GREAT LEARNING – PGPDSBA

LI_BFSI_01 - Life Insurance Sales

Source: https://www.istockphoto.com/photo/insurance-protecting-family-health-live-house-and-car-concept-gm1199060494-342911158

Batch: PGPDSBA.O.AUG22.B

Name: Saurabh Girpunje

Contents:

Index:

Table of Figures:

1. Introduction

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.

1.1. Defining problem statement:

The life insurance company wants to predict how much bonus to give its agents. This will help them create special activities for agents who perform well and provide training for agents who do not perform as well. We have a dataset with information about the agents' performance, sales history, customer satisfaction, and other relevant details. Overall, the company's goal is to improve agent performance, increase sales, make customers happier, and make more money by using the model's predictions to allocate resources and investments effectively.

1.2. Need of the study/project:

Overall, the study/project of predicting bonuses for agents in a leading life insurance company helps evaluate performance, motivate, and engage high-performing agents, improve the performance of low-performing agents, manage talent effectively, and make data-driven decisions for fair compensation and rewards.

1.3. Understanding business/social opportunity:

Life insurance companies help people, organizations, and the economy in many ways. They keep our money safe, help us maintain a good life, and give us a feeling of safety and calmness. They also teach us how to prevent losses, make us more successful, and help us understand the dangers and results of risks through education.

2. EDA and Business Implication:

2.1. Data Dictionary:

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

AgentBonus is the target variable and it is of continuous datatype. This problem belongs to the Regression analysis.

2.2. Checking rows and columns in dataset:

```
Total number of rows in dataset are 4520 and columns are 20
```

2.3. Getting top 5 rows in dataset:

	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	Month
0	7000000	4409	22.0	4.0	Agent	Salaried	Graduate	Female	3	Manager	2.0	Single	
1	7000001	2214	11.0	2.0	Third Party Partner	Salaried	Graduate	Male	4	Manager	4.0	Divorced	
2	7000002	4273	26.0	4.0	Agent	Free Lancer	Post Graduate	Male	4	Exe	3.0	Unmarried	
3	7000003	1791	11.0	NaN	Third Party Partner	Salaried	Graduate	Fe male	3	Executive	3.0	Divorced	
4	7000004	2955	6.0	NaN	Agent	Small Business	UG	Male	3	Executive	4.0	Divorced	

2.4. Getting last 5 rows in dataset:

	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	Mc
4515	7004515	3953	4.0	8.0	Agent	Small Business	Graduate	Male	4	Senior Manager	2.0	Single	
4516	7004516	2939	9.0	9.0	Agent	Salaried	Under Graduate	Female	2	Executive	2.0	Married	
4517	7004517	3792	23.0	23.0	Agent	Salaried	Engineer	Female	5	AVP	5.0	Single	
4518	7004518	4816	10.0	10.0	Online	Small Business	Graduate	Female	4	Executive	2.0	Single	
4519	7004519	4764	14.0	10.0	Agent	Salaried	Under Graduate	Female	5	Manager	2.0	Married	

2.5. Getting information about the dataset:

```
<class 'pandas.core.frame.DataFrame'>
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)
```

2.6. Describing dataset:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
CustID	4520.00	NaN	NaN	NaN	7002259.50	1304.96	7000000.00	7001129.75	7002259.50	7003389.25	7004519.00
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

Insights:

- There are 7 variables of float datatype, 5 variables of int datatype and 8 variables of object datatype.
- CustID variable can be drop as this only includes customer id and will be of no use in building model and further analysis.
- Age of customer varies from 2 years to 58 years. This shows that usually customer with 2 years cannot take insurance by himself/herself. It must be insurance taken by the customer for their children and that should be age of children.
- There are null values in some of the columns. Need to treat them.
- There are no duplicate values in the dataset.
- The AgentBonus includes information about the bonus amounts for agents. The mean bonus is approximately 4,077.84, with a standard deviation of 1,403.32. The minimum bonus is 1,605, and the maximum bonus is 9,608.
- There are 3 Gender unique values present. Need to verify the values.
- The values present varies in scale. So, need of scaling before building model.

2.7. Univariate Analysis:

1] Agent Bonus:

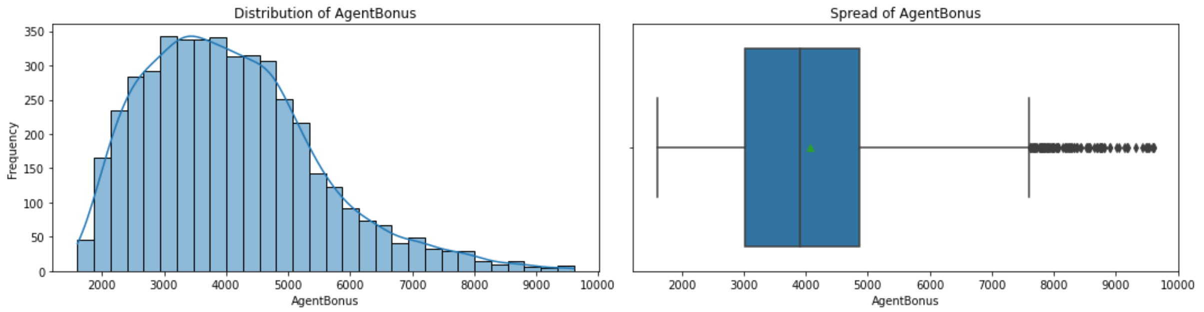


Figure 1: Univariate Analysis of Agent Bonus

It is right skewed distribution. Here, mean value is greater than median value. Most of the Agent got bonus in the median range. Outliers are also present in a large extent. Some of the Agent are performing well so they are receiving higher amount of Bonus value.

2] Age:

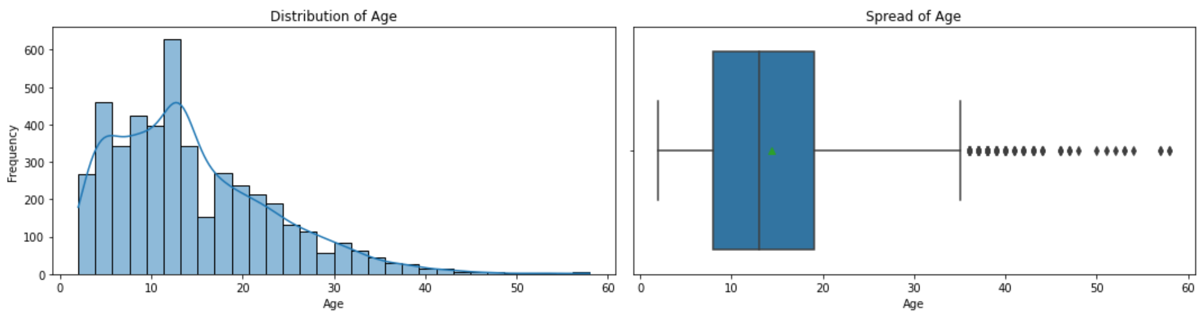


Figure 2: Univariate Analysis of Age

Most of the age of the customer are seen to be between 8 years to 20 years. This seems that customers are taking insurance for their children. Agent should focus on this strategy to sell more insurance or target the customer by introducing new schemes for children insurance.

3] CustTenure:

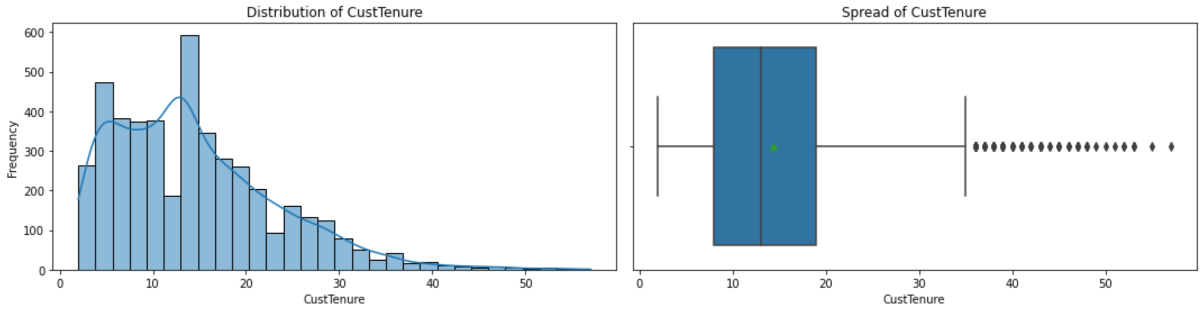


Figure 3: Univariate Analysis of CustTenure

It is been observed that most of the customers have opted for the insurance for maximum of the 10-20 years. It is seen that customers are avoiding to take insurance for long duration of time.

