4/11/2024

Life Insurance Sales – Capstone Report

PGP-DSBA

Karthick Raj S

Table of Contents

List of Tables

List of Figures

Abstract

The Objective of the problem statement is to predict the Agents Bonus. By predicting the agent's bonus, the Life Insurance Company can implement engagement and upskilling programs for its agents. The secondary objective is to get an optimum bonus for each employee so that will reduce the employee compensation for the company. The Dataset contains the Life Insurance Policy Details and the last month's Bonus Amount given to the agents. It's a monthly data. The Data Cleaning is done in four parts i.e., Treatment for Spelling Mistakes, Special Characters & other, Missing Values Treatment, Outlier Treatment and Removal of Unwanted Variables. Cluster Analysis is performed as a part of EDA. The Customer segmentations is providing valuable insights about the customers to understand the existing customers. To Predict the agent's bonus, 14 models are built, tuned and validated. The most optimum model is stacking regressor with 6 base model and linear regression as final predictor. For ease use of the prediction, CART-Bagging without any base models can be used for the prediction as it has nearly the R-Square value with stacking regressor. The Future scope of the study can be done on finding the most important agents that are required for the company development if the agent's details are also provided.

Introduction

Defining problem statement

Problem Statement:

Life Insurance Sales

"The dataset belongs to a leading life insurance company. The company wants to predict the *bonus for its agents so that it may design appropriate engagement activity for their high performing agents and upskill programs for low performing agents."*

The Objective of the above problem statement is to predict the monthly bonus of the insurance company to design engagement & upskilling programs for their agents.

Need of the project

By predicting the bonus for its agents, the insurance company can classify the agents into different bonus level category such as High, Medium & Low.

For High Category, it can set some rewards to get even higher targets for insurance sales.

The agents in the Medium & Low category can be given training to move to the next category.

For Designing & Implementing this Engagement Strategy, The Bonus amount is important that makes prediction of bonus amount as great significance.

Understanding business

The Company is selling Life Insurance Financial Products to its customers through agents mostly. The Agents sells the product to the customers and the amount paid by the customers is the main source of income for the company. If the customer's policy gets matured or the customer's claims the policy for some other valid reasons on the policy, the company will be paying back the amount as discussed in the policy document. The Company will have many ventures & investments also while they have the customer's money on their hands.

Data Report

Understanding how data was collected in terms of time, frequency and methodology

The Dataset contains the Life Insurance Policy Details and the last month's Bonus Amount given to the agents.

The Dataset is collected at the end of each month. It is a monthly data.

Visual inspection of data

The Data has 4520 Rows and 20 Columns.

The Data has details of 4520 Insurance Policies.

The Top Five Rows of Data is

*View has been changed to fit the Report**

	0	1	2	3	4
CustID	7000000	7000001	7000002	7000003	7000004
AgentBonus	4409	2214	4273	1791	2955
Age	22	11	26	11	6
CustTenure	4	2	4		
Channel	Agent	Third Party Partner	Agent	Third Party Partner	Agent
Occupation	Salaried	Salaried	Free Lancer	Salaried	Small Business
EducationField	Graduate	Graduate	Post Graduate	Graduate	UG
Gender	Female	Male	Male	Fe male	Male
ExistingProdType	3	4	4	3	3
Designation	Manager	Manager	Exe	Executive	Executive
NumberOfPolicy	2	4	3	3	4
MaritalStatus	Single	Divorced	Unmarried	Divorced	Divorced
MonthlyIncome	20993	20130	17090	17909	18468
Complaint	1	0	1	1	0
ExistingPolicyTenure	2	3	2	2	4
SumAssured	806761	294502		268635	366405
Zone	North	North	North	West	West
PaymentMethod	Half Yearly	Yearly	Yearly	Half Yearly	Half Yearly
LastMonthCalls	5	7	0	0	2
CustCareScore	2	3	3	5	5

Table 1 Top Five rows

Descriptive Statistics

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
AgentBonus	4520.00	NaN	NaN	NaN	4077.84	1403.32	1605.00	3027.75	3911.50	4867.25	9608.00
Age	4251.00	NaN	NaN	NaN	14.49	9.04	2.00	7.00	13.00	20.00	58.00
CustTenure	4294.00	NaN	NaN	NaN	14.47	8.96	2.00	7.00	13.00	20.00	57.00
Channel	4520	3	Agent	3194	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Occupation	4520	5	Salaried	2192	NaN	NaN	NaN	NaN	NaN	NaN	NaN
EducationField	4520	7	Graduate	1870	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	4520	3	Male	2688	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ExistingProdType	4520.00	NaN	NaN	NaN	3.69	1.02	1.00	3.00	4.00	4.00	6.00
Designation	4520	6	Manager	1620	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NumberOfPolicy	4475.00	NaN	NaN	NaN	3.57	1.46	1.00	2.00	4.00	5.00	6.00
MaritalStatus	4520	4	Married	2268	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthlyIncome	4284.00	NaN	NaN	NaN	22890.31	4885.60	16009.00	19683.50	21606.00	24725.00	38456.00
Complaint	4520.00	NaN	NaN	NaN	0.29	0.45	0.00	0.00	0.00	1.00	1.00
ExistingPolicyTenure	4336.00	NaN	NaN	NaN	4.13	3.35	1.00	2.00	3.00	6.00	25.00
SumAssured	4366.00	NaN	NaN	NaN	619999.70	246234.82	168536.00	439443.25	578976.50	758236.00	1838496.00
Zone	4520	4	West	2566	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PaymentMethod	4520	4	Half Yearly	2656	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LastMonthCalls	4520.00	NaN	NaN	NaN	4.63	3.62	0.00	2.00	3.00	8.00	18.00
CustCareScore	4468.00	NaN	NaN	NaN	3.07	1.38	1.00	2.00	3.00	4.00	5.00

Table 2 Descriptive Statistics

The Minimum age of the Customer can’t be 2.

There are also missing values in the data based on the count of the statistics.

The Data has to be cleaned before EDA.

Data Description

Variable	Description
CustID	Unique customer ID
AgentBonus	Bonus amount given to each agents in last month
Age	Age of customer
CustTenure	Tenure of customer in organization
Channel	Channel through which acquisition of customer is done
Occupation	Occupation of customer
EducationField	Field of education of customer
Gender	Gender of customer
ExistingProdType	Existing product type of customer
Designation	Designation of customer in their organization
NumberOfPolicy	Total number of existing policy of a customer
MaritalStatus	Marital status of customer
MonthlyIncome	Gross monthly income of customer
Complaint	Indicator of complaint registered in last one month by customer
ExistingPolicyTenure	Max tenure in all existing policies of customer
SumAssured	Max of sum assured in all existing policies of customer
Zone	Customer belongs to which zone in India. Like East, West, North and South
PaymentMethod	Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly
LastMonthCalls	Total calls attempted by company to a customer for cross sell
CustCareScore	Customer satisfaction score given by customer in previous service call

Table 3 Data Description

Understanding of attributes

The Info of the data is

```
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CustID                4520 non-null   int64
 1   AgentBonus            4520 non-null   int64
 2   Age                   4251 non-null   float64
 3   CustTenure            4294 non-null   float64
 4   Channel               4520 non-null   object
 5   Occupation            4520 non-null   object
 6   EducationField        4520 non-null   object
 7   Gender                4520 non-null   object
 8   ExistingProdType      4520 non-null   int64
 9   Designation           4520 non-null   object
 10  NumberOfPolicy        4475 non-null   float64
 11  MaritalStatus         4520 non-null   object
 12  MonthlyIncome         4284 non-null   float64
 13  Complaint             4520 non-null   int64
 14  ExistingPolicyTenure  4336 non-null   float64
 15  SumAssured            4366 non-null   float64
 16  Zone                  4520 non-null   object
 17  PaymentMethod         4520 non-null   object
 18  LastMonthCalls        4520 non-null   int64
 19  CustCareScore         4468 non-null   float64
dtypes: float64(7), int64(5), object(8)
```

Table 4 Data info

There are 12 Numerical Variables & 8 Object variables.

Based on the Minimum, Maximum and Quartiles, three variables have to be categorical instead of float/int(numerical).

The three variables are

- CustCareScore
- ExistingProdType
- Complaint

Exploratory data analysis and Business Implication

Data Cleaning and Pre-Processing is done as a part of EDA.

Data Cleaning and Pre-processing

The Data Cleaning is done in four parts

- ✓ Treatment for Spelling Mistakes, Special Characters and other
- ✓ Missing Values Treatment
- ✓ Outlier Treatment
- ✓ Removal of Unwanted Variables

Data Cleaning for Special Characters, Spelling Mistakes & Others

There is no special character in the objects columns.

There are some Category that has to be cleaned in the object columns for spelling mistakes and wrong category classification.

The Category in object variables are

```
['Agent' 'Third Party Partner' 'Online']
**********************************************************************************************************
['Salaried' 'Free Lancer' 'Small Business' 'Laarge Business'
 'Large Business']
**********************************************************************************************************
['Graduate' 'Post Graduate' 'UG' 'Under Graduate' 'Engineer' 'Diploma'
 'MBA']
**********************************************************************************************************
['Female' 'Male' 'Fe male']
**********************************************************************************************************
['Manager' 'Exe' 'Executive' 'VP' 'AVP' 'Senior Manager']
**********************************************************************************************************
['Single' 'Divorced' 'Unmarried' 'Married']
**********************************************************************************************************
['North' 'West' 'East' 'South']
**********************************************************************************************************
['Half Yearly' 'Yearly' 'Quarterly' 'Monthly']
**********************************************************************************************************
```

1. In Occupation - Laarge Business has to be cleaned and changed to Large Business.
2. In Gender - Fe male has to be cleaned and changed to Female.
3. In Designation - Exe & Executive can be combined as Single category as 'Executive'.
4. In Marital Status - Single and Unmarried can be combined into 'Unmarried' Category.
5. In Educationfield
   - Graduate, Under Graduate & UG can be combined as Single category - 'UG'.
   - Engineer can be considered as UG, it is not a high-level education field category. Only 9%*(Refer the table below*)* are Engineers. If the Engineers have done master's degree, they would have selected Post Graduate or MBA Categories.

- Post Graduate & MBA can be combined as Single category - 'PG'.

Table 5 EducationField %

EducationField	%
Graduate	41.37
Under Graduate	26.33
Diploma	10.97
Engineer	9.03
Post Graduate	5.58
UG	5.09
MBA	1.64

The Cleaned Category of the object variables are

```
CHANNEL
['Agent', 'Third Party Partner', 'Online']
**********************************************************************************************************
OCCUPATION
['Salaried', 'Small Business', 'Large Business', 'Free Lancer']
**********************************************************************************************************
EDUCATIONFIELD
['UG', 'Diploma', 'PG']
**********************************************************************************************************
GENDER
['Male', 'Female']
**********************************************************************************************************
DESIGNATION
['Executive', 'Manager', 'Senior Manager', 'AVP', 'VP']
**********************************************************************************************************
MARITALSTATUS
['Married', 'Unmarried', 'Divorced']
**********************************************************************************************************
ZONE
['West', 'North', 'East', 'South']
**********************************************************************************************************
PAYMENTMETHOD
['Half Yearly', 'Yearly', 'Monthly', 'Quarterly']
**********************************************************************************************************
```

There are No Duplicates in the data. CustID is used for checking the duplicates with dataset.

The CustID is dropped as it doesn’t have any significance in the prediction model.

All the categorical Variables are changed to Category Datatype.

The New Info of data is

```
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   AgentBonus            4520 non-null   int64
 1   Age                   4251 non-null   float64
 2   CustTenure            4294 non-null   float64
 3   Channel               4520 non-null   category
 4   Occupation            4520 non-null   category
 5   EducationField        4520 non-null   category
 6   Gender                4520 non-null   category
 7   ExistingProdType      4520 non-null   category
 8   Designation           4520 non-null   category
 9   NumberOfPolicy        4475 non-null   float64
 10  MaritalStatus         4520 non-null   category
 11  MonthlyIncome         4284 non-null   float64
 12  Complaint             4520 non-null   category
 13  ExistingPolicyTenure  4336 non-null   float64
 14  SumAssured            4366 non-null   float64
 15  Zone                  4520 non-null   category
 16  PaymentMethod         4520 non-null   category
 17  LastMonthCalls        4520 non-null   int64
 18  CustCareScore         4468 non-null   category
dtypes: category(11), float64(6), int64(2)
```

Table 6 New Data Info

Missing Value treatment

There is 1.35% of missing data to the total dataset.

The Missing Value % for variables is

Variable Name	%
Age	5.95
CustTenure	5
NumberOfPolicy	1
MonthlyIncome	5.22
SumAssured	3.41
ExistingPolicyTenure	4.07
CustCareScore	1.15

Table 7 Missing Value Table

(All other variables have no missing values)*

The Missing Value Plot shows the missing values in the dataset.

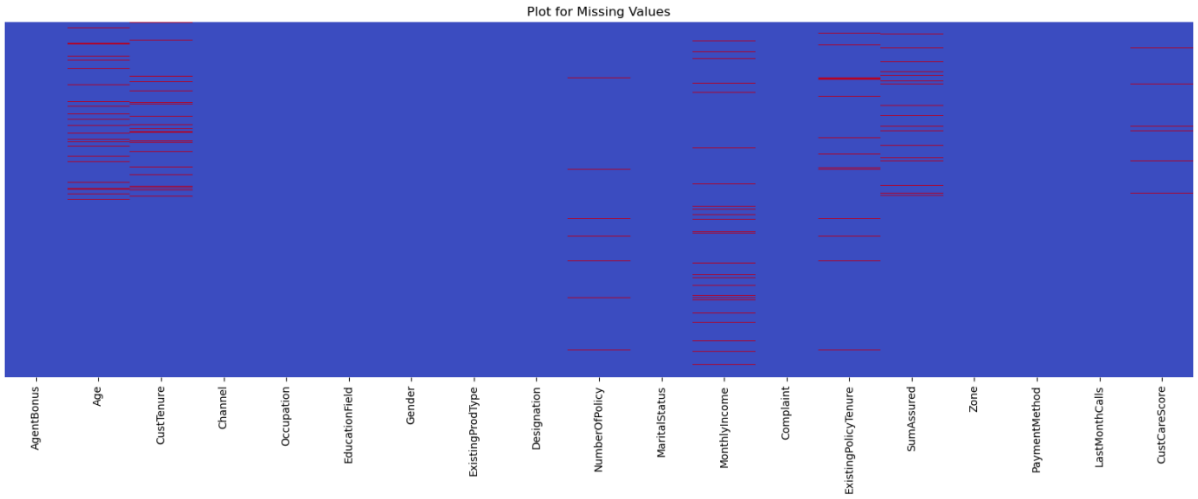


Figure A Missing Value Plot

Only few data are missing.

Categorical Variable Treatment:

CustCareScore : 52 Missing Values

The Null values are imputed with mode for the categorical variable.

The Mode for CustCareScore is 3.0.

