Capstone Project Insurance

By

Gaurav Akotkar

Contents

1. Introduction of the Business Problem

a) Defining 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.

Below is Data Dictionary of dataset Sales.xlsx

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

Table 1: Data Dictionary

b) Need of the study/project

In India there are 24 life insurance companies, among Life Insurance companies, Life Insurance Corporation is only public sector company. Yearly life insurance business is growing by 10.73% year on year. Claim settlement ratio of insurance for 2021-2022 is 98.64 % from 98.39% in previous year.

In this Project we have to predict the bonus amount for the Agents working in the Life Insurance Company. This would help to understand performance of the agents and would help us organise upskill program for underperforming agent.

c) Understanding business/social opportunity

This project will help us to encourage Agents to perform well. Giving awards for performers and providing upskill programs will help Agents to bring customers for the companies thus increases sales of company. Large amount would be received by company to invest in various schemes and would perform financially and in stock markets which would certainly make their investors happy by providing good dividends and would attract new investors.

Life insurance is necessary financial aid. It helps family of person financially if there is sudden demise of the insurer hence, they can have a safe future. In recent past we have gone through COVID pandemic where many people have lost their dear ones and some of them were only bread earner of their family. Thus, those who had got their Life Insurance has insured their family with certain sum which they can use for various important purpose like education, daily survival etc. According to latest Economy survey 3 out of 100 Indian are having Life insurance which is very low compared to developed country like America where 52% of people having life insurance. Hence, having life insurance is very important and should be initial step in investing money to avoid individual's family fall into financial menace.

2. Data Report

a) Visual inspection of data (rows, columns, descriptive details)

There are 4520 rows and 20 Columns in data. The is data is having numerical columns as well as string columns.

Below table shows type of variables in dataset.

Data	Variable	Data Type
Sales	AgentBonus	Continous
Sales	Age	Continous
Sales	CustTenure	Continous
Sales	Channel	Categorical
Sales	Occupation	Categorical
Sales	EducationField	Categorical
Sales	Gender	Categorical
Sales	ExistingProdType	Categorical
Sales	Designation	Categorical
Sales	NumberOfPolicy	Categorical
Sales	MaritalStatus	Categorical
Sales	MonthlyIncome	Continous
Sales	Complaint	Categorical
Sales	ExistingPolicyTenure	Continous
Sales	SumAssured	Continous
Sales	Zone	Categorical
Sales	PaymentMethod	Categorical
Sales	LastMonthCalls	Continous
Sales	CustCareScore	Categorical

Table 2: Data Type

b) Understanding of attributes (variable info, renaming if required)

Below table show Info of the variable in the dataset.

Sr.No	Column Name	Non-Null Count	Data Type
1	CustID	4520 non-null	int64
2	AgentBonus	4520 non-null	int64
3	Age	4251 non-null	float64
4	CustTenure	4294 non-null	float64
5	Channel	4520 non-null	object
6	Occupation	4520 non-null	object
7	EducationField	4520 non-null	object
8	Gender	4520 non-null	object
9	ExistingProdType	4520 non-null	int64
10	Designation	4520 non-null	object
11	NumberOfPolicy	4475 non-null	float64
12	MaritalStatus	4520 non-null	object

13	MonthlyIncome	4284 non-null	float64
14	Complaint	4520 non-null	int64
15	ExistingPolicyTenure	4336 non-null	float64
16	SumAssured	4366 non-null	float64
17	Zone	4520 non-null	object
18	PaymentMethod	4520 non-null	object

Table 3: Variable Info

From above table we can see that there is variable of object, int64, and float64. Also, there are some null values present in data.

Below table shows Descriptive analysis of continuous variables present in data.

	count	mean	std	min	25%	50%	75%	max
AgentBonus	4520	4077.838	1403.322	1605	3027.75	3911.5	4867.25	9608
Age	4251	14.49471	9.037629	2	7	13	20	58
CustTenure	4294	14.46903	8.963671	2	7	13	20	57
MonthlyIncome	4284	22890.31	4885.601	16009	19683.5	21606	24725	38456
SumAssured	4366	619999.7	246234.8	168536	439443.3	578976.5	758236	1838496
LastMonthCalls	4520	4.626991	3.620132	0	2	3	8	18
ExistingPolicyTenure	4336	4.130074	3.346386	1	2	3	6	25

Table 4: Descriptive Analysis of Continuous Variable

From above table we can see that none of continuous variable is normally distributed as median and mean are not equal. In age column we can see that minimum value is 2 years is very rare in Life Insurance company.

Below image shows the unique values of object data type Categorical variable.

```
Channel = ['Agent' 'Third Party Partner' 'Online']


Occupation = ['Salaried' 'Free Lancer' 'Small Business' 'Laarge Business'
 'Large Business']


EducationField = ['Graduate' 'Post Graduate' 'UG' 'Under Graduate' 'Engineer' 'Diploma'
 'MBA']


Gender = ['Female' 'Male' 'Fe male']


Designation = ['Manager' 'Exe' 'Executive' 'VP' 'AVP' 'Senior Manager']


MaritalStatus = ['Single' 'Divorced' 'Unmarried' 'Married']


Zone = ['North' 'West' 'East' 'South']
```

Figure 1: Unique values in Categorical Variable

From image Figure 1 we can see that in Occupation column two values named Laarge Business and Large Business as no such word 'Laarge' is there, maybe it is typing error. Similarly in

Gender column two values named Female and Fe male. Female is correct and later is wrong. The correct values are replaced in the dataset.

Below table shows numeric categorical variable.

Variable	Count	Max	Min
ExistingProdType	4520	6	1
NumberOfPolicy	4475	6	1
Complaint	4520	1	0
CustCareScore	4468	5	1

Table 5: Categorical Variable (Numeric)

From counts we can see that there are null values present in the data. Columns Complaint is having two levels 0 and 1. Other’s columns are 5-6 levels.

3. Exploratory data analysis

a) Removal of unwanted variables (if applicable)

For analysis Columns CustID i.e., Customer ID is removed as the columns is not used for analysis.

b) Correction in Values of Variables.