4] Monthly Income:

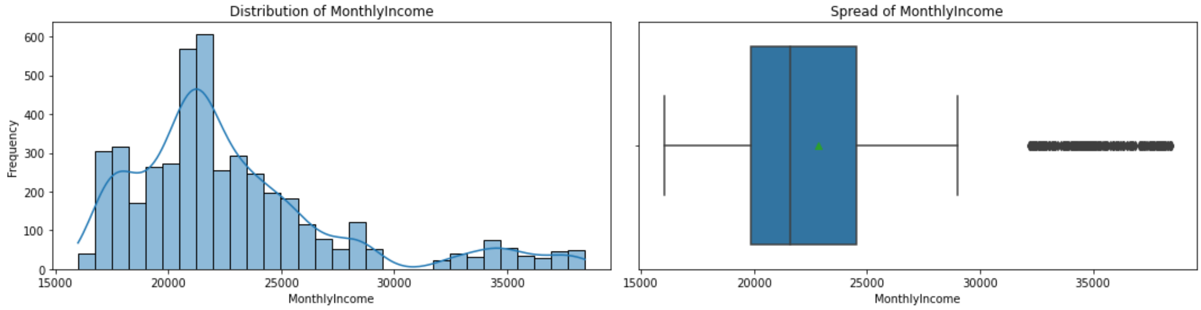
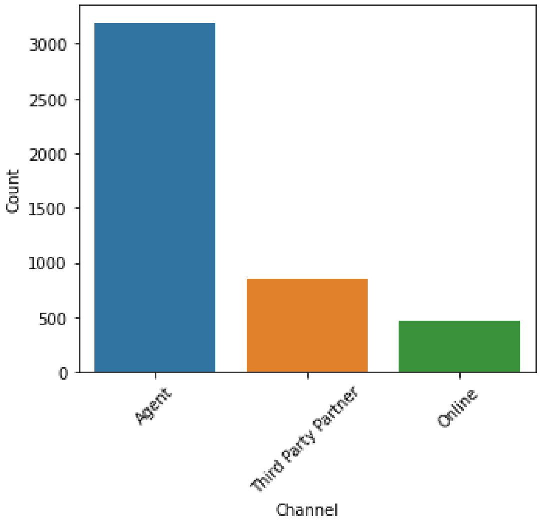


Figure 4:Univariate Analysis of Monthly Income

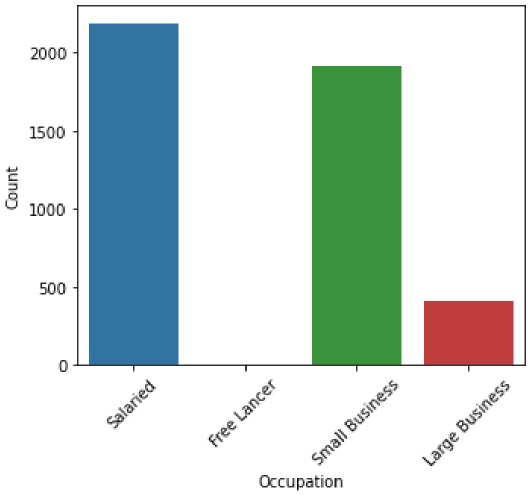
Most of the customers have the monthly income between 20000 to 25000.

5] Channel:



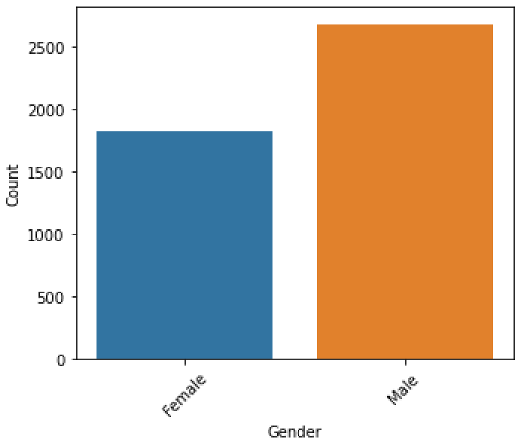
Most Channel used for selling insurance is through Agent only. Online is the least Channel used by Customer's. So, company should focus on giving bonus to Agent's so to acquire more and more customers.

6] Occupation:



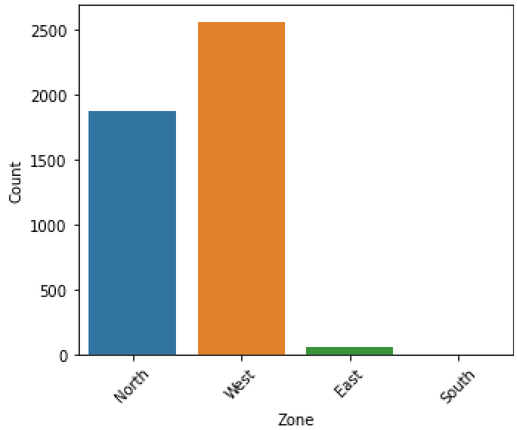
It is seen that Customers who are Salaried are taking more number of Insurance. Free Lancer customers are almost negligible. Customers with Small Business are also taking Insurance.

7] Gender:



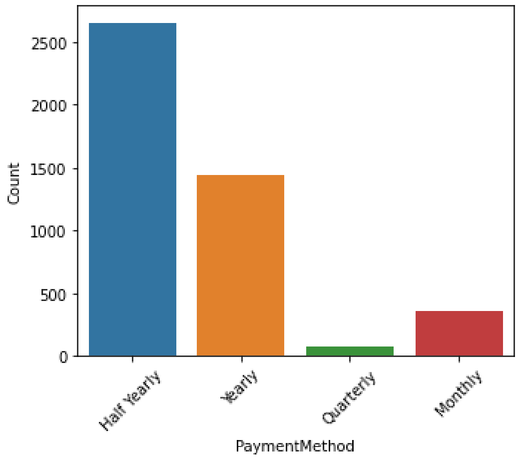
Male are dominating.

8] Zone:



It is clearly seen that South zone is neglected as there are almost negligible customers who are taking insurance. West zone is highest is taking insurance. The company should increase number of Agent's in the East and South zone and need to tell benefits of the taking insurance.

9] Payment Method:



Most of the Customers are doing their payment method as Half Yearly. Different schemes and benefits should be introduced on the Yearly as well as Half yearly payments.

2.8. Bivariate Analysis:

1] Agent Bonus vs Occupation:

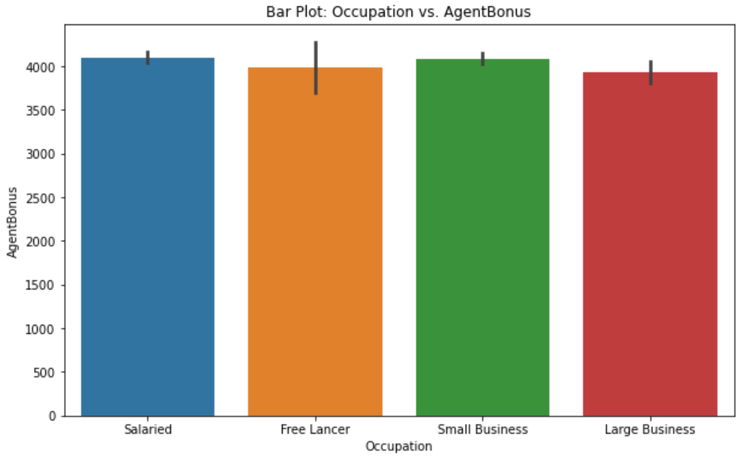


Figure 5: Bivariate Analysis of Agent Bonus vs Occupation

Agent receives more Bonus for subscription of customers with Large Business. Almost bonus is common from all type of Occupation.

2] Agent Bonus vs Designation:

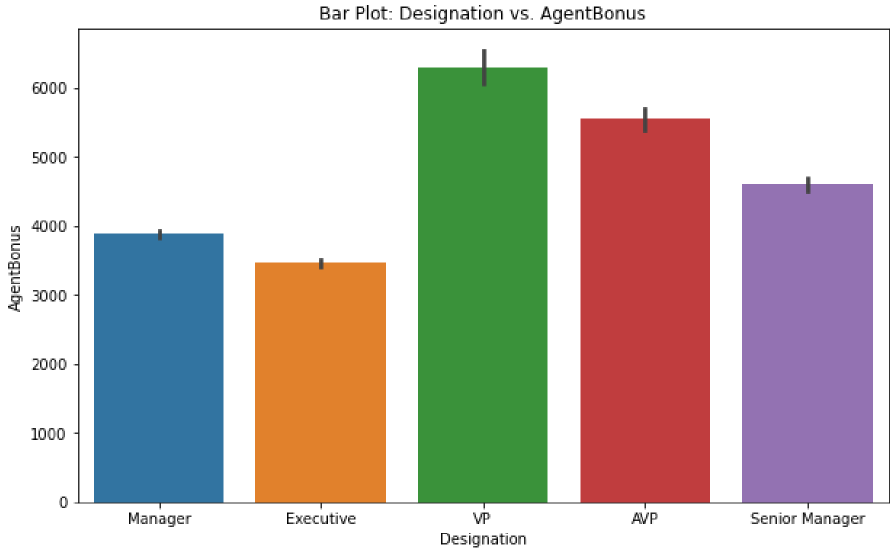


Figure 6: Bivariate Analysis of Agent Bonus vs Designation

Most of the Bonus received to Agent is through VP and AVP. Might due to high salary also increases rating of the Agent and might increases Bonus.

