI Purchase) under investigation.

The significance of identifying the most important variables lies in their ability to contribute significantly to the prediction or classification task at hand. By focusing on these key variables, researchers and data scientists can allocate their resources and efforts more efficiently. These variables provide valuable insights into the underlying patterns and relationships that influence the outcome of interest.

Information gain is a measure used in decision tree algorithms to determine the importance of a variable in predicting the outcome. It quantifies the reduction in uncertainty or entropy achieved by incorporating a particular variable into the model. Variables with higher information gain provide more relevant and valuable information for prediction. By selecting variables with high information gain, decision trees can make more accurate and informed splits, leading to better classification or regression results.

		#	Info. gain	Gain ratio	Gini
1	N Value_of_Property		0.023	0.011	0.015
2	N Mosaic		0.020	0.010	0.014
3	N Credit_Score		0.017	0.008	0.011
4	N Total_outstanding_balance__mortg		0.013	0.007	0.009
5	N Term		0.013	0.008	0.009
6	C Residential_Status	4	0.011	0.011	0.008
7	C Final_Grade	10	0.011	0.004	0.007
8	N Income_Range		0.007	0.004	0.005
9	C Employment_Status	9	0.006	0.004	0.004
10	N Age		0.004	0.002	0.003

Figure 12: Top 10 features based on Information Gain

ANALYSIS TO FIND THE CUSTOMER BASE

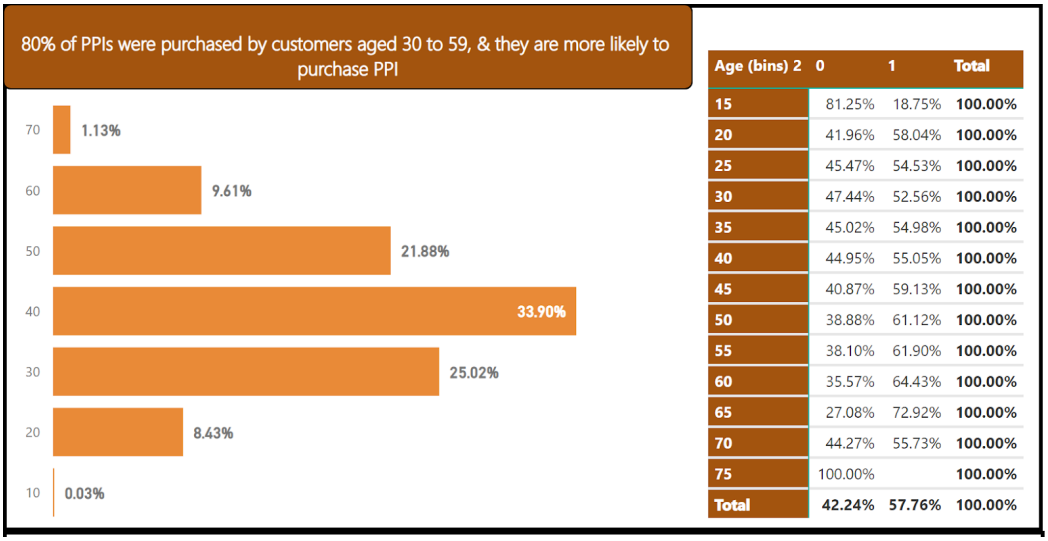
Using EDA:

To identify the target audience among customers who currently do not have PPI, a detailed analysis was conducted on the customers who already have PPI. This analysis involved exploratory data analysis (**EDA**) of the top significant variables as per **Information Gain**, through which the characteristics and attributes of the PPI-holding customers were thoroughly examined. By scrutinizing their demographic information, consumer bank data, and other relevant variables, valuable insights were gained about the characteristics that make individuals more likely to purchase PPI.

The objective of this analysis was to draw conclusions and infer patterns that could aid in identifying the target audience among the non-PPI customers. By understanding the key characteristics exhibited by PPI buyers, it becomes possible to develop a profile or set of criteria for potential customers who are most likely to be interested in obtaining PPI coverage. This information is invaluable for creating targeted marketing strategies and tailored campaigns to attract the attention and interest of the desired target audience, ultimately increasing the chances of converting non-PPI customers into PPI buyers.

The following are the outcomes of the Exploratory Data Analysis –

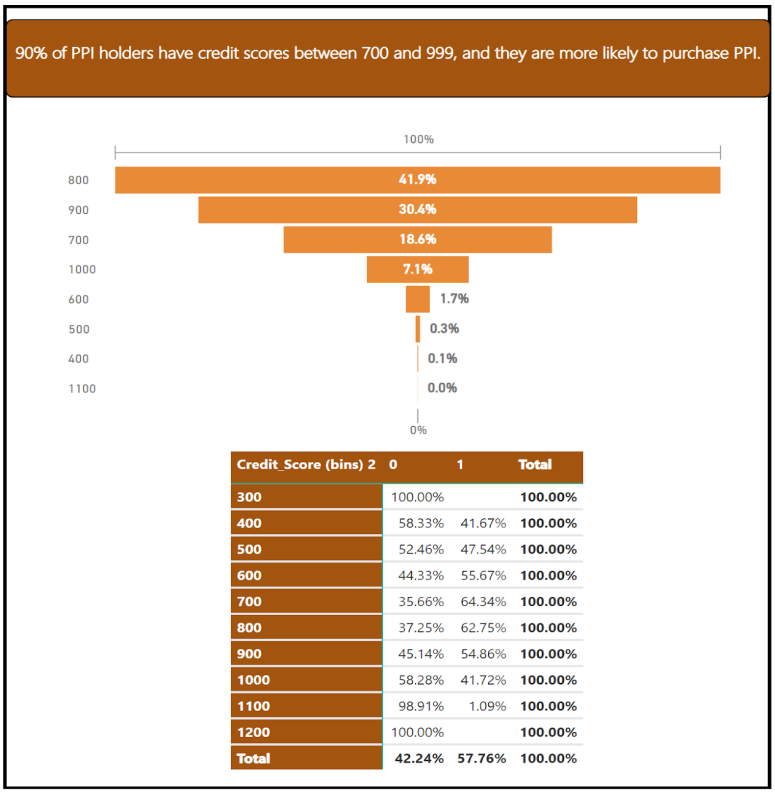
In this exploratory data analysis (EDA) section, every figure consists of a bar chart and a table. The bar chart provides insights into the attributes and traits of customers who have previously purchased PPI. On the other hand, the table offers a comparison between the percentage distribution of customers who have PPI and those who do not (PPI=0). By examining both the chart and the table, we can gain a comprehensive understanding of the differences and similarities between these two customer groups in relation to PPI ownership.