Variables Occupation and Gender have some misspell values such as Large Business is misspelt as Laarge Business and Female misspelt as Fe male. These values are replaced by their correct values.

Similarly in Education UG is replaced by Under Graduate and in Designation Exe is replaced by Executive.

c) Univariate Analysis

Below is the distribution of continuous variables.

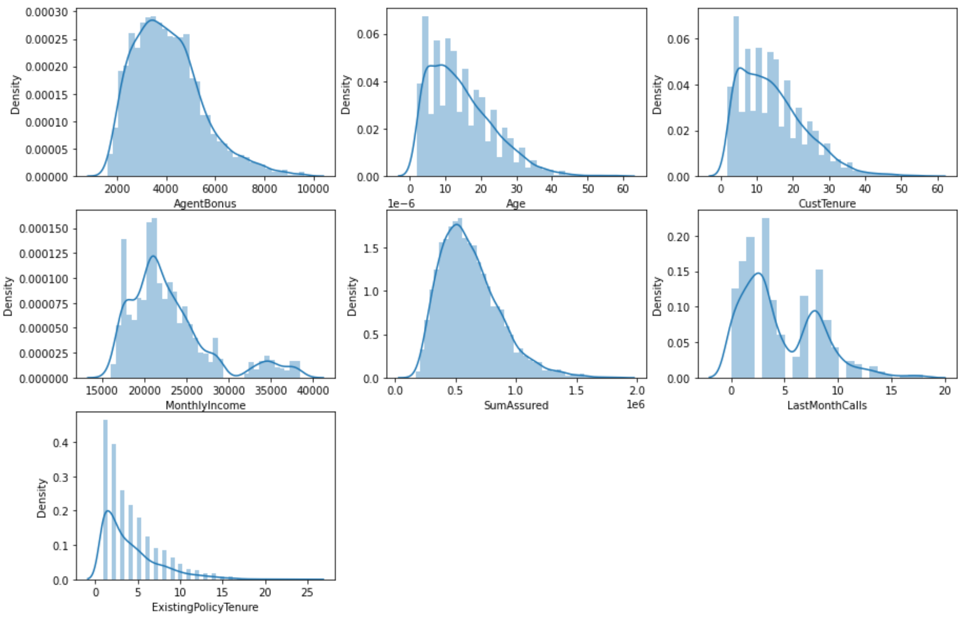

Figure 2: Continuous Variable Histogram

From the Histogram shown in above images we can see that data is right skewed i.e., positive skewed. Below table shows skewness values.

Continuous Variable	Skewness
AgentBonus	0.822348
Age	0.941341
CustTenure	0.93371
MonthlyIncome	1.363615
SumAssured	0.96932
LastMonthCalls	0.810417
ExistingPolicyTenure	1.539933

Table 6: Skewness of Continuous Variables

Below is the Boxplot for the continuous variables.

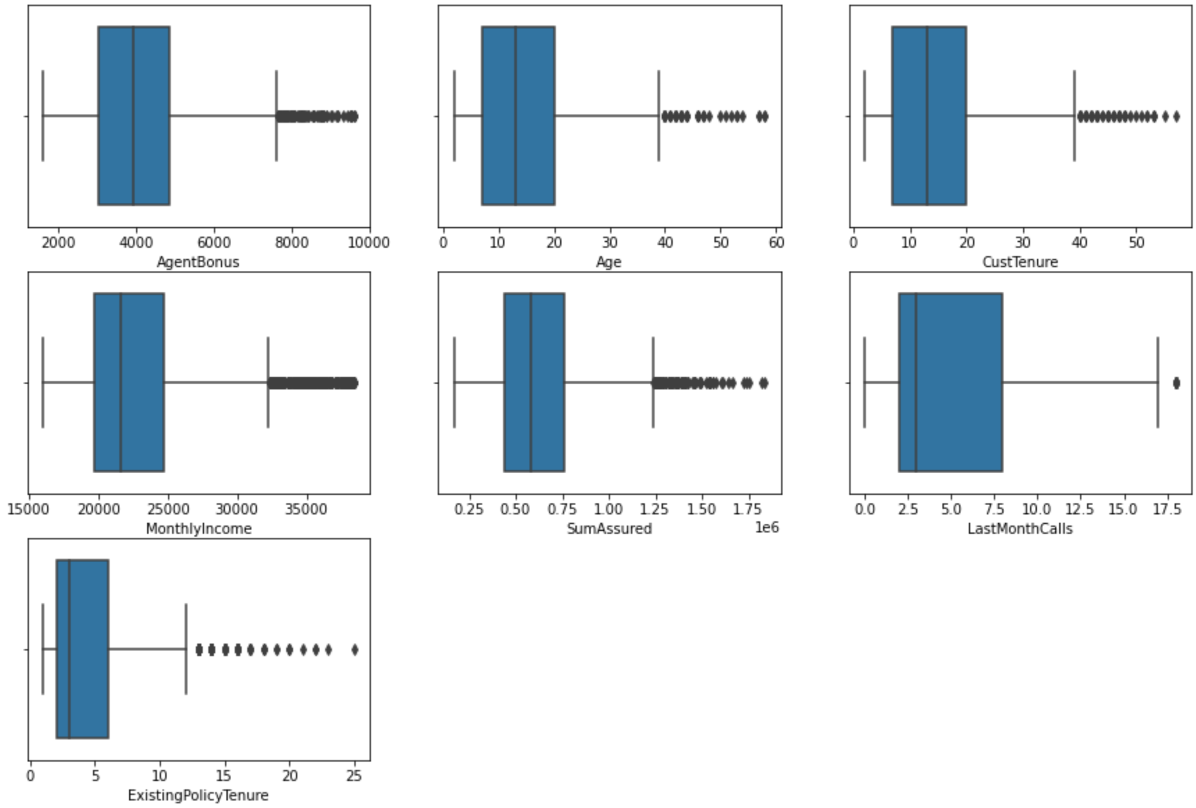

Figure 3: Boxplot of Continuous Variables

From above boxplots we can see that each continuous variable has outlier present in the variable. Outliers are those values which are greater or less than1.5 x Interquartile Range. Interquartile Range + 75 percentile of data or - 25 percentile of data respectively. Also, we can see that data is right skewed.