Numerical Variable Treatment:

The count for Numerical variables missing values is

```
Age : 269
CustTenure : 226
NumberOfPolicy : 45
MonthlyIncome : 236
ExistingPolicyTenure : 184
SumAssured : 154
```

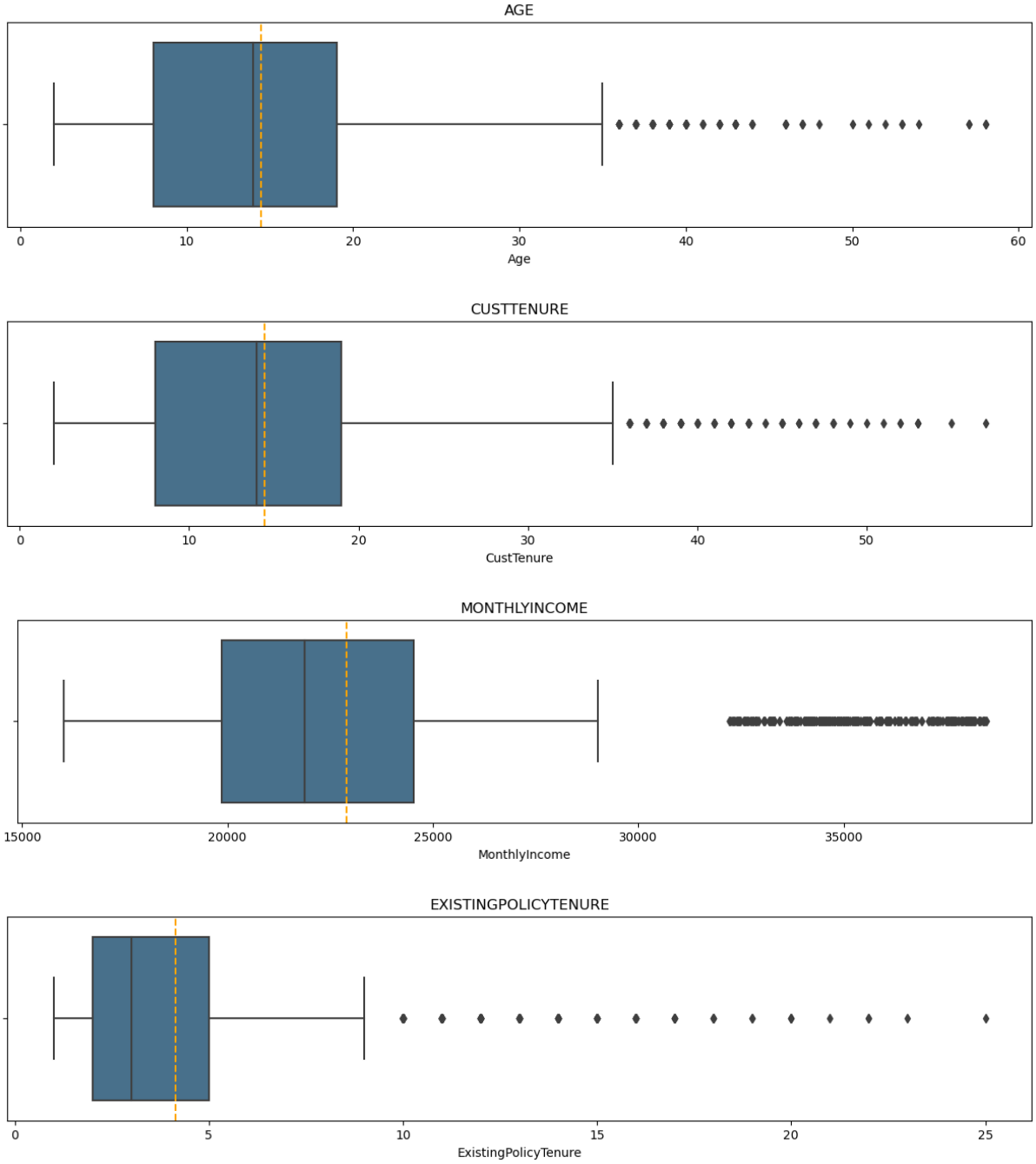
The Null values are imputed with mean for the numerical variable.

The Mean values are

```
Age : 14.494707127734651
CustTenure : 14.469026548672566
NumberOfPolicy : 3.56536312849162
MonthlyIncome : 22890.309990662932
ExistingPolicyTenure : 4.130073800738008
SumAssured : 619999.6992670636
```

Outlier treatment

Boxplot:



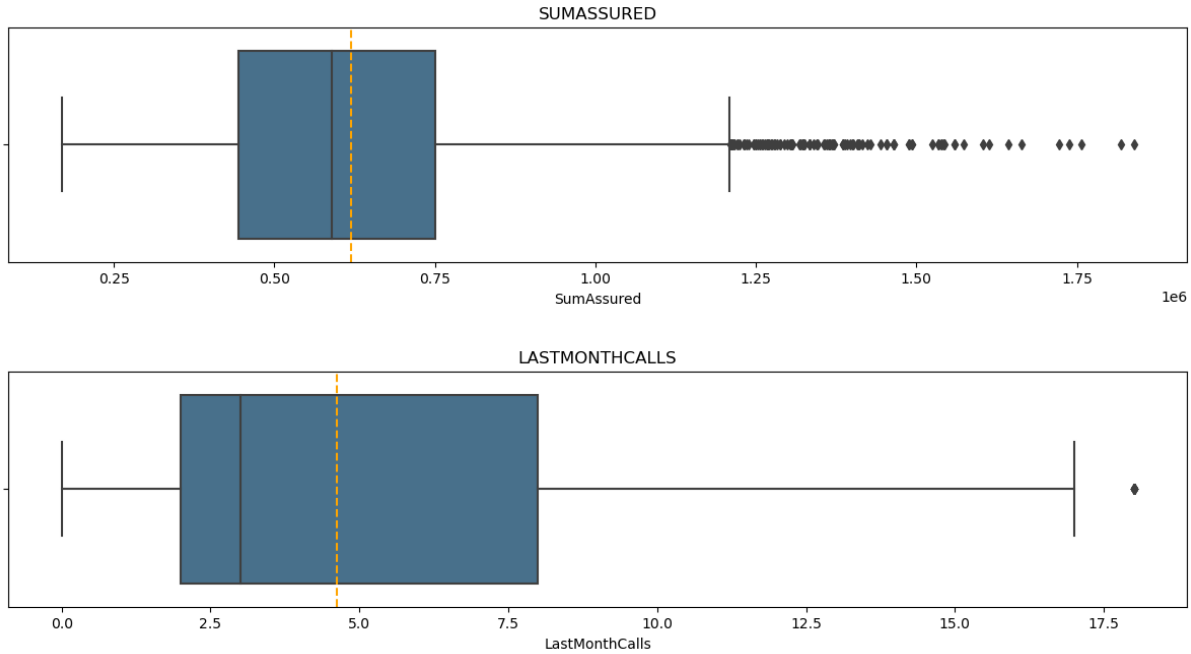


Figure B BoxPlot

The Mean age of the customer can't be 14. There is also clear indication of outliers in the above plots.

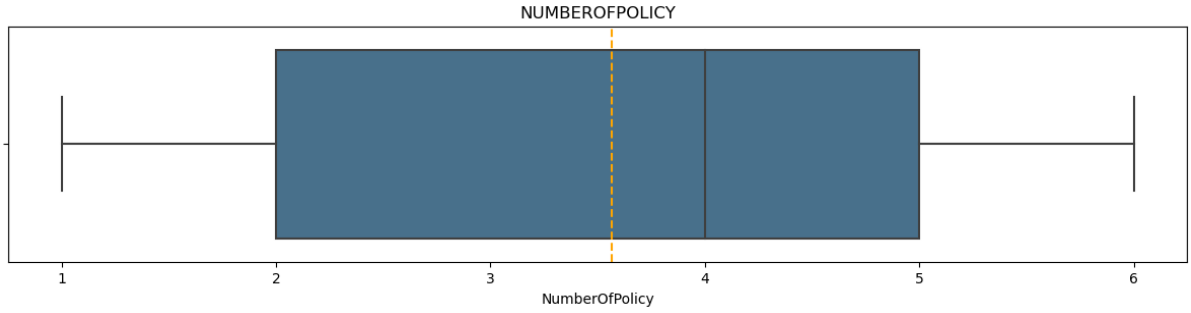


Figure C Box Plot Of Numerofpolicy

The NumberofPolicy doent have any outliers.

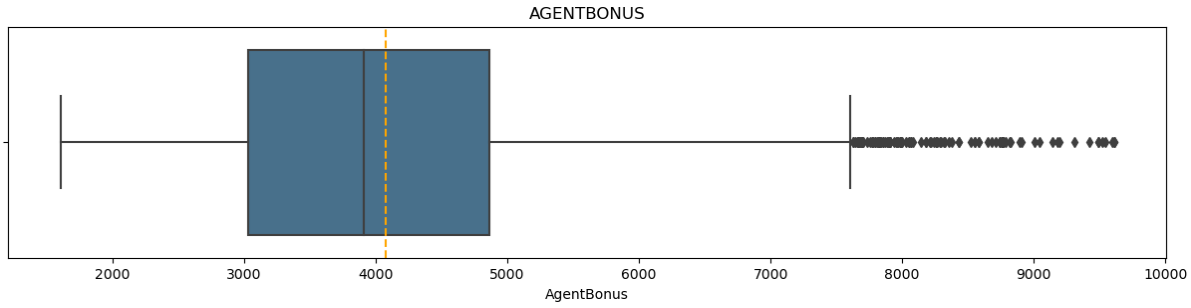


Figure D Box Plot For Agent Bonus

The AgentBonus is the target variable which has been eliminated from the outlier treatment.

The Upper limit for the Numerical Variables is

```
Age                      35.50
CustTenure               35.50
NumberOfPolicy            9.50
MonthlyIncome         31542.38
ExistingPolicyTenure      9.50
SumAssured          1208311.88
LastMonthCalls           17.00
```

The Lower Limit for the Numerical Variable is

```
Age                      -8.50
CustTenure               -8.50
NumberOfPolicy           -2.50
MonthlyIncome         12847.38
ExistingPolicyTenure     -2.50
SumAssured          -13825.12
LastMonthCalls           -7.00
```

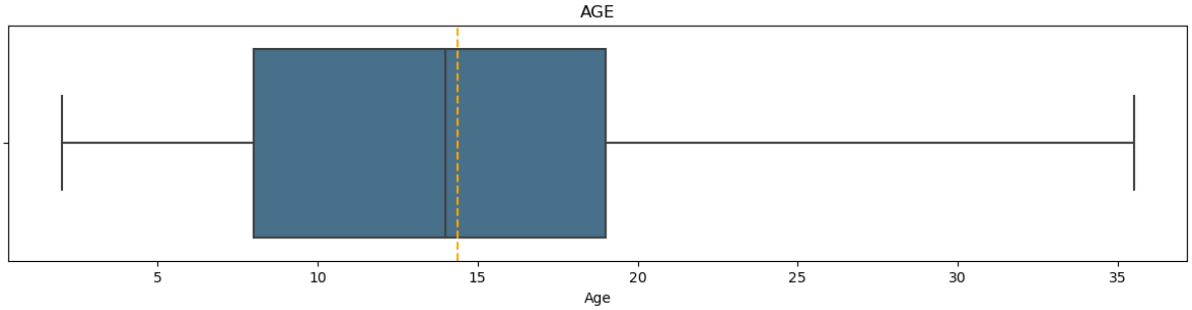
The Total Outlier to the dataset is 3.32%

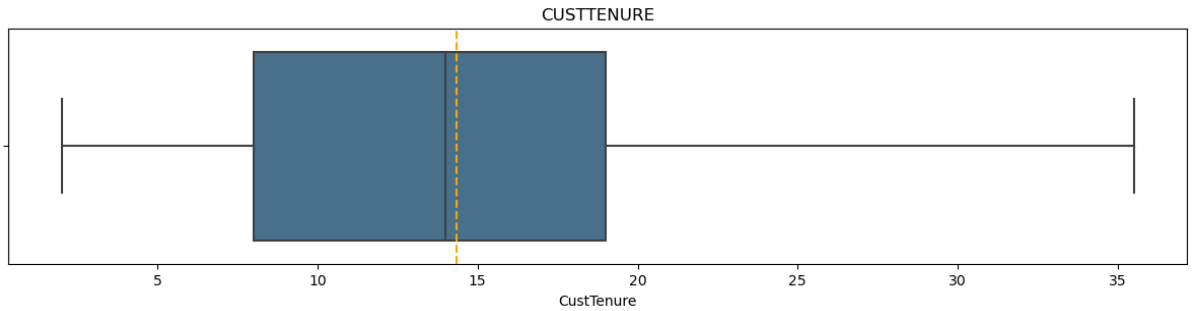
The Outlier percentage for each variable is

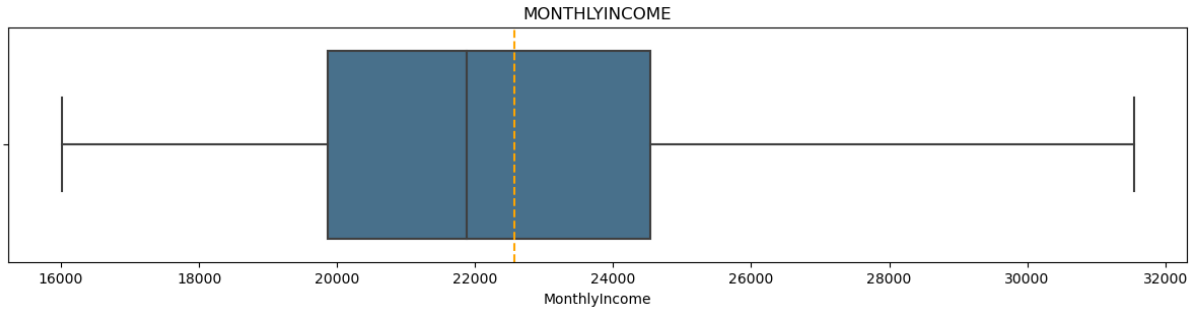
Variable Name	%
Age	2.32
CustTenure	2.15
MonthlyIncome	8.5
ExistingPolicyTenure	7.63
SumAssured	2.43
LastMonthCalls	0.27