Age (bins) 2	0	1	Total
15	81.25%	18.75%	100.00%
20	41.96%	58.04%	100.00%
25	45.47%	54.53%	100.00%
30	47.44%	52.56%	100.00%
35	45.02%	54.98%	100.00%
40	44.95%	55.05%	100.00%
45	40.87%	59.13%	100.00%
50	38.88%	61.12%	100.00%
55	38.10%	61.90%	100.00%
60	35.57%	64.43%	100.00%
65	27.08%	72.92%	100.00%
70	44.27%	55.73%	100.00%
75	100.00%		100.00%
Total	42.24%	57.76%	100.00%

Figure 13: Age-wise distribution of PPI Holders and their percentage proportion with non-buyers

Figure 14: Credit Score-wise distribution of PPI Holders and their percentage proportion with non-buyers



Credit_Score (bins) 2	0	1	Total
300	100.00%		100.00%
400	58.33%	41.67%	100.00%
500	52.46%	47.54%	100.00%
600	44.33%	55.67%	100.00%
700	35.66%	64.34%	100.00%
800	37.25%	62.75%	100.00%
900	45.14%	54.86%	100.00%
1000	58.28%	41.72%	100.00%
1100	98.91%	1.09%	100.00%
1200	100.00%		100.00%
Total	42.24%	57.76%	100.00%

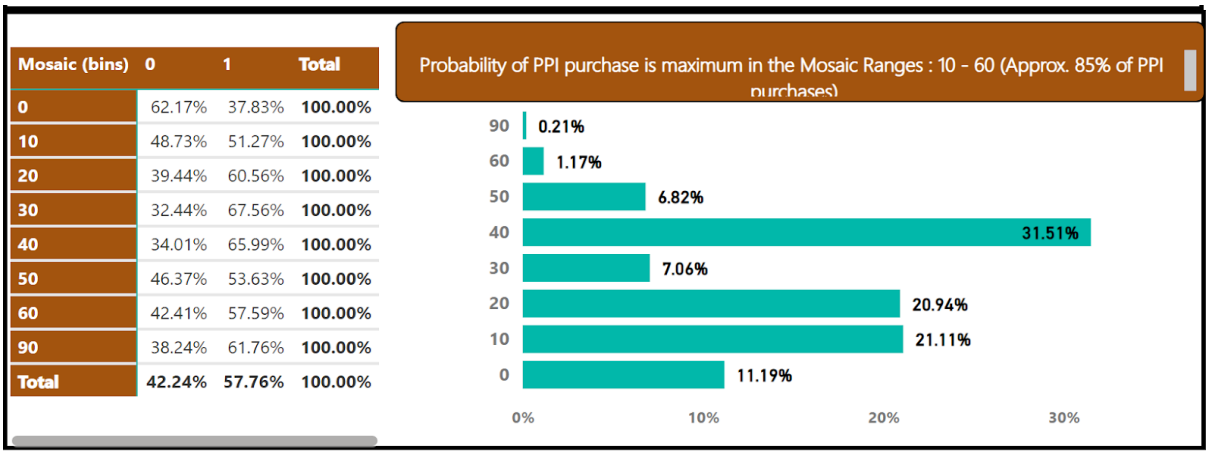
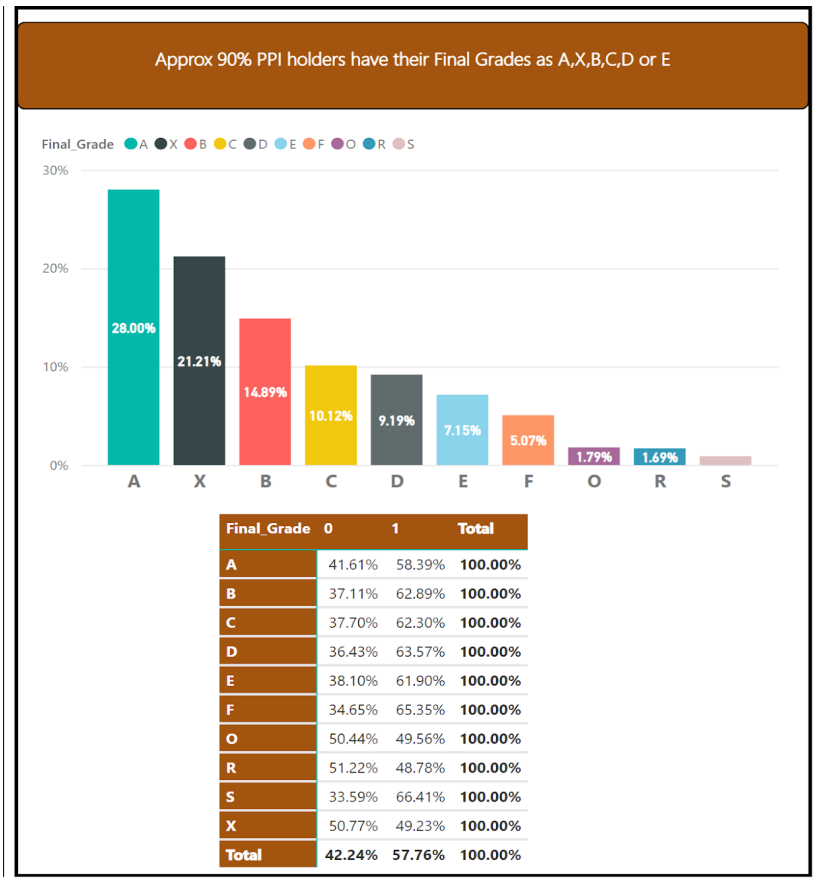


Figure 15: Mosaic-wise distribution of PPI Holders and their percentage proportion with non-buyers



Final_Grade	0	1	Total
A	41.61%	58.39%	100.00%
B	37.11%	62.89%	100.00%
C	37.70%	62.30%	100.00%
D	36.43%	63.57%	100.00%
E	38.10%	61.90%	100.00%
F	34.65%	65.35%	100.00%
O	50.44%	49.56%	100.00%
R	51.22%	48.78%	100.00%
S	33.59%	66.41%	100.00%
X	50.77%	49.23%	100.00%
Total	42.24%	57.76%	100.00%

Figure 16: Grade-wise distribution of PPI Holders and their percentage proportion with non-buyers

Income_Range	0	1	Total
0	83.52%	16.48%	100.00%
1	37.10%	62.90%	100.00%
2	30.97%	69.03%	100.00%
3	32.21%	67.79%	100.00%
4	35.27%	64.73%	100.00%
5	39.12%	60.88%	100.00%
6	48.39%	51.61%	100.00%
Total	42.24%	57.76%	100.00%