3] Agent Bonus vs Zone:

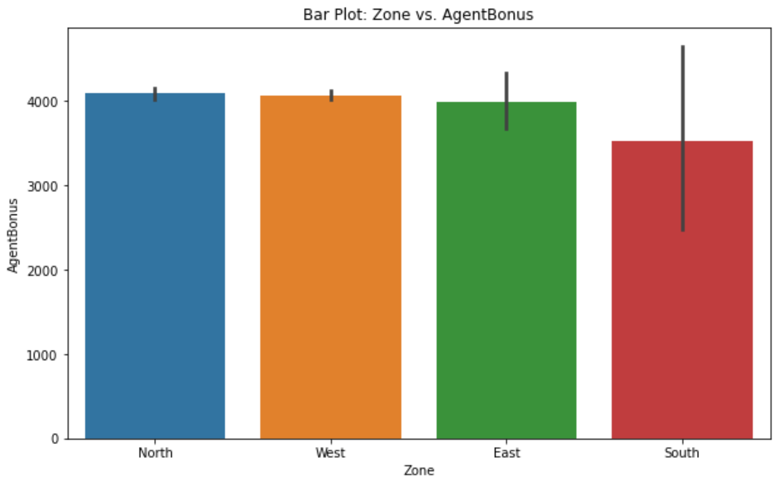


Figure 7: Bivariate Analysis of Agent Bonus vs Zone

Most of the bonus achieved is from North, West and East zone.

4] AgentBonus, Designation and Payment method:

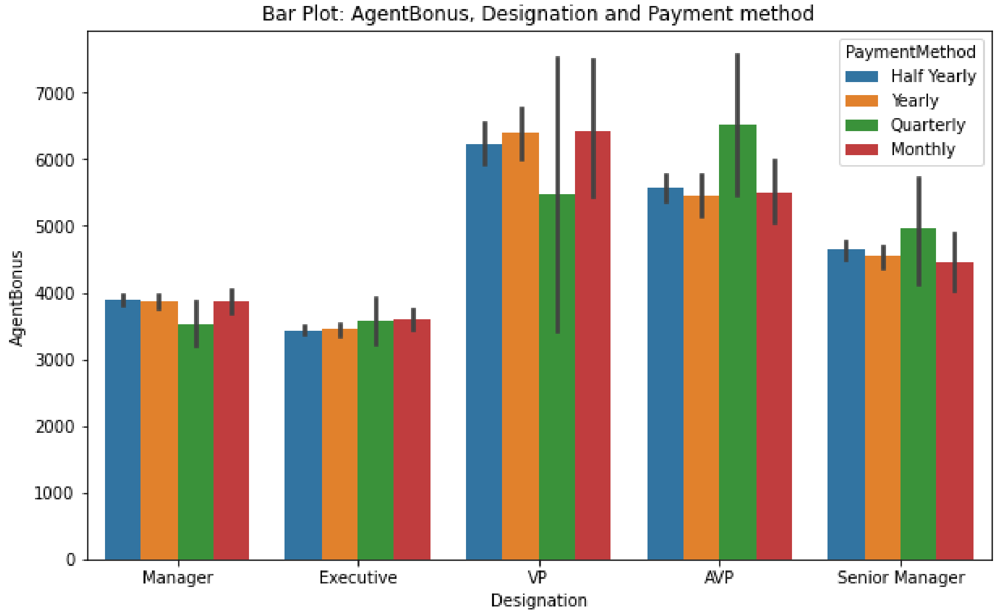


Figure 8: AgentBonus, Designation and Payment method

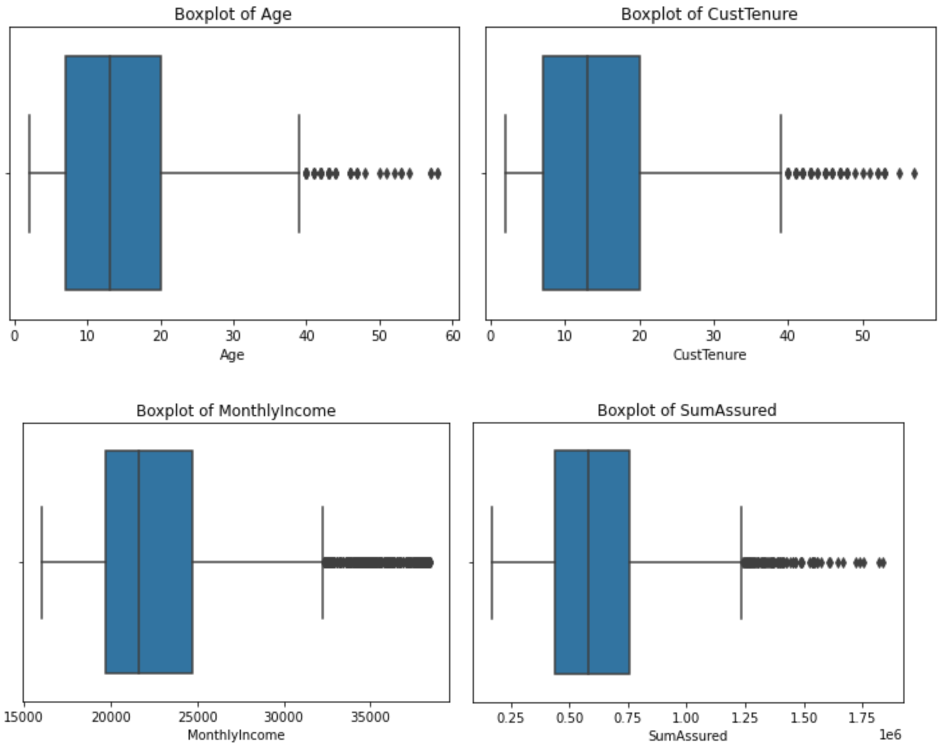
Most bonus income are from VP who pay monthly and yearly, and AVP who pay quarterly.

3. Data Cleaning and Pre-processing

3.1. Outlier Treatment:

We had used Inter Quantile Range (IQR) for the outlier treatment. The values which are below 25th percentile of the data, are treated as LL (Lower limit) and the values which are above the 75th percentile of the data, are treated as UL (Upper limit).

Outlier's are present in some of the columns:



- It is possible to have the age more than 52, as this is the true outliers, so we are not removing outliers. Also, changing the data is also not relevant as this may cause discrepancy in the dataset.
- It is possible to have Income more than 33000 per month. So, changing data is not relevant.

3.2. Irrelevant variable:

We had removed CustID column as this column is of no use. It only contains ID of the customer which may have issue while building model.

3.3. Filling Null values:

Only 1.35% data is missing, so we can fill this data. So, we are using Mean, Median, Mode method to treat missing values here. As, missing values are present in the continuous variables are most of that variables have outliers so it is relevant to fill these null values with using median. After treating Null values, now there are 0 null values.

```
AgentBonus              0
Age                     0
CustTenure              0
Channel                 0
Occupation              0
EducationField          0
Gender                  0
ExistingProdType        0
Designation             0
NumberOfPolicy          0
MaritalStatus           0
MonthlyIncome           0
Complaint               0
ExistingPolicyTenure    0
SumAssured              0
Zone                    0
PaymentMethod           0
LastMonthCalls          0
CustCareScore           0
dtype: int64
```

3.4. Data Cleaning:

We had segregated data into Numerical and Categorical datatypes:

```
Agent                   3194
Third Party Partner      858
Online                   468
Name: Channel, dtype: int64
 ----------------------
Salaried            2192
Small Business      1918
Large Business       255
Laarge Business      153
Free Lancer            2
Name: Occupation, dtype: int64
 ----------------------
Graduate            1870
Under Graduate      1190
Diploma              496
Engineer             408
Post Graduate        252
UG                   230
MBA                   74
Name: EducationField, dtype: int64
 ----------------------
Male        2688
Female      1507
Fe male      325
Name: Gender, dtype: int64
 ----------------------
```

```
Manager           1620
Executive         1535
Senior Manager     676
AVP                336
VP                 226
Exe                127
Name: Designation, dtype: int64
 ----------------------
Married      2268
Single       1254
Divorced      804
Unmarried     194
Name: MaritalStatus, dtype: int64
 ----------------------
West     2566
North    1884
East       64
South       6
Name: Zone, dtype: int64
 ----------------------
Half Yearly    2656
Yearly         1434
Monthly         354
Quarterly        76
Name: PaymentMethod, dtype: int64
 ----------------------
```

- It is observed that many categorical columns have incorrect data.
- In Occupation column, Large Business is repeated two times and can be combined into same column.
- In Education field, UG and Under Graduate can be combined into one.
- In Gender, Female category is incorrectly spelled so it is also combined into one.
- In Designation, Exe and Executive can be combined into one.
- In MaritalStatus, Single and Unmarried can be combined into one as both meaning same.