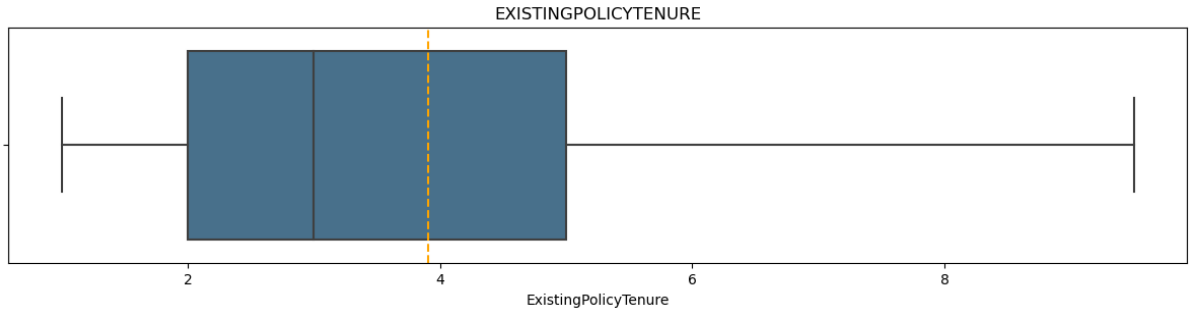
Table 8 Outlier % Table

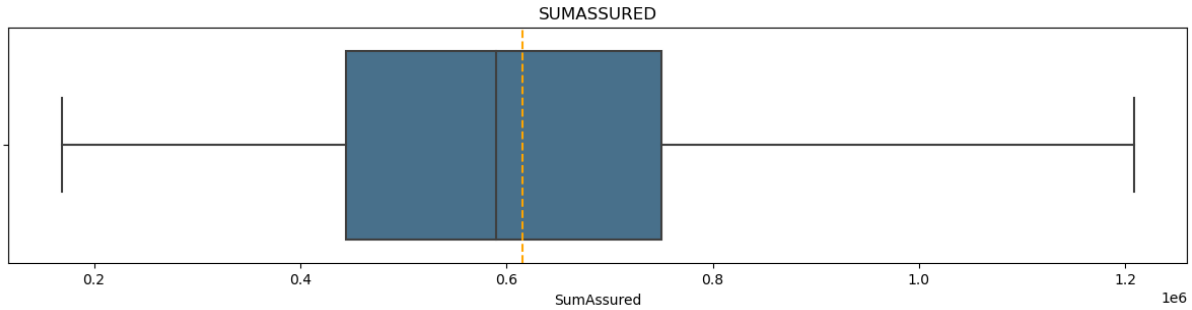
Boxplot (After Outlier Treatment):











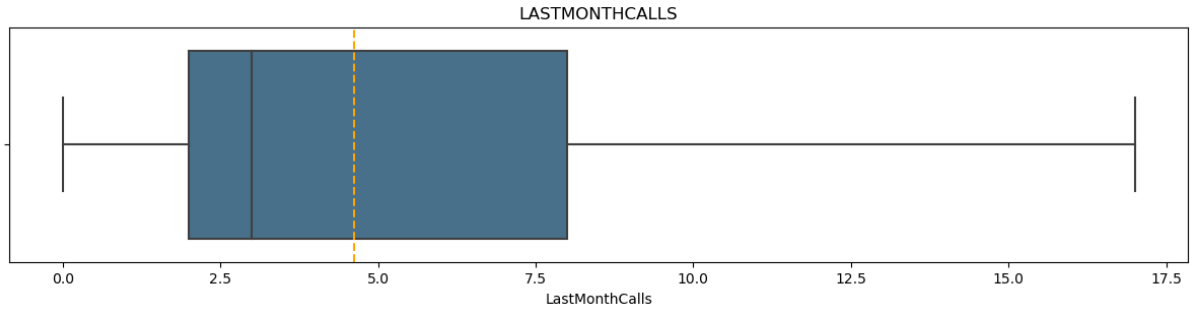


Figure E Boxplot After Outlier Treatment

Removal of unwanted variables

The CustID has been removed as it has no meaning.

The Boxplot for Age with respective to Designation

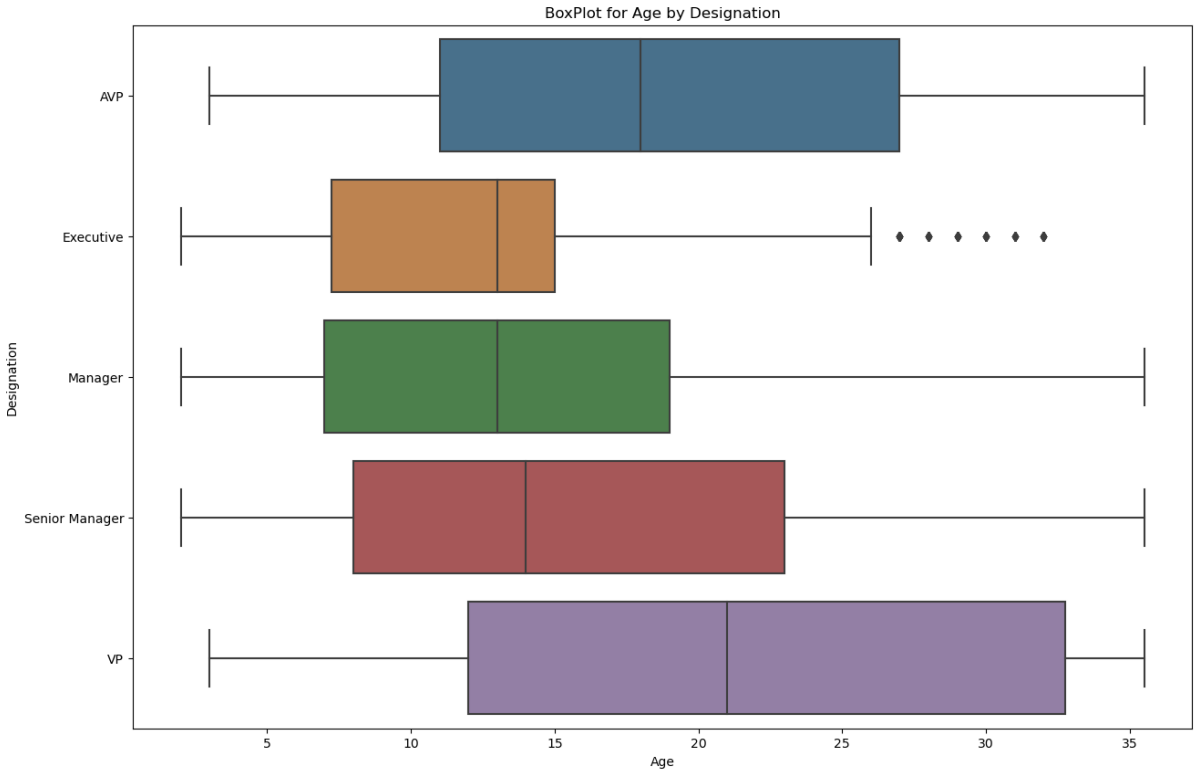


Figure F Boxplot by Designation

The Mean Age of customers when compared to their designation doesn't give any valid inferences.

Based on the Descriptive statistics & Boxplot, Age has some Invalid data.

There is a total of 1918 data with customers age less than the customer tenure in their organisation.

42.43% of the age data has these incorrect details.

The Age column has been dropped as it doesn't give any significant meaning to the model and inferences.

Descriptive Statistics:

The Descriptive Statistics after data cleaning is

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
AgentBonus	4520.00	NaN	NaN	NaN	4077.84	1403.32	1605.00	3027.75	3911.50	4867.25	9608.00
CustTenure	4520.00	NaN	NaN	NaN	14.34	8.33	2.00	8.00	14.00	19.00	35.50
Channel	4520	3	Agent	3194	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Occupation	4520	4	Salaried	2192	NaN	NaN	NaN	NaN	NaN	NaN	NaN
EducationField	4520	3	UG	3698	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	4520	2	Male	2688	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ExistingProdType	4520.00	6.00	4.00	1916.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Designation	4520	5	Executive	1662	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NumberOfPolicy	4520.00	NaN	NaN	NaN	3.57	1.45	1.00	2.00	4.00	5.00	6.00
MaritalStatus	4520	3	Married	2268	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthlyIncome	4520.00	NaN	NaN	NaN	22574.03	3948.15	16009.00	19858.00	21877.00	24531.75	31542.38
Complaint	4520.00	2.00	0.00	3222.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ExistingPolicyTenure	4520.00	NaN	NaN	NaN	3.91	2.68	1.00	2.00	3.00	5.00	9.50
SumAssured	4520.00	NaN	NaN	NaN	615902.26	229255.42	168536.00	444476.25	590012.50	750010.50	1208311.88
Zone	4520	4	West	2566	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PaymentMethod	4520	4	Half Yearly	2656	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LastMonthCalls	4520.00	NaN	NaN	NaN	4.62	3.61	0.00	2.00	3.00	8.00	17.00
CustCareScore	4520.00	5.00	3.00	1419.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 9 Descriptive statistics after cleaning

Observations:

The AgentBonus is in the range of 1,400 to 9,600.

The Customer work experience ranges from 2 to 35 years.

Most of the Customers are salaried people and has UG Degree.

The Most used PaymentMethod is Half Yearly payments.

The West has the greatest number of Insurance Policies.

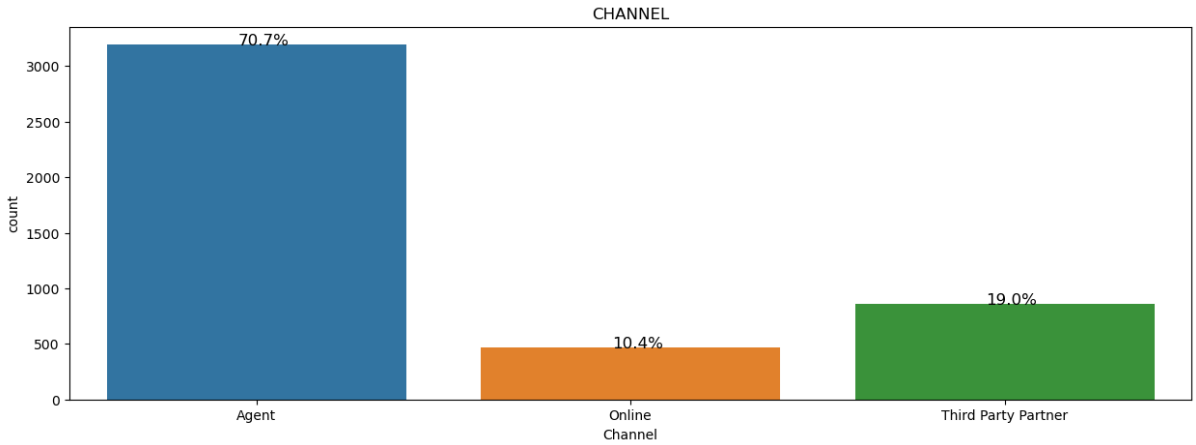
Most of the Customers are coming through agents and their sumassured ranges from 6,15,902.26 to 12,08,311.88.

The Most bought product type is 4.

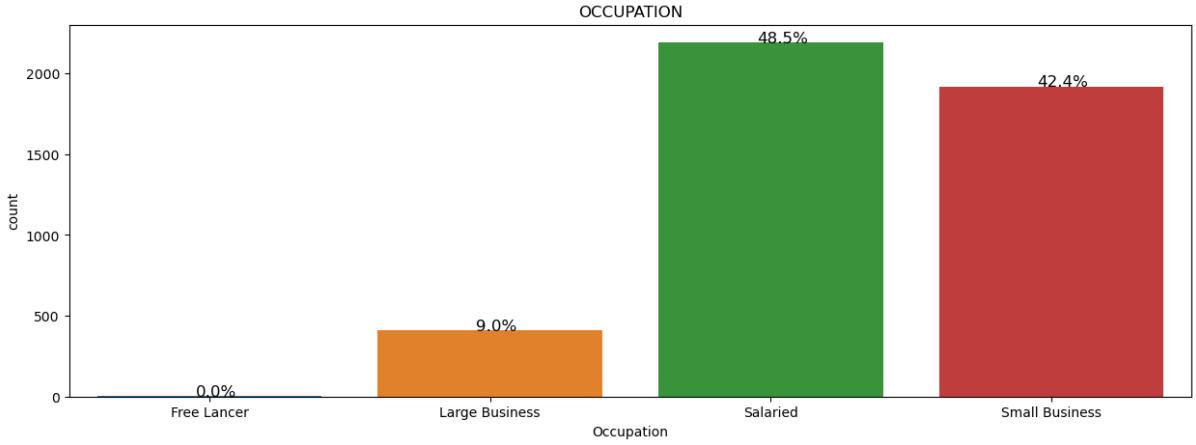
Univariate analysis

CountPlot:

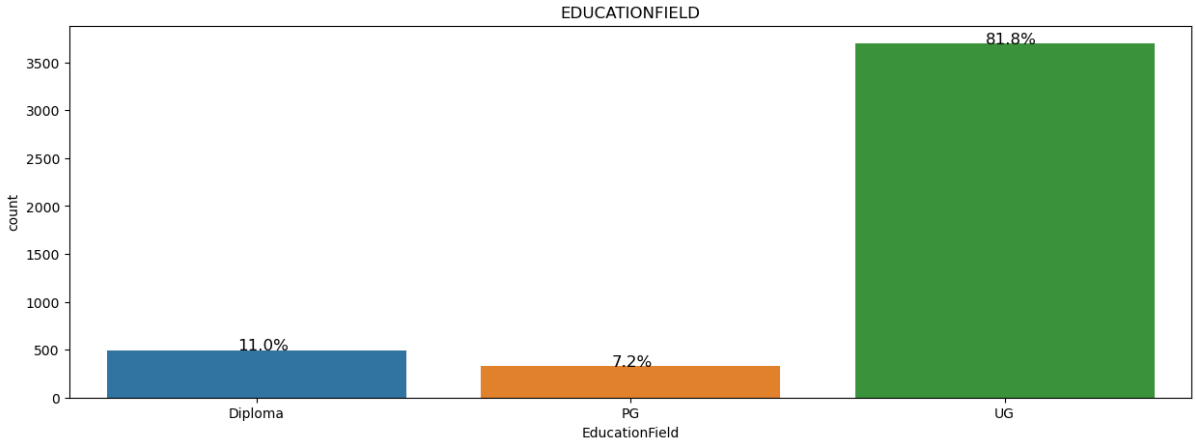
Figure G CountPlot



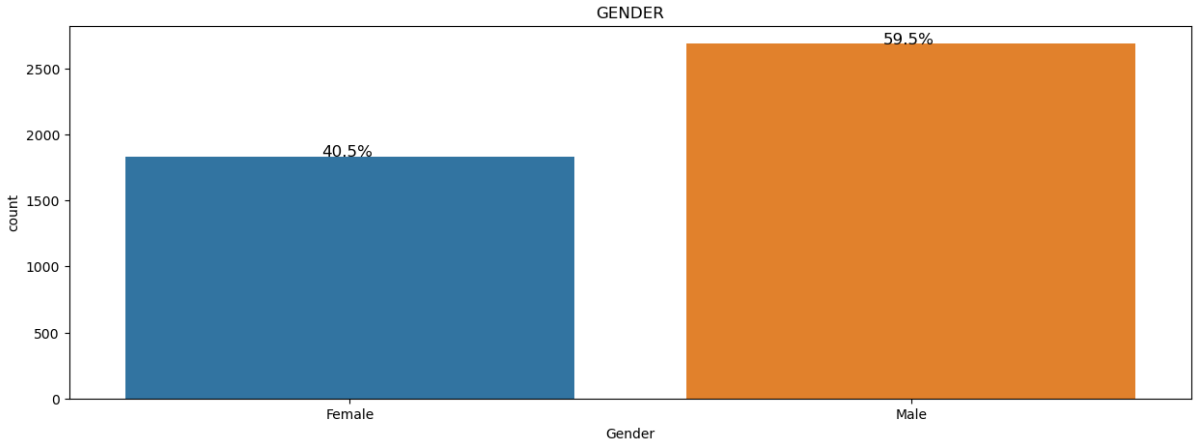
70% of the Customer bought the insurance from agent channel.



