Table of Contents

Table Of Figure

1. Introduction of the business problem

1.1. Problem Statement:

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.

1.2. Need of the study/project:

- Company wants to know how the agents are performing and if appropriate bonuses are being provided to them.
- Company wants insights on their sales channels like where most policies are being sold, what is the most popular channel, study of the demographics to better design insurance policies. Company can check which channels and regions are performing better and focus on weak areas.
- Company can predict bonuses for agents beforehand which will enable company to better manage their finances and avoid bankruptcy.

1.3. Understanding business/social opportunity:

- Company wants to look at their weak regions and channels to improve their services and expand their business in future.
- Company wants to study their customer to provide them with services better suited for them.
- Company wants to know how their customer services are being looked at by the customers.
- Company wants to avoid future bankruptcy and solvency issues.

2. Data Report

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

Data was provided by the insurance company as part of their data collection drive to better understand their business.

2.2. Visual inspection of data (rows, columns, descriptive details):

Variable	Discerption
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

Fig 2.2.1 Data Dictionary

Data dictionary is attached to better understand the data.

	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncon
0	7000000	4409	22.0	4.0	Agent	Salaried	Graduate	Female	3	Manager	2.0	Single	20993
1	7000001	2214	11.0	2.0	Third Party Partner	Salaried	Graduate	Male	4	Manager	4.0	Divorced	20130
2	7000002	4273	26.0	4.0	Agent	Free Lancer	Post Graduate	Male	4	Exe	3.0	Unmarried	17090
3	7000003	1791	11.0	NaN	Third Party Partner	Salaried	Graduate	Fe male	3	Executive	3.0	Divorced	17909
4	7000004	2955	6.0	NaN	Agent	Small Business	UG	Male	3	Executive	4.0	Divorced	18468

Fig 2.2.2 Data Snippet

Here is a quick look at the data. It was 20 columns and 4520 rows of customer data.

```
 #   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)
```

Fig 2.2.3 Data Info

Data has 8 object type columns and 12 numerical data consisting of integer and float type data.

	CustID	AgentBonus	Age	CustTenure	ExistingProdType	NumberOfPolicy	MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured	LastMonthCalls	CustCareScore
count	4.520000e+03	4520.000000	4251.000000	4294.000000	4520.000000	4475.000000	4284.000000	4520.000000	4336.000000	4.366000e+03	4520.000000	4468.000000
mean	7.002260e+06	4077.838274	14.494707	14.469027	3.688938	3.565363	22890.309991	0.287168	4.130074	6.199997e+05	4.626991	3.067592
std	1.304956e+03	1403.321711	9.037629	8.963671	1.015769	1.455926	4885.600757	0.452491	3.346386	2.462348e+05	3.620132	1.382968
min	7.000000e+06	1605.000000	2.000000	2.000000	1.000000	1.000000	16009.000000	0.000000	1.000000	1.685360e+05	0.000000	1.000000
25%	7.001130e+06	3027.750000	7.000000	7.000000	3.000000	2.000000	19683.500000	0.000000	2.000000	4.394432e+05	2.000000	2.000000
50%	7.002260e+06	3911.500000	13.000000	13.000000	4.000000	4.000000	21606.000000	0.000000	3.000000	5.789765e+05	3.000000	3.000000
75%	7.003389e+06	4867.250000	20.000000	20.000000	4.000000	5.000000	24725.000000	1.000000	6.000000	7.582360e+05	8.000000	4.000000
max	7.004519e+06	9608.000000	58.000000	57.000000	6.000000	6.000000	38456.000000	1.000000	25.000000	1.838496e+06	18.000000	5.000000

Fig 2.2.4 Data Description

As we can see from the data that average agent bonus is 4077, average age of the customer is 14 years, mean income is 22890, average tenure for customers is 14.46 years.

2.3 Understanding of attributes (variable info, renaming if required):

Data has few null values. We will later impute them with KNN imputer.

Here is a list and count of missing values.

```
CustID                    0
AgentBonus                0
Age                     269
CustTenure              226
Channel                   0
Occupation                0
EducationField            0
Gender                    0
ExistingProdType          0
Designation               0
NumberOfPolicy           45
MaritalStatus             0
MonthlyIncome           236
Complaint                 0
ExistingPolicyTenure    184
SumAssured              154
Zone                      0
PaymentMethod             0
LastMonthCalls            0
CustCareScore            52
dtype: int64
```

Fig 2.3.1 List of Missing Values

We can see that there are few data points which needs reassigning as they are either mistyped or duplicates with different names.

- In occupation columns we change 'Laarge Business' and rename it to 'Large Business' and merge with existing.
- We do the same for 'Graduate' and 'UG' which are renamed to 'Undergraduate' as they all mean the same.
- We change 'Fe male' to 'Female' and correct the typing mistake.
- We change designation 'Exe' to 'Executive' and merge with existing as they mean the same.
- We also merge 'unmarried' and 'single' as they both mean the same thing.

3. Exploratory data analysis:

3.1 Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

```
Channel
Agent                  0.706637
Third Party Partner    0.189823
Online                 0.103540
Name: proportion, dtype: float64
Occupation
Salaried          0.484956
Small Business    0.424336
Large Business    0.090265
Free Lancer       0.000442
Name: proportion, dtype: float64
EducationField
Under Graduate    0.727876
Diploma           0.109735
Engineer          0.090265
Post Graduate     0.055752
MBA               0.016372
Name: proportion, dtype: float64
Gender
Male      0.59469
Female    0.40531
Name: proportion, dtype: float64
Designation
Executive         0.367699
Manager           0.358407
Senior Manager    0.149558
AVP               0.074336
VP                0.050000
Name: proportion, dtype: float64
MaritalStatus
Married     0.501770
Single      0.320354
Divorced    0.177876
Name: proportion, dtype: float64
```

```
Zone
West     0.567699
North    0.416814
East     0.014159
South    0.001327
Name: proportion, dtype: float64
PaymentMethod
Half Yearly    0.587611
Yearly         0.317257
Monthly        0.078319
Quarterly      0.016814
Name: proportion, dtype: float64
CustCareScore
3.0    0.305953
1.0    0.207699
5.0    0.199866
4.0    0.184870
2.0    0.101611
Name: proportion, dtype: float64
ExistingProdType
4    0.423894
3    0.302876
5    0.156637
2    0.048894
1    0.040487
6    0.027212
Name: proportion, dtype: float64
NumberOfPolicy
4.0    0.244469
3.0    0.209832
5.0    0.191285
2.0    0.158883
1.0    0.097877
6.0    0.097654
Name: proportion, dtype: float64
Complaint
0    0.712832
1    0.287168
Name: proportion, dtype: float64
```

Fig 3.1.1 Variable Value Counts

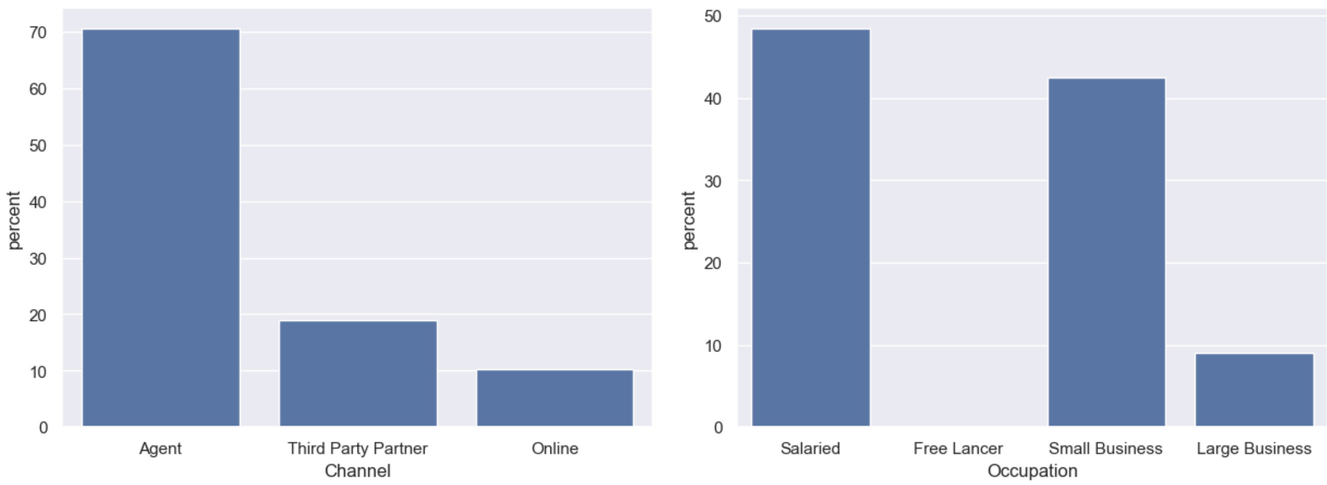


Fig 3.1.2

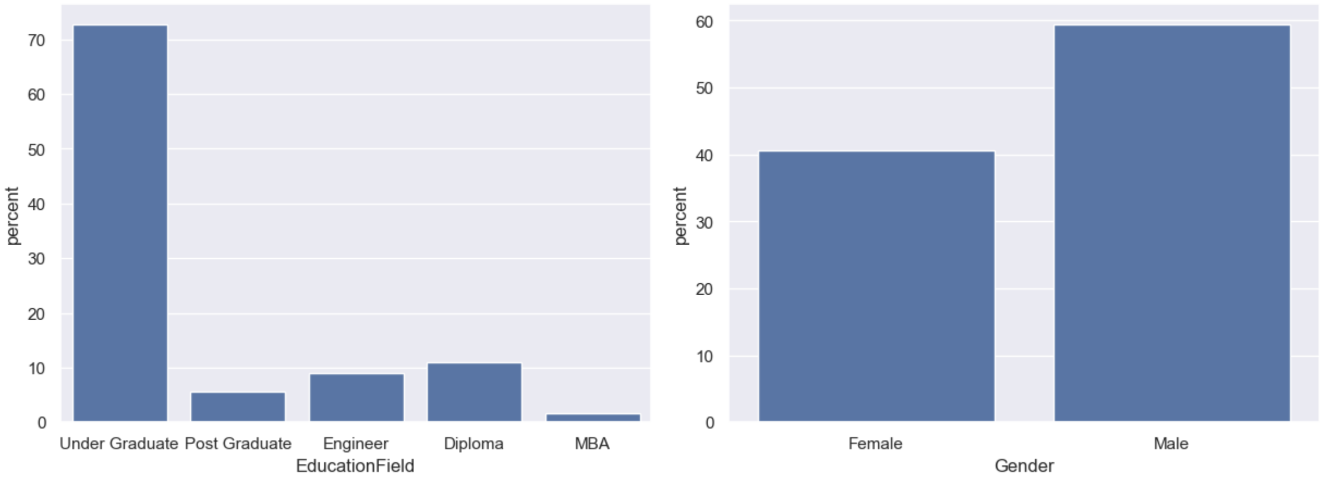


Fig 3.1.3

Fig 3.1.4

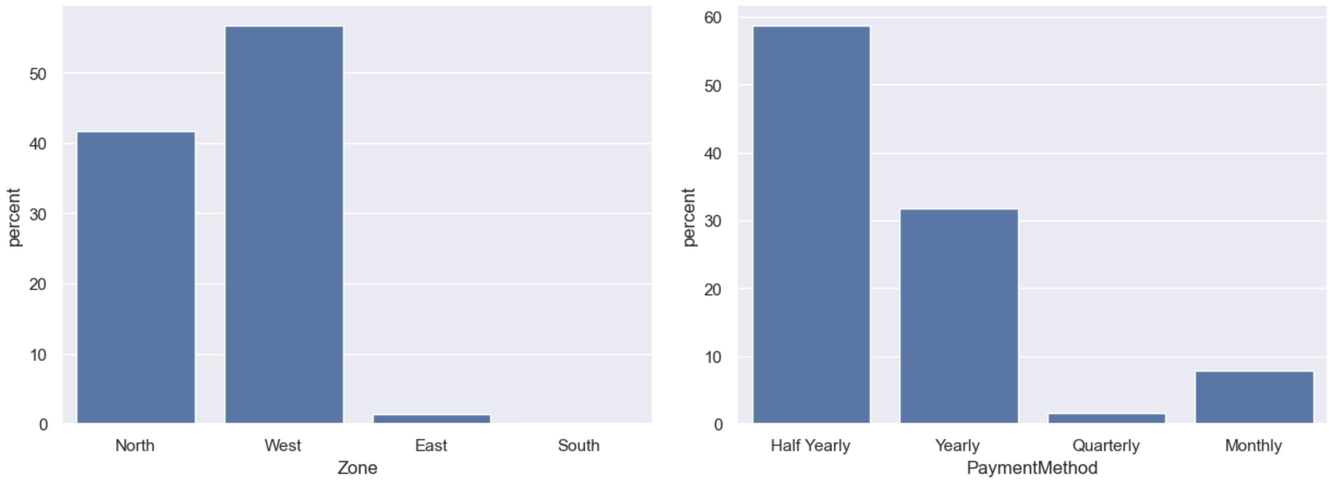


Fig 3.1.5

Distribution of different column values are shown above. We can see from the value counts that-

- 70% of the people prefer buying insurance through Agents.
- Most people either are salaried or have small businesses.
- 73% of the customers are undergraduates.
- Ratio of Male: Female in customer data is 60:40.
- Most of the people are either executive or managers.