After fixing columns data, it is shown as below.

```
Manager           1620
Executive         1535
Senior Manager     676
AVP                336
VP                 226
Exe                127
Name: Designation, dtype: int64
 ----------------------
Married      2268
Single       1254
Divorced      804
Unmarried     194
Name: MaritalStatus, dtype: int64
 ----------------------
West     2566
North    1884
East       64
South       6
Name: Zone, dtype: int64
 ----------------------
Half Yearly    2656
Yearly         1434
Monthly         354
Quarterly        76
Name: PaymentMethod, dtype: int64
 ----------------------
```

```
Agent                  3194
Third Party Partner     858
Online                  468
Name: Channel, dtype: int64
 ----------------------
Salaried          2192
Small Business    1918
Large Business     255
Laarge Business    153
Free Lancer          2
Name: Occupation, dtype: int64
 ----------------------
Graduate         1870
Under Graduate   1190
Diploma           496
Engineer          408
Post Graduate     252
UG                230
MBA                74
Name: EducationField, dtype: int64
 ----------------------
Male      2688
Female    1507
Fe male    325
Name: Gender, dtype: int64
 ----------------------
```

3.5. Encoded Data:

We had used One-Hot Encoding here as One-Hot encoding is used to represent categorical data in a binary format, where each category is represented by a unique binary vector with all elements as zeros except for one, indicating the category's presence.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 34 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   AgentBonus                   4520 non-null   int64  
 1   Age                          4520 non-null   float64
 2   CustTenure                   4520 non-null   float64
 3   ExistingProdType             4520 non-null   int64  
 4   NumberOfPolicy               4520 non-null   float64
 5   MonthlyIncome                4520 non-null   float64
 6   Complaint                    4520 non-null   int64  
 7   ExistingPolicyTenure         4520 non-null   float64
 8   SumAssured                   4520 non-null   float64
 9   LastMonthCalls               4520 non-null   int64  
 10  CustCareScore                4520 non-null   float64
 11  Channel_Online               4520 non-null   uint8  
 12  Channel_Third Party Partner  4520 non-null   uint8  
 13  Occupation_Large Business    4520 non-null   uint8  
 14  Occupation_Salaried          4520 non-null   uint8  
 15  Occupation_Small Business    4520 non-null   uint8  
 16  EducationField_Engineer      4520 non-null   uint8  
 17  EducationField_Graduate      4520 non-null   uint8  
 18  EducationField_MBA           4520 non-null   uint8  
 19  EducationField_Post Graduate 4520 non-null   uint8  
 20  EducationField_Under Graduate 4520 non-null  uint8  
 21  Gender_Male                  4520 non-null   uint8  
 22  Designation_Executive        4520 non-null   uint8  
 23  Designation_Manager          4520 non-null   uint8  
 24  Designation_Senior Manager   4520 non-null   uint8  
 25  Designation_VP               4520 non-null   uint8  
 26  MaritalStatus_Married        4520 non-null   uint8  
 27  MaritalStatus_Unmarried      4520 non-null   uint8  
 28  Zone_North                   4520 non-null   uint8  
 29  Zone_South                   4520 non-null   uint8  
 30  Zone_West                    4520 non-null   uint8  
 31  PaymentMethod_Monthly        4520 non-null   uint8  
 32  PaymentMethod_Quarterly      4520 non-null   uint8  
 33  PaymentMethod_Yearly         4520 non-null   uint8  
dtypes: float64(7), int64(4), uint8(23)
```

We had 7 float variable, 5 int64 datatype variable and remaining 23 encoded variables. We had drop first while using one-hot encoding.

3.6. Heatmap:

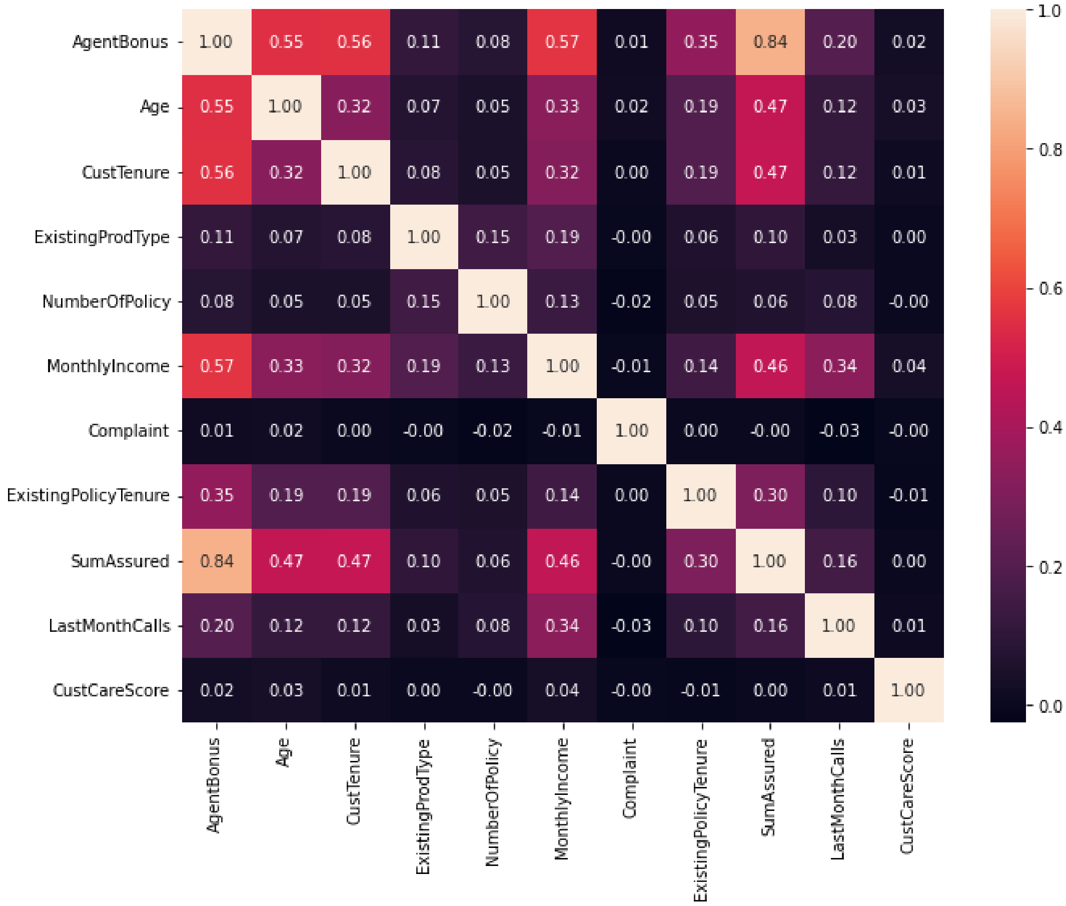


Figure 9: Correlation Heatmap

- There is no much correlation between the variables in the dataset.
- Only SumAssured is highly correlated with AgentBonus with ratio value of 0.84
- There is no correlation between Complaint and CustTenure, ExistingProdType.

3.7. Variable transformation:

We are using One-Hot encoding here to transform the categorical variables into numerical to build model. We are using one hot encoding here because data is not ordinal. The variables provided in the dataset, such as "Channel," "Occupation," "EducationField," "Gender," "Designation," "MaritalStatus," "Zone," and "PaymentMethod," do not have a natural ordering or inherent numerical relationship. These variables represent different categories or classes without any specific rank or order. One-hot encoding will expand the dimensionality of the dataset, creating additional binary columns for each unique category within each variable. We are also dropping first here.

```
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   AgentBonus                      4520 non-null   int64  
 1   Age                             4520 non-null   float64
 2   CustTenure                      4520 non-null   float64
 3   ExistingProdType                4520 non-null   int64  
 4   NumberOfPolicy                  4520 non-null   float64
 5   MonthlyIncome                   4520 non-null   float64
 6   Complaint                       4520 non-null   int64  
 7   ExistingPolicyTenure            4520 non-null   float64
 8   SumAssured                      4520 non-null   float64
 9   LastMonthCalls                  4520 non-null   int64  
 10  CustCareScore                   4520 non-null   float64
 11  Channel_Online                  4520 non-null   uint8  
 12  Channel_Third Party Partner     4520 non-null   uint8  
 13  Occupation_Large Business       4520 non-null   uint8  
 14  Occupation_Salaried             4520 non-null   uint8  
 15  Occupation_Small Business       4520 non-null   uint8  
 16  EducationField_Engineer         4520 non-null   uint8  
 17  EducationField_Graduate         4520 non-null   uint8  
 18  EducationField_MBA              4520 non-null   uint8  
 19  EducationField_Post Graduate    4520 non-null   uint8  
 20  EducationField_Under Graduate   4520 non-null   uint8  
 21  Gender_Male                     4520 non-null   uint8  
 22  Designation_Executive           4520 non-null   uint8  
 23  Designation_Manager             4520 non-null   uint8  
 24  Designation_Senior Manager      4520 non-null   uint8  
 25  Designation_VP                  4520 non-null   uint8  
 26  MaritalStatus_Married           4520 non-null   uint8  
 27  MaritalStatus_Unmarried         4520 non-null   uint8  
 28  Zone_North                      4520 non-null   uint8  
 29  Zone_South                      4520 non-null   uint8  
 30  Zone_West                       4520 non-null   uint8  
 31  PaymentMethod_Monthly           4520 non-null   uint8  
 32  PaymentMethod_Quarterly         4520 non-null   uint8  
 33  PaymentMethod_Yearly            4520 non-null   uint8  
dtypes: float64(7), int64(4), uint8(23)
```

Insights:

The data is unbalanced. It is seen that South zone has almost negligible data that means Customers are also very few from South zone. More data should be collected from the South zone.
The data for the "Zone" variable is not evenly distributed. Most of the data is from the West and North zones, while there are very few observations from the East and South zones. This imbalance can make it harder to accurately analyse or model the data. To address this, we can try techniques like creating more samples for the underrepresented zones, collecting additional data for those zones, or considering the unique characteristics of each zone in the analysis.
We can try techniques like creating more samples of the less common category, reducing the samples of the more common categories, or using special methods that handle imbalanced data. We also need to choose evaluation metrics that consider the imbalance.

4. Model building

4.1. Train and Test Dataset

We had divided the dataset into train and test with 70-30 train-test split.

After splitting dataset into Train and Test, we have 3164 rows and 34 columns in Train dataset and 1356 rows and 34 columns in Test dataset.

4.2. RFE (Recursive feature elimination):

Recursive feature elimination (RFE) is a feature selection algorithm that works by iteratively removing the least important features from a dataset. The importance of each feature is determined by a scoring function, such as the coefficient of determination (R2) or the mean squared error (MSE).
Applying RFE on Linear Regression on whole 33 features initially to observe p-value and VIF values.

Ranking is given based on the RFE.