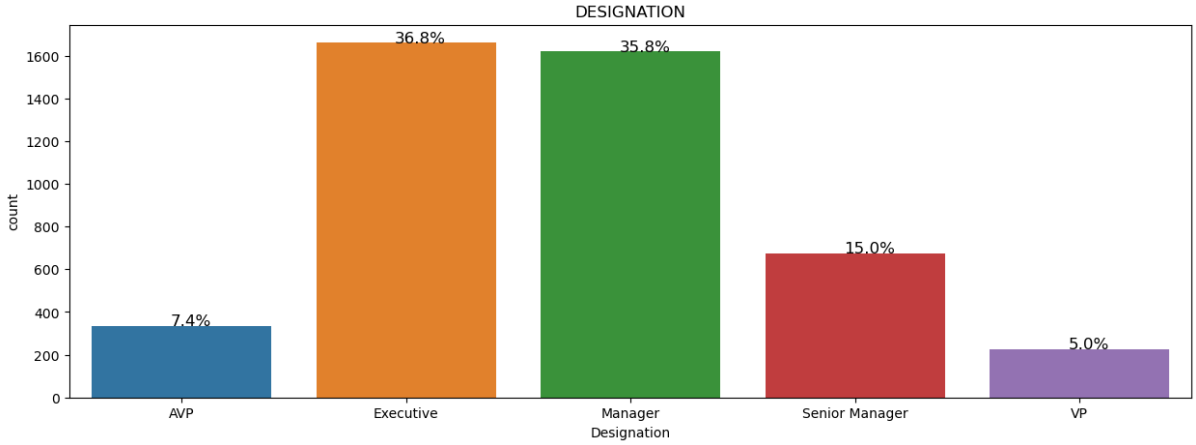
90% of the customers are salaried and small business people.



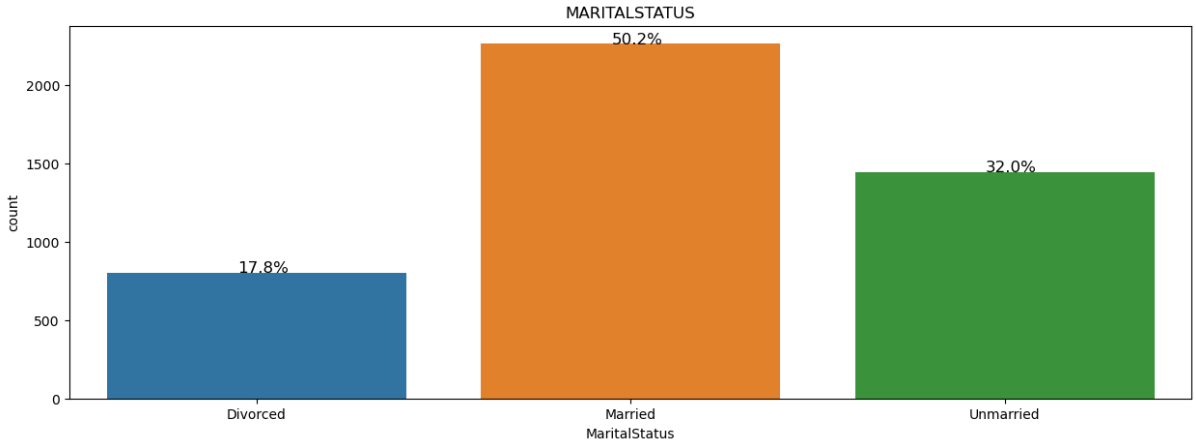
Only 7% of the people has done masters and 81% of the customers has UG degree.



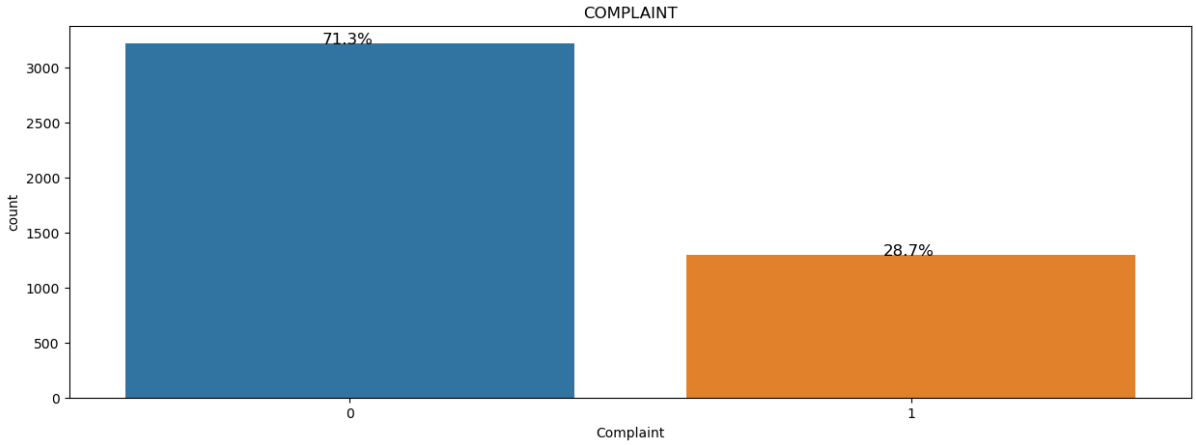
60% of the customers are male.



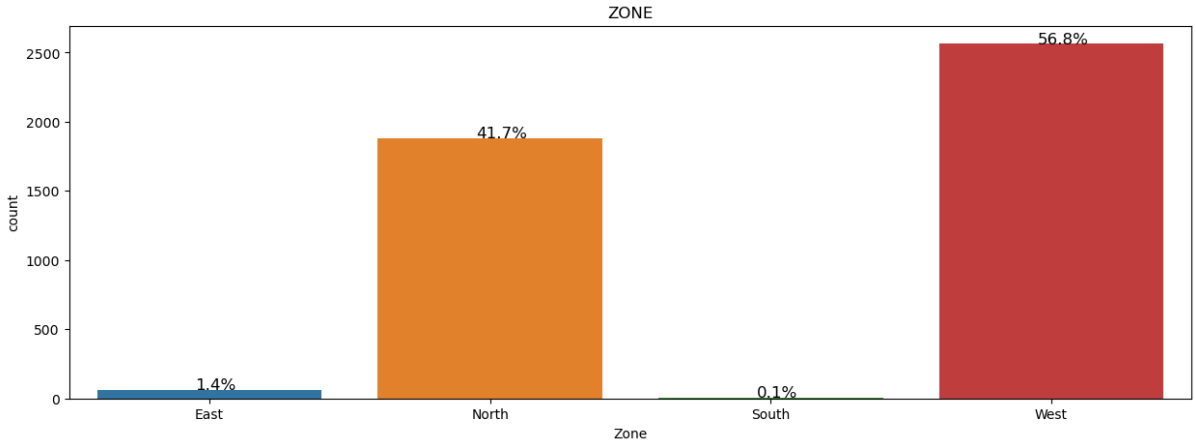
72% of the Customers designation in their organisation as Executive or Managers.



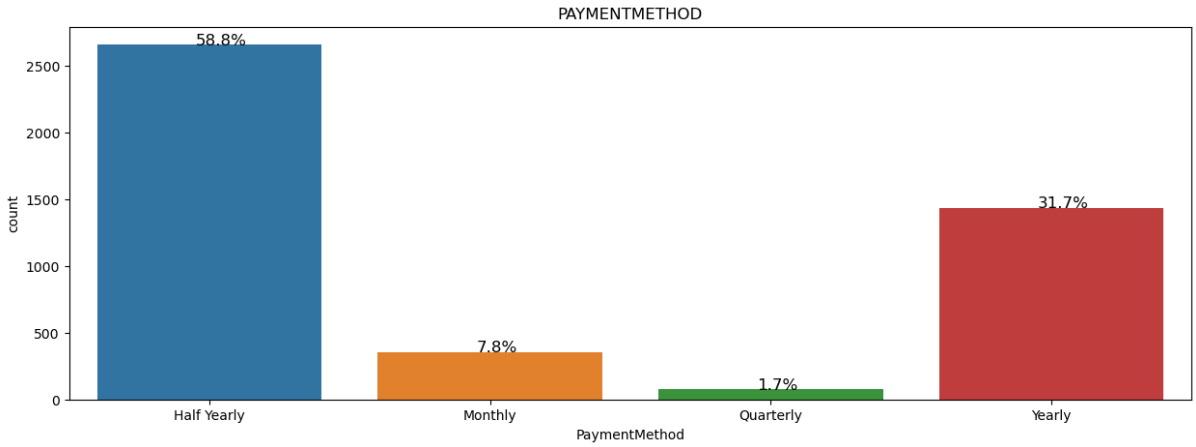
Half of the customers are married.



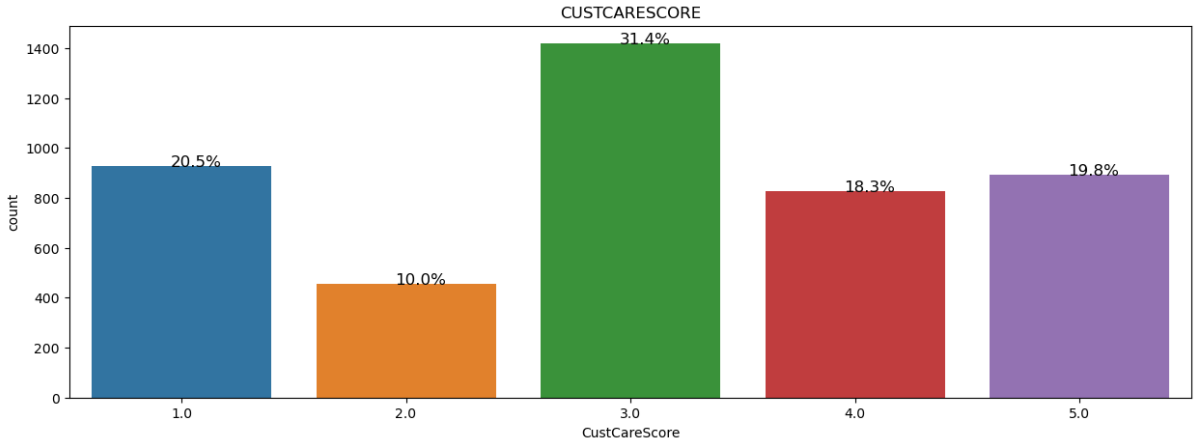
Only 29% of the customers has complaint.



97% of the customers are from North & west.



The Most preferred policy payment tenure is Half yearly or Yearly.



38% of the customers are satisfied from the service call

31% feels neutral and 31% feels dissatisfied.

Histogram:

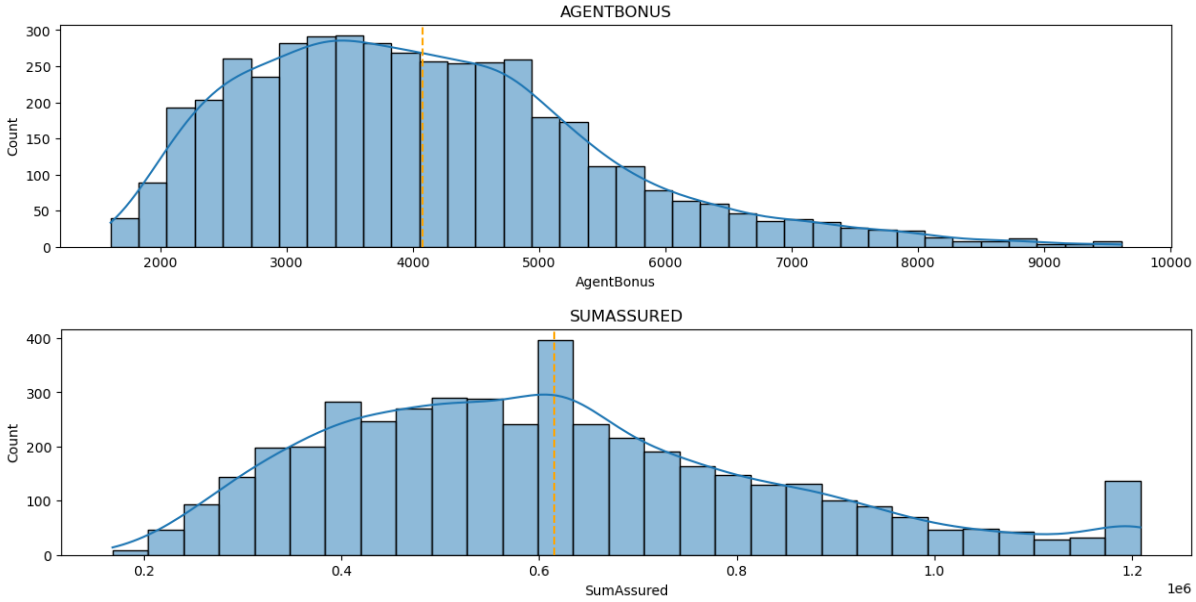


Figure H Histogram

Both SumAssured and AgentBonus are Nearly positively Skewed.

Bivariate analysis

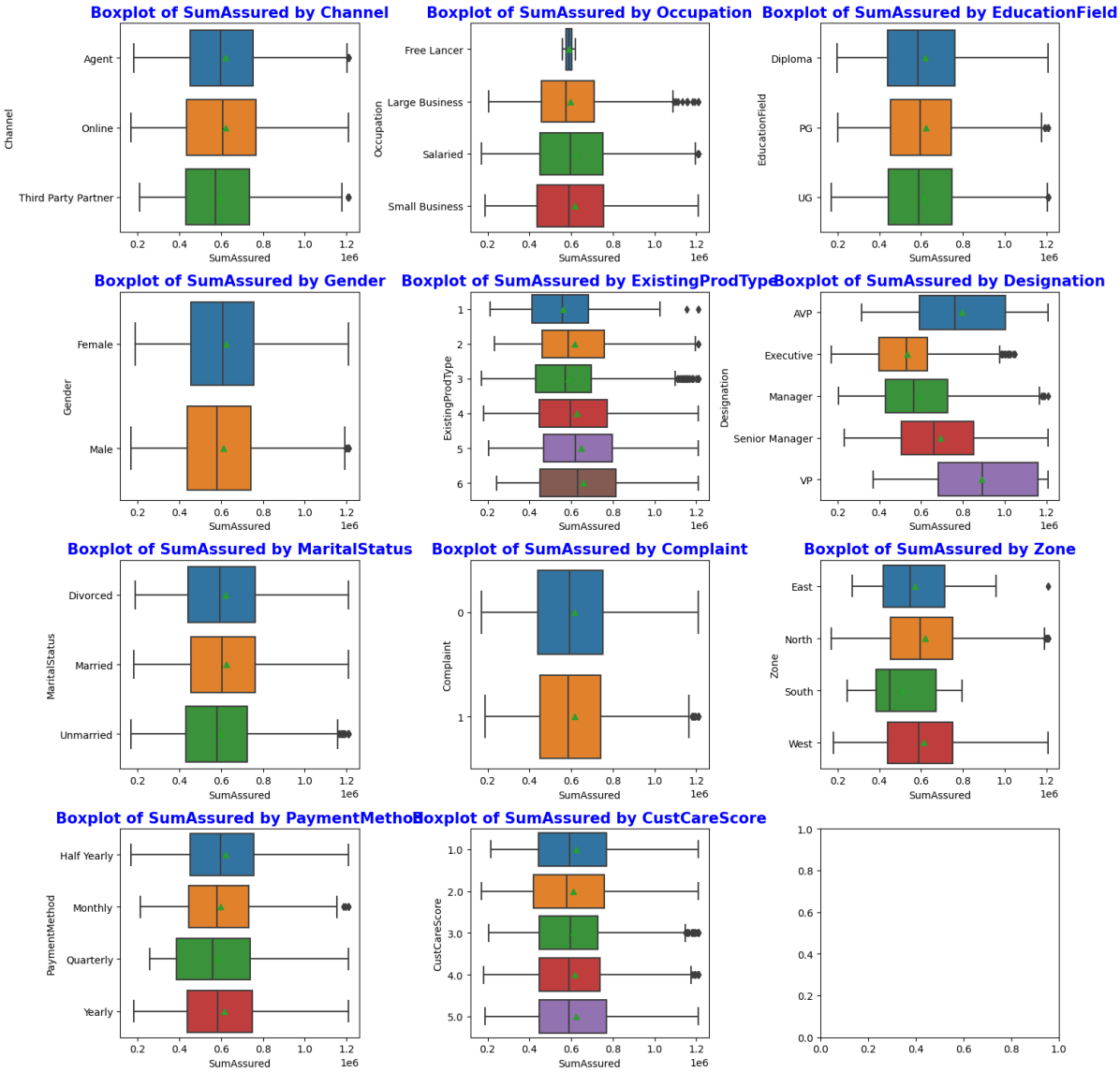


Figure I Boxplot of SumAssured by Categorical Variables

The Average SumAssured is nearly same for all categories except designation.