- 50% of the customers are married.
- West and North zone count for 98% of the sales. Sales from East and South is almost nil.
- 59% of the customers opt for half-yearly payment method.
- Customers tend to rate 3 for the customer service but other ratings are almost equally distributed.
- 43% of the people opt for product type 4, product type 2 ,1,6 have little or no takers.
- Most people have more than 1 policy and very few people have 6 policies.
- 72% of people did not register a complaint last month, while 28% did register complaints.

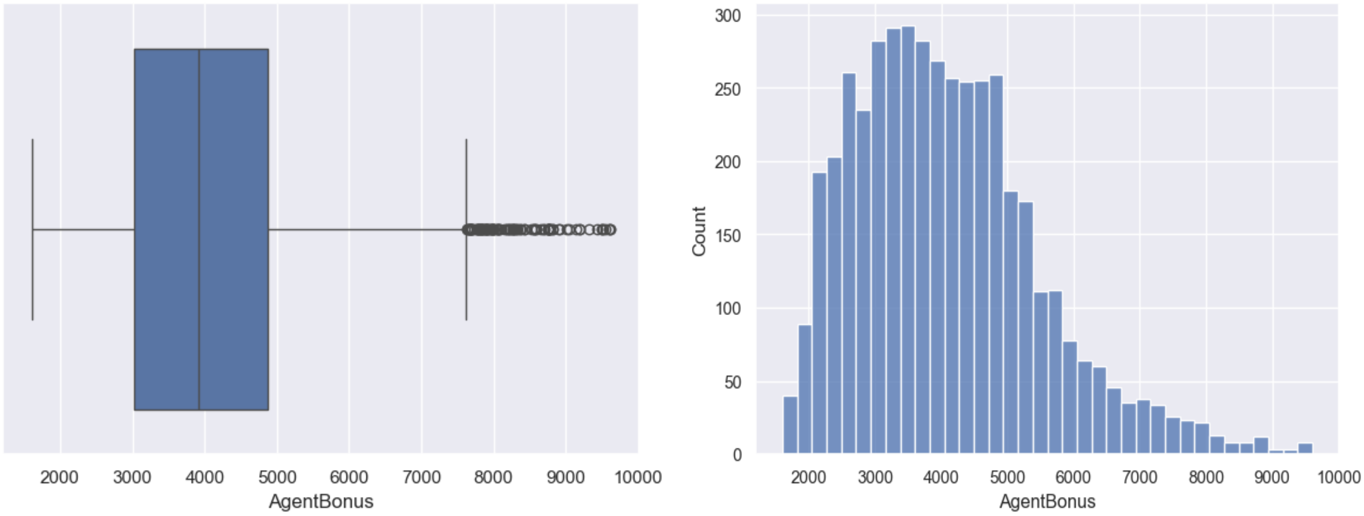


Fig 3.1.6 Displot/ Histplot Agent Bonus

- Distribution of Agent bonus data is right skewed.
- There are many outlier bonuses which are higher than 7.8k.

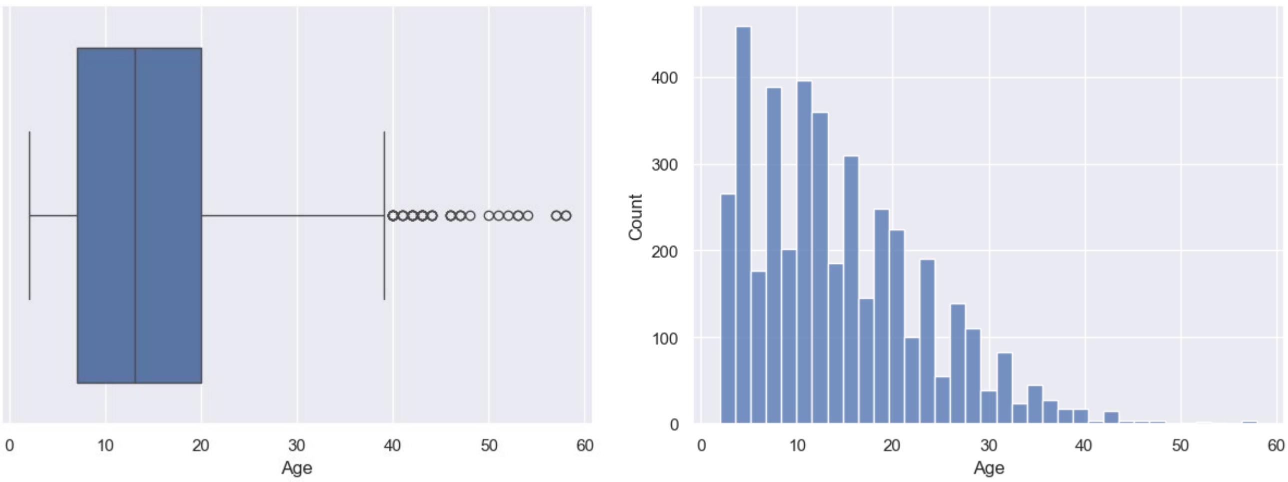


Fig 3.1.7 Displot/ Histplot Age

- Distribution of Age of age is right skewed.
- Age data has few outliers with customers who have age higher than 40 years.

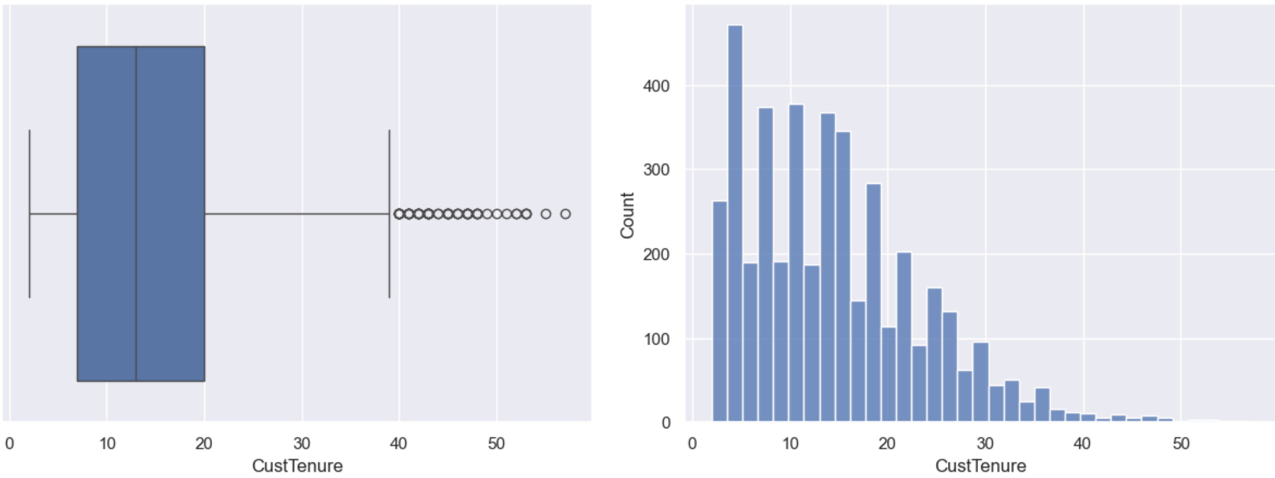


Fig 3.1.8 Displot/ Histplot CustTenure

- Distribution of customer tenure data is right skewed.
- There are few customers who are associated with insurance company for more than 40 years.

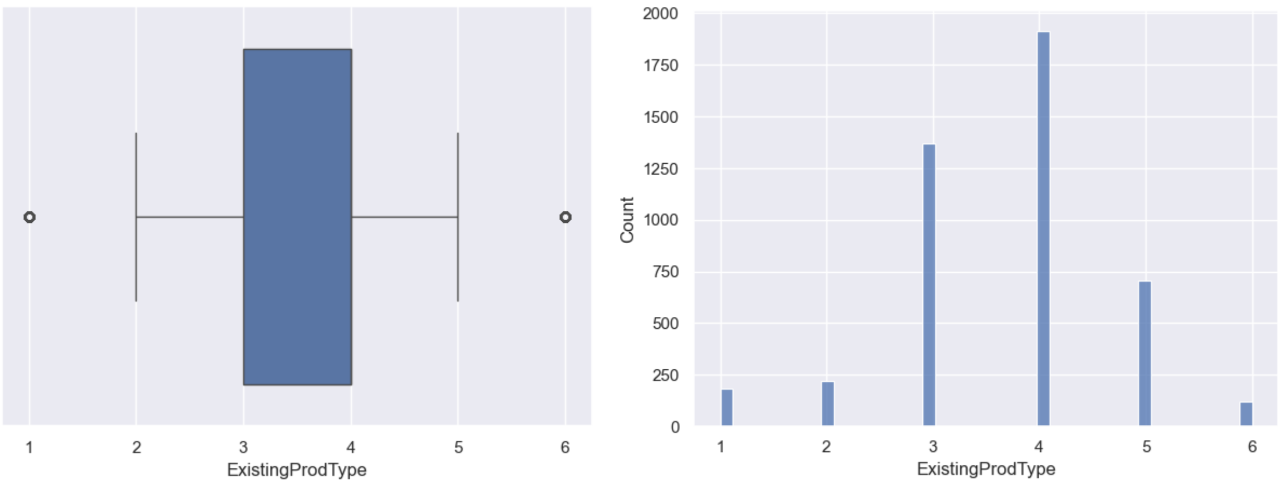


Fig 3.1.9 Displot/ Histplot ExistingProdType

- Distribution of product type is normally distributed.
- Product type 3 and 4 have higher frequency.