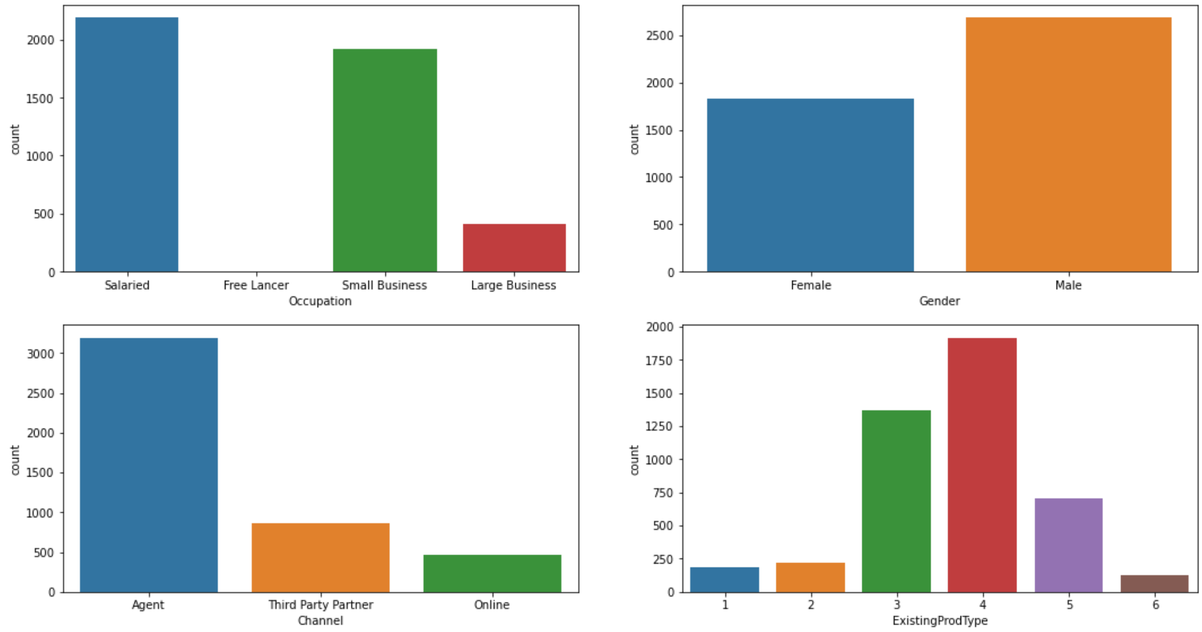

Figure 4: Count Plot of Categorical Variables1

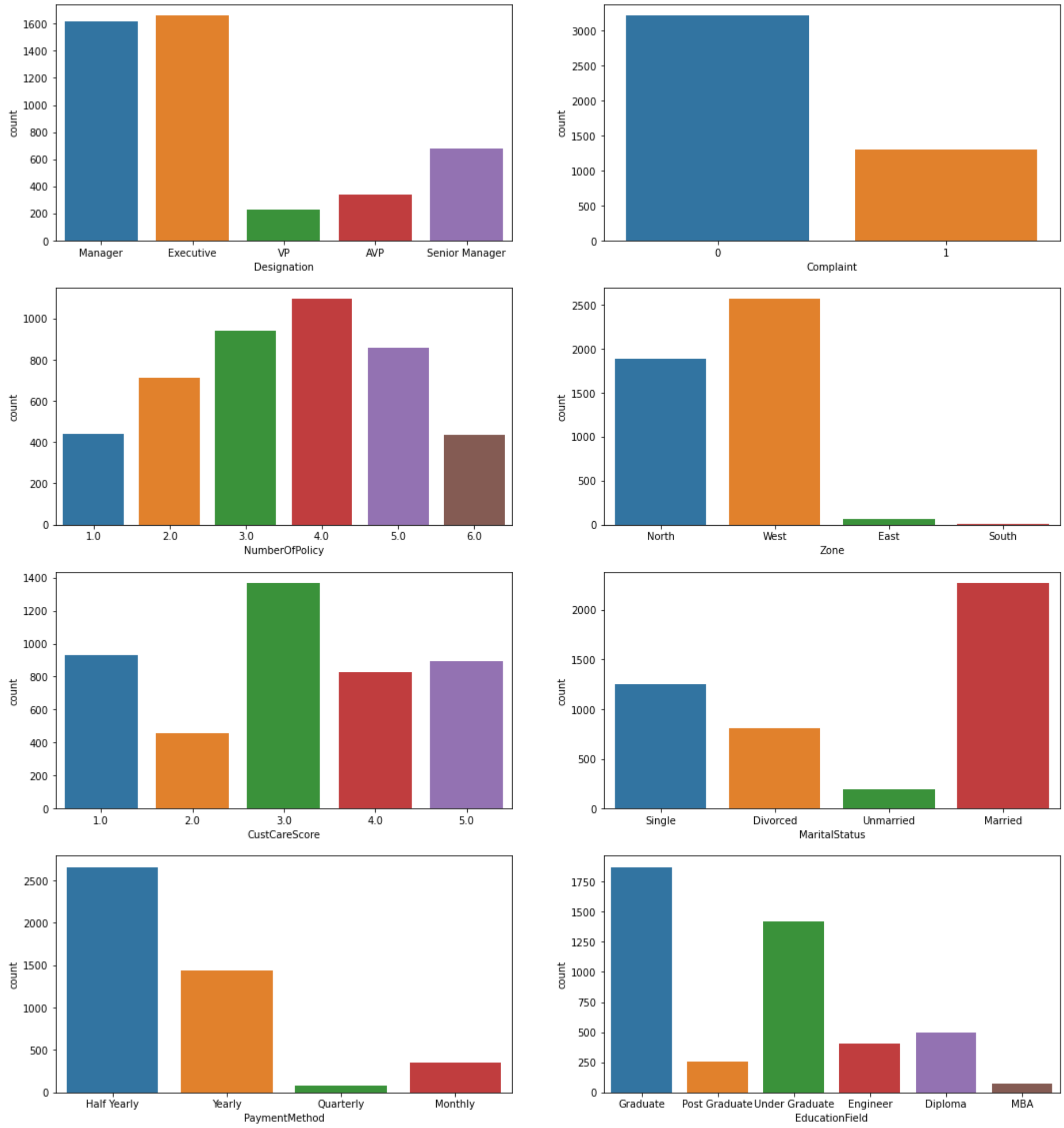

Figure 5: Count Plot of Categorical Variable 2

From above image we can see some data imbalance in categorical variables. But all these are independent variable so treatment not required.