The SumAssured Varies with Designation.

VP and AVP has the highest mean SumAssured policies.

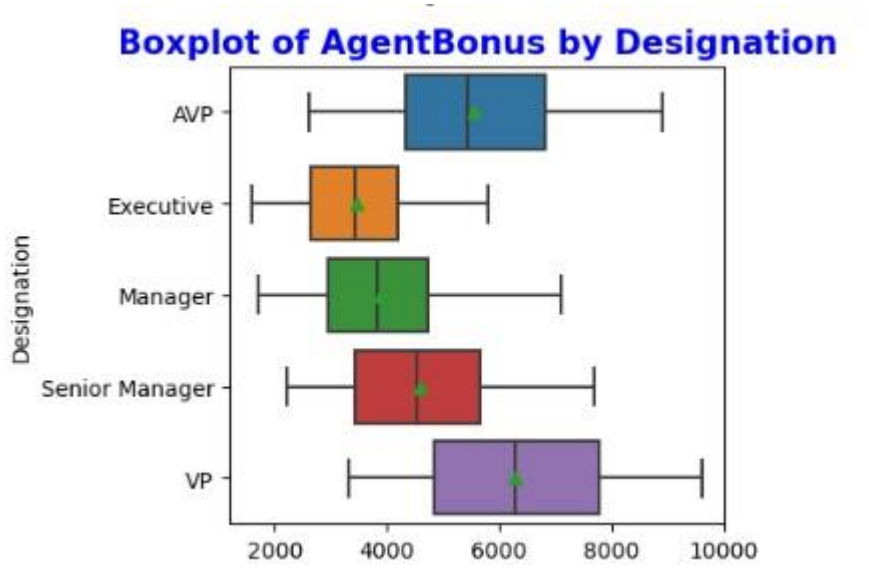


Figure J Boxplot of SumAssured by Designation

Because of the SumAssured, AgentBonus also varies with the customer sumassured.

Correlation Plot:

All the variables are positively correlated.

AgentBonus and SumAssured has High Positive Correlation.

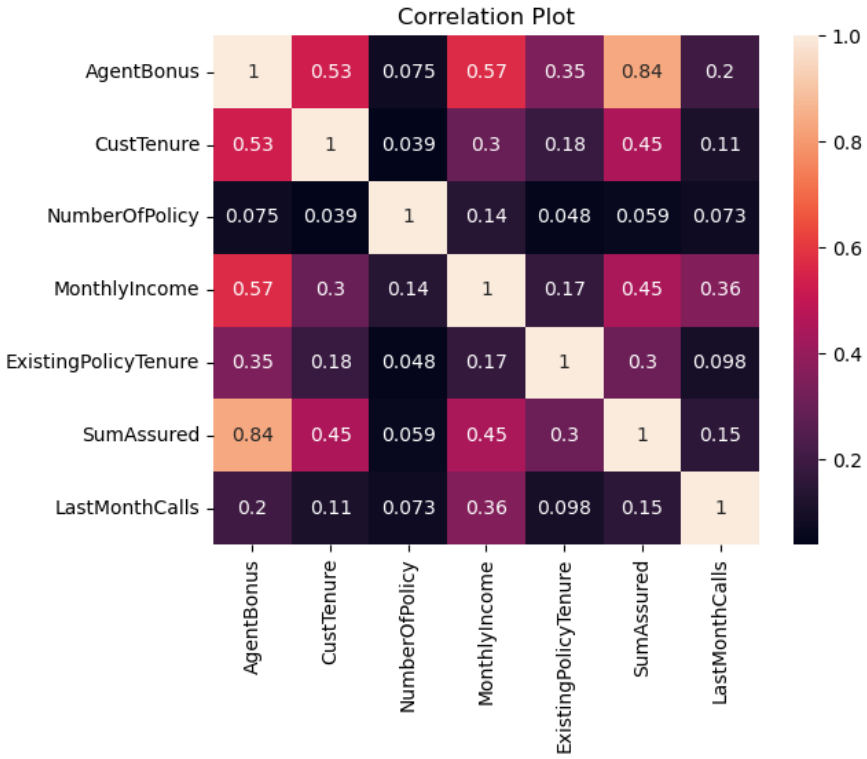


Figure K Correlation Plot

ScatterPlot:

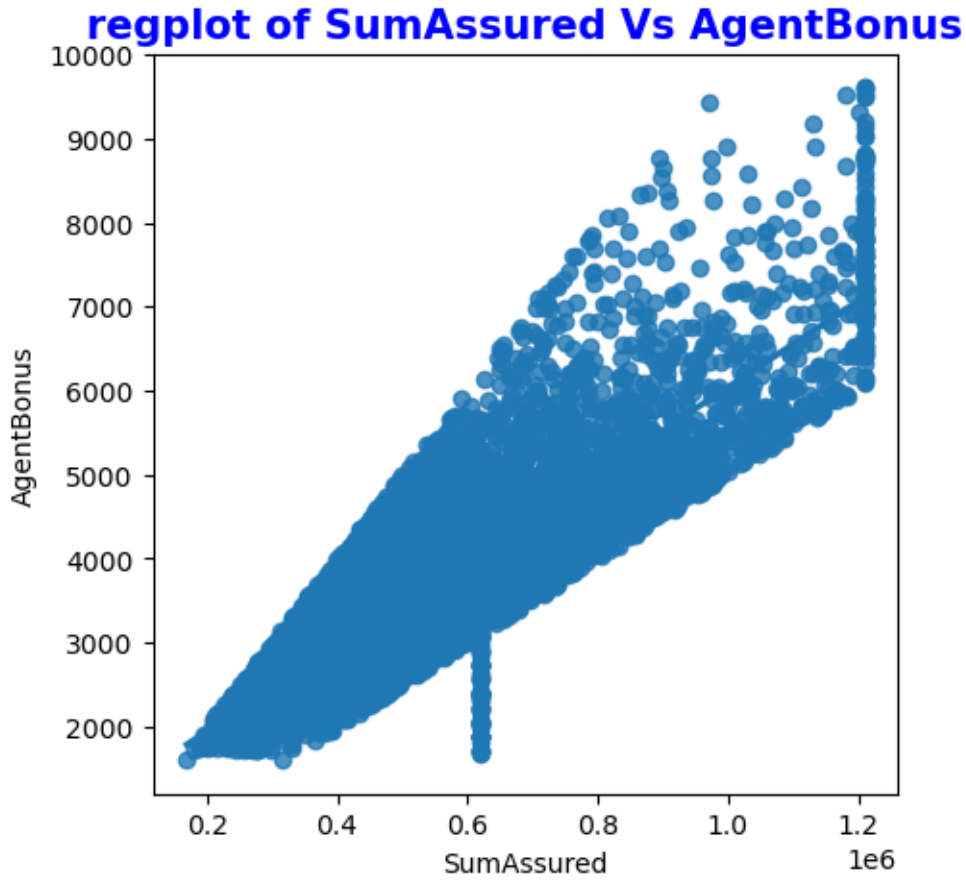


Figure L ScatterPlot of SumAssured and AgentBonus

As the SumAssured increases, Agent Bonus also increases.

MultiVariate Analysis

	mean		
MaritalStatus	Divorced	Married	Unmarried
Gender			
Female	628180.52	638665.28	603916.71
Male	614470.48	617831.88	593425.44

	mean		
EducationField	Diploma	PG	UG
Gender			
Female	639525.42	621241.71	623015.21
Male	605559.81	625305.60	608730.38

	mean					
ExistingProdType	1	2	3	4	5	6
Gender						
Female	557513.40	621467.11	591178.84	638696.47	657557.97	692528.36
Male	566762.88	612175.38	581852.41	619783.12	643274.99	638965.38

Table 10 Multivariate Analysis by gender with Marital Status, Education & Prod Type

There is no difference between the gender with the sumAssured

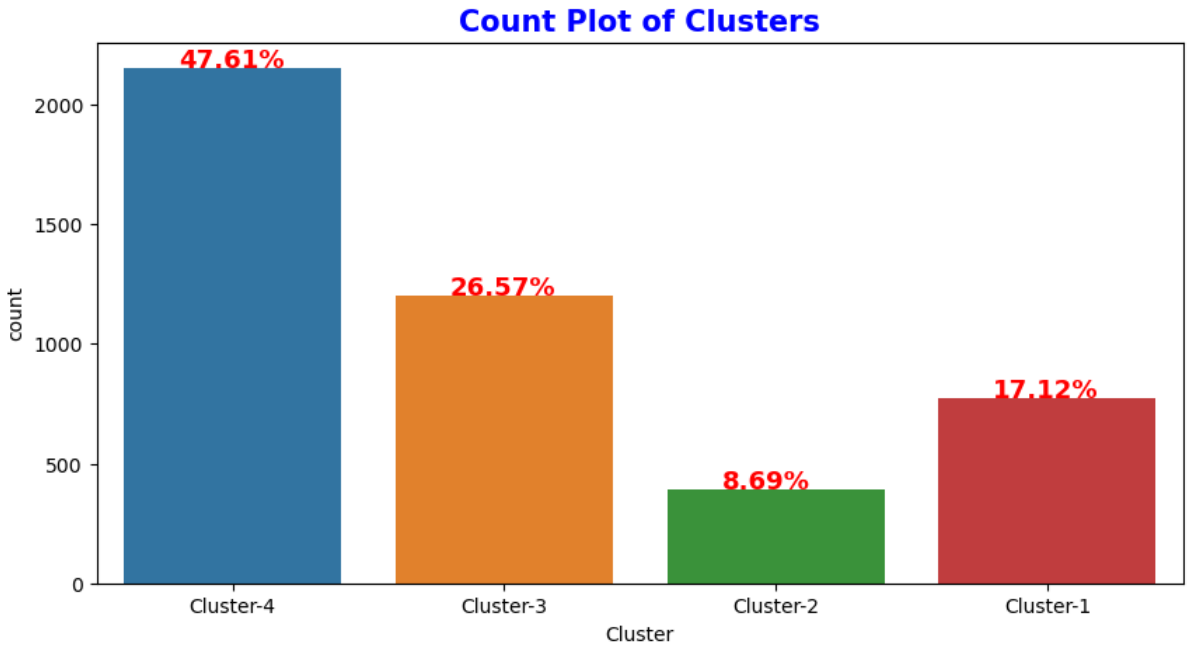
Business Insights from EDA

Cluster Analysis

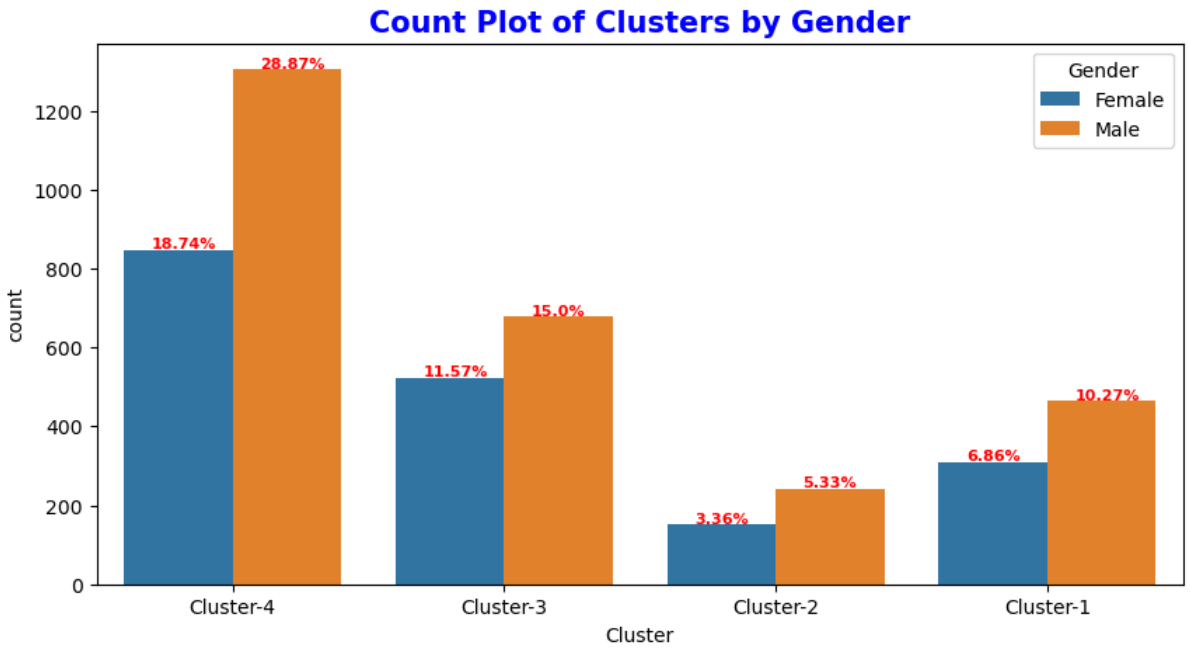
Cluster Analysis is done for finding the customer segmentation.

There are four clusters.