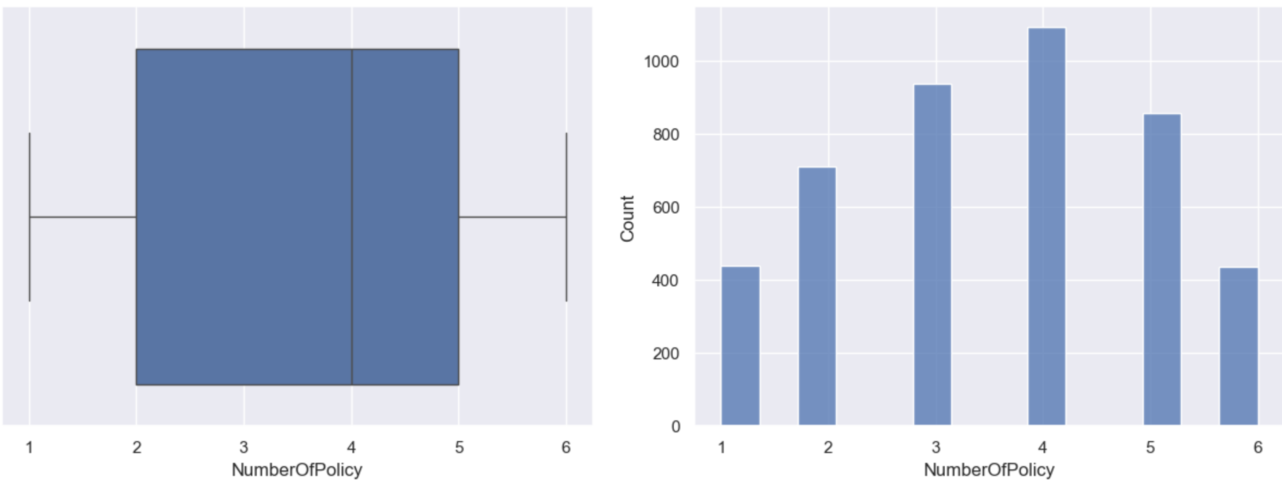


Fig 3.1.10 Displot/ Histplot NumberOfPolicy

- Distribution of number of policies is normally distributed.
- Most people have 3 or 4 policies.

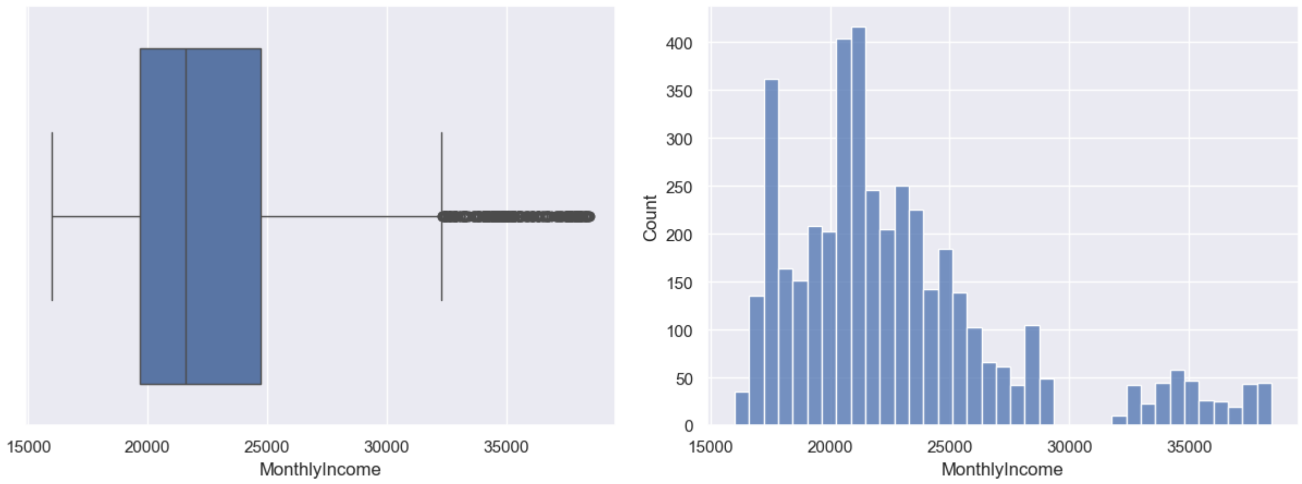


Fig 3.1.11 Displot/ Histplot MonthlyIncome

- Distribution of monthly income is two tailed distributed.
- Data has many outliers with some people having income more than 30000.

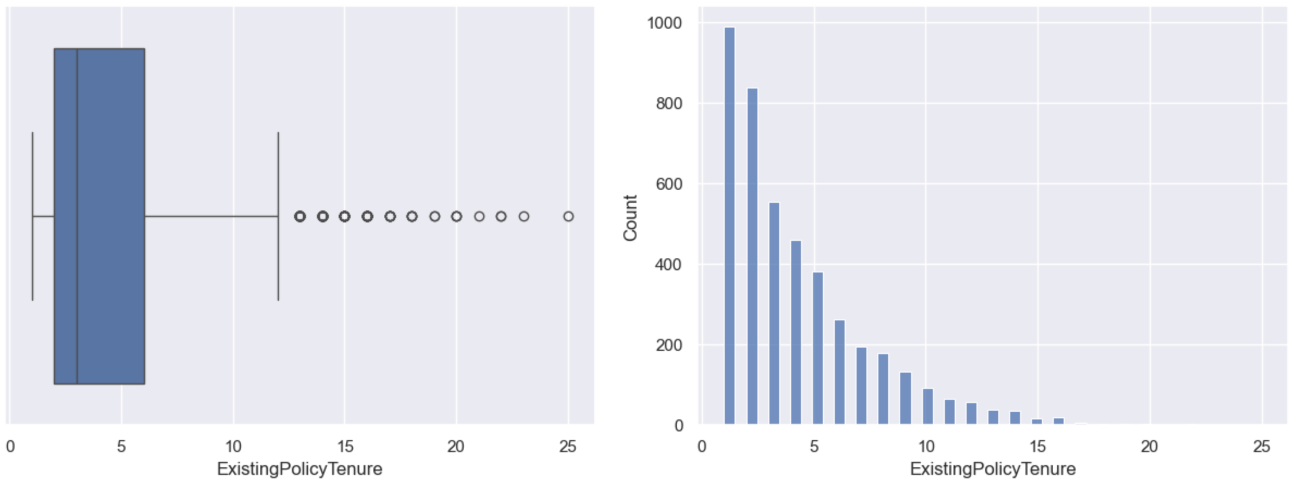


Fig 3.1.12 Displot/ Histplot ExistingPolicyTenure

- Distribution of policy tenure is right skewed.
- Data has few outliers as some customers have policy tenure of more than 12 years.

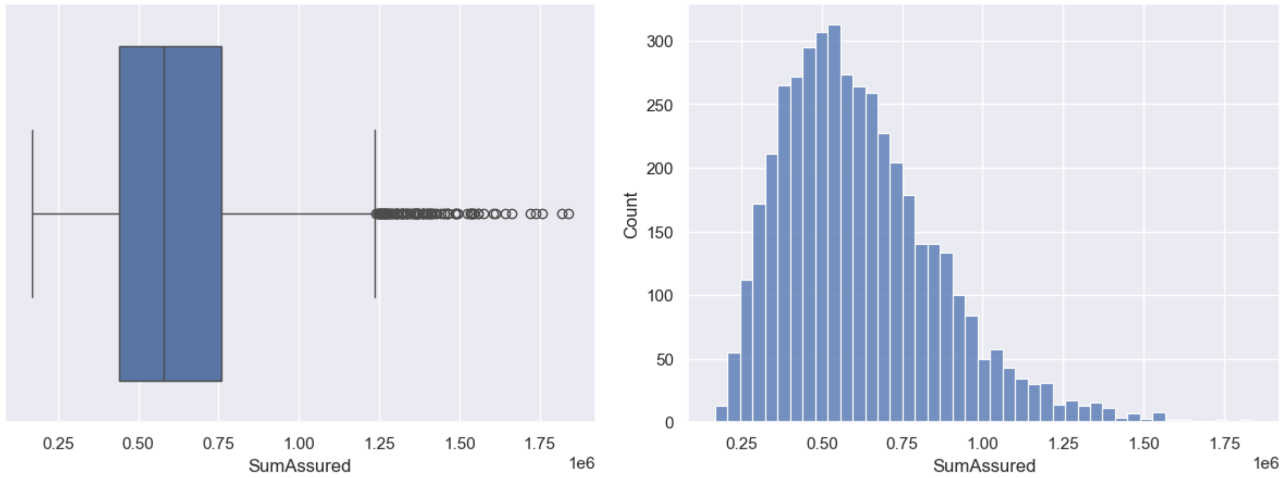


Fig 3.1.13 Displot/ Histplot SumAssured

- Distribution of sum assured is almost normally distributed.
- Data has many high outlier sum assured above 1000000.

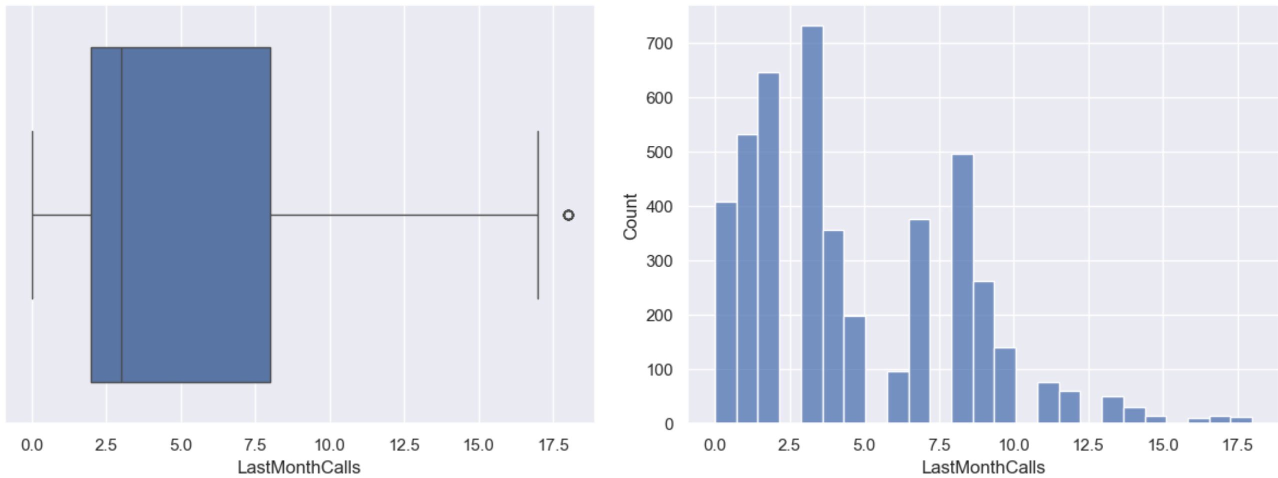


Fig 3.1.14 Displot/ Histplot LastMonthCalls

- Distribution of last month calls data type is not normally distributed with no clear skew on any side.

3.2 Bivariate analysis (relationship between different variables, correlations):

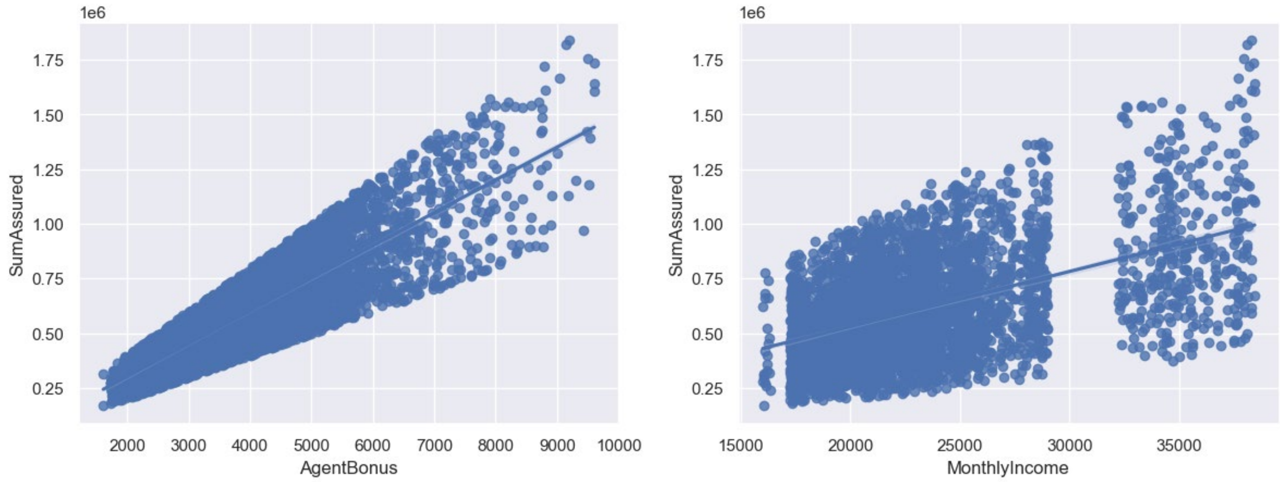


Fig 3.2.1 Scatterplot Agentbonus/MonthlyIncome wrt SumAssured

- We can see from the scatter plot that with the increase of sum assured, agent bonus is increasing. The trend line is almost perfect.
- We can see from the scatter plot that with the increase of monthly income, sum assured is increasing. We can say that with higher income customers are buying policies with higher sum assured.