Insights:

1. More male candidates are taking insurance.
2. People prefer Agent than any other channel.
3. People from west and North prefer insurance from this company. Very from other regions.
4. People prefer Half yearly payment and rarely prefer quarterly or monthly option.
5. Married people take more insurance may be because married people have more responsibility.

6. Most of the insurers are Graduates and working as Managers.
7. Customer care score is mostly 3 shows that customer executives should work on their communication skills.

d) Bivariate Analysis

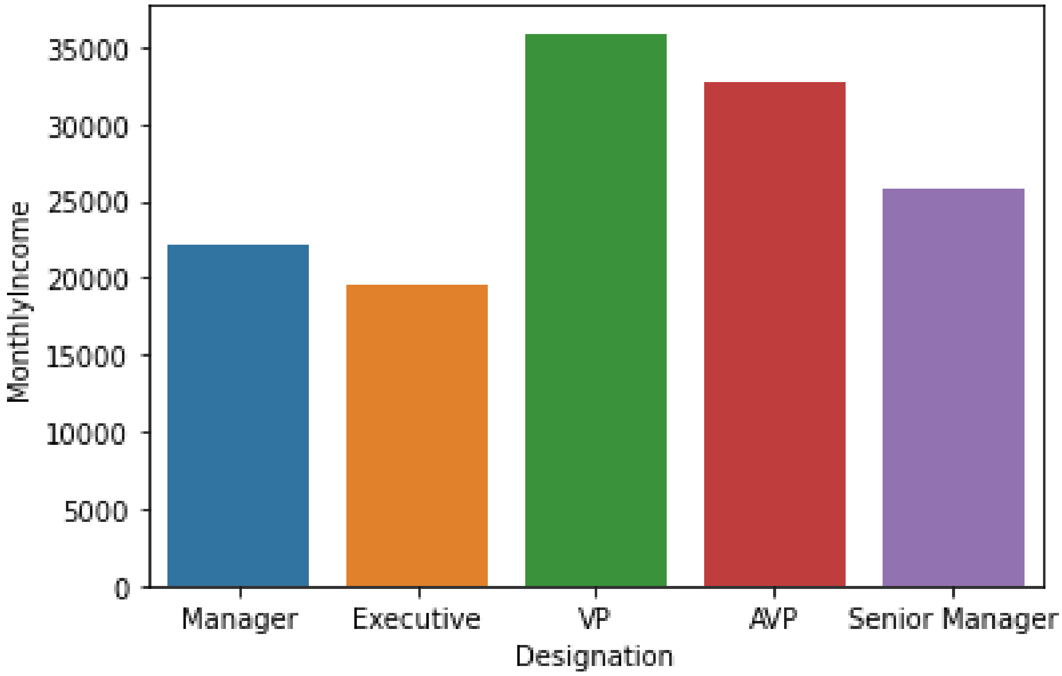

Figure 6: Designation vs Monthly Income

From above image we can see that income of VPs is highest and executive are lowest. Hence shows the higher the position more is the salary.

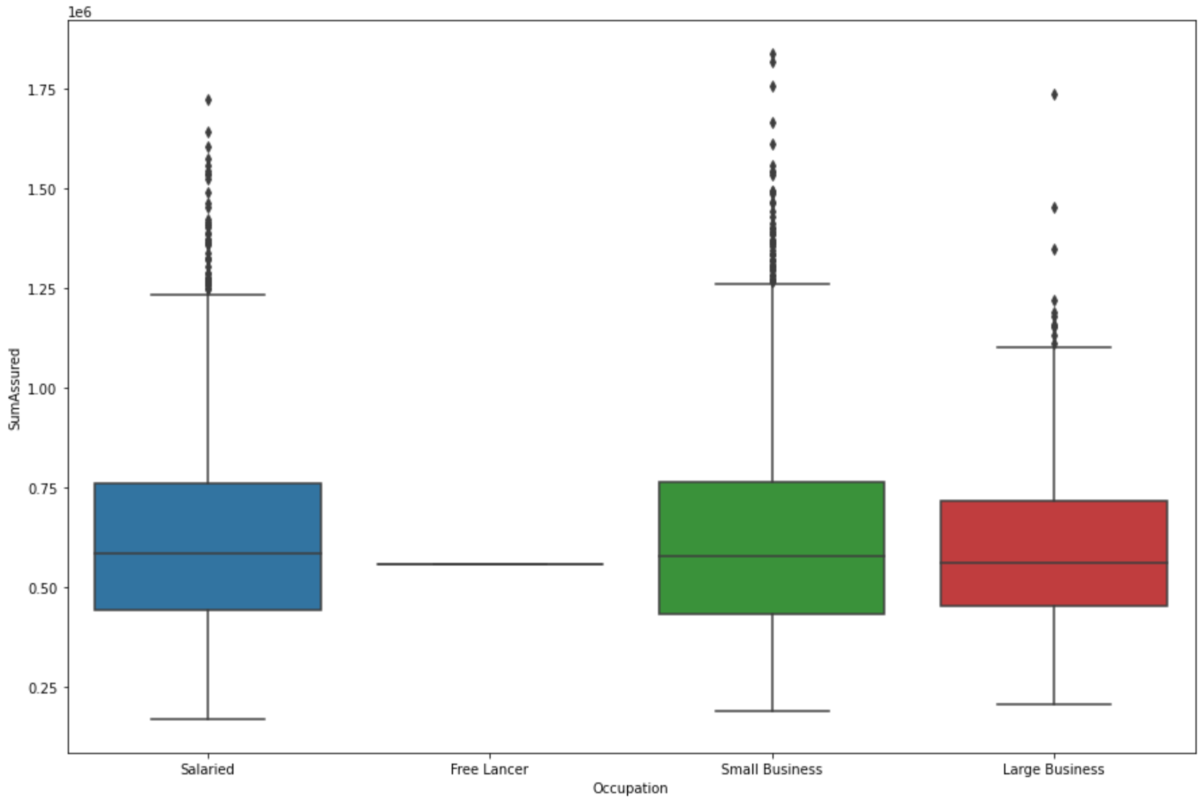

Figure 7: Box Plot Occupation vs Sum Assured

From Box Plot shown we can see that median are almost equal. Sum assured to small business is highest. Freelance have assured less amount.