Figure M CountPlot by Clusters

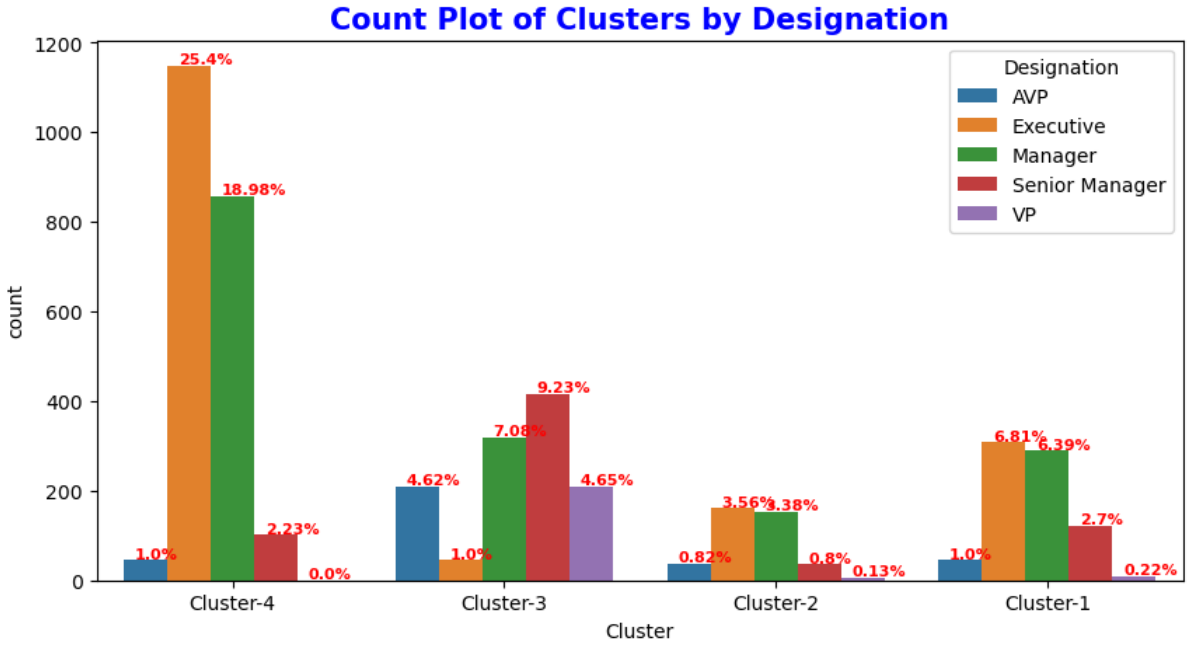


Cluster 4 has the highest Insurance sold.



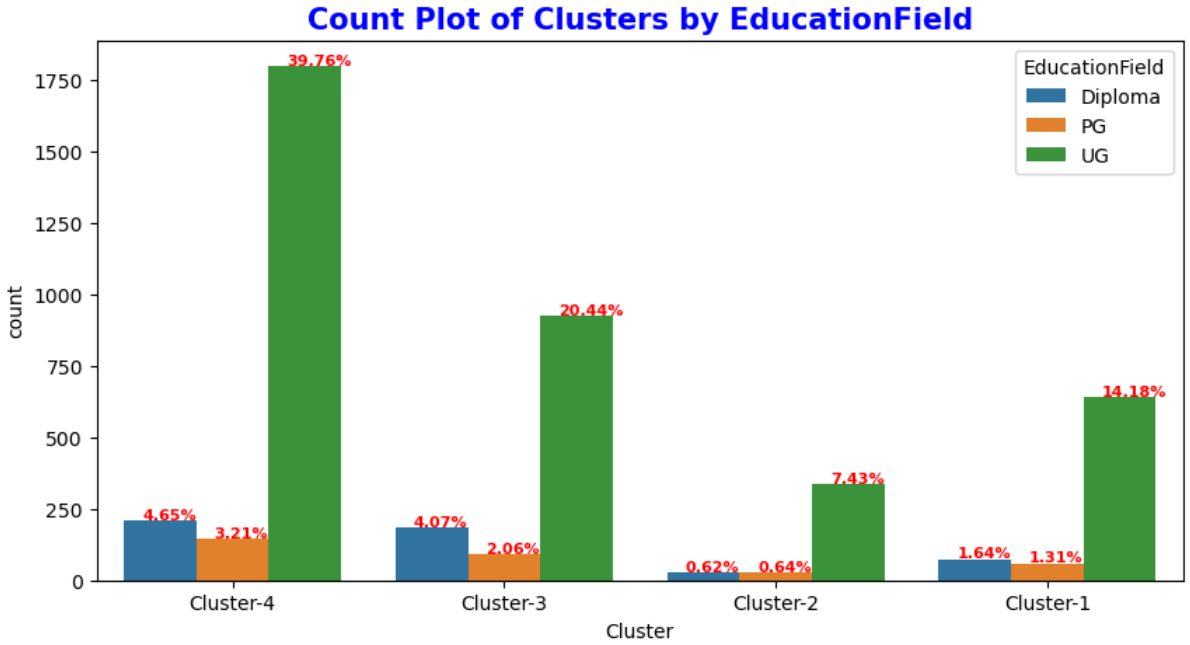
28% of the total insurance sales is bought by the cluster 4 male customers.

Cluster 4 females has the second highest insurance products bought.



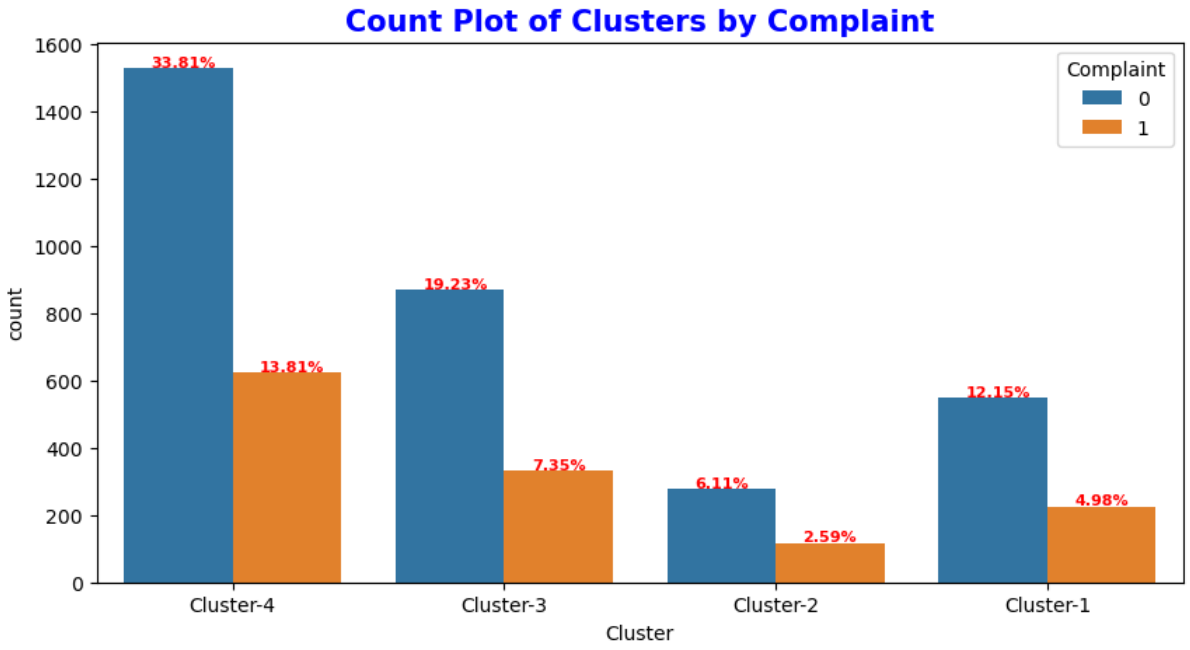
The Customers from the cluster 4 are mostly executive and Managers.

The VP are in the cluster 3.



Cluster 4 and 3 are similar in degree background.

Cluster 4 has highest in UG, PG and Diploma customers.



Cluster 4 has the highest complaint raised.

SumAssured by Cluster

Cluster	max	mean	min
Cluster-1	1208311.88	624377.66	204950.00
Cluster-2	1208311.88	588727.38	212720.00
Cluster-3	1208311.88	813048.29	290924.00
Cluster-4	1047880.00	507792.31	168536.00

Table 11 SumAssured by Cluster

AgentBonus by Cluster

Cluster	max	mean	min
Cluster-1	8771	4124.97	1729
Cluster-2	8380	3976.69	1898
Cluster-3	9608	5390.22	2533
Cluster-4	6823	3346.94	1605

Table 12 AgentBonus by Cluster

The Cluster 3 has the highest SumAssured and Bonus.

Second highest is Cluster 1 and then Cluster 4.

Business Implications

Based on the 4 clusters, Cluster 4 has the highest customers and their average sumassured is moderate. They are the most common customers to buy insurance and they are not investing much.

These cluster can be more focused on selling top up to increase the SumAssured.

The Premium Customers are in the Cluster 3. They are mostly VP by designation.

Only the services has to be maintained properly for the premium cluster.

Cluster 1 can be classified as Silver.

Cluster 1 and 4 can also more focued on top up and selling other products to move them to the next cluster.

Model building and Tuning

The Dataset is Split into Independent and Dependent variables. The Independent Variables are scaled with standardscalar.

The Scaled data of independent variables and dependent variables are merged and they are split into train and test data with 70:30 ratio respectively.

The below models are trained to find the best model for the prediction of agent bonus.

Linear Regression

The Linear Regression is modelled using the Statsmodel Package.

Both Independent and Dependent variable are merged for this Linear Regression.

Model 1:

Formula:

AgentBonus ~ CustTenure + NumberOfPolicy + MonthlyIncome + Complaint + ExistingPolicyTenure + SumAssured + LastMonthCalls + CustCareScore + Channel_Online + Channel_Third_Party_Partner + Occupation_Large_Business + Occupation_Salaried + Occupation_Small_Business + EducationField_PG + EducationField_UG + Gender_Male + ExistingProdType_2 + ExistingProdType_3 + ExistingProdType_4 + ExistingProdType_5 + ExistingProdType_6 + Designation_Executive + Designation_Manager + Designation_Senior_Manager + Designation_VP + MaritalStatus_Married + MaritalStatus_Unmarried + Zone_North + Zone_South + Zone_West + PaymentMethod_Monthly + PaymentMethod_Quarterly + PaymentMethod_Yearly

Most of the variables are not significant and there is correlation between the independent variable (Inferred from EDA).

The correlated variables are removed using VIF method. The VIF greater than 5 are removed individually and the VIF of others are calculated. The below table shows the VIF values for the 29 variables that are not correlated.

S.no	variables	VIF
1	MonthlyIncome	2.43
2	PaymentMethod_Yearly	2.42
3	MaritalStatus_Unmarried	1.95
4	EducationField_PG	1.94
5	MaritalStatus_Married	1.94
6	EducationField_UG	1.88
7	PaymentMethod_Monthly	1.86
8	ExistingProdType_5	1.85
9	ExistingProdType_2	1.85
10	Designation_VP	1.79
11	ExistingProdType_3	1.63

12	SumAssured	1.55
13	Designation_Senior_Manager	1.53
14	Occupation_Small_Business	1.36
15	CustTenure	1.28
16	Designation_Manager	1.27
17	PaymentMethod_Quarterly	1.19
18	LastMonthCalls	1.19
19	ExistingProdType_6	1.19
20	Occupation_Large_Business	1.12
21	ExistingPolicyTenure	1.12
22	NumberOfPolicy	1.1
23	Channel_Online	1.05
24	Channel_Third_Party_Partner	1.04
25	Gender_Male	1.02
26	CustCareScore	1.02
27	Zone_West	1.01
28	Zone_South	1.01
29	Complaint	1.01

Table 13 VIF Values

Model 2:

Formula:

AgentBonus ~ CustTenure + NumberOfPolicy + MonthlyIncome + Complaint + ExistingPolicyTenure + SumAssured + LastMonthCalls + CustCareScore + Channel_Online + Channel_Third_Party_Partner + Occupation_Large_Business + Occupation_Small_Business + EducationField_PG + EducationField_UG + Gender_Male + ExistingProdType_2 + ExistingProdType_3 + ExistingProdType_5 + ExistingProdType_6 + Designation_Manager + Designation_Senior_Manager + Designation_VP + MaritalStatus_Married + MaritalStatus_Unmarried + Zone_South + Zone_West + PaymentMethod_Monthly + PaymentMethod_Quarterly + PaymentMethod_Yearly

Most of the independent variables are not significant in the model.

The Significant variables are,

CustTenure

MonthlyIncome

ExistingPolicyTenure

SumAssured

CustCareScore

Designation_Manager

Designation_VP

From the EDA, Sumassured was highly positively correlated with the dependent variable – AgentBonus.

After removing the insignificant variables, the final model is as below

Model 3:

OLS Regression Results

Dep. Variable:	AgentBonus	**R-squared:**	0.783
Model:	OLS	**Adj. R-squared:**	0.783
Method:	Least Squares	**F-statistic:**	1630.
Date:	Sun, 24 Mar 2024	**Prob (F-statistic):**	0.00
Time:	12:28:45	**Log-Likelihood:**	-25003.
No. Observations:	3164	**AIC:**	5.002e+04
Df Residuals:	3156	**BIC:**	5.007e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4073.5441	11.648	349.726	0.000	4050.706	4096.382
CustTenure	225.3250	13.152	17.132	0.000	199.537	251.113
MonthlyIncome	287.9973	14.856	19.386	0.000	258.869	317.125
ExistingPolicyTenure	121.6540	12.267	9.917	0.000	97.602	145.706
SumAssured	888.4446	14.341	61.951	0.000	860.326	916.563
CustCareScore	24.1449	11.652	2.072	0.038	1.298	46.992
Designation_Manager	-40.5992	11.875	-3.419	0.001	-63.882	-17.316
Designation_VP	51.2924	13.880	3.695	0.000	24.077	78.507

Omnibus:	128.913	**Durbin-Watson:**	1.971
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	149.440
Skew:	0.475	**Prob(JB):**	3.54e-33
Kurtosis:	3.482	**Cond. No.**	2.32

Formula:

AgentBonus ~ CustTenure + MonthlyIncome + ExistingPolicyTenure + SumAssured + CustCareScore + Designation_Manager + Designation_VP

The model strength is 78.3% (R-Squared).

The hypotheses of the F test are as follows:

H0: $\beta1=\beta2=\ldots=\beta k=0$

H1: At least one of $\beta1,\beta2,\ldots,\beta k\neq0$

Prob (F-statistic) is less than 0.05. Rejects the null Hypothesis.

All the predictor variables are significant.

Omnibus Test : Tests the skewness and kurtosis of the residuals, this value should be as low as possible. Further JB P value also proves the normality of the data.

Durbin Watson Test : Tests the homoscedasticity of the residuals the value should be betweeen 2. It is nearly 2(1.9). There is no autocorrelation.

Jarque-Bera Test : Confirmatory test in addtition to the Omnibus Test to check the skewness and kurtosis of the residuals. P-value is less than 0.05. The dataset is normally distributed.

Condition Number : High condition number means presence of multi-collinearity. The comdition number is less than 10. There is no Multicollinearity.

The Final Equation for prediction of AgentBonus is

AgentBonus = (4073.54) * Intercept + (225.32) * CustTenure + (288.0) * MonthlyIncome + (121.65) * ExistingPolicyTenure + (888.44) * SumAssured + (24.14) * CustCareScore + (-40.6) * Designation_Manager + (51.29) * Designation_VP.