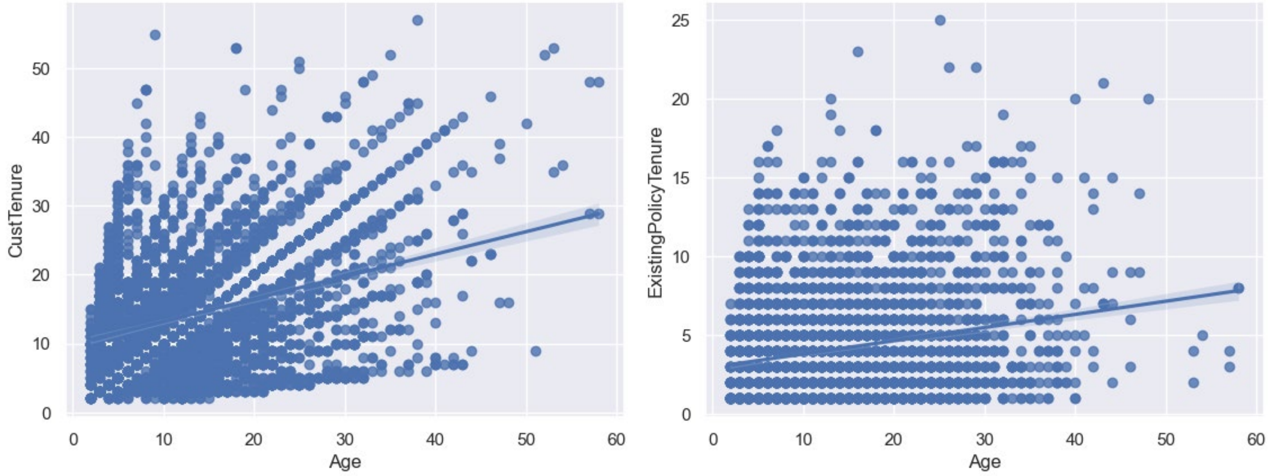


Fig 3.2.2 Scatterplot Age wrt CustTenure/ExistingPolicyTenure

- With higher age policy tenure of customers are increasing.
- With higher age the existing policy tenure is also increasing.

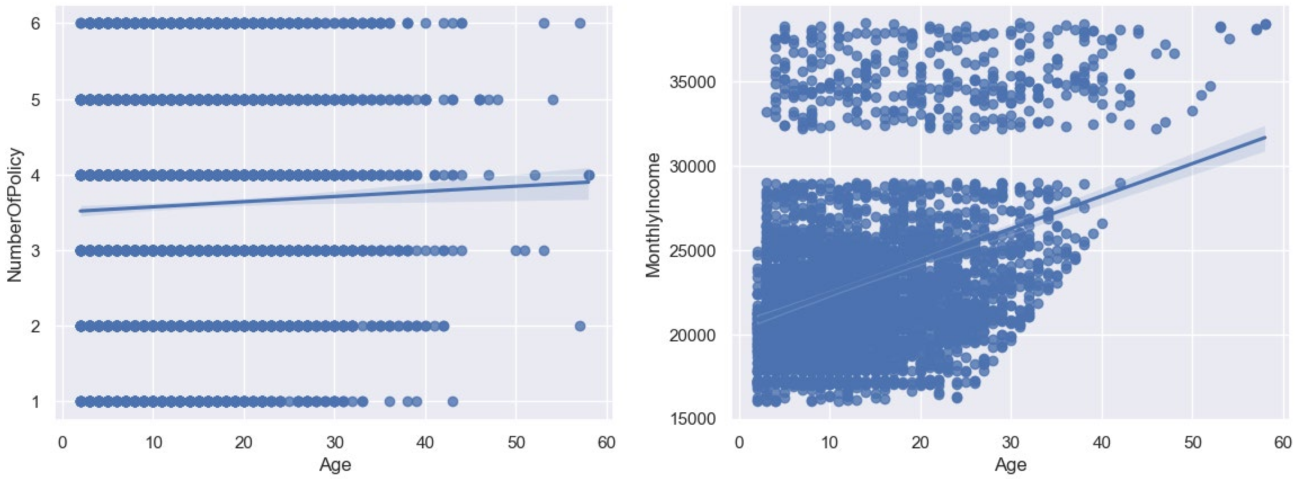


Fig 3.2.3 Scatterplot Age wrt NumberOfPolicy/MonthlyIncome

- With the increase of age the number of policy is increasing.
- Monthly income is also increasing with the increase of age.

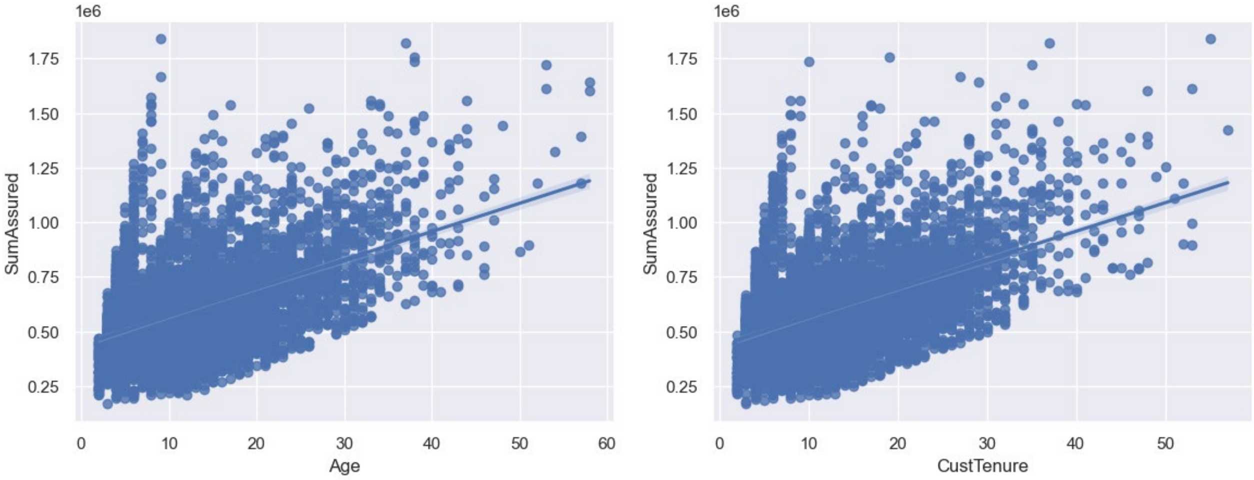


Fig 3.2.4 Scatterplot Age/CustTenure wrt SumAssured

- Sum assured of policies is increasing with increase of age.
- As Customer tenure increases, sum assured also increases.

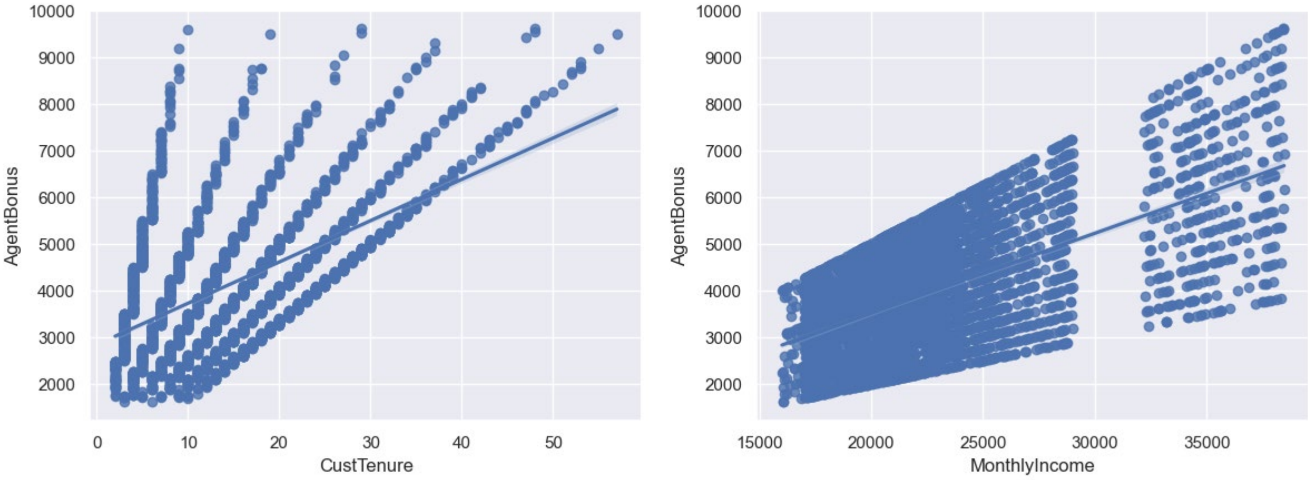


Fig 3.2.5 Scatterplot CustTenure/MonthlyIncome wrt AgentBonus

- As tenure of customer increase the bonus, they provide to agents also increase.
- People with higher income also provide better bonus to agents.