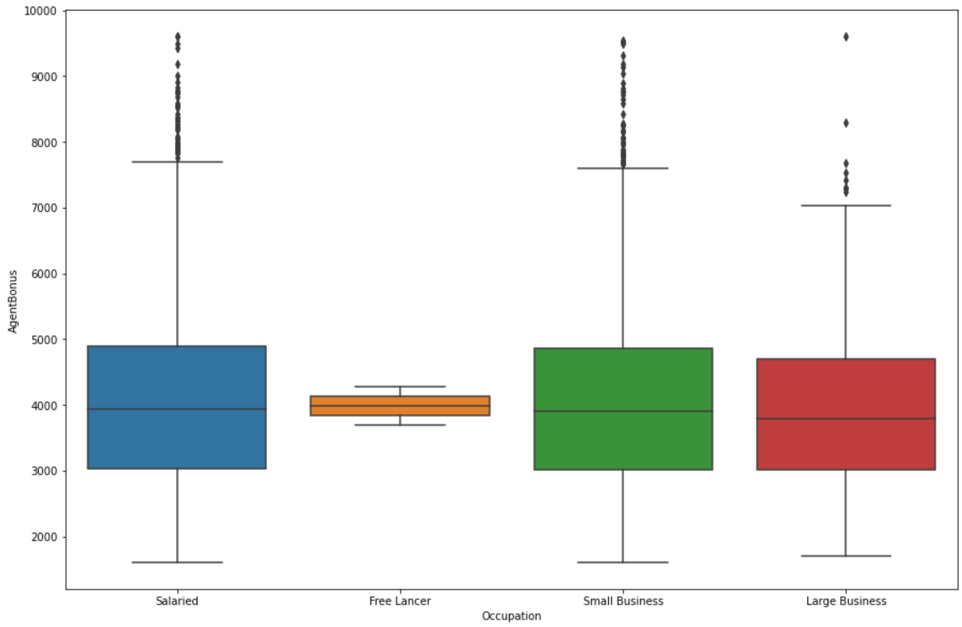

Figure 8: Box Plot Agent Bonus vs Occupation

From above image we can see that median of Agent Bonus for all Occupations are almost same. There are very few Freelancers who are insured.

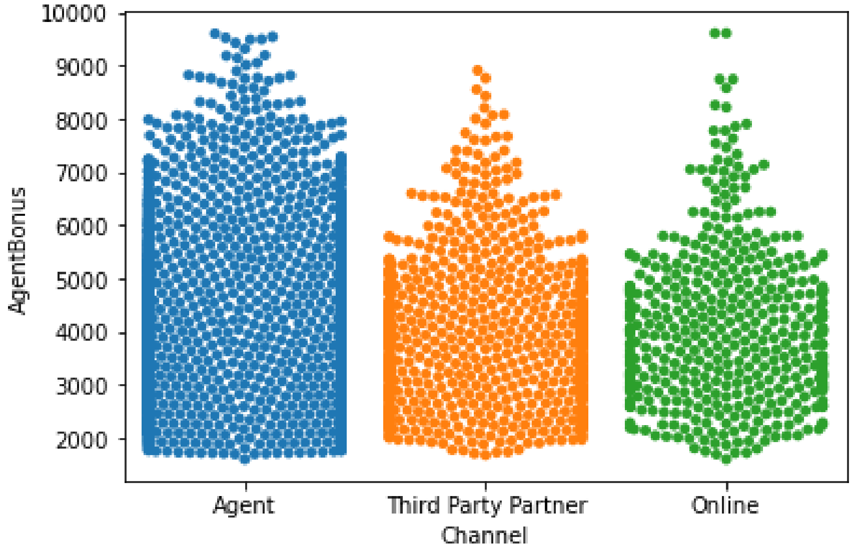

Figure 9: Swarm Plot Agent Bonus vs Channel

From above swarm plot if agent channel is preferred agent gets more Bonus. Also, people prefer Agent channel than other.

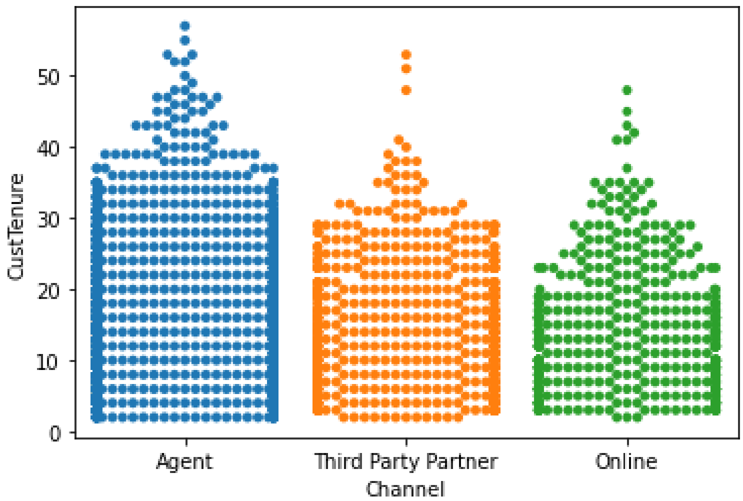

Figure 10: Swarm Plot Channel vs Customer Tenure

From above plot we can see that customers stays longer if they are taking insurance form agent. Online platform has less tenure. Shows that agents explain their policies well than the data available on internet.