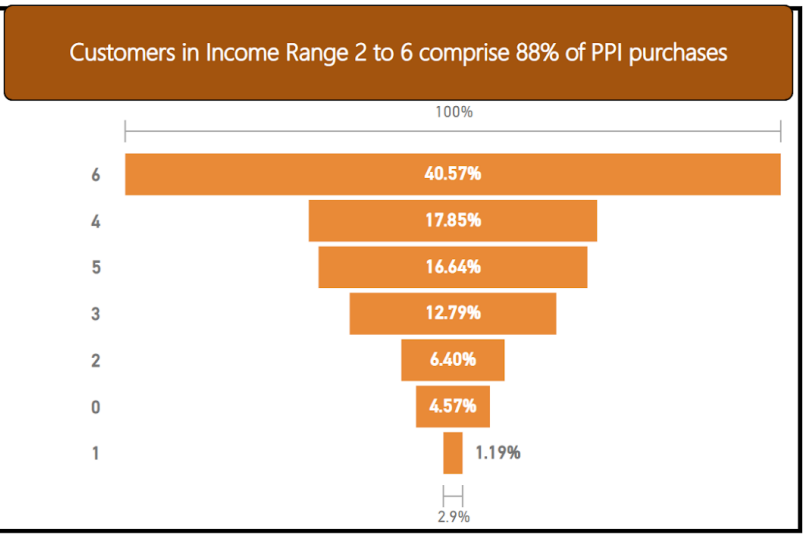
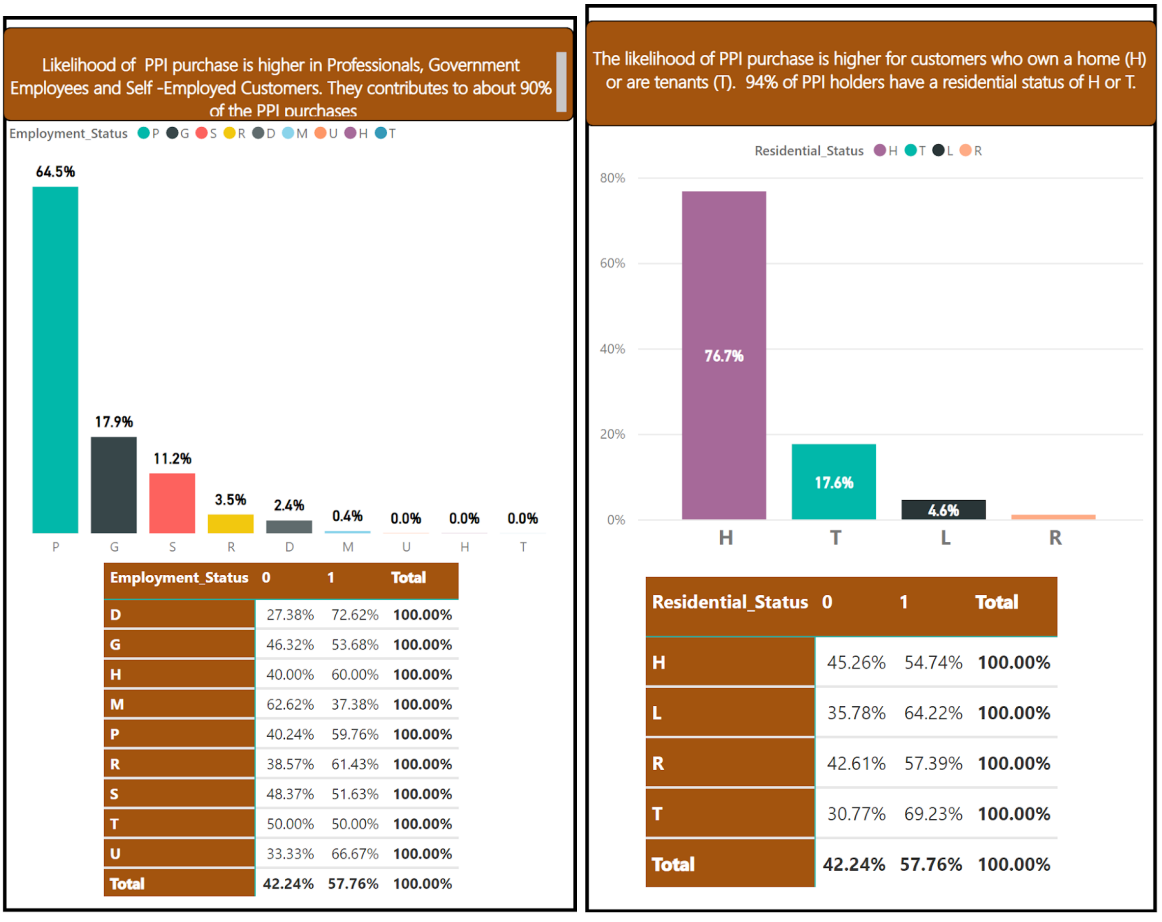


Figure 17: Income-wise distribution of PPI Holders and their percentage proportion with non-buyers

Figure 18: Distribution of PPI Holders based on their Residential & Employment Statuses and their percentage proportion with non-buyers.



Employment_Status	0	1	Total
D	27.38%	72.62%	100.00%
G	46.32%	53.68%	100.00%
H	40.00%	60.00%	100.00%
M	62.62%	37.38%	100.00%
P	40.24%	59.76%	100.00%
R	38.57%	61.43%	100.00%
S	48.37%	51.63%	100.00%
T	50.00%	50.00%	100.00%
U	33.33%	66.67%	100.00%
Total	42.24%	57.76%	100.00%

Residential_Status	0	1	Total
H	45.26%	54.74%	100.00%
L	35.78%	64.22%	100.00%
R	42.61%	57.39%	100.00%
T	30.77%	69.23%	100.00%
Total	42.24%	57.76%	100.00%

Figure 19: Distribution of PPI Holders based on their Marital Status & their percentage proportion with non-buyers.

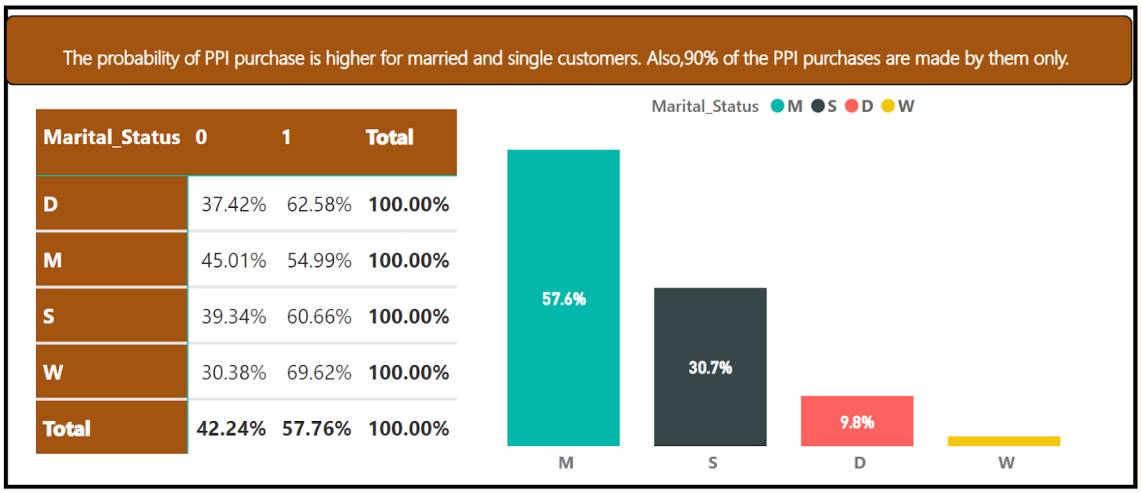
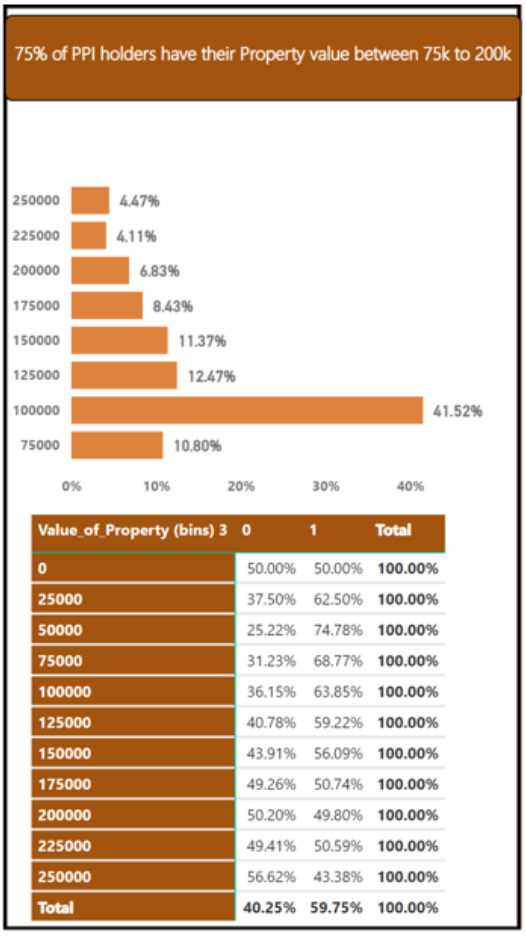
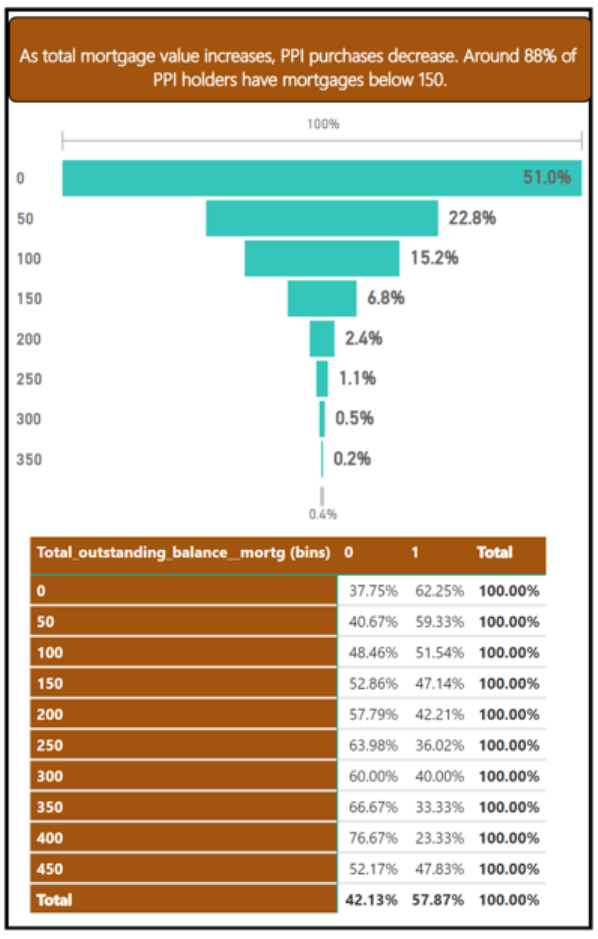