Validation Against Test:

The Final Regression model is used for prediction.

The R-squared value for Test data is 0.769

The RMSE for Linear Regression model is 671.51

Lasso regression and ridge regression are both known as regularization methods because they both attempt to minimize the sum of squared residuals (RSS) along with some penalty term.

Ridge Regression

Ridge regression, the L2 penalty shrinks coefficients towards zero but never to absolute zero.

The Equation for prediction of AgentBonus is

AgentBonus = (220.88) * CustTenure + (10.19) * NumberOfPolicy + (167.12) * MonthlyIncome + (19.51) * Complaint + (121.17) * ExistingPolicyTenure + (877.51) * SumAssured + (-9.72) * LastMonthCalls + (22.61) * CustCareScore + (10.06) * Channel_Online + (-2.3) * Channel_Third_Party_Partner + (-42.61) * Occupation_Large_Business + (-53.32) * Occupation_Salaried + (-61.33) * Occupation_Small_Business + (-2.38) * EducationField_PG + (-3.27) * EducationField_UG + (13.93) * Gender_Male + (16.82) * ExistingProdType_2 + (-70.59) * ExistingProdType_3 + (-23.97) * ExistingProdType_4 + (0.49) * ExistingProdType_5 + (22.84) * ExistingProdType_6 + (-229.82) * Designation_Executive + (-210.53) * Designation_Manager + (-92.98) * Designation_Senior_Manager + (43.25) * Designation_VP + (-28.57) * MaritalStatus_Married + (6.05) * MaritalStatus_Unmarried + (14.56) * Zone_North + (6.08) * Zone_South + (18.79) * Zone_West + (-13.24) * PaymentMethod_Monthly + (9.64) * PaymentMethod_Quarterly + (-49.26) * PaymentMethod_Yearly.

Validation Against Test:

The Ridge Regression model is used for prediction.

The R-squared value for Test data is 0.771

The RMSE for Ridge Regression model is 668.43.

The Ridge performs slightly better than Earlier Model.

Lasso Regression

Lasso regression, also known as L1 regularization, brings balance between simplicity and accuracy. It can provide interpretable models while effectively managing the risk of overfitting. It shrinks the coefficient to zero.

The Equation for prediction of AgentBonus is

AgentBonus = (219.82) * CustTenure + (9.43) * NumberOfPolicy + (170.79) * MonthlyIncome + (18.45) * Complaint + (120.04) * ExistingPolicyTenure + (881.27) * SumAssured + (-7.8) * LastMonthCalls + (22.0) * CustCareScore + (9.01) * Channel_Online + (-1.47) * Channel_Third_Party_Partner + (-9.02) * Occupation_Large_Business + (3.36) * Occupation_Salaried + (-3.46) * Occupation_Small_Business + (-0.0) * EducationField_PG + (-0.81) * EducationField_UG + (13.04) * Gender_Male + (16.58) * ExistingProdType_2 + (-48.29) * ExistingProdType_3 + (-2.64) * ExistingProdType_4 + (11.7) * ExistingProdType_5 + (27.32) * ExistingProdType_6 + (-221.21) * Designation_Executive + (-203.7) * Designation_Manager + (-89.32) * Designation_Senior_Manager + (42.41) * Designation_VP + (-28.41) * MaritalStatus_Married + (4.91) * MaritalStatus_Unmarried + (-0.0) * Zone_North + (4.14) * Zone_South + (3.6) * Zone_West + (-0.0) * PaymentMethod_Monthly + (11.15) * PaymentMethod_Quarterly + (-43.4) * PaymentMethod_Yearly.

Validation Against Test:

The Lasso Regression model is used for prediction.

The R-squared value for Test data is 0.771

The RMSE for Lasso Regression model is 668.02.

The Ridge and Lasso mostly performs similar in this case.

SGD Regression

Linear model fitted by minimizing a regularized empirical loss with SGD.

SGD stands for Stochastic Gradient Descent.

The Equation for prediction of AgentBonus is

AgentBonus = (217.35) * CustTenure + (7.08) * NumberOfPolicy + (159.05) * MonthlyIncome + (17.25) * Complaint + (117.63) * ExistingPolicyTenure + (880.21) * SumAssured + (-9.63) * LastMonthCalls + (13.35) * CustCareScore + (15.63) * Channel_Online + (-9.73) * Channel_Third_Party_Partner + (-20.95) * Occupation_Large_Business + (-13.1) * Occupation_Salaried + (-19.08) * Occupation_Small_Business + (3.76) * EducationField_PG + (-9.66) * EducationField_UG + (9.34) * Gender_Male + (13.76) * ExistingProdType_2 + (-71.3) * ExistingProdType_3 + (-20.84) * ExistingProdType_4 + (3.17) * ExistingProdType_5 + (27.34) * ExistingProdType_6 + (-231.72) * Designation_Executive + (-218.33) * Designation_Manager + (-98.5) * Designation_Senior_Manager + (35.13) * Designation_VP + (-25.08) * MaritalStatus_Married + (4.47) * MaritalStatus_Unmarried + (18.46) * Zone_North + (2.0) * Zone_South + (15.8) * Zone_West + (-17.54) * PaymentMethod_Monthly + (10.78) * PaymentMethod_Quarterly + (-46.37) * PaymentMethod_Yearly

Validation Against Test:

The SGD Regression model is used for prediction.

The R-squared value for Test data is 0.7703

The RMSE for SGD Regression model is 669.63.

CART

The CART model uses the Decision Tree Regressor. It has the feature Importance Plot.

The Feature Importance chart for CART model is

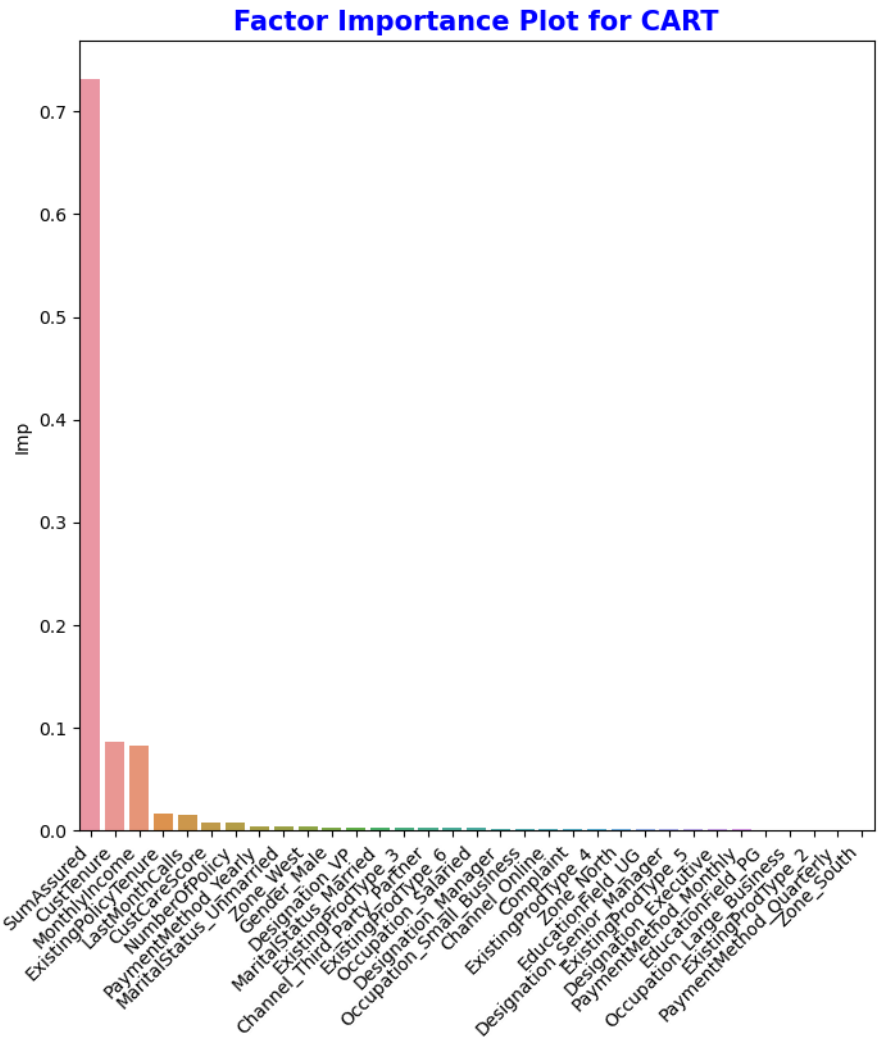


Figure N Feature Importance Plot - CART

Sumassured is the most important feature.

The CART model is overfitting the training data.

The R-squared for training data is 1 and RMSE is 0.

Validation Against Test:

The CART model is used for prediction.

The R-squared value for Test data is 0.65

The RMSE for CART model is 821.64.

The CART Model has to be tuned to get better performance.

Model Tuning

The Model tuning is done for getting better models.

CART Tuned/Pruned

RandomizedSearchCV is used for tuning the model.

The Parameters used in the gridsearch are

'criterion':['squared_error','friedman_mse', 'absolute_error', 'poisson'],

'max_depth':list(range(10,100)),

'min_samples_split':list(range(2,5)),

'min_samples_leaf':list(range(1,100)),

'min_impurity_decrease':list(np.arange(0.0001,0.01,0.001))

The tuned CART Model is

DecisionTreeRegressor(criterion='friedman_mse', max_depth=32,

min_impurity_decrease=0.0051, min_samples_leaf=24,

min_samples_split=4).

The Variables - SumAssured, MonthlyIncome and CustTenure are the important factors for tuned CART Model.

The Feature Importance chart for CART model is

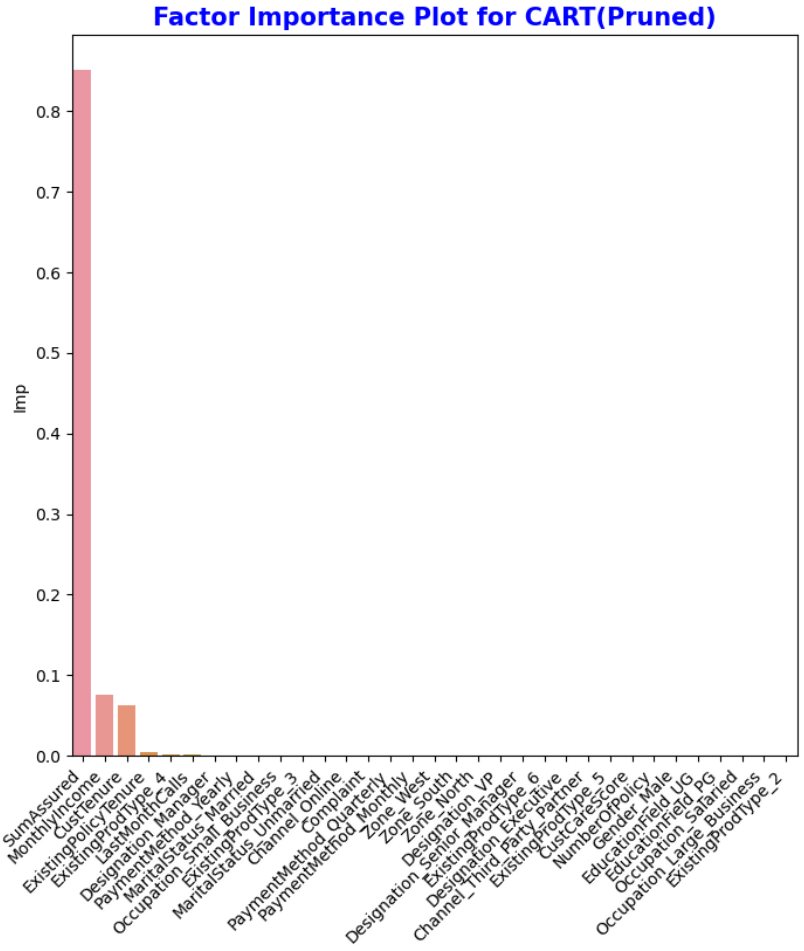


Figure O Feature Importance Plot - CART(Pruned)

Validation Against Test:

The CART (Pruned/Tuned) model is used for prediction.

The R-squared value for Test data is 0.78

The RMSE for CART Tuned model is 652.38.

The RMSE for tuned CART is decreased compared to the CART Model.

The tuned CART Model gives an better prediction than the base CART Model.

K-Neighbor Regressor

The K-Neighbor Regressor is tuned with GridSearchCV.

The Parameters are

'n_neighbors':list(range(1,21,2)),

'weights':['uniform','distance'],

'metric':['euclidean','chebyshev','manhattan','minkowski']

The Optimum tuned model is

KNeighborsRegressor(metric='manhattan', n_neighbors=17, weights='distance')

Even after tuning, K-Neighbor regressor is overfitting the train data.

The R-Squared value for train data is 1.

The RMSE for trained data model is 0.

Validation Against Test:

The K-neighbor Regressor model is used for prediction.

The R-squared value for Test data is 0.54

The RMSE for K-neighbor Regressor model is 940.38.

This is the least preferred model. It performs worse than the other models.

Random Forest Regressor

The Random Forest Regressor is used to predict the agent bonus.

The Parameters for Random Forest Regressor are

'criterion':['squared_error','friedman_mse', 'absolute_error', 'poisson'],

'n_estimators':list(range(100,1000,2)),'min_samples_leaf':list(range(1,10)),

'max_samples':list(np.arange(0.1,1))

The Optimum model after RandomizedSearchCV is

RandomForestRegressor(max_samples=0.1, min_samples_leaf=2, n_estimators=144)

The Feature Importance Plot for RF is similar to the CART Model

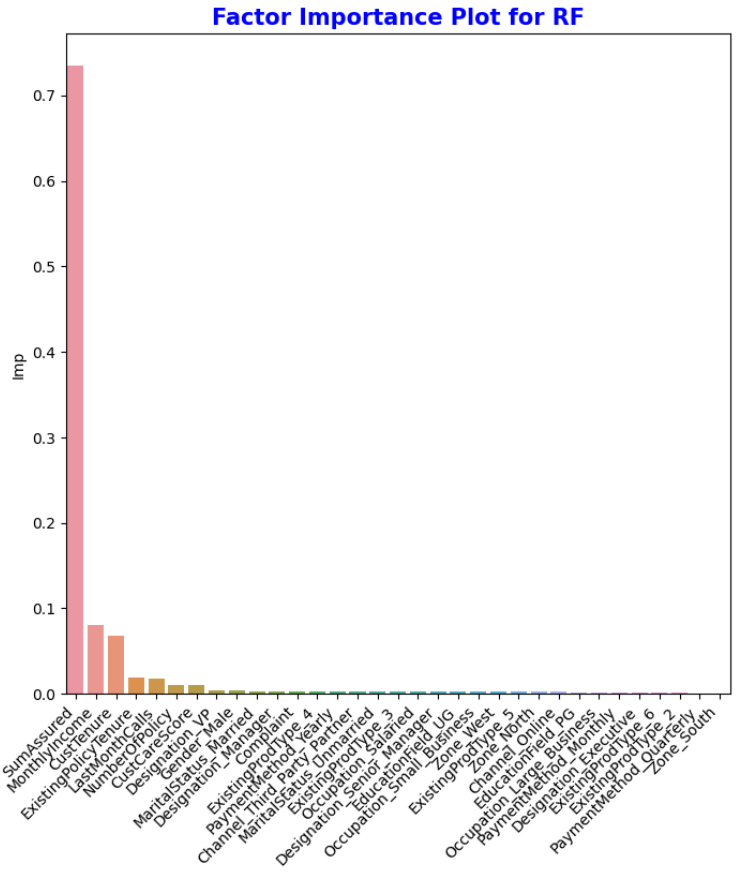


Figure P Feature Importance Plot - RF Regressor

Validation Against Test:

The Random Forest Regressor model is used for prediction.