```
[('const', False, 2),
 ('Age', True, 1),
 ('CustTenure', True, 1),
 ('ExistingProdType', True, 1),
 ('NumberOfPolicy', True, 1),
 ('MonthlyIncome', True, 1),
 ('Complaint', True, 1),
 ('ExistingPolicyTenure', True, 1),
 ('SumAssured', True, 1),
 ('LastMonthCalls', True, 1),
 ('CustCareScore', True, 1),
 ('Channel_Online', True, 1),
 ('Channel_Third Party Partner', True, 1),
 ('Occupation_Large Business', True, 1),
 ('Occupation_Salaried', True, 1),
 ('Occupation_Small Business', True, 1),
 ('EducationField_Engineer', True, 1),
 ('EducationField_Graduate', True, 1),
 ('EducationField_MBA', True, 1),
 ('EducationField_Post Graduate', True, 1),
 ('EducationField_Under Graduate', True, 1),
 ('Gender_Male', True, 1),
 ('Designation_Executive', True, 1),
 ('Designation_Manager', True, 1),
 ('Designation_Senior Manager', True, 1),
 ('Designation_VP', True, 1),
 ('MaritalStatus_Married', True, 1),
 ('MaritalStatus_Unmarried', True, 1),
 ('Zone_North', True, 1),
 ('Zone_South', True, 1),
 ('Zone_West', True, 1),
 ('PaymentMethod_Monthly', True, 1),
 ('PaymentMethod_Quarterly', True, 1),
 ('PaymentMethod_Yearly', True, 1)]
```

4.3. VIF:

VIF stands for "Variance Inflation Factor." It is a metric used in statistical analysis and regression modeling to assess the severity of multicollinearity among the predictor variables. VIF measures how much the variance of an estimated regression coefficient increases when the predictor variable is added to a model compared to when it is not included, helping to identify potential issues of multicollinearity that can affect the reliability of regression results.

After passing the arbitrary selected columns by RFE we will manually evaluate each models p-value and VIF value. Unless we find the acceptable range for p-values and VIF we keep dropping the variables one at a time based on below criteria.

High p-value High VIF : Drop the variable
High p-value Low VIF or Low p-value High VIF : Drop the variable with high p-value first
Low p-value Low VIF : accept the variable

Checking VIF
Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. We will initially checking vif for continuous features only.

	Features	VIF
7	SumAssured	1.71
4	MonthlyIncome	1.47
0	Age	1.34
1	CustTenure	1.31
8	LastMonthCalls	1.15
6	ExistingPolicyTenure	1.11
2	ExistingProdType	1.05
3	NumberOfPolicy	1.04
5	Complaint	1.00
9	CustCareScore	1.00

4.4. Model Building:

4.4.1. Linear Regression Model:

So, we had build first model with Linear Regression and observed p-value and vif values

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             AgentBonus   R-squared:                       0.807
Model:                            OLS   Adj. R-squared:                  0.805
Method:                 Least Squares   F-statistic:                     397.6
Date:                Thu, 10 Aug 2023   Prob (F-statistic):               0.00
Time:                        19:09:05   Log-Likelihood:                -24793.
No. Observations:                3164   AIC:                         4.965e+04
Df Residuals:                    3130   BIC:                         4.986e+04
Df Model:                          33
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                           5020.8320    462.836     10.848      0.000    4113.339    5928.325
Age                              197.0576     12.725     15.485      0.000     172.107     222.009
CustTenure                       215.1578     12.602     17.073      0.000     190.449     239.867
ExistingProdType                  57.0775     22.799      2.504      0.012      12.375     101.780
NumberOfPolicy                    -4.0600     11.515     -0.353      0.724     -26.638      18.518
MonthlyIncome                    152.8857     24.021      6.365      0.000     105.787     199.985
Complaint                         16.7452     11.006      1.521      0.128      -4.835      38.325
ExistingPolicyTenure             118.3257     11.568     10.228      0.000      95.643     141.008
SumAssured                       816.2145     14.453     56.474      0.000     787.876     844.553
LastMonthCalls                   -13.4599     12.090     -1.113      0.266     -37.166      10.246
CustCareScore                      4.3513     11.068      0.393      0.694     -17.350      26.052
Channel_Online                    50.1147     36.885      1.359      0.174     -22.206     122.435
Channel_Third Party Partner        4.9278     28.487      0.173      0.863     -50.927      60.783
Occupation_Large Business       -505.3527    468.929     -1.078      0.281   -1424.793     414.087
Occupation_Salaried             -473.0792    439.184     -1.077      0.281   -1334.198     388.039
Occupation_Small Business       -558.1736    449.071     -1.243      0.214   -1438.678     322.331
EducationField_Engineer          -41.2778    177.964     -0.232      0.817    -390.217     307.661
EducationField_Graduate          -63.9115    100.069     -0.639      0.523    -260.119     132.296
EducationField_MBA                22.9842    135.252      0.170      0.865    -242.207     288.176
EducationField_Post Graduate    -116.1063    110.119     -1.054      0.292    -332.019      99.807
EducationField_Under Graduate     11.2879     39.590      0.285      0.776     -66.337      88.913
Gender_Male                       12.8507     22.614      0.568      0.570     -31.490      57.191
Designation_Executive           -464.9181     64.783     -7.177      0.000    -591.940    -337.896
Designation_Manager             -440.9625     55.340     -7.968      0.000    -549.468    -332.457
Designation_Senior Manager      -270.3271     52.431     -5.156      0.000    -373.130    -167.524
Designation_VP                   -30.5612     72.188     -0.423      0.672    -172.101     110.979
MaritalStatus_Married            -45.1682     30.746     -1.469      0.142    -105.452      15.115
MaritalStatus_Unmarried           -2.8193     32.929     -0.086      0.932     -67.384      61.746
Zone_North                        -6.5860     93.942     -0.070      0.944    -190.781     177.609
Zone_South                       204.4239    322.036      0.635      0.526    -426.999     835.847
Zone_West                         -2.4516     93.488     -0.026      0.979    -185.755     180.852
PaymentMethod_Monthly            205.5532     61.222      3.358      0.001      85.514     325.592
PaymentMethod_Quarterly          134.1416     89.511      1.499      0.134     -41.365     309.648
PaymentMethod_Yearly             -78.2722     35.106     -2.230      0.026    -147.105      -9.439
==============================================================================
Omnibus:                      139.518   Durbin-Watson:                   1.992
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              160.457
Skew:                           0.508   Prob(JB):                     1.44e-35
Kurtosis:                       3.429   Cond. No.                         151.
==============================================================================
```

Here we can see that many of the features has p-value > 0.05. So, we will remove those features with p-value > 0.05 as they are not significant.

So, we will again run RFE by considering less number of features and drop features one-by-one those have p-value > 0.05. Also, we will observe R-squared and Adj. R-squared.