Figure 20: Distribution of PPI Holders based on their Property Value & Outstanding Mortgage Balance and their percentage proportion with non-buyers.

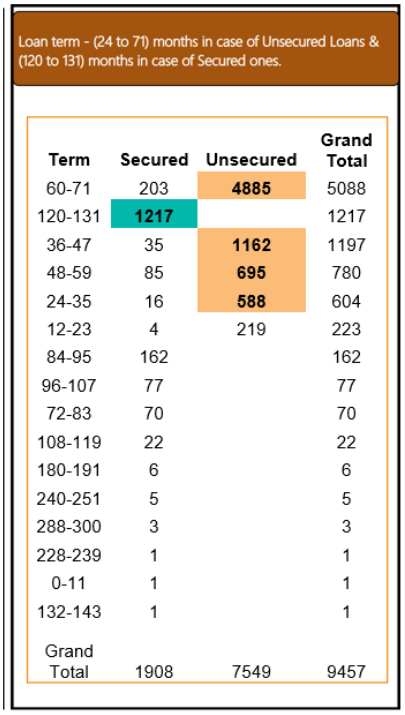


Value_of_Property (bins) 3	0	1	Total
0	50.00%	50.00%	100.00%
25000	37.50%	62.50%	100.00%
50000	25.22%	74.78%	100.00%
75000	31.23%	68.77%	100.00%
100000	36.15%	63.85%	100.00%
125000	40.78%	59.22%	100.00%
150000	43.91%	56.09%	100.00%
175000	49.26%	50.74%	100.00%
200000	50.20%	49.80%	100.00%
225000	49.41%	50.59%	100.00%
250000	56.62%	43.38%	100.00%
Total	40.25%	59.75%	100.00%



Total_outstanding_balance__mortg (bins)	0	1	Total
0	37.75%	62.25%	100.00%
50	40.67%	59.33%	100.00%
100	48.46%	51.54%	100.00%
150	52.86%	47.14%	100.00%
200	57.79%	42.21%	100.00%
250	63.98%	36.02%	100.00%
300	60.00%	40.00%	100.00%
350	66.67%	33.33%	100.00%
400	76.67%	23.33%	100.00%
450	52.17%	47.83%	100.00%
Total	42.13%	57.87%	100.00%

Figure 21: Distribution of PPI Holders based on their Term Value



Term	Secured	Unsecured	Grand Total
60-71	203	**4885**	5088
120-131	**1217**		1217
36-47	35	**1162**	1197
48-59	85	**695**	780
24-35	16	**588**	604
12-23	4	219	223
84-95	162		162
96-107	77		77
72-83	70		70
108-119	22		22
180-191	6		6
240-251	5		5
288-300	3		3
228-239	1		1
0-11	1		1
132-143	1		1
Grand Total	1908	7549	9457

Using Clustering Approach:

This approach aims to enhance our understanding of the customer profile by employing the following steps:

1. Employing Hierarchical clustering on customers with a PPI =1:

To accomplish this, a Hierarchical clustering model was constructed using 10 significant variables. The model yielded six distinct clusters, and the Silhouette scores for all six clusters were determined to be positive. This clustering process enables the grouping of customers based on their similarities, helping us gain valuable insights into their behaviours and preferences.

2. Constructing a Decision Tree using the cluster labels as the target variable:

By utilizing the cluster labels obtained from the Hierarchical clustering, a Decision Tree model was built. This Decision Tree provides us with an explanation of the key factors driving each cluster's characteristics. Furthermore, it offers us a breakdown of the population distribution within each cluster specifically for customers with a PPI value of 1.

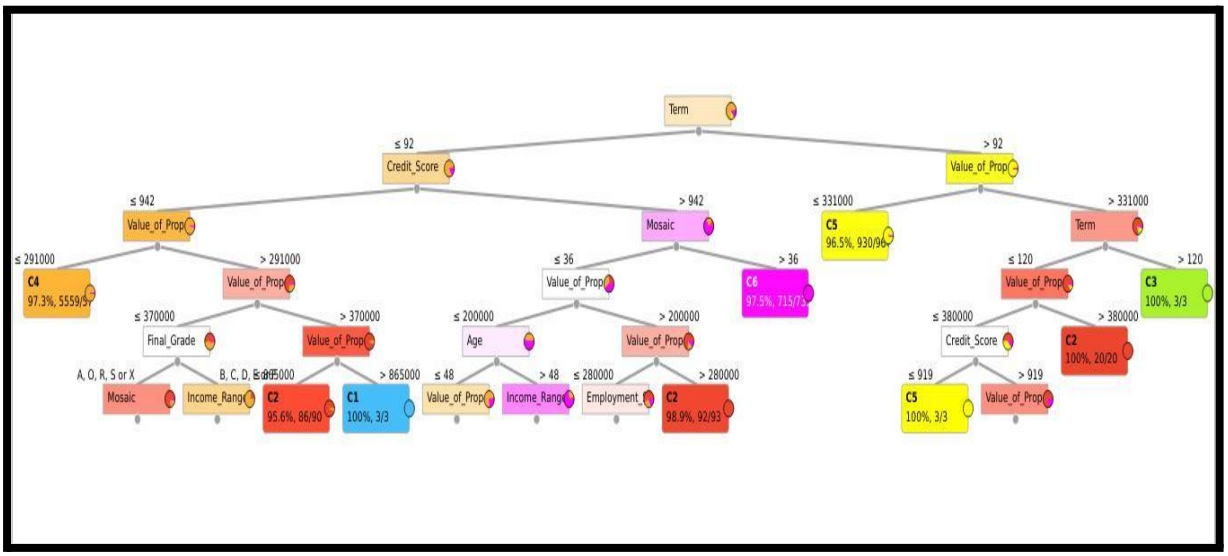


Figure 22: Decision Tree Results

3. Clubbed clusters further for PPI=1 based on the decision tree:

In order to enhance the analysis and gain a deeper understanding of the data, we proceeded to clubbed clusters together based on a decision tree approach, taking into consideration their

common attributes. This process allowed us to create more cohesive and meaningful groups of data points with a PPI (Point of Interest) value of 1.