The R-squared value for Test data is 0.80. The RMSE for Random Forest Regressor model is 623.79. The RF Regressor works better than the tuned CART model for prediction of dependent variable.

Ensemble modelling

Ensemble techniques are used to create a much better and accurate model to predict the agent bonus.

Bagging, Boosting and Heterogenous Ensemble models are trained below.

Bagging

The CART model is used as the base model for Bagging. The model is CART Bagging.

The parameter for CART Bagging with RandomizedSearchCV are

'n_estimators':list(range(100,500,2)),

'max_samples':list(np.arange(0.01,1,0.01)),

'max_features':list(np.arange(0.01,1,0.01))

Validation Against Test:

The CART Bagging model is used for prediction.

The R-squared value for Test data is 0.81

The RMSE for CART Bagging Regressor model is 597.75.

The CART Bagging improves the prediction.

XG Boosting Regressor

GradientBoostingRegressor is used for the XG boosting model.

The Parameters for RandomizedSearchCV are

'criterion':['friedman_mse', 'squared_error'],
'learning_rate':list(np.arange(0.01,0.02,0.0000005)),
'n_estimators':list(np.arange(100,1000,2))

The Optimim XGB model is

GradientBoostingRegressor(learning_rate=0.01817900000000818, n_estimators=844)

The Feature Plot is similar to RF Regressor with SumAssured, MonthlyIncome and CustTenure as most important.

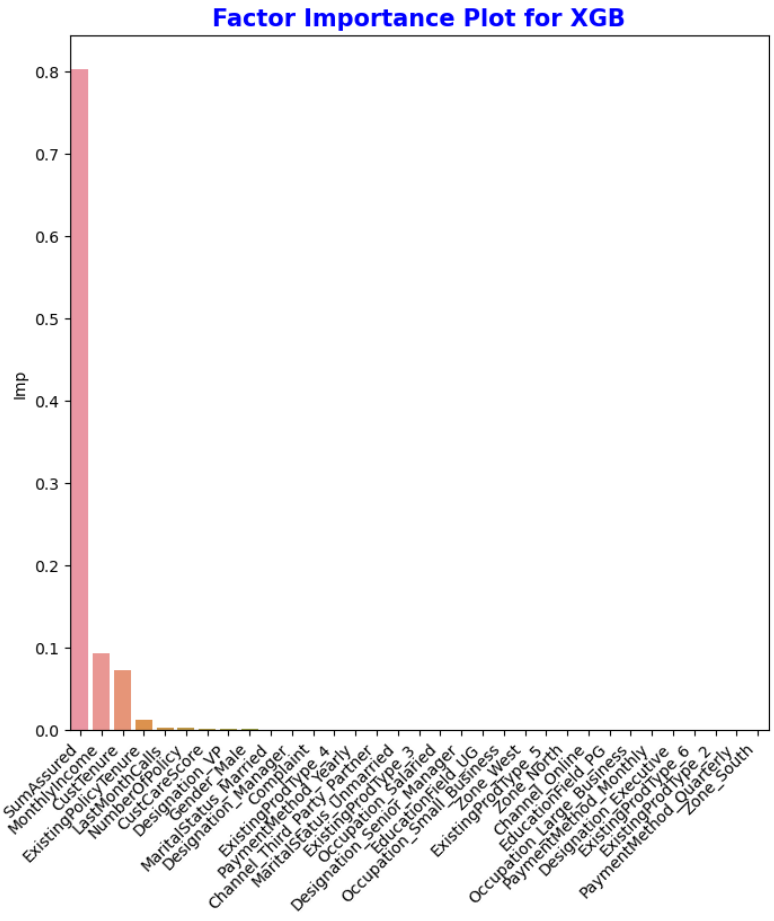


Figure Q Feature Importance Plot - XGB

Validation Against Test:

The XGB model is used for prediction. The R-squared value for Test data is 0.81. The RMSE for XG Boosting Regressor model is 608.45. XGB has the same R-Square value as CART Bagging but the RMSE is lower for CART Bagging.

ADA Boosting Regressor

AdaBoostRegressor is used for ADA Boosting.

The Parameter for RandomizedSearchCV are

'n_estimators':list(np.arange(100,500)),

'learning_rate':list(np.arange(0.1,0.5,0.00001))

The Tuned Model parameter is

AdaBoostRegressor(learning_rate=0. 16381999999997526, n_estimators=204)

Validation Against Test:

The ADA Boosting model is used for prediction.

The R-squared value for Test data is 0.75

The RMSE for ADA Boosting Regressor model is 694.25.

ADA Boosting is not as good as other Ensemble models.

Voting Regressor

The Voting Regressor is used to create heterogenous ensemble model.

The base model used for this are

Lasso regression, CART (Pruned/Tuned), Random Forest Regressor, XG Boosting Regressor, CART Bagging and ADA Boosting.

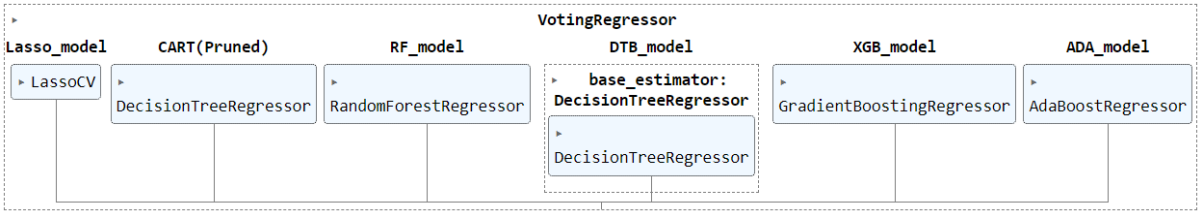


Figure R Voting Regressor Model

The Parameter for Voting Ensemble Regressor is

VotingRegressor(estimators=

[('Lasso_model',LassoCV(alphas=[0.0001, 0.001, 0.01, 0.1, 1, 10])), ('CART(Pruned)',DecisionTreeRegressor(criterion='friedman_mse', max_depth=32, min_impurity_decrease=0.0051, min_samples_leaf=24, min_samples_split=4)),

('RF_model', RandomForestRegressor(max_samples=0.1, min_samples_leaf=2, n_estimators=144)),

('DTB_model', BaggingRegressor(base_estimator=DecisionTreeRegressor(),max_features=0.97, max_samples=0.68, n_estimators=378)),

('XGB_model', GradientBoostingRegressor(learning_rate=0.01817900000000818, n_estimators=844)),

('ADA_model', AdaBoostRegressor(learning_rate=0.16381999999997526, n_estimators=204))],n_jobs=-1)

Validation Against Test:

The Voting Regressor model is used for prediction.

The R-squared value for Test data is 0.80

The RMSE for Voting Regressor model is 612.05.

The Weighted Voting regressor is used for improving the model performance.

Weighted Voting Regressor

The base model used for this are

Lasso regression, CART (Pruned/Tuned), Random Forest Regressor, XG Boosting Regressor, CART Bagging and ADA Boosting.

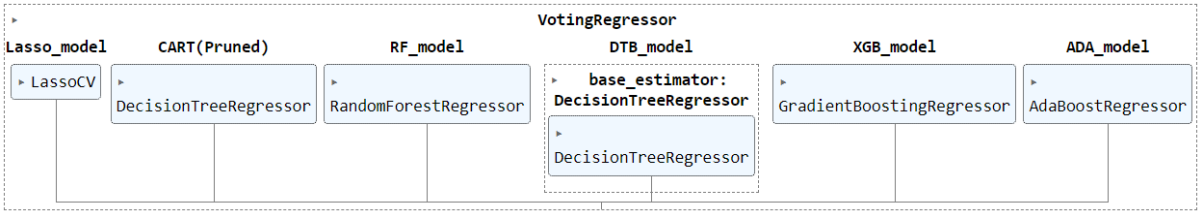


Figure S Weighted Voting Regressor Model

The Above Base model has an weightage of [3,3,5,5,5,3].
The Parameter for Weighted Voting regressor is

VotingRegressor(estimators=

[('Lasso_model',LassoCV(alphas=[0.0001, 0.001, 0.01, 0.1, 1, 10])),
('CART(Pruned)',DecisionTreeRegressor(criterion='friedman_mse', max_depth=32, min_impurity_decrease=0.0051, min_samples_leaf=24, min_samples_split=4)),

('RF_model', RandomForestRegressor(max_samples=0.1, min_samples_leaf=2, n_estimators=144)),

('DTB_model', BaggingRegressor(base_estimator=DecisionTreeRegressor(),max_features=0.97, max_samples=0.68, n_estimators=378)),

('XGB_model', GradientBoostingRegressor(learning_rate=0.01817900000000818, n_estimators=844)),

('ADA_model', AdaBoostRegressor(learning_rate=0.16381999999997526, n_estimators=204))],n_jobs=-1, weights=[3, 3, 5, 5, 5, 3])

Validation Against Test:

The Weighted Voting Regressor model is used for prediction.

The R-squared value for Test data is 0.81

The RMSE for Weighted Voting Regressor model is 608.44.

The Weighted Voting regressor improves the model performance but not with lower RMSE compared with other model (Mainly with CART Bagging).

Stacking Regressor

The Stacking Regressor is used to get better ensemble model.

The Base model used are

Lasso regression, CART (Pruned/Tuned), Random Forest Regressor, XG Boosting Regressor, CART Bagging and ADA Boosting.

The Final Estimater model is Linear Regression.

It stacks the output of individual estimator and use a regressor to compute the final prediction.

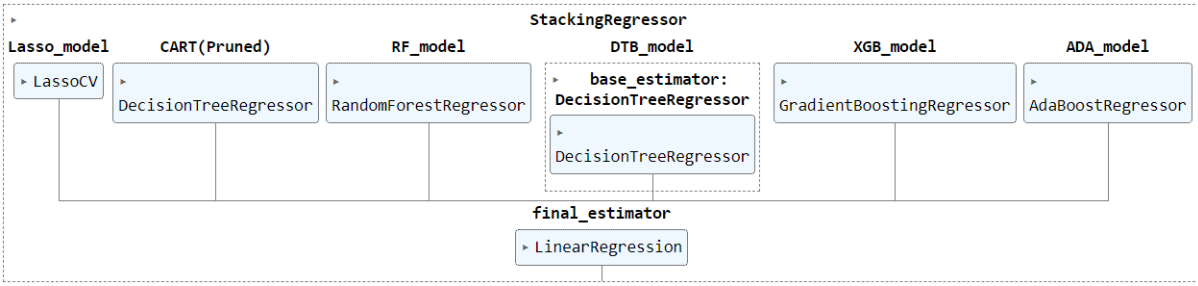


Figure T Stacking Regressor Model

The Parameter for Stacking Ensemble Regressor are

StackingRegressor(estimators=

[('Lasso_model', LassoCV(alphas=[0.0001, 0.001, 0.01, 0.1, 1, 10])),

('CART(Pruned)', DecisionTreeRegressor(criterion='friedman_mse', max_depth=32, min_impurity_decrease=0.0051, min_samples_leaf=24, min_samples_split=4)),

('RF_model',RandomForestRegressor(max_samples=0.1,min_samples_leaf=2, n_estimators=144)),

('DTB_model',BaggingRegressor(base_estimator=DecisionTreeRegressor(),max_features=0.97, max_samples=0.68, n_estimators=378)),

('XGB_model',GradientBoostingRegressor(learning_rate=0.01817900000000818, n_estimators=844)),

('ADA_model',AdaBoostRegressor(learning_rate=0.16381999999997526, n_estimators=204))],

final_estimator=LinearRegression(), n_jobs=-1)

Validation Against Test:

The Stacking Regressor model is used for prediction.

The R-squared value for Test data is 0.81

The RMSE for Stacking Regressor model is 595.57.

Interpretation of the most optimum model

The R-Squared and RMSE are compared across the model to find the best optimum model.

	Model	R-Squared(%)	MAE	MSE	MAPE	RMSE	Max Error
0	Linear Regression	76.90	524.81	450935.45	0.13	671.52	2610.78
1	Lasso Regression	77.14	523.55	446259.25	0.14	668.03	2642.07
2	Ridge Regression	77.11	523.89	446805.61	0.14	668.44	2640.16
3	SGD Regression	77.03	524.36	448407.12	0.14	669.63	2698.12
4	CART	65.42	602.07	675099.60	0.16	821.64	4313.00
5	CART(Pruned)	78.20	504.03	425608.81	0.13	652.39	3379.93
6	K-Neighbors Regressor	54.70	763.75	884328.55	0.20	940.39	3691.89
7	Random Forest Regressor	80.07	490.59	389114.07	0.13	623.79	2754.00
8	CART-Bagging	81.70	462.61	357314.97	0.12	597.76	2666.33
9	XG-Boosting	81.04	474.57	370214.41	0.12	608.45	2618.39
10	ADA Boosting	75.31	573.02	481994.88	0.15	694.26	2268.86
11	Voting Regressor	80.81	483.15	374605.96	0.13	612.05	2675.87
12	Voting Regressor(Weighted)	81.04	479.80	370204.38	0.13	608.44	2681.54
13	Stacking Regressor	81.83	459.97	354706.89	0.12	595.57	2630.49

Table 14 Performance Metrics

The Stacking Regressor with 6 Base model as Lasso regression, CART (Pruned/Tuned), Random Forest Regressor, XG Boosting Regressor, CART Bagging, ADA Boosting and Final Estimator as Linear Regression has low RMSE and High R-Squared Value.

81.81% of all movements of dependent variable are completely explained by movements in the other independent variables.

CART Can be used as it the second best model with 81.70% R-Squared value and it wont require any base models as stacking regressor.

From the models, the most important features are

SumAssured, MonthlyIncome And CustTenure.

Implication & Recommendation on the business

- The Agent Bonus is determined from SumAssured of the Policy, Monthly income of the Customer and Customer Tenure with their organisation.
- The Agents sells the product to the customers and the amount paid by the customers is the main source of income for the company. Higher the SumAssured, Higher will be bonus and profit for the company to invest that amount elsewhere till the policy matures. This benefits both agents and company.
- To Motivate agents to sell high Sum Assured policy, The Sum Assured of a policy and their respective Agent bonus table can be created from the models and that can be used to boost the agents to sell higher sum assured policies.
- By predicting the bonus for its agents, the insurance company can classify the agents into different bonus level category such as High, Medium & Low for Training.
- It will also help in finding the most important agents that are required for the company development if the agent's details are provided which can be done as future scope of the study.
- The Insurance Company can also strategize itself to find the best optimum bonus for its employees and reducing the cost to increase its profit.