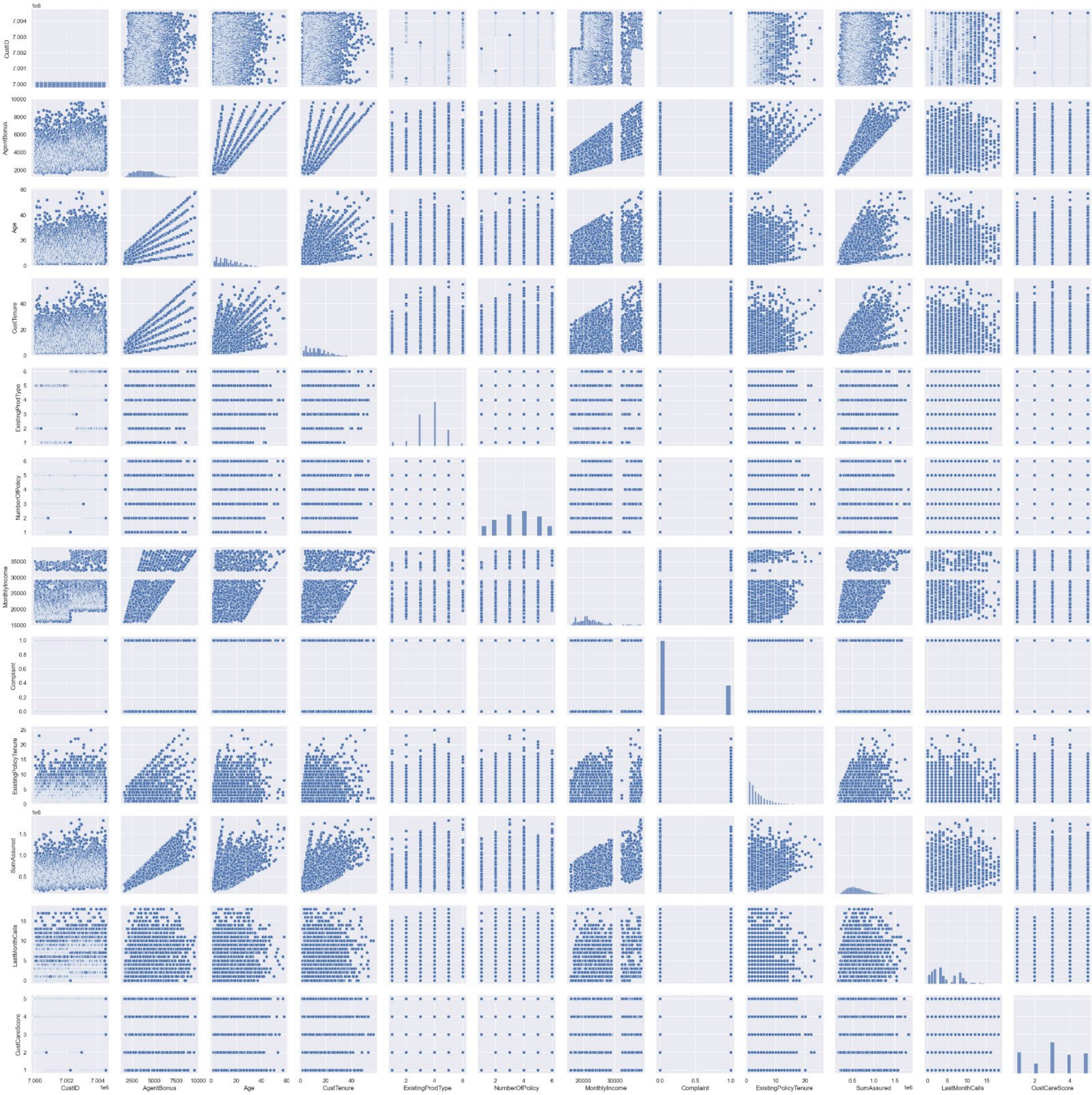


Fig 3.2.6 Pair Plot

- Here is the pair plot of each variable. We can see the positive and negative trends between variables here.

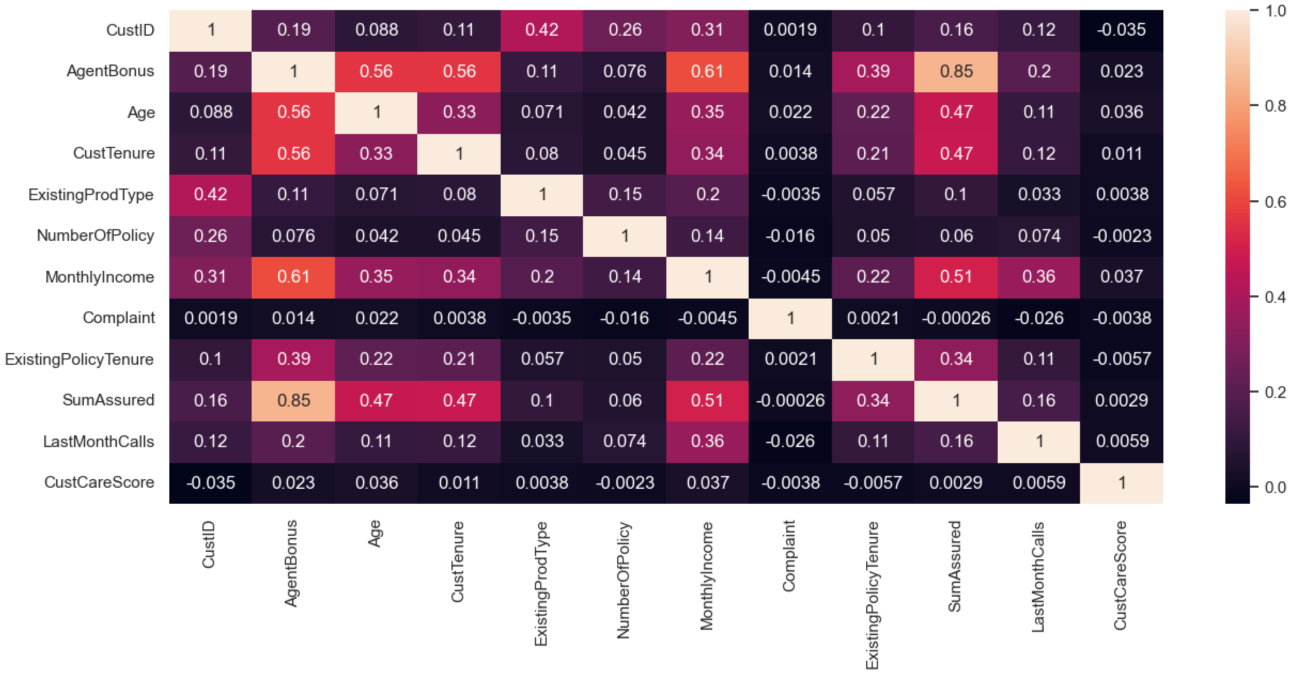


Fig 3.2.7 HeatMap

- Here is the heatmap. We can see correlation coefficients between each variable and understand if they are positively correlated or negatively.

3.3 Removal of unwanted variables:

We have removed 'customer id' column from our data as that column was not necessary in our analysis.

We also removed 'Age' column as the data has severe entry errors. There is no possibility that a person has lower age in years than they have been associated with the company.

3.4 Missing Value treatment:

We have divided our data into categorical and numerical data to better optimise it for later analysis.

```
Channel            0
Occupation         0
EducationField     0
Gender             0
Designation        0
MaritalStatus      0
Zone               0
PaymentMethod      0
dtype: int64
```

```
AgentBonus                0
CustTenure              226
ExistingProdType          0
NumberOfPolicy           45
MonthlyIncome           236
Complaint                 0
ExistingPolicyTenure    184
SumAssured              154
LastMonthCalls            0
CustCareScore            52
dtype: int64
```

Fig 3.4.1 MissingValues Before Treatment

We can see that our categorical columns have no null values, but our numerical columns have few null values.

We will now use KNN-imputer to impute those null values in numerical columns.

```
Channel          0
Occupation       0
EducationField   0
Gender           0
Designation      0
MaritalStatus    0
Zone             0
PaymentMethod    0
dtype: int64
```

```
AgentBonus            0
CustTenure            0
ExistingProdType      0
NumberOfPolicy        0
MonthlyIncome         0
Complaint             0
ExistingPolicyTenure  0
SumAssured            0
LastMonthCalls        0
CustCareScore         0
dtype: int64
```

Fig 3.4.2 Missing Values After Treatment

We can now see that both categorical and numerical columns have no null values now and our data is free from null values.

3.5 Outlier treatment:

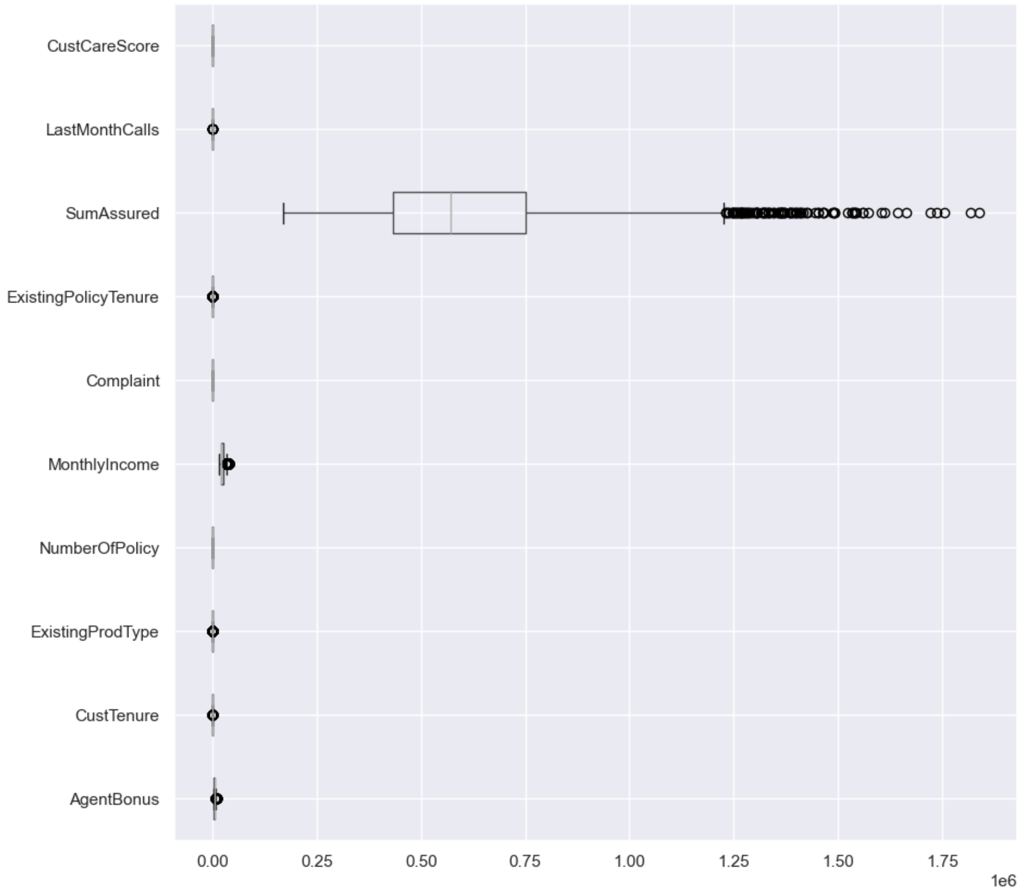


Fig 3.5.1 Outliers Before Treatment

We can see that the data has outliers and now will treat those outliers with 1.5*IQR range.

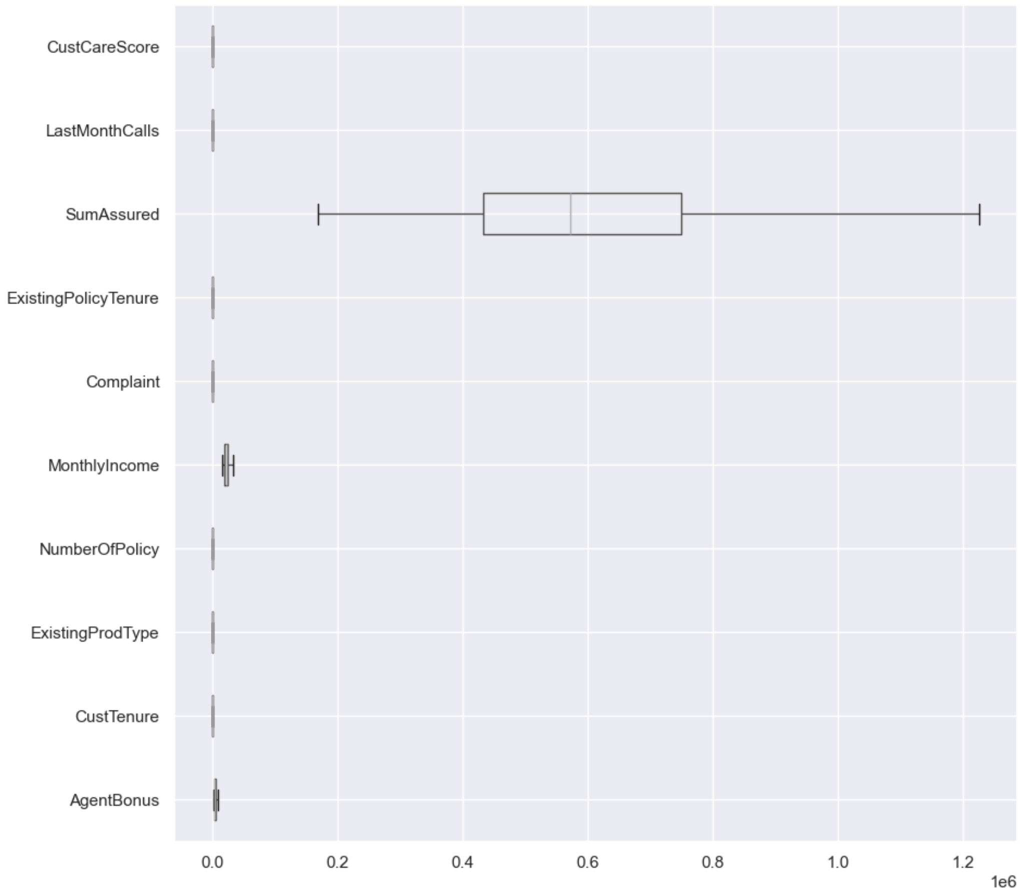


Fig 3.5.2 Outliers After Treatment

After treating the outliers, we can see that data is free from outliers.