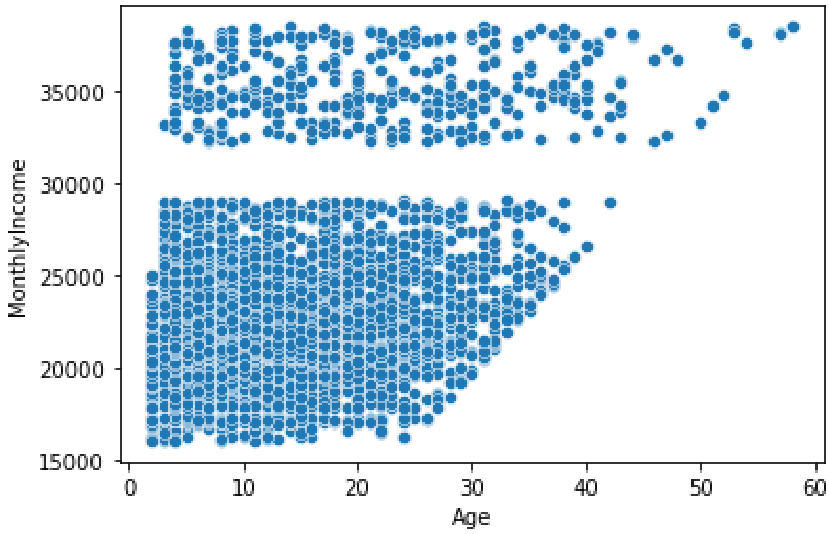

Figure 11: Age vs Monthly Income

From above image we can see that after age of 25 Monthly income increases with increase in age may be with age people gets promoted and their income increases.

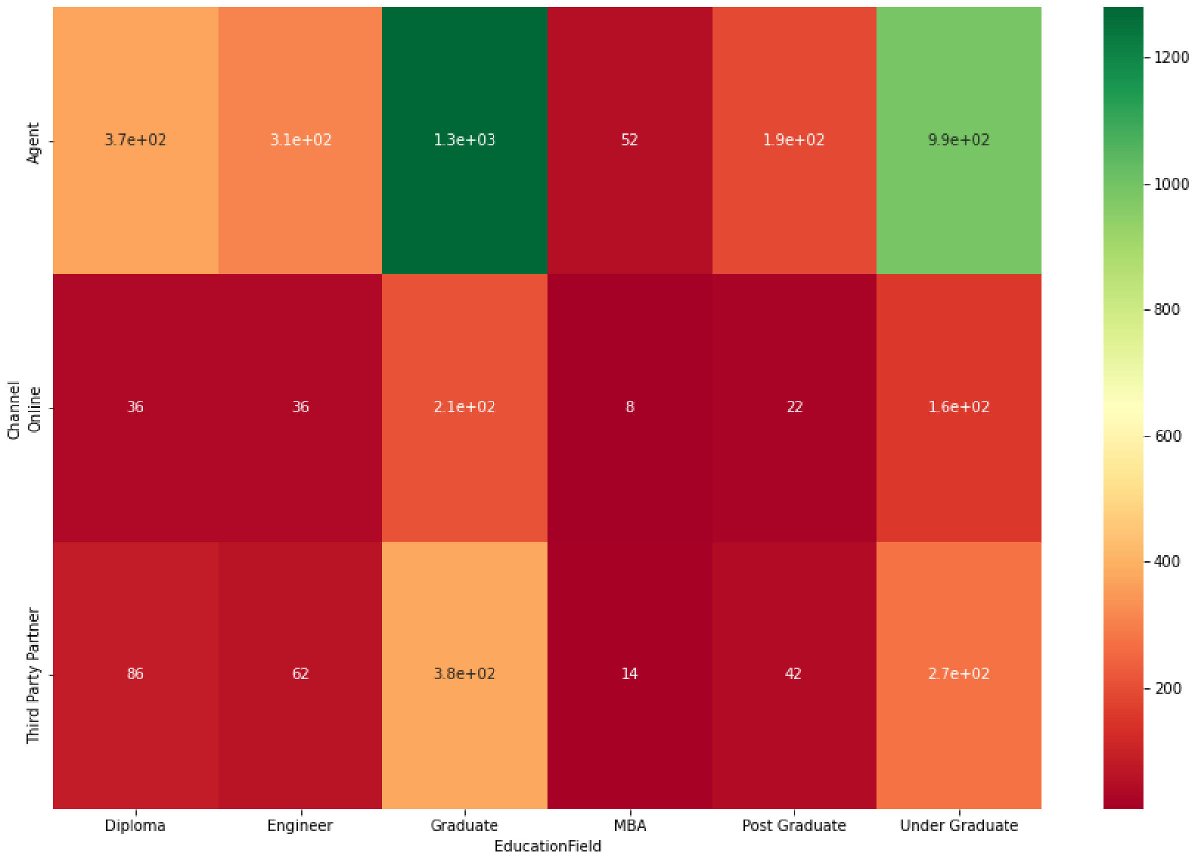

Figure 12: Heatmap Education vs Channel

From Figure11 we can see that insurance is mostly taken from Agents shows that working people have less time to think about insurance and hand over their insurance work to their agents.