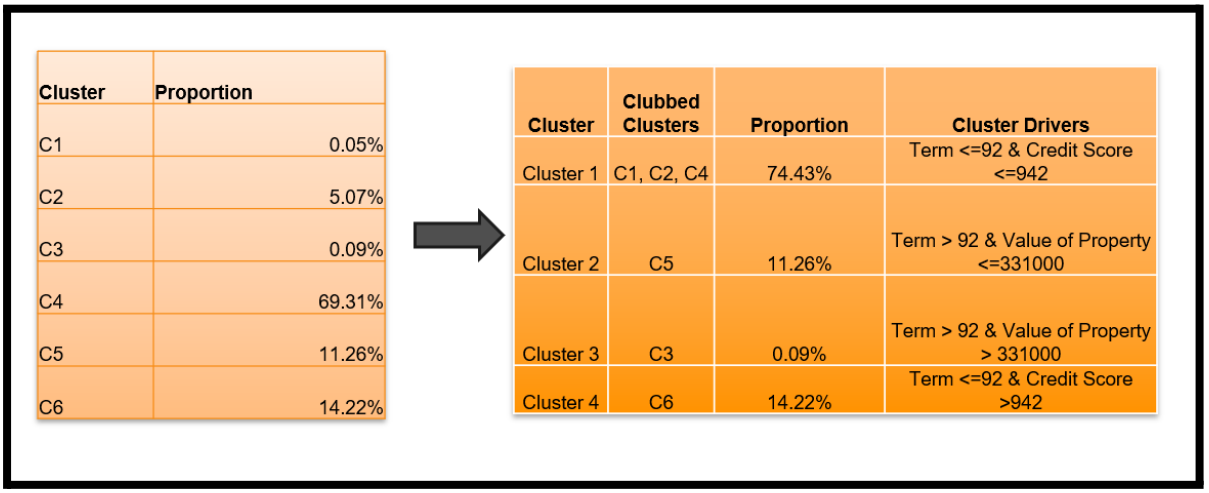
By leveraging the decision tree algorithm, we were able to identify similarities and patterns among the clusters. This involved examining various attributes shared by the clusters and utilizing their collective information to form new consolidated groups.

The decision tree methodology proved to be a valuable technique in this context, as it provided a systematic and logical approach to cluster aggregation. By considering the shared characteristics among the clusters, we could identify relationships and dependencies that might have otherwise been overlooked.

Through this process, we aimed to enhance the interpretability and coherence of the data, allowing for a more comprehensive analysis. By clubbing clusters together based on their common attributes, we could gain insights into the underlying factors that contribute to a PPI value of 1.

This approach not only helped to simplify the data representation but also facilitated the identification of overarching trends and patterns. It allowed us to uncover meaningful relationships within the data, leading to a more nuanced understanding of the factors influencing the PPI values.

Overall, the clubbing of clusters based on the decision tree technique proved to be a valuable step in our analysis, enabling us to create consolidated groups that better captured the essence of the data and enhanced our ability to draw insightful conclusions.



Cluster	Proportion
C1	0.05%
C2	5.07%
C3	0.09%
C4	69.31%
C5	11.26%
C6	14.22%

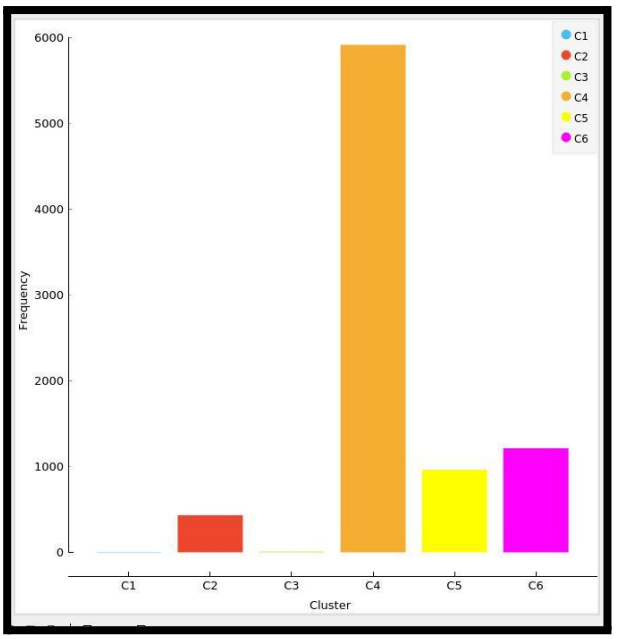
Cluster	Clubbed Clusters	Proportion	Cluster Drivers
Cluster 1	C1, C2, C4	74.43%	Term <=92 & Credit Score <=942
Cluster 2	C5	11.26%	Term > 92 & Value of Property <=331000
Cluster 3	C3	0.09%	Term > 92 & Value of Property > 331000
Cluster 4	C6	14.22%	Term <=92 & Credit Score >942

Figure 23: Clusters clubbed for PPI=1

4. Applying the cluster ML model to predict the cluster assignment of customers with a PPI value of 0: The trained cluster ML model is now utilized to classify each customer with a PPI value of 0 into one of the six previously formed clusters. This step allows us to effectively

predict the target audience for customers who possess a different PPI value. By leveraging the insights gained from the identified clusters, we can tailor our marketing strategies and communication efforts to suit the specific preferences and needs of each target audience, thereby enhancing the overall effectiveness of our campaigns.

PPI =1



PPI =0

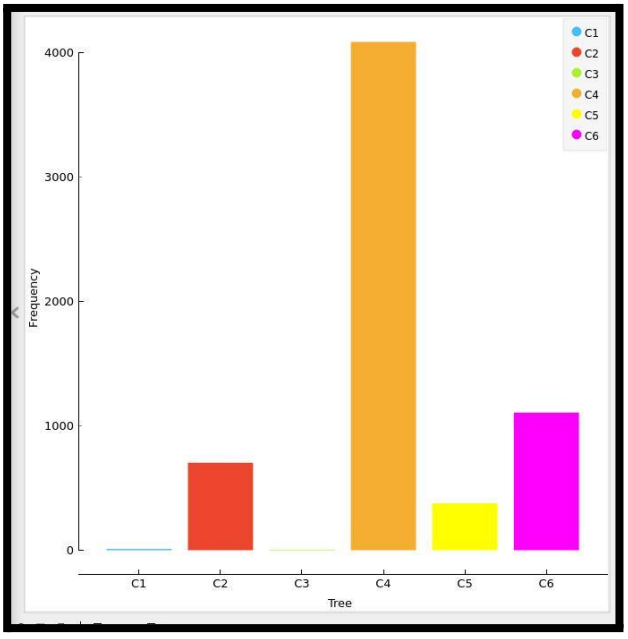
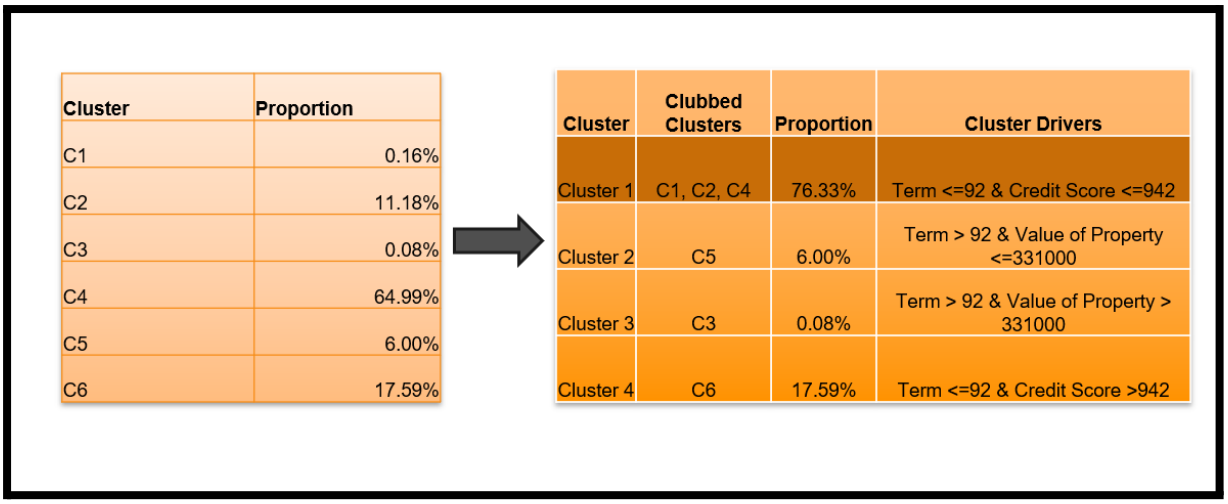