3.6 Variable transformation:

We now transform our variables so that it is ready for further processing. We have created dummy variables for our categorical data with 0,1 encoding. Then we join our categorical and numerical data to form the whole dataset. Now we scale the dataset using standardscaler. Here is a snippet of the dataset after scaling.

	Channel_Online	Channel_Third Party Partner	Occupation_Large Business	Occupation_Salaried	Occupation_Small Business	EducationField_Engineer	EducationField_MBA	EducationField_Post Graduate	EducationField_Under Graduate	Gender_Male	...	AgentBonus	CustTenure
0	-0.339851	-0.484044	-0.314995	1.030555	-0.858560	-0.314995	-0.129012	-0.242990	0.611441	-1.211301	...	0.254928	-1.202108
1	-0.339851	2.065930	-0.314995	1.030555	-0.858560	-0.314995	-0.129012	-0.242990	0.611441	0.825559	...	-1.361260	-1.438151
2	-0.339851	-0.484044	-0.314995	-0.970351	-0.858560	-0.314995	-0.129012	4.115399	-1.635481	0.825559	...	0.154790	-1.202108
3	-0.339851	2.065930	-0.314995	1.030555	-0.858560	-0.314995	-0.129012	-0.242990	0.611441	-1.211301	...	-1.672717	-0.871647
4	-0.339851	-0.484044	-0.314995	-0.970351	1.164741	-0.314995	-0.129012	-0.242990	0.611441	0.825559	...	-0.815659	-0.375956

Fig 3.6.1 Data After Scaling

3.7 Addition of new variables:

We did not need to add any extra variable to the dataset at this point of time.

4. Business insights from EDA

4.1 Is the data unbalanced? If so, what can be done? Please explain in the context of the business:

We can not say anything about the unbalanced nature of the data as the problem is not a classification problem but a regression problem. We do not need to classify the data, but we have to predict the bonuses of agents.

4.2 Any business insights using clustering:

We clustered the data using K-means clustering and determined the appropriate number of clusters using silhouette score. We take the no of cluster of which silhouette score is the highest. In our case the optimal number of clusters is 3 which has the highest silhouette score of 0.1116.

```
Clus_kmeans3
1    0.599336
0    0.311062
2    0.089602
Name: proportion
```

Here is the distribution of customers in each cluster. We can see that cluster 1 has highest number of customers.

Here is a snippet which compares important variables across all the clusters.

Clus_kmeans3	CustTenure	MonthlyIncome	SumAssured	CustCareScore	ExistingPolicyTenure	NumberOfPolicy	LastMonthCalls	AgentBonus
0	20.0	27836.305929	834120.626600	3.0	5.0	4.0	7.0	5428.995733
1	11.0	20614.663126	506732.699961	3.0	3.0	4.0	3.0	3400.782577
2	12.0	22216.472585	595146.896641	3.0	3.0	4.0	4.0	3915.903704

Fig 4.2.1 Clusters

- We can see that cluster 2 comprises of 8.9 % of all customers. They are associated with the insurance company for 12 years, have mean monthly income of 22216, mean sum assured of 595146. They take on an average of 4 policies, each policy having tenure of 3 years, they also receive average of 4 calls for cross selling. Mean agent bonus they provide is 3915.
- We can see that cluster 1 comprises of 59.9% of all customers. They are associated with the insurance company for 11 years, have mean monthly income of 20614, mean sum assured of 506732. They take on an average of 4 policies, each policy having tenure of 3 years, they also receive average of 3 calls for cross selling. Mean agent bonus they provide is 3400.
- We can see that cluster 0 comprises of 31.1% of all customers. They are associated with the insurance company for 20 years, have mean monthly income of 27836, mean sum assured of 834120. They take on an average of 4 policies, each policy having tenure of 5 years, they also receive average of 7 calls for cross selling. Mean agent bonus they provide is 5428.
- We can see from the clusters that customers from cluster 0 are more years of association with company, they earn more and have policies which have higher sum assured. As a result, they also contribute higher agent bonuses. In terms of value to the company we can see that Cluster 0>Cluster 2>Cluster 1.

4.3 Any other business insights:

- As we can see sales from agents take the higher share of revenues. Company needs to improve online and third-party policy sales.
- While North and West region sales are high, East and South sales are close to none and company needs to do aggressive marketing in those areas and penetrate the market.
- Company needs to do focus more on Cluster 0 which has 31% share in sales. They are high income, high expenditure group who buy higher sum assured policies for longer tenures and subsequently contribute to higher agent bonuses.
- Company needs to improve their customer service as 3, which can be termed as average, is the most common rating customers provide.
- While 72% of customers did not register a complaint last month,28% registered complaint and it is a big number. Company needs to improve services so that these complaints come down drastically.

5. Model building and interpretation.

5.1 Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)

We now build various regression models and compare various metrics for regression like RMSE, MSE,MAE and R-squared.

- We prepare the data for model building. Since regression only takes numerical data as input, we had changed the categorical data to numerical data. We used dummy encoding to change the data to numerical data.

- We now define and divide the data into predictor and target variable.
- After that we divided the data into train and test data with a ratio of 75:25.

Our data is now ready for modelling. We will build various regression models and compare their test data performance against each other.

Linear Regression: -

We will use statsmodel library to build our linear regression model as it gives us better understanding of the features.

Linear regression assumes that there is no collinearity between variables. So, we have to remove collinearity from the variables first. We use VIF calculator to check VIF values and we will remove those features who have VIF values more than 5.

Here is a snippet of the VIF values sorted in deceneidng order.

	variables	VIF
3	Occupation_Salaried	109.303069
25	MonthlyIncome	100.605565
4	Occupation_Small Business	91.310275
23	ExistingProdType	66.337068
18	Zone_West	41.699919
2	Occupation_Large Business	40.104908
16	Zone_North	31.181761
5	EducationField_Engineer	20.823367
28	SumAssured	14.655058
31	Clus_kmeans3	12.246520
10	Designation_Executive	10.513577
8	EducationField_Under Graduate	9.109204
24	NumberOfPolicy	7.878140
11	Designation_Manager	7.650407
30	CustCareScore	5.972521
22	CustTenure	5.332561
14	MaritalStatus_Married	3.903527
27	ExistingPolicyTenure	3.386278
29	LastMonthCalls	3.219996

As we can check from the snippet that many variables have VIF values more than 5. We will remove those variables one by one as removing all at once will cause corruption in the data.

We have removed 9 varibales one by one and created final list of variables which will be used to build initial linear regression model. Here is a snippet of the VIF chart after removing all collinearity.

	variables	VIF
22	CustCareScore	4.992657
7	Designation_Executive	3.862980
18	CustTenure	3.819376
8	Designation_Manager	3.762367
11	MaritalStatus_Married	3.549114
20	ExistingPolicyTenure	2.940310
21	LastMonthCalls	2.893097
12	MaritalStatus_Single	2.619798
6	Gender_Male	2.418984
9	Designation_Senior Manager	2.302948
3	Occupation_Small Business	1.938090
13	Zone_North	1.709226
10	Designation_VP	1.550093
17	PaymentMethod_Yearly	1.539778
19	Complaint	1.390485
1	Channel_Third Party Partner	1.285436
2	Occupation_Large Business	1.230518
0	Channel_Online	1.151522

This is the final list of VIF checked variables which we will be using for linear regression model building.

We will now merge X and y as its required for the Statsmodel model building. We create a formula with VIF checked variables and pass that onto our regression model. Here is the formula we used .

f_1= 'AgentBonus ~ Channel_Online + Q("Channel_Third Party Partner") + Q("Occupation_Large Business") + Q("Occupation_Small Business") + EducationField_MBA + Q("EducationField_Post Graduate") + Gender_Male + Designation_Executive + Designation_Manager + Q("Designation_Senior Manager") + Designation_VP + MaritalStatus_Married + MaritalStatus_Single + Zone_North + Zone_South + PaymentMethod_Monthly + PaymentMethod_Quarterly + PaymentMethod_Yearly + CustTenure + Complaint + ExistingPolicyTenure + LastMonthCalls + CustCareScore'

We now build the model and here is the model summary.

OLS Regression Results			
Dep. Variable:	AgentBonus	**R-squared:**	0.511
Model:	OLS	**Adj. R-squared:**	0.508
Method:	Least Squares	**F-statistic:**	153.0
Date:	Sat, 16 Dec 2023	**Prob (F-statistic):**	0.00
Time:	13:09:35	**Log-Likelihood:**	-28029.
No. Observations:	3390	**AIC:**	5.611e+04
Df Residuals:	3366	**BIC:**	5.625e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3594.2034	102.277	35.142	0.000	3393.671	3794.735
Channel_Online	79.2272	55.776	1.420	0.156	-30.132	188.586
Q("Channel_Third Party Partner")	-22.6463	41.688	-0.543	0.587	-104.383	59.091
Q("Occupation_Large Business")	-27.9603	59.322	-0.471	0.637	-144.272	88.351
Q("Occupation_Small Business")	-48.1707	35.696	-1.349	0.177	-118.159	21.817
EducationField_MBA	70.9749	134.850	0.526	0.599	-193.422	335.372
Q("EducationField_Post Graduate")	14.9026	73.583	0.203	0.840	-129.369	159.174
Gender_Male	-15.5566	33.400	-0.466	0.641	-81.043	49.929
Designation_Executive	-1325.2277	69.881	-18.964	0.000	-1462.242	-1188.214
Designation_Manager	-1076.3397	67.449	-15.958	0.000	-1208.584	-944.095
Q("Designation_Senior Manager")	-513.2190	73.780	-6.956	0.000	-657.877	-368.561
Designation_VP	464.1806	94.149	4.930	0.000	279.585	648.776

We can check from the model summary that there are few variables which have p value more than 0.05. those variables are not significant in our model and we can remove them from our model to strip down our model.

After removing them from our formula here is the final formula.

f_2= 'AgentBonus ~ Designation_Executive + Designation_Manager + Q("Designation_Senior Manager") + Designation_VP + MaritalStatus_Single + Zone_North + PaymentMethod_Quarterly + CustTenure + ExistingPolicyTenure'

We now build the model and here is the model summary.

OLS Regression Results

Dep. Variable:	AgentBonus	**R-squared:**	0.509
Model:	OLS	**Adj. R-squared:**	0.508
Method:	Least Squares	**F-statistic:**	583.4
Date:	Sun, 17 Dec 2023	**Prob (F-statistic):**	0.00
Time:	15:29:56	**Log-Likelihood:**	-28038.
No. Observations:	3390	**AIC:**	5.609e+04
Df Residuals:	3383	**BIC:**	5.613e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	3638.8910	74.172	49.060	0.000	3493.464	3784.318
Designation_Executive	-1314.9767	67.396	-19.511	0.000	-1447.119	-1182.835
Designation_Manager	-1072.6283	66.655	-16.092	0.000	-1203.316	-941.941
Q("Designation_Senior Manager")	-521.2436	73.264	-7.115	0.000	-664.891	-377.596
Designation_VP	446.2691	93.455	4.775	0.000	263.036	629.502
CustTenure	64.6308	2.045	31.611	0.000	60.622	68.640
ExistingPolicyTenure	98.7090	5.514	17.900	0.000	87.897	109.521

Omnibus:	121.414	**Durbin-Watson:**	1.979
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	134.293
Skew:	0.487	**Prob(JB):**	6.90e-30
Kurtosis:	2.960	**Cond. No.**	151.