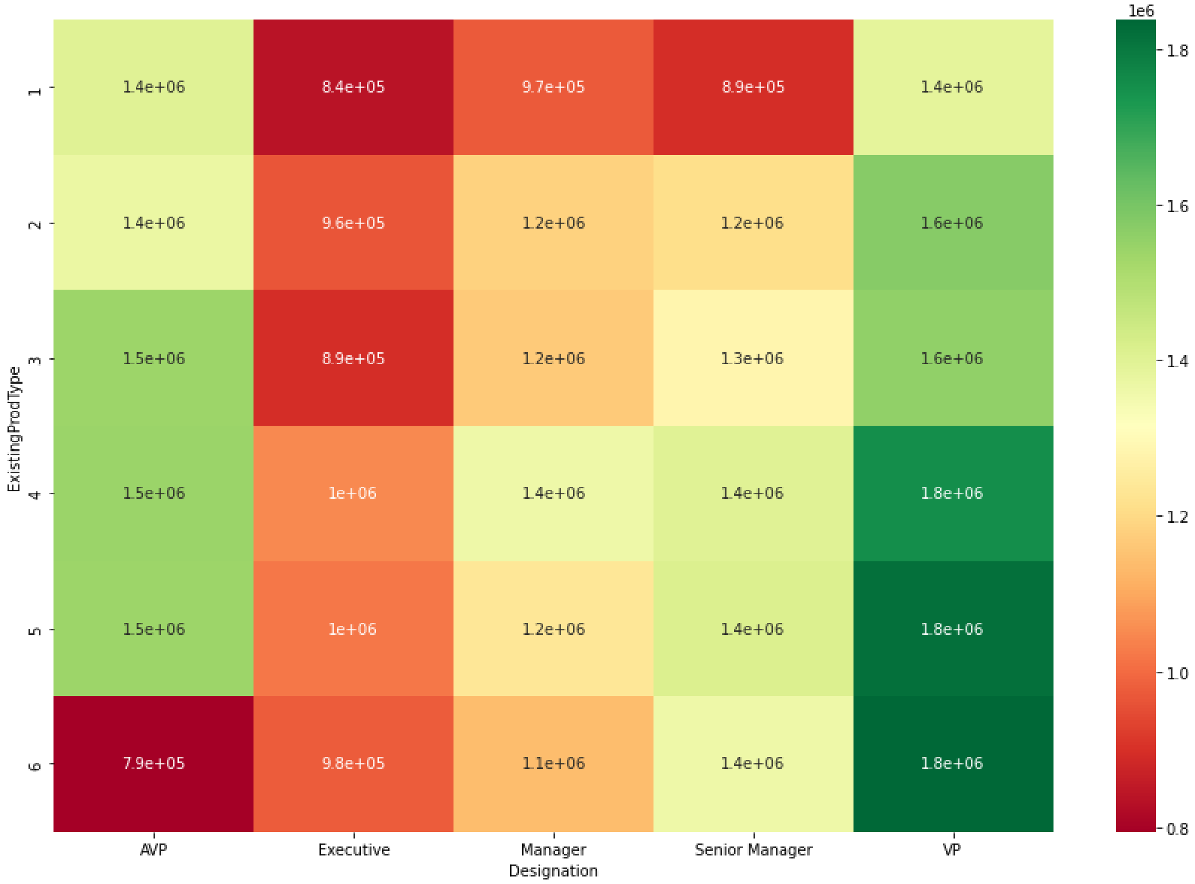

Figure 13: Sum Assured in Policy Type

From above plot we can see that maximum sum assured in policy type 6. Premium for such products are high as such policy are taken by VP and their monthly income is high.

e) Muti Variate Plots

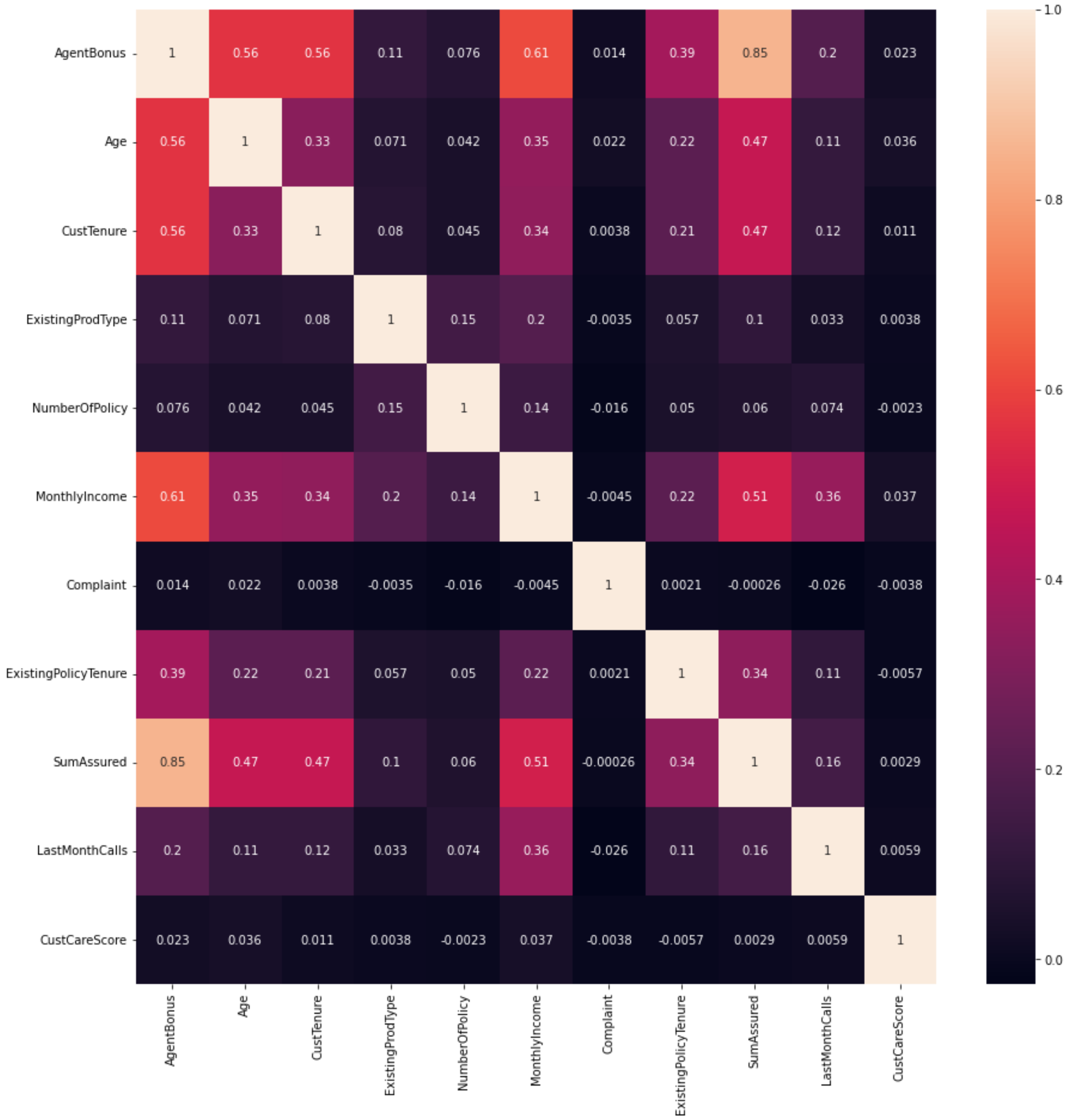

Figure 14: Correlation Plot

From above correlation plots we can see that most of the variables are independent. Agent Bonus is mostly depended on other variables.