Figure 24: Cluster-wise Population distribution for PPI=1 & PPI=0 after running DT model

5. Clubbing the clusters for PPI = 0 and interpreting the results: The same clubbing was done for the PPI=0 clusters as well. These clusters along with the decision tree predictions is now analysed and interpreted. Since, 74.43% of the PPI holders lie in the Cluster 1 (See PPI=1 case), this cluster is of upmost importance to determine the target audience. Upon running

this same model on customers not having PPI, the customers falling into cluster 1 will be having the attributes which can possibly drive them to buy a PPI. This clubbing and clustering approach also gives information about whom should the bank target first. Their preference order must be CLUSTER 1 > CLUSTER4> CLUSTER2 > CLUSTER3.



Cluster	Proportion
C1	0.16%
C2	11.18%
C3	0.08%
C4	64.99%
C5	6.00%
C6	17.59%

Cluster	Clubbed Clusters	Proportion	Cluster Drivers
Cluster 1	C1, C2, C4	76.33%	Term <=92 & Credit Score <=942
Cluster 2	C5	6.00%	Term > 92 & Value of Property <=331000
Cluster 3	C3	0.08%	Term > 92 & Value of Property > 331000
Cluster 4	C6	17.59%	Term <=92 & Credit Score >942

Figure 25: Clusters clubbed for PPI=1

Recommendations:

By conducting exploratory data analysis (EDA) and utilizing decision tree analysis, we were able to identify the target audience. Through feature selection techniques, we identified the most significant variables, which enabled us to uncover the characteristics of customers who have purchased PPI.

Using this this information, the bank can now determine their potential target audience. While there is no definitive rule that guarantees a person will purchase PPI, the EDA technique allows us to identify common attributes among customers that are highly likely to convert non-buyers into buyers of PPI.

EDA is a manual process that provides an initial understanding of the data. To gain a more precise understanding of the target audience, Clustering and Decision tree approach was used.

It clearly defines the drivers for a customer to purchase the PPI. After running the model on PPI=0 customers, we get well defined clusters to whom we should target. These clusters can be prioritized based on the percentage of customers purchasing the PPI in that cluster.

ANALYSIS TO IDENTIFY THE SUITABLE INSURANCE TYPE TO SELL

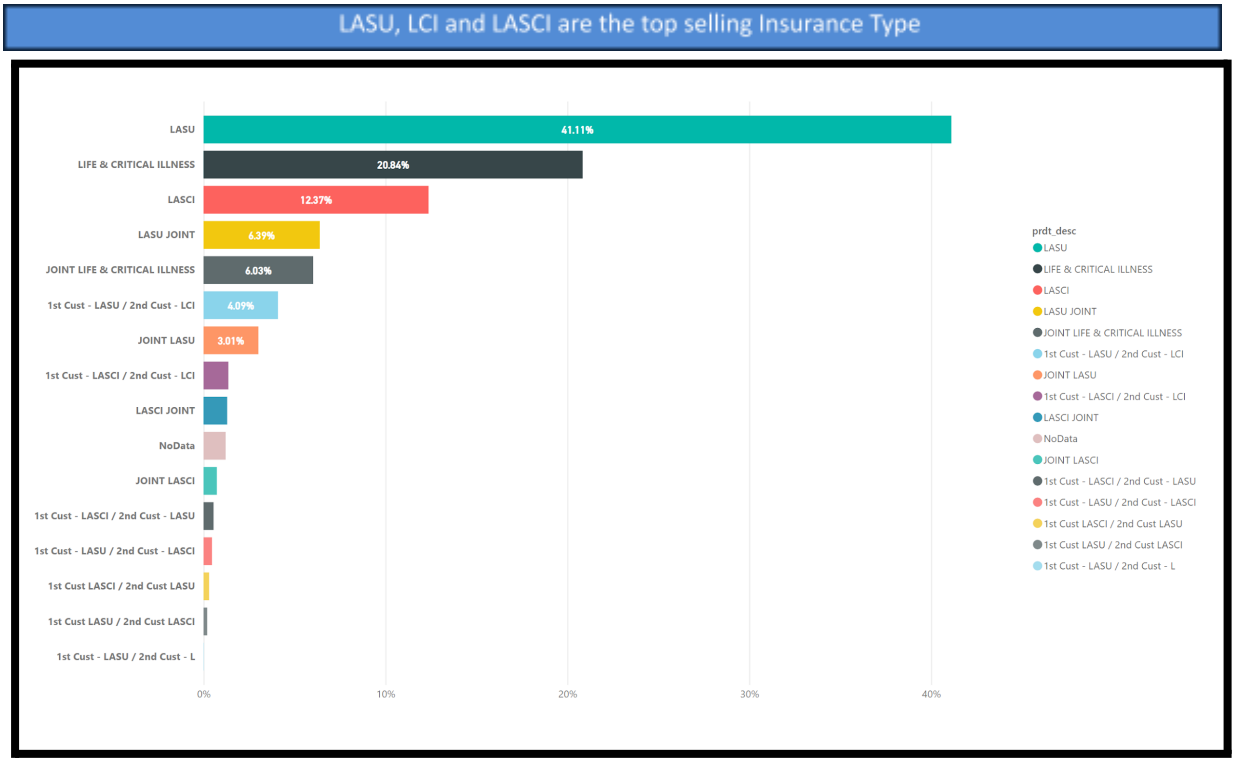


Figure 26: Insurance Product Distribution

Upon analysing the product descriptions of customers who have purchased PPI, it becomes evident that the "LASU" product emerges as the top-selling item. Approximately 42% of the customers have opted to buy this product, indicating its high demand. Following closely behind is the "Life and Critical Illness" product, which has garnered significant popularity, with over 20% of the customers selecting it. Another noteworthy product is "LASCI," which has attracted a considerable customer base, with over 12% of customers opting for it.

When considering the overall picture, these three products collectively contribute to around 75% of the total PPI sales. This implies that the majority of customers who purchase PPI show a strong preference for these specific offerings. Understanding the preferences and trends related to these top-selling products can provide valuable insights for the bank's marketing and sales strategies. By focusing on promoting and tailoring offerings related to "LASU," "Life and Critical Illness," and "LASCI," the bank can potentially maximize its sales and capitalize on the popularity of these products among its customer base.

By EDA:

Having acquired a comprehensive understanding of the attributes and characteristics of our target audience, we conducted exploratory data analysis (EDA) on a subset of significant variables. The objective was to determine the type of product that our target audience is most inclined to purchase. By scrutinizing these variables through EDA, we aimed to identify the specific product that resonates most strongly with our target audience's preferences and needs.