We can see that our model doesn't have any insignificant variable.

We finally predict our test data using this model.

```
The MSE of the model is 919855.6831659729
The RMSE of the model is 959.0910713618248
The MAE of the model is 775.6911195409459
The R2 of the model is: 0.5202917318276443
```

As we can see from this metrics that our model did not perform well and we have to move to more complex regression models.

<u>Random Forest: -</u>

We will now build our first ensemble model which is Random Forest. This model uses bagging technique to grow trees. We build the model without any tuning and random_state=42.

Here is the result of our vanilla randomforest model.

```
The MSE of the model is 315162.51731048676
The RMSE of the model is 561.3933712740886
The MAE of the model is 429.4654336283186
The R2 of the model is: 0.8356415379731099
```

As we can see from the snippet that the model performance improved significantly .

XGBoost:-

We will now build our fist model using boosting technique. We will build vanilla XGBoost model with random_state=42 and check performance.

```
The MSE of the model is 351309.51837777055
The RMSE of the model is 592.713690054288
The MAE of the model is 453.5307895896709
The R2 of the model is: 0.8167907382238166
```

As we can see the model performed good on our data. WE will still build few more models to check if we can get better results.

LightGBM:-

This is arelaitively new model built by Microsoft which uses boosting . We will build vanilla LightGBM model with random_state=42 and check performance.

```
The MSE of the model is 312044.38689102884
The RMSE of the model is 558.6093329788081
The MAE of the model is 430.6758921850999
The R2 of the model is: 0.8372676549508355
```

We can see vanilla LightGBM model performs better on our data than other models. R-squared and RMSE has improved significantly.

5.2 Test your predictive model against the test set using various appropriate performance metrics.

We will now check our vanilla model performance using the test data. The evaluation metrics used are MSE,RMSE,MAE and R-squared.

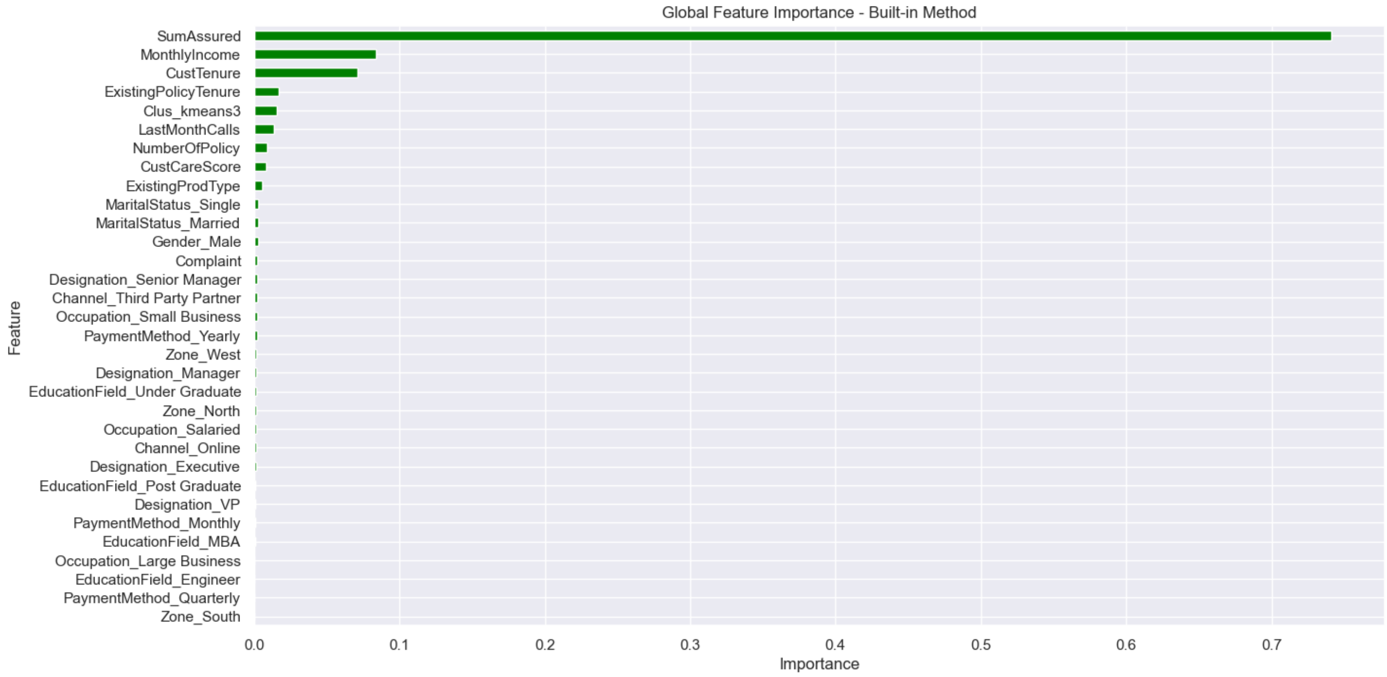
	Model	MSE	MAE	RMSE	R2
0	Linear regression	919855.683166	775.691120	959.091071	0.520292
1	Random_Forest	315162.517310	429.465434	561.393371	0.835642
2	XGBoost	351309.518378	453.530790	592.713690	0.81679
3	LightGBM	312044.386891	430.675892	558.609333	0.837268

As we can see from the comparison table of test data performance of our models, Linear Regression could not explain our data . Vanilla random forest and LightGBM did a better job reducing RMSE,MSE and MAE and had better R-squared which indicates how much variation in our data can be explained by our model.

5.3 Interpretation of the model(s)

We will now try to interpret two of our better performing models with feature importance and see which features played significantly in our model.

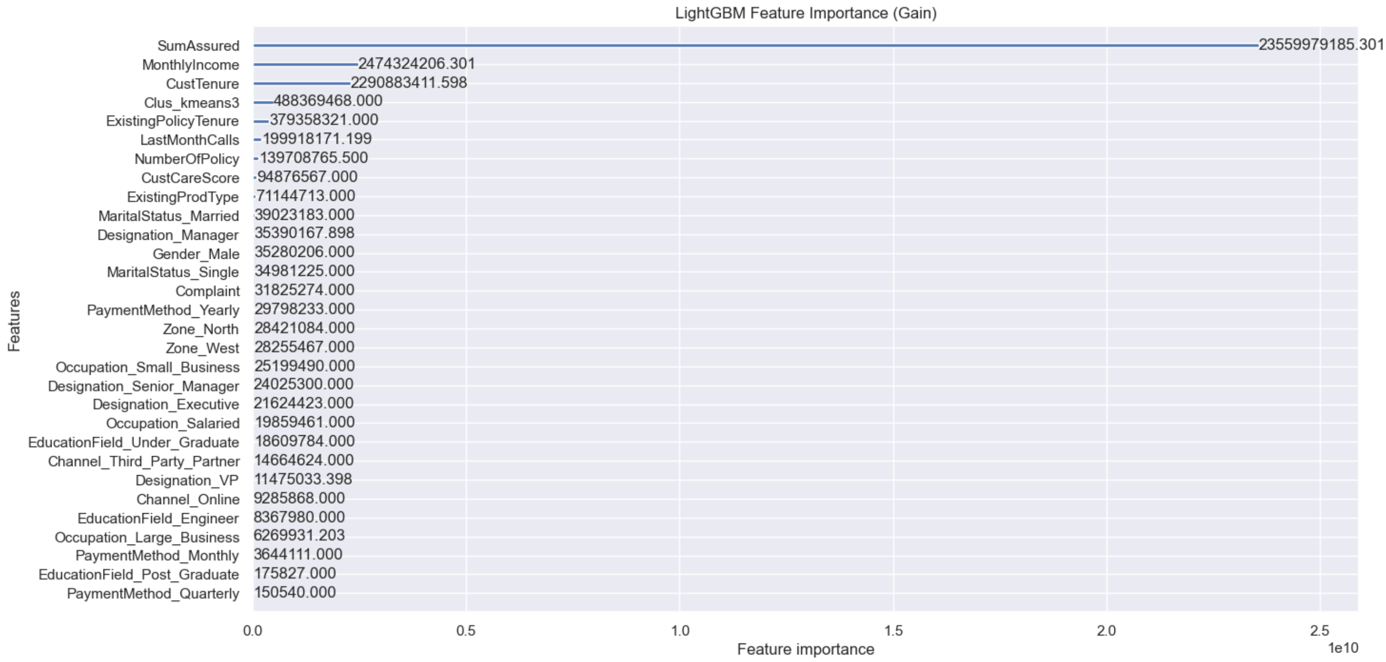
Random_Forest:-



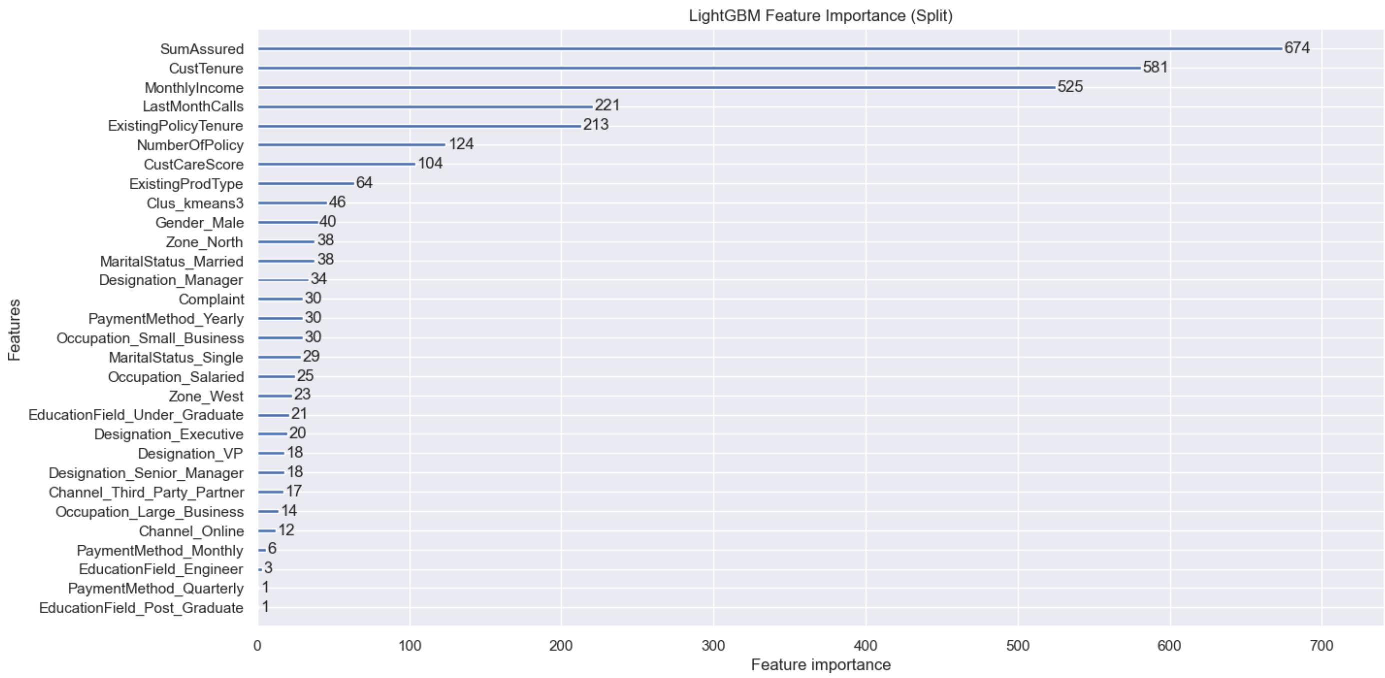
As we can see from the above feature importance plot , SumAssured played the most important role in our model. Zone_south is our least important feature.

LightGBM:-

We check both gain and split feature importance for our LightGBm model.



We can see from our gain feature importance that Sum_assured is still our most important feature and Paymentmethod_quarterly is our least important feature.



We can see from our split feature importance that Sum_assured is still our most important feature and educationField_Post_graduate is our least important feature.

Combining these two interpretations we can safely say that Sum_assured is our most important feature and educationField_Post_graduate & Paymentmethod_quarterly are our least important features.

6. Model Tuning and business implication

6.1 Ensemble modelling, wherever applicable

As we can see Ensemble modelling gives us the most accurat predictions with lowest RMSE and highest R-squared.

Among all the models LightGBM has the highest R-squared and lowest RMSE value. So we will take this model as our preferred model and will tune it further.

Our vanilla LightGBM model has R-sqaured value 0. 837268 and RMSE value 558.609333.

6.2 Any other model tuning measures(if applicable)

We will tune our LightGBM model with bayesian optimization instead of GridSearch or RandomSearch. RandomSearch randomly picks up values from parameter grid and tries to provide lower RMSE and Gridsearch checks all the values in the grid. These tuning techniques are good but they take a lot of compute and unnecessary parameter searches. Whereas Bayesian Optimization is an approach that uses Bayes Theorem to direct the search in order to find the minimum or maximum of an objective function which is a more efficient way of searching hyperparameters. We will use Optuna library for our Bayesian Optimization and direct our model to search for lowest RMSE.

After searching for hyperparameter with lowest RMSE we have got our optimal hyperparameter.

'max_depth': 7, 'lambda_l1': 2.1182744379105e-08, 'lambda_l2': 9.308098753900692, 'num_leaves': 156, 'feature_fraction': 0.9942536778538993, 'bagging_fraction': 0.992955657382847, 'bagging_freq': 9, 'min_child_samples': 1

These are our optimal hyperparameters. We will run our LightGBM model with these hyperparameters and check for evaluation metrics.

```
The MSE of the model is 306341.3353369492
The RMSE of the model is 553.4811065763214
The MAE of the model is 426.919466756832
The R2 of the model is: 0.8402418182183702
```

We can see that our tuned LightGBM model has lower RMSE and higher R-sqaured than before. Hence this is our optimal model for the given data.

6.3 Interpretation of the most optimum model and its implication on the business

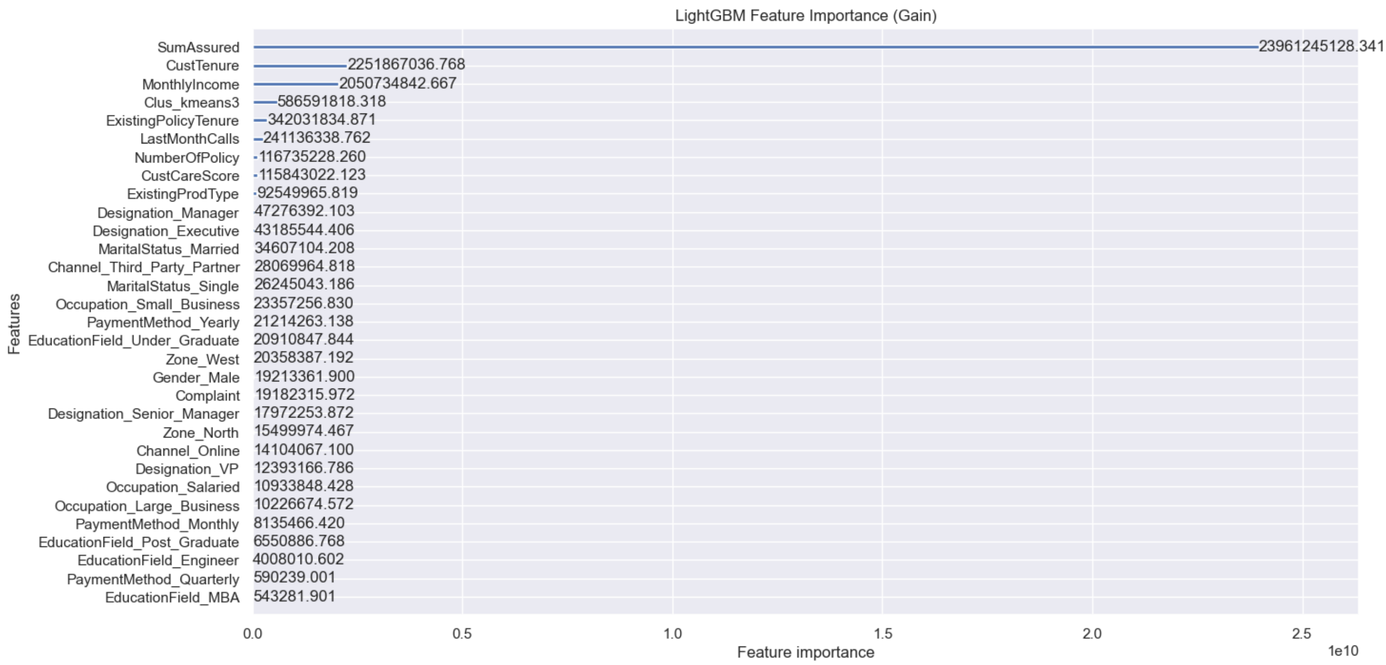
We can see that LightGBM with hypertuned parameters is our best model for our data.

Before delving into the specific methods for calculating feature importance in LightGBM, it's crucial to understand that there are two primary methods to do so: gain and split.

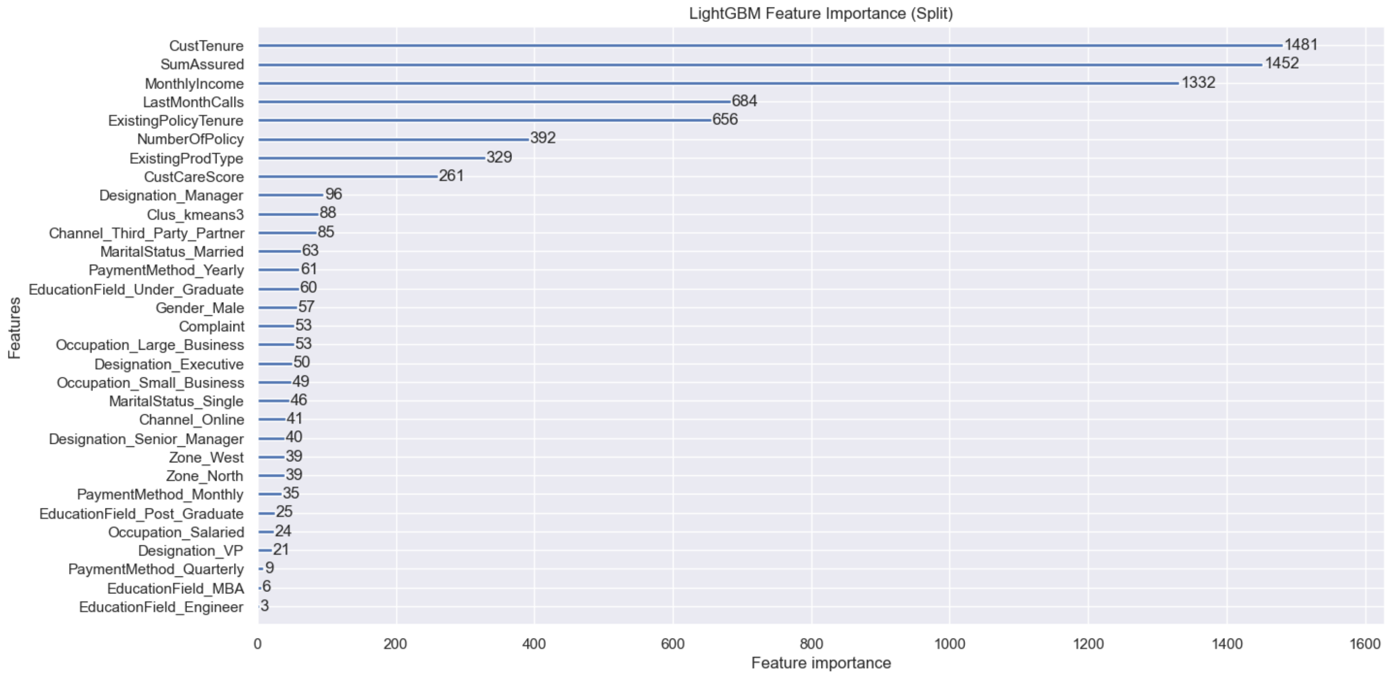
Each method looks at different aspects of the decision trees within the model to derive the importance of features. Here's a brief overview of each method:

- **Gain**: The gain method calculates feature importance based on the improvement in the splitting criterion (e.g., Gini impurity, information gain, squared error) that results from using a feature in a tree's split. In other words, it measures how much a feature contributes to reducing the overall error or increasing the purity of nodes in the trees.
- **Split**: The split method, on the other hand, calculates feature importance by counting how many times a feature is used to split nodes across all the trees in the model. This method focuses on the frequency of a feature being used in the trees, as it assumes that more frequently used features are more important.

Both methods have their assumptions and are not perfect. However, they can provide valuable insights into the features of your model.



We can see from our gain feature importance that Sum_assured is still our most important feature,CustTenure being the second important and Paymentmethod_quarterly and EducationField_MBA are our least important features.



As We can see from our split feature importance that CustTenure ,which is used 1481 times to split nodes ,is still our most important feature, Sum_assured ,which is used 1452 times to split nodes,being the second important and Paymentmethod_quarterly , EducationField_MBA and EducationField_Engineer are our least important features.

So by comparing and combining both split and gain feature importance we can safely say that CustTenure, Sum_assured are our most important features and Paymentmethod_quarterly , EducationField_MBA and EducationField_Engineer are our least important features.

<u>Insights from Analysis:-</u>

- Company wants to predict the ideal bonus and what is the analysis for high and low performing agents
- respectively.
- From the model, the high performing agent we will find variable significance, for eg, Sum Assured and CustTenure are highly significant here and highly correlated to our target variable.

- SumAssured is highly significant as the agents who perform good are the ones who are getting more profit for the company selling more high value policies.
- If the Designation is VP the person buys more number of policies and higher valued policies.
- Therefore, for high and low performing agents, we will train them, suggesting them to purchase or get policies with high sum assured.
- Another important feature is Customer tenure where the agents need to focus on retaining customers for longer periods of time.
- Focusing on customers with greater monthly incomes as greater the monthly income, greater is the possibility of the customer buying a higher valued policy.
- From our models we can find insights and remove all the least significant variables.

Recommendations:-

- For High Performing Agents we can create a healthy contest with a threshold. Where, if they achieve the desired sum assured, they are eligible for certain incentives like latest gadgets, exotic family vacation packages and some extra perks as well.
- For low performing agents, we can introduce certain feedback upskill programs to train them into closing higher sum assured policies, reaching certain people to ultimately becoming top/high performers.
- Apart from this, we need more data/predictors like Premium Amount, this will help us to solve the business problem even better as well have more variables to test upon thereby having more accurate results in real time problems like this.
- I also feel another predictor can be added as customers geographical location or Region and not just the zones as people living in rural areas are less likely to buy a policy whereas those living in a highly developed location are likely to be belonging to the upper class and should be targeted.
- As we can see sales from agents take the higher share of revenues. Company needs to improve online and third-party policy sales.
- While North and West region sales are high, East and South sales are close to none and company needs to do aggressive marketing in those areas and penetrate the market.
- Company needs to do focus more on Cluster 0 which has 31% share in sales. They are high income, high expenditure group who buy higher sum assured policies for longer tenures and subsequently contribute to higher agent bonuses.
- Company needs to improve their customer service as 3, which can be termed as average, is the most common rating customers provide.
- While 72% of customers did not register a complaint last month,28% registered complaint and it is a big number. Company needs to improve services so that these complaints come down drastically.