Now again running RFE with considering 23 features and training the data with OLS.

```
[('const', False, 12),
 ('Age', True, 1),
 ('CustTenure', True, 1),
 ('ExistingProdType', True, 1),
 ('NumberOfPolicy', False, 9),
 ('MonthlyIncome', True, 1),
 ('Complaint', False, 2),
 ('ExistingPolicyTenure', True, 1),
 ('SumAssured', True, 1),
 ('LastMonthCalls', False, 3),
 ('CustCareScore', False, 8),
 ('Channel_Online', True, 1),
 ('Channel_Third Party Partner', False, 6),
 ('Occupation_Large Business', True, 1),
 ('Occupation_Salaried', True, 1),
 ('Occupation_Small Business', True, 1),
 ('EducationField_Engineer', True, 1),
 ('EducationField_Graduate', True, 1),
 ('EducationField_MBA', True, 1),
 ('EducationField_Post Graduate', True, 1),
 ('EducationField_Under Graduate', False, 5),
 ('Gender_Male', False, 4),
 ('Designation_Executive', True, 1),
 ('Designation_Manager', True, 1),
 ('Designation_Senior Manager', True, 1),
 ('Designation_VP', True, 1),
 ('MaritalStatus_Married', True, 1),
 ('MaritalStatus_Unmarried', False, 10),
 ('Zone_North', False, 7),
 ('Zone_South', True, 1),
 ('Zone_West', False, 11),
 ('PaymentMethod_Monthly', True, 1),
 ('PaymentMethod_Quarterly', True, 1),
 ('PaymentMethod_Yearly', True, 1)]
```

OLS Regression Results

Dep. Variable:	AgentBonus	R-squared:	0.807
Model:	OLS	Adj. R-squared:	0.806
Method:	Least Squares	F-statistic:	571.2
Date:	Thu, 10 Aug 2023	Prob (F-statistic):	0.00
Time:	19:09:05	Log-Likelihood:	-24795.
No. Observations:	3164	AIC:	4.964e+04
Df Residuals:	3140	BIC:	4.978e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5054.6523	450.494	11.220	0.000	4171.360	5937.944
Age	197.6011	12.691	15.570	0.000	172.718	222.485
CustTenure	215.2949	12.580	17.114	0.000	190.628	239.962
ExistingProdType	52.9548	22.226	2.383	0.017	9.376	96.533
MonthlyIncome	151.7814	23.855	6.363	0.000	105.008	198.555
ExistingPolicyTenure	117.0854	11.512	10.171	0.000	94.514	139.657
SumAssured	816.0348	14.410	56.628	0.000	787.780	844.289
Channel_Online	48.3413	36.186	1.336	0.182	-22.609	119.291
Occupation_Large Business	-522.1215	467.464	-1.117	0.264	-1438.688	394.445
Occupation_Salaried	-503.2033	437.927	-1.149	0.251	-1361.856	355.449
Occupation_Small Business	-585.2828	447.904	-1.307	0.191	-1463.498	292.932
EducationField_Engineer	-60.8603	174.639	-0.348	0.727	-403.278	281.558
EducationField_Graduate	-69.2517	94.969	-0.729	0.466	-255.460	116.956
EducationField_MBA	18.7067	131.111	0.143	0.887	-238.365	275.778
EducationField_Post Graduate	-117.8661	105.425	-1.118	0.264	-324.576	88.844
Designation_Executive	-454.4367	63.950	-7.106	0.000	-579.825	-329.049
Designation_Manager	-439.7889	55.020	-7.993	0.000	-547.668	-331.909
Designation_Senior Manager	-273.4243	52.204	-5.238	0.000	-375.781	-171.068
Designation_VP	-34.0341	71.866	-0.474	0.636	-174.943	106.875
MaritalStatus_Married	-43.1914	22.089	-1.955	0.051	-86.502	0.120
Zone_South	184.0944	308.014	0.598	0.550	-419.835	788.024
PaymentMethod_Monthly	196.4268	60.020	3.273	0.001	78.745	314.108
PaymentMethod_Quarterly	129.5773	88.738	1.460	0.144	-44.412	303.567
PaymentMethod_Yearly	-73.8148	34.704	-2.127	0.033	-141.860	-5.770

Omnibus:	141.398	Durbin-Watson:	1.995
Prob(Omnibus):	0.000	Jarque-Bera (JB):	163.081
Skew:	0.511	Prob(JB):	3.87e-36
Kurtosis:	3.438	Cond. No.	126.

Here, we observe that, the 10 features are removed but R-squared values is stable. So, we can proceed by dropping that variable. Now dropping the features which has p-value > 0.05 one-by-one and observing R2 squared and Adj R-squared.

So, after performing this process repeatedly, we now have with 12 features left in the train dataset.

```
                             OLS Regression Results
==============================================================================
Dep. Variable:             AgentBonus   R-squared:                       0.807
Model:                            OLS   Adj. R-squared:                  0.806
Method:                 Least Squares   F-statistic:                     1095.
Date:                Thu, 10 Aug 2023   Prob (F-statistic):               0.00
Time:                        19:09:05   Log-Likelihood:                -24800.
No. Observations:                3164   AIC:                         4.963e+04
Df Residuals:                    3151   BIC:                         4.970e+04
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                       4475.4023     52.540     85.181      0.000    4372.387    4578.418
Age                          198.7119     12.657     15.700      0.000     173.895     223.529
CustTenure                   215.2153     12.566     17.127      0.000     190.577     239.853
ExistingProdType              49.2265     20.846      2.361      0.018       8.353      90.100
MonthlyIncome                148.5145     20.788      7.144      0.000     107.756     189.273
ExistingPolicyTenure         117.3797     11.494     10.213      0.000      94.844     139.915
SumAssured                   815.5805     14.397     56.648      0.000     787.352     843.809
Designation_Executive       -450.3848     63.460     -7.097      0.000    -574.813    -325.957
Designation_Manager         -432.5758     54.209     -7.980      0.000    -538.863    -326.288
Designation_Senior Manager  -267.1051     50.065     -5.335      0.000    -365.269    -168.941
MaritalStatus_Married        -44.4243     21.969     -2.022      0.043     -87.500      -1.349
PaymentMethod_Monthly        189.0994     57.367      3.296      0.001      76.618     301.580
PaymentMethod_Yearly         -70.9304     33.855     -2.095      0.036    -137.311      -4.550
==============================================================================
Omnibus:                      135.765   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              155.495
Skew:                           0.501   Prob(JB):                     1.72e-34
Kurtosis:                       3.418   Cond. No.                         15.2
==============================================================================
```

After dropping features one-by-one, R-squared is stable that means it seems to be liable to drop features with high p-value.

	Features	VIF
2	ExistingProdType	3.00
11	PaymentMethod_Yearly	2.56
10	PaymentMethod_Monthly	2.03
6	Designation_Executive	2.01
9	MaritalStatus_Married	1.93
3	MonthlyIncome	1.77
5	SumAssured	1.72
7	Designation_Manager	1.56
0	Age	1.34
1	CustTenure	1.32
8	Designation_Senior Manager	1.26
4	ExistingPolicyTenure	1.11

We have VIF values less than 5. So, these features can be used in building the model's.

<u>Interpretation:</u>

- This output represents the results of an Ordinary Least Squares (OLS) regression model with the "AgentBonus" variable as the dependent variable and 12 predictor variables.
- R-squared (R^2): The coefficient of determination is 0.807, indicating that approximately 80.7% of the variance in the dependent variable (AgentBonus) can be explained by the predictor variables in the model.
- Adjusted R-squared (Adj. R^2): The adjusted R-squared is 0.806, which is a modified version of R-squared that considers the number of predictor variables in the model, providing a more accurate measure of the model's goodness of fit
- P-values: A p-value less than 0.05 is typically considered statistically significant, suggesting that the predictor variable has a significant impact on the dependent variable. In this output, several variables (e.g., Age, CustTenure, MonthlyIncome, etc.) have p-values less than 0.05, indicating they are statistically significant predictors.
- Overall, the model appears to have a good fit with a high R-squared value, and several predictor variables show significant associations with the "AgentBonus" variable. However, further analysis required before finalize the model.

Predicting the Linear Regression model:

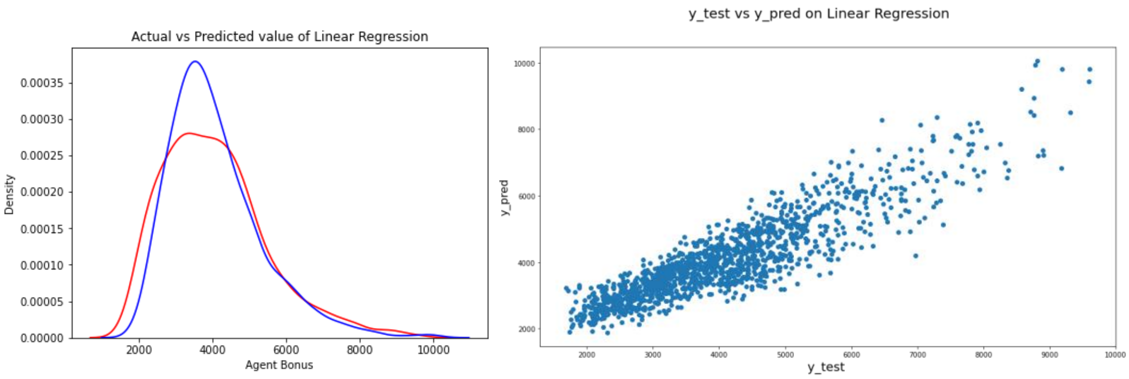


Figure 10: y_test vs y_pred on Linear Regression

Equation of the Line:

(4475.4) * const + (198.71) * Age + (215.22) * CustTenure + (49.23) * ExistingProdType + (148.51) * MonthlyIncome + (117.38) * ExistingPolicyTenure + (815.58) * SumAssured + (-450.38) * Designation_Executive + (-432.58) * Designation_Manager + (-267.11) * Designation_Senior Manager + (-44.42) * MaritalStatus_Married + (189.1) * PaymentMethod_Monthly + (-70.93) * PaymentMethod_Yearly

4.4.2.Lasso Regression Model:

Model evaluation:

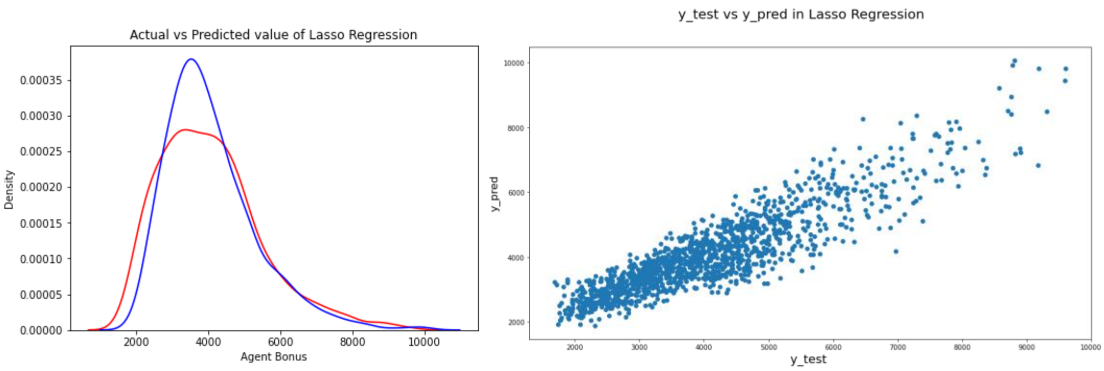


Figure 11: y_test vs y_pred in Lasso Regression

We had calculated the R2-squared and RSME values also.

Equation of line is:

Agent_Bonus = 4400.95 + (199.02) * Age + (216.31) * CustTenure + (26.72) * ExistingProdType + (171.60) * MonthlyIncome + (117.11) * ExistingPolicyTenure + (816.81) * SumAssured + (-363.45) * Designation_Executive + (-358.83) * Designation_Manager + (-205.72) * Designation_Senior Manager + (-39.88) * MaritalStatus_Married + (134.35) * PaymentMethod_Monthly + (-42.38) * PaymentMethod_Yearly

4.4.3.Ridge Regression Model:

Model Evaluation:

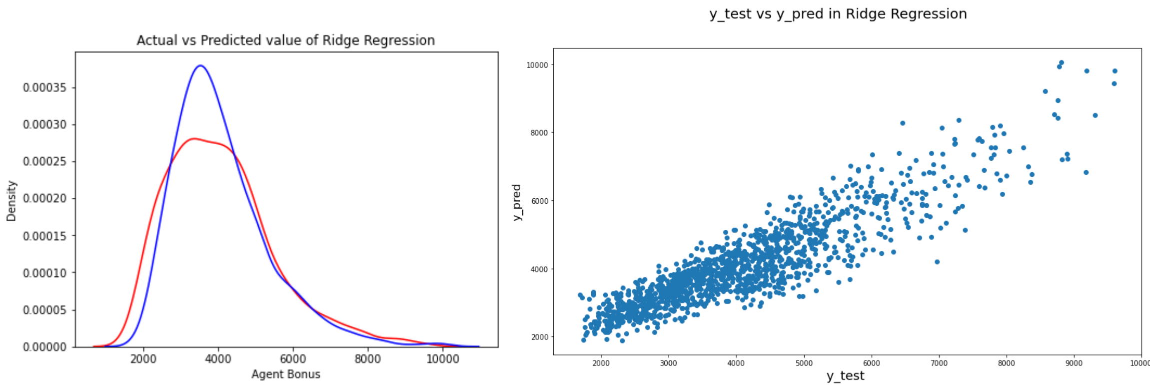


Figure 12: y_test vs y_pred in Ridge Regression

Equation of line is:

AgentBonus = 4467.48 + 198.84 * Age + 215.41 * CustTenure + 47.57 * ExistingProdType + 151.12 * MonthlyIncome + 117.43 * ExistingPolicyTenure + 815.41 * SumAssured - 440.60 * Designation_Executive - 424.28 * Designation_Manager - 260.29 * Designation_Senior Manager - 44.31 * MaritalStatus_Married + 185.24 * PaymentMethod_Monthly - 69.06 * PaymentMethod_Yearly

4.4.4. Elastic Net Regression Model:

Model Evaluation:

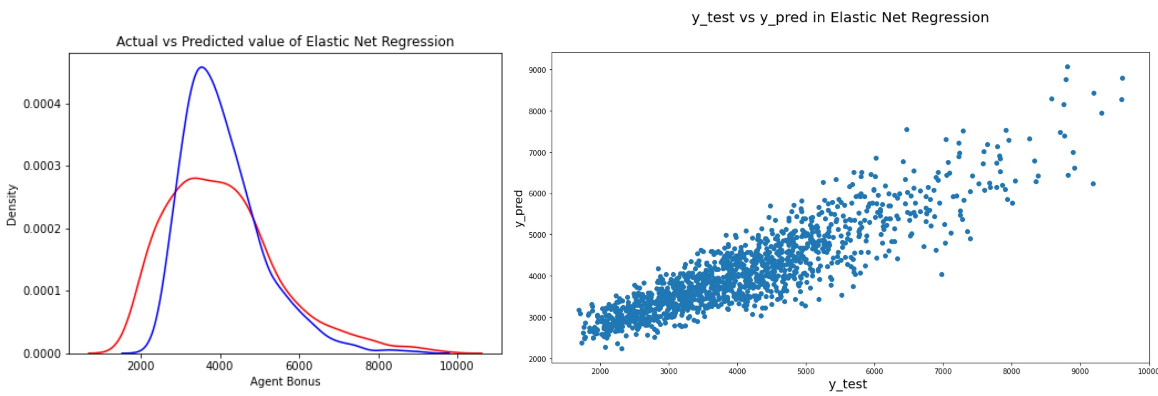


Figure 13: y_test vs y_pred in Elastic Net Regression

Equation of line is:

AgentBonus = 4130.80 + 222.28 * Age + 229.27 * CustTenure + 5.87 * ExistingProdType + 241.02 * MonthlyIncome + 128.64 * ExistingPolicyTenure + 537.59 * SumAssured - 65.94 * Designation_Executive - 52.16 * Designation_Manager + 19.14 * Designation_Senior Manager - 3.27 * MaritalStatus_Married + 8.49 * PaymentMethod_Monthly - 11.92 * PaymentMethod_Yearly

4.4.5.Decision Tree:

We had used GridSearchCV to optimize the best hyper-parameters.

We got best hyper-parameters as:

Best Hyperparameters: {'max_depth': 10, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 10}

Model Evaluation:

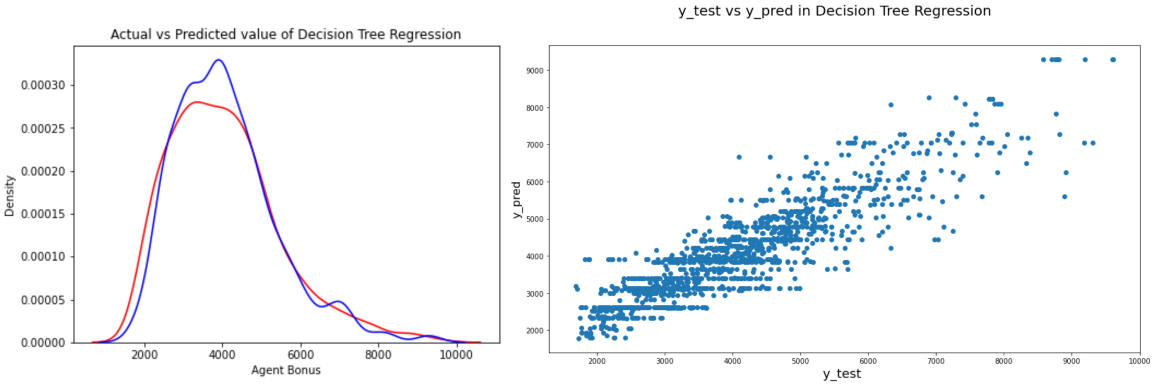


Figure 14: y_test vs y_pred in Decision Tree Regression

4.4.6.Random Forest:

We had trained and fit this Random Forest model with Best Hyper-parameters with the help of GridSearchCV library.