Following is the conclusion of the EDA performed:

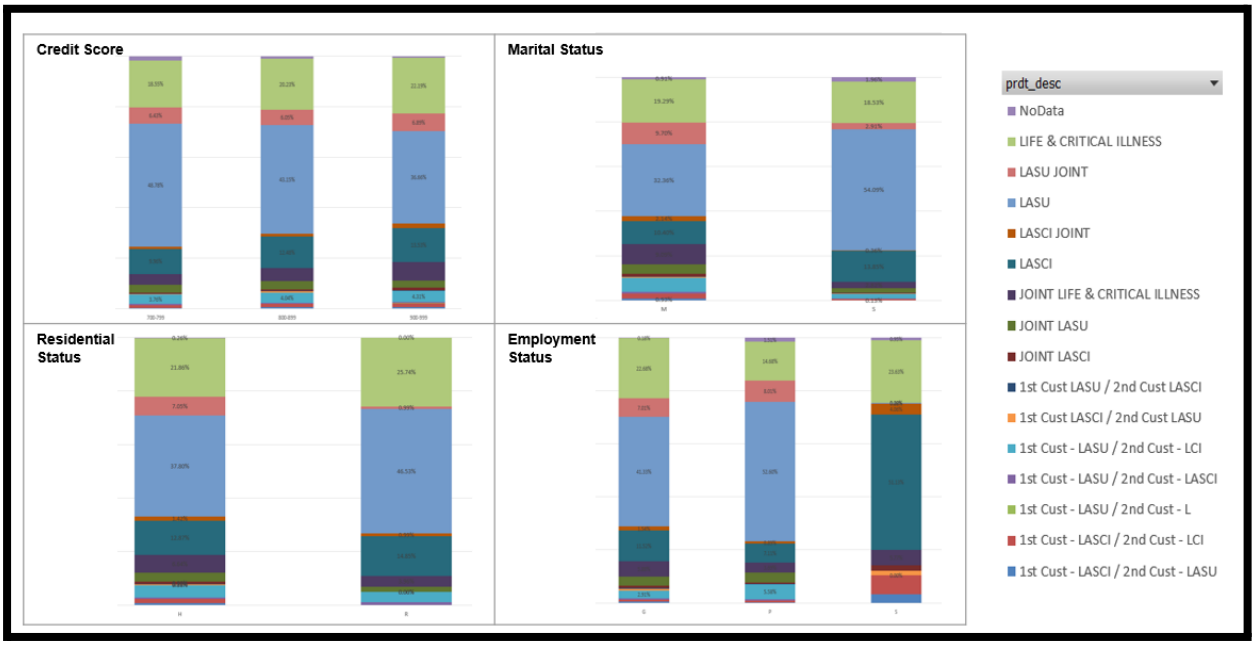


Figure 27: EDA chart-1 to find top selling Insurance Product

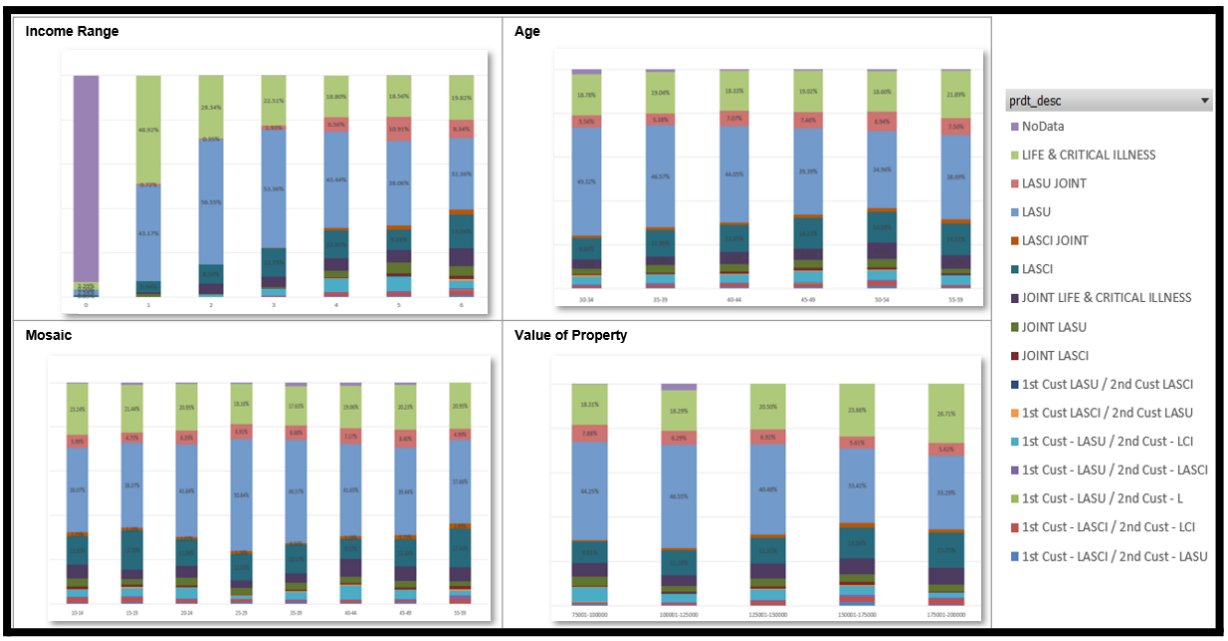


Figure 28: EDA chart-2 to find top selling Insurance Product

Upon conducting exploratory data analysis (EDA) on a subset of significant variables related to our target audience, it was reaffirmed that the product "LASU" maintains its position as the top-selling choice across all categories. Following "LASU," "LCI" consistently emerges as the second most popular product in most cases. It is worth noting that "LASCI" closely trails behind "LCI" and, in fact, in certain instances, its sales are nearly equal to those of "LCI."

This consistent pattern observed in the EDA findings further emphasizes the prominence of "LASU" as the preferred product among our target audience. It signifies the high demand and appeal of "LASU" in capturing their attention and meeting their needs. However, the noteworthy performance of "LCI" and "LASCI" suggests that these products hold significant market potential as well, with a substantial customer base showing interest in them.

By Random Forest:

In order to identify the top selling insurance products, we employed the Random Forest algorithm. The model was trained on a dataset that included customers who have purchased PPI, and subsequently tested on a separate dataset where the PPI value was equal to 0.

Model Accuracy: 68.2%

After running the Random Forest model, we obtained the following results for the top selling insurance products:

1) LASU - This product emerged as the clear frontrunner, accounting for a substantial 52.73% of the total sales.

2) LCI - With a significant share of 15.34% in the sales, LCI secures the second spot in the list of top selling insurance products.

3) LASCI - Not far behind LCI, LASCI captured a notable market share of 13.86%. Its competitive performance further emphasizes its significance and potential as an insurance product.

These findings, obtained from the Random Forest model, provide valuable insights into the preferences and purchasing behaviour of customers. By leveraging the identified top selling insurance products, the bank can strategically focus its marketing efforts, optimize sales strategies, and cater to the specific needs and interests of its target audience.

Moreover, it is crucial to note that the model achieved an accuracy rate of 68.2%, indicating its effectiveness in predicting customer preferences and determining the top selling insurance products.

What is a Random Forest Model?

The Random Forest model is a powerful machine learning algorithm used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to make predictions. Each decision tree in the Random Forest is constructed using a different subset of the training data and a random subset of the input features.

The model works by generating a multitude of decision trees, where each tree independently makes predictions based on a subset of features. During the training process, the trees are grown by splitting the data into various subsets using different criteria, such as Gini impurity or information gain. The final prediction of the Random Forest model is determined through a majority vote or averaging of the predictions made by individual trees.