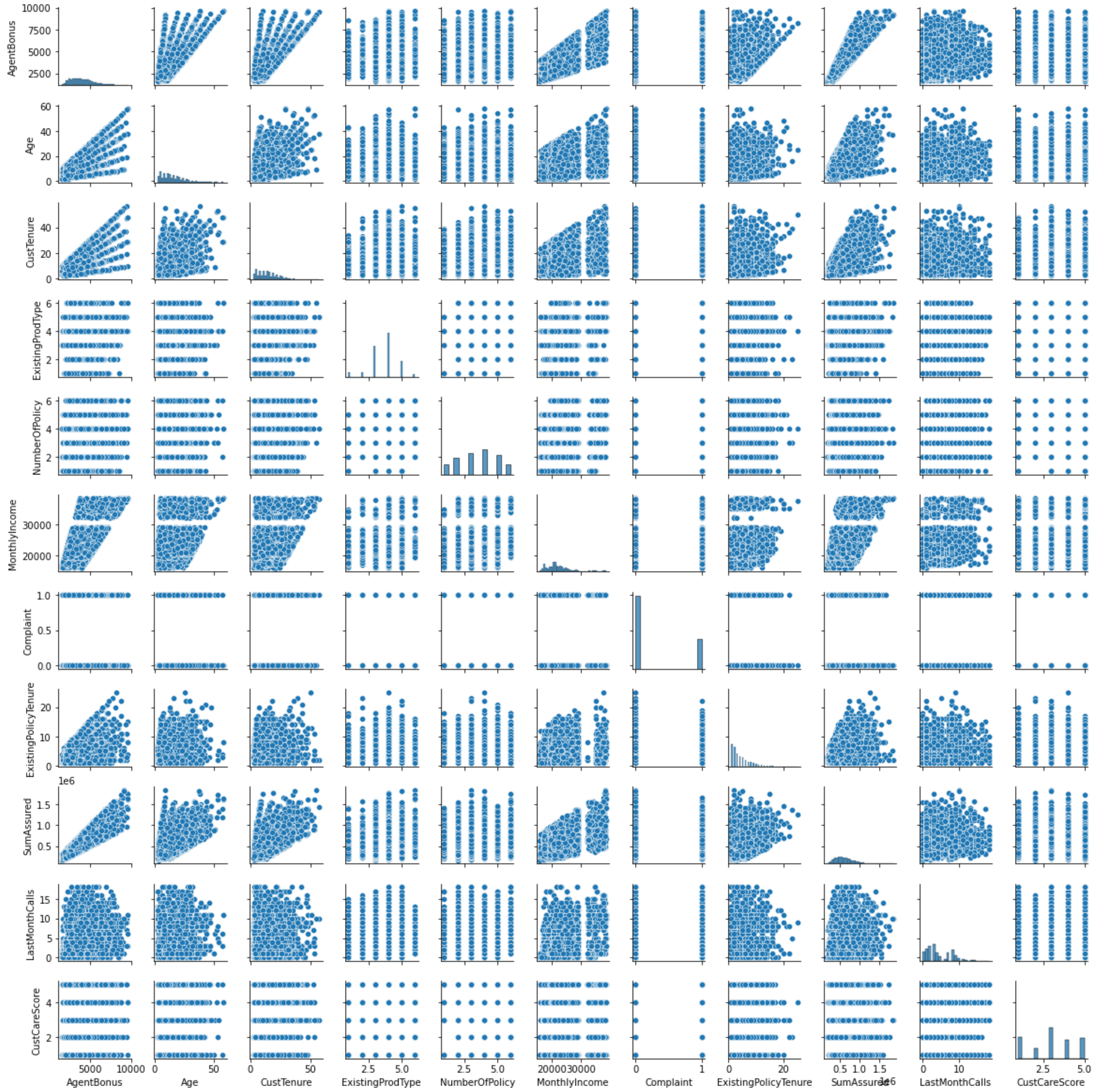

Figure 15: Pair plot

As seen in correlation heatmap in Figure 13 we can see how data is correlated with other variables. Agent Bonus is mostly correlated which is our target variable. The are few correlations exist between other variable which will be treated further.

f) Missing Value Treatment

If we drop null values the shape of dataset is now 3447 rows and 20 columns. Out of 4520 rows, 1073 rows are dropped almost 23% of data is dropped so dropping is not a good idea.

Replacing with median values seems better option here. Below shown table shows how data info.

Sr.no	Columns	Non Null Count	Data Type
1	AgentBonus	4520 non-null	int64
2	Age	4520 non-null	float64
3	CustTenure	4520 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	4520 non-null	float64
11	MaritalStatus	4520 non-null	object
12	MonthlyIncome	4520 non-null	float64
13	Complaint	4520 non-null	int64
14	ExistingPolicyTenure	4520 non-null	float64
15	SumAssured	4520 non-null	float64
16	Zone	4520 non-null	object
17	PaymentMethod	4520 non-null	object
18	LastMonthCalls	4520 non-null	int64
19	CustCareScore	4520 non-null	float64

Table 7: Data Info After Null Value Treatment

g) Outlier Treatment

Outliers above 75 percentiles + 1.5 Inter Quartile Range is replaced by 75 percentiles + 1.5 Inter Quartile Range and Outliers lower than 1.5 Inter Quartile Range – 25 Percentile are replaced by 1.5 Inter Quartile Range – 25 Percentile.

Below are Boxplots of Continuous Variables after outlier treatment.