Model Evaluation:

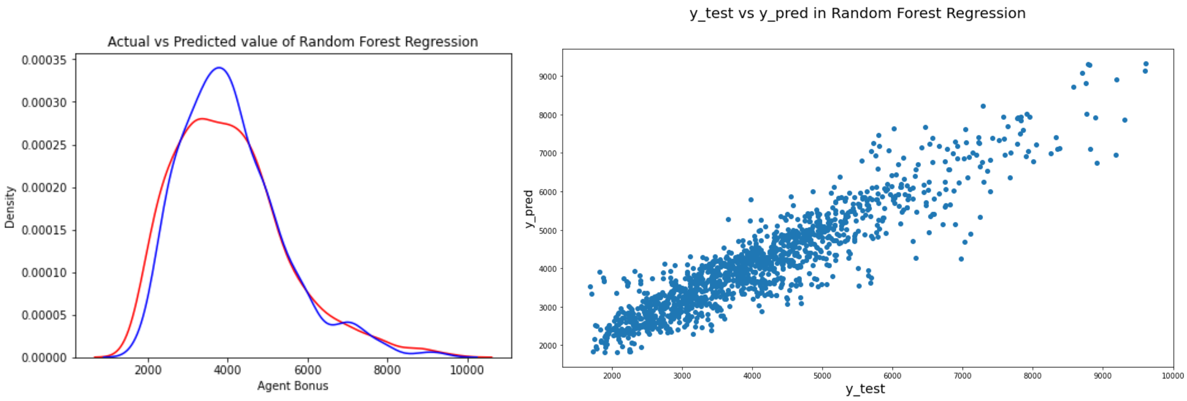


Figure 15: y_test vs y_pred in Random Forest Regression

Also,

We had found the Feature Importance:

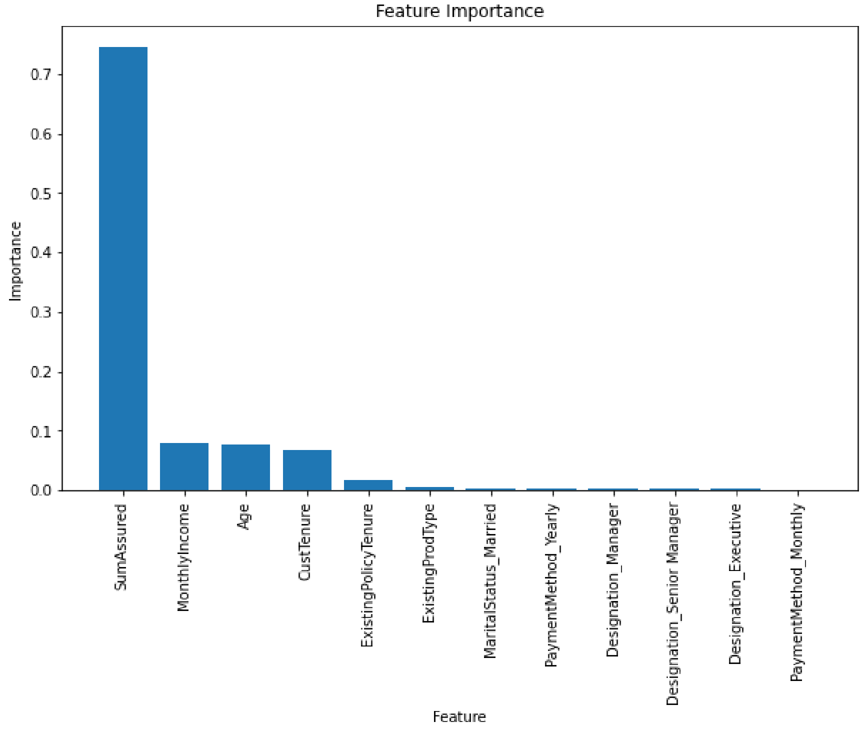


Figure 16: Feature Importance

Interpretation:

- From the above graph, it is found that SumAssured is the most importance feature in the dataset followed by MonthlyIncome and Age and then CustTenure.

5. Model validation

Model Name	Train RSME	Test RSME	Train R2	Test R2
Linear Regression Model	613.47	623.56	0.806	0.807
Lasso Regression Model	613.72	623.72	0.806	0.808
Ridge Regression Model	613.47	623.57	0.807	0.808
Elastic Net Regression Model	675.72	688.87	0.765	0.764
Decision Tree Regression Model	461.93	645.18	0.89	0.794
Random Forest Regression Model	252.97	562.32	0.967	0.843

- Linear Regression, Lasso Regression, and Ridge Regression models have similar train RSME and test RSME values, indicating that they are generalizing well on unseen data. Their R-squared values are also relatively high, suggesting a good fit to the data.
- Elastic Net Regression Model has a slightly higher test RSME compared to train RSME, indicating some overfitting. The R-squared values are relatively lower, suggesting that this model may not fit the data as well as the other models.

- Decision Tree Regression Model shows a significant difference between train RSME and test RSME, indicating potential overfitting. While the train R-squared is high (0.89), the test R-squared is lower (0.794), suggesting that the model may not generalize well to new data.
- Random Forest Regression Model shows a significant difference between train RSME and test RSME, indicating overfitting.
- Lasso Regression Model may be the most appropriate model to finalize. It shows similar performance to the other models but has the advantage of performing feature selection due to its L1 regularization. By setting some coefficients to zero, Lasso Regression effectively identifies and excludes irrelevant features, leading to a potentially more interpretable and simpler model.
- Additionally, the Lasso Regression model has slightly better Test RSME and Test R2 compared to the other models, which indicates better generalization to unseen data. The slight improvement in performance on the test set is a positive sign that the Lasso Regression model may be more robust in making predictions on new data.

5.1. Ensemble Techniques:

We had applied Ada-Boosting on Lasso Regression Model and Ridge Regression model. We also applied Bagging on Lasso Regression Model and Ridge Regression model. Model has same output as below compared to the original model.

Model Name	Train RSME	Test RSME	Train R2	Test R2
Linear Regression Model	613.47	623.56	0.806	0.807
Lasso Regression Model	613.72	623.72	0.806	0.808
Ridge Regression Model	613.47	623.57	0.807	0.808
Elastic Net Regression Model	675.72	688.87	0.765	0.764
Decision Tree Regression Model	461.93	645.18	0.89	0.794
Random Forest Regression Model	252.97	562.32	0.967	0.843
Lasso+AdaBoosting Regression Model	623.63	637.61	0.8	0.799
Lasso+Bagging Regression Model	613.64	623.91	0.806	0.807
Ridge+AdaBoosting Regression Model	623.95	638.12	0.8	0.799
Ridge+Bagging Regression Model	613.63	623.9	0.806	0.808

Interpretation:

- Overall, the ensemble techniques (AdaBoosting and Bagging) did not provide a substantial improvement in performance over the individual Lasso and Ridge Regression models in this case.
- Both Ridge Regression and Lasso Regression models have similar performance and generalization on the test data. They also exhibit a good balance between model complexity and performance. Either of these models can be considered a suitable choice for training in the future, depending on the specific requirements and interpretability needs.

6. Final interpretation / recommendation

- Company should focus more on Large Business customers
- Company should focus more on Unmarried customers by introducing them with benefits plans
- Company should focus more on Female customers

- Should focus more on Salaried persons especially Managers and Executive
- Most customers were acquired through agents, indicating their significant role in customer acquisition
- Graduate is the most common education field among customers
- There are more male customers than female customers, and there is a category called Female that requires further investigation
- Most customers are married, followed by single and divorced individuals
- The dataset represents customers primarily from the West zone, followed by the North zone
- Half Yearly is the most common payment frequency chosen by customers
- Based on these insights, the life insurance company can make data-driven decisions regarding bonus allocation to agents. For example, they may consider introducing specific bonus schemes based on different product types, tenure, or payment methods.
- Additionally, they could focus on rewarding higher-performing agents with higher sum assured policies or those who have been with the company for a longer time. The company can use this information to design appropriate engagement activities and incentives to motivate agents and improve overall performance.
- Ridge Regression Model and the Lasso Regression Model are the most optimum models for predicting the bonus for agents. Both models have very similar Train RSME and Test RSME values, indicating that they are generalizing well to unseen data. Additionally, their R-squared values for both the training and testing sets are quite high, which means they explain a significant portion of the variance in the bonus amounts.
- The high R-squared values indicate that the model can capture the relationships between the input features and the bonus amounts reasonably well. This means that the company can confidently use the model's predictions to allocate bonuses to agents with a high level of accuracy.
- With a reliable bonus prediction model, the company can easily identify high-performing agents who deserve higher bonuses based on their predicted bonus amounts. This allows the company to recognize and reward agents who bring in significant business or demonstrate exceptional performance.
- By using the model's predictions, the company can optimize its bonus allocation strategy. They can allocate bonuses based on factors such as the agent's tenure, product type, income, age, etc., as indicated by the feature importance analysis.
- Fair and accurate bonus allocation can motivate agents to perform better and increase their retention within the company. When agents feel valued and rewarded appropriately, they are more likely to remain committed to their roles and contribute to the company's success.
- By using data-driven bonus allocation, the company can avoid overpaying bonuses to low-performing agents and ensure that the allocated budget is utilized effectively.
- The company can continuously update and refine the model as new data becomes available. This ensures that the model adapts to changing patterns and remains relevant over time.
- Ridge and Lasso Regression, empowers the life insurance company to make informed decisions regarding bonus allocation, agent motivation, and business strategy. It enhances the company's ability to retain high-performing agents, optimize resource allocation, and ultimately improve overall business performance.
- When bonuses are precisely calculated based on agent performance metrics, such as tenure, sales achievements, and customer satisfaction, it ensures that rewards are distributed equally. High-performing agents receive rewards that match their contributions, and this, in turn, motivates them to continue excelling.
- Agents with outstanding customer satisfaction scores might receive specialized training to capitalize on their strengths. On the other hand, those who face challenges in certain areas can benefit from targeted training to improve their performance. As a result, engagement initiatives and training programs become not only effective but also efficient, aligning closely with organizational goals.

The End