One of the key advantages of the Random Forest model is its ability to handle high-dimensional datasets with a large number of input features. It can also handle missing data and outliers without significant loss of performance. Additionally, Random Forest models are robust against overfitting, as the aggregation of multiple trees helps to reduce variance and increase generalization.

The Random Forest algorithm provides several important benefits, including feature importance ranking, as it measures the contribution of each feature in the prediction process. It can also handle both categorical and continuous input variables without the need for extensive data pre-processing. Furthermore, it can handle imbalanced datasets by using techniques such as balanced subsampling.

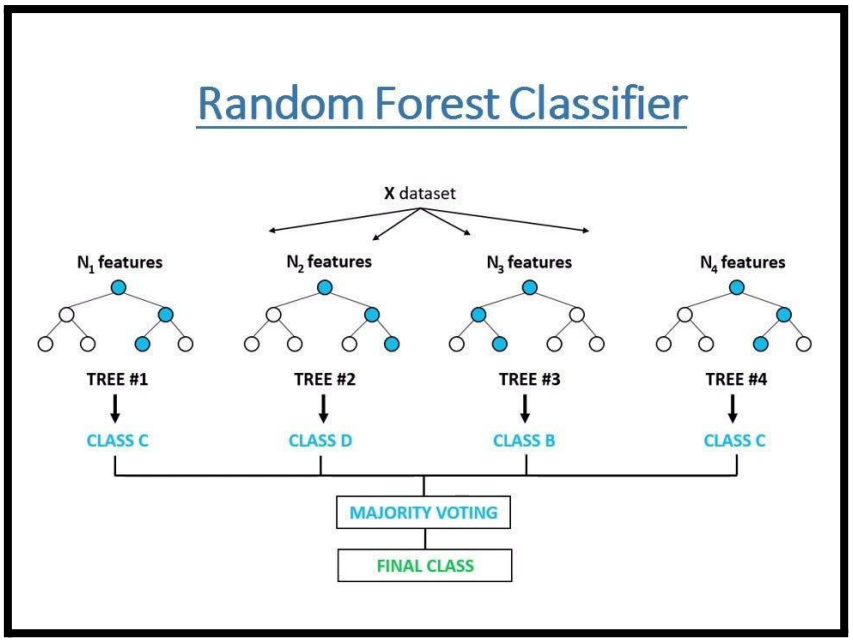


Figure 29: Random Forest Classifier

In summary, the Random Forest model is a versatile and powerful algorithm that leverages the collective decision-making of multiple trees to provide accurate predictions. Its robustness, scalability, and ability to handle complex datasets make it a popular choice for various machine learning tasks.

Recommendations:

Both exploratory data analysis (EDA) and the Random Forest algorithm reinforce a consistent finding: LASU is the leading product in terms of sales, closely followed by LCI and LASCI. These results indicate that the bank should prioritize its efforts and resources towards these three insurance products.

PPI CROSS SELL STRATEGY

Strategy 1: Provide Incentives-

One approach to implementing a cross-sell strategy for Payment Protection Insurance (PPI) is through incentives. Banks have some flexibility in adjusting loan interest rates or Annual Percentage Rates (APRs) without incurring losses, as there is usually a margin available to them. In this case, the bank can offer a concession in APR to their existing loan customers who haven't purchased the PPI product yet. This incentive is specifically targeted towards non-buyers of PPI.

By reducing the APR, the bank may incur a minimal loss, but this can be easily offset by the purchase of the PPI product. Thus, it becomes a profitable sale for the bank rather than a loss-bearing one.

It's important to consider that interest rates or APRs have long-term implications. Analysing the dataset reveals that customers without PPI have an average loan term of 59.15 months (approximately 5 years), with the majority (around 49% of customers) having loan terms ranging from 60 to 71 months (5 to 6 years). Therefore, the bank would typically receive the total interest on their loan at the end of this 5–6-year period.

However, by offering a slight concession in APR that leads to the purchase of a PPI product, the bank can obtain the amount for the PPI product immediately. This provides the bank with the opportunity to reinvest this money and potentially multiply it. Customers will also be pleased to pay a reduced amount of interest on their loan, and in case of any unfortunate events, they will have the added security of the PPI coverage.

Hence, this strategy brings the additional benefit of increasing liquidity for the bank, resulting in a win-win situation for both the customer and the bank.

Strategy 2: Corporate-Tie Ups-

Typically, banks establish partnerships with various organizations. However, if a bank lacks such partnerships, this strategy proposes that the bank should pursue tie-ups with a select group of companies. These companies can be categorized into tiers and based on the bank's customer base and their interests; the bank can approach these companies to form corporate tie-ups.

Through this corporate tie-up, both parties involved can benefit. The bank can offer additional concessions in loan Annual Percentage Rates (APRs) or provide other advantages exclusively to employees of the partnered companies who are willing to purchase a Payment Protection Insurance (PPI) product from the bank. Furthermore, the option to purchase the PPI product can also be extended to the company itself, allowing them to include the cost of the PPI in the employee's Cost-to-Company (CTC) package, provided the customer willingly agrees to it.

One compelling reason to prioritize corporate tie-ups is the significant potential within the customer base without PPI. Currently, 58% of these customers are professionals or individuals employed in private firms. Therefore, the opportunity for expanding PPI coverage in this particular segment is substantial. Refer to the table below showing the employment status-wise distribution.

Employment Statuses	Percentage Distribution of customers
D	1.29%
G	21.01%
H	0.03%
M	0.97%
P	58.83%
R	3.51%
S	14.33%
T	0.01%
U	0.03%
Grand Total	**100.00%**

Table 2: Percentage Distribution of the Customers based on Employment Status

This strategy creates a win-win scenario for all parties involved. Employees of the partnered companies gain the advantage of receiving special concessions or benefits on their loans if they opt for the PPI product. Simultaneously, the bank benefits from increased PPI product sales and the potential for long-term customer relationships through these corporate tie-ups. Additionally, companies benefit by offering their employees an added benefit of financial protection through the PPI product, enhancing employee satisfaction and loyalty.

By establishing these corporate tie-ups and incentivizing PPI product purchases, the bank not only expands its customer base but also strengthens its relationships with partner companies. This collaborative approach can lead to mutual growth and create a positive brand image for the bank as a reliable and supportive financial institution.

Strategy 3: Government Promotion of PPIs-

Insurance serves two vital purposes in society: risk protection and the cultivation of behaviours that mitigate risk. The insurance industry relies significantly on government involvement, as governments formulate policies and regulations that govern the insurance market. It is well-known that governments worldwide actively promote various types of insurance, such as health insurance and vehicle insurance.

As part of an innovative initiative, a bank or a group of banks can approach the government with a proposal to promote Payment Protection Insurance (PPI) on a larger scale. While pursuing government support may present challenges, given that governments already promote other types of insurance, adding one more form of insurance should not be overly difficult for them.

Upon gaining government agreement, the government can implement certain tax relaxations for individuals who have purchased PPI. This incentive is expected to significantly boost PPI sales nationwide. Assuming that the bank's existing customers would prefer to buy insurance from their own bank, this strategy opens up a vast market for PPI products, as 42% of the bank's customers currently do not possess PPI coverage.

Government promotion of such insurance products will enhance their image as an entity genuinely concerned about citizens' well-being, particularly in situations where they may be unable to repay their loans. By actively supporting PPI, the government demonstrates its commitment to safeguarding the financial security of individuals and reinforcing responsible borrowing behaviour.

This strategy not only benefits the bank by expanding its PPI market share and increasing sales but also contributes to the government's objectives of promoting financial stability and protecting citizens from unforeseen circumstances. The partnership between the bank and the government in promoting PPI reinforces a collaborative approach towards enhancing the overall welfare of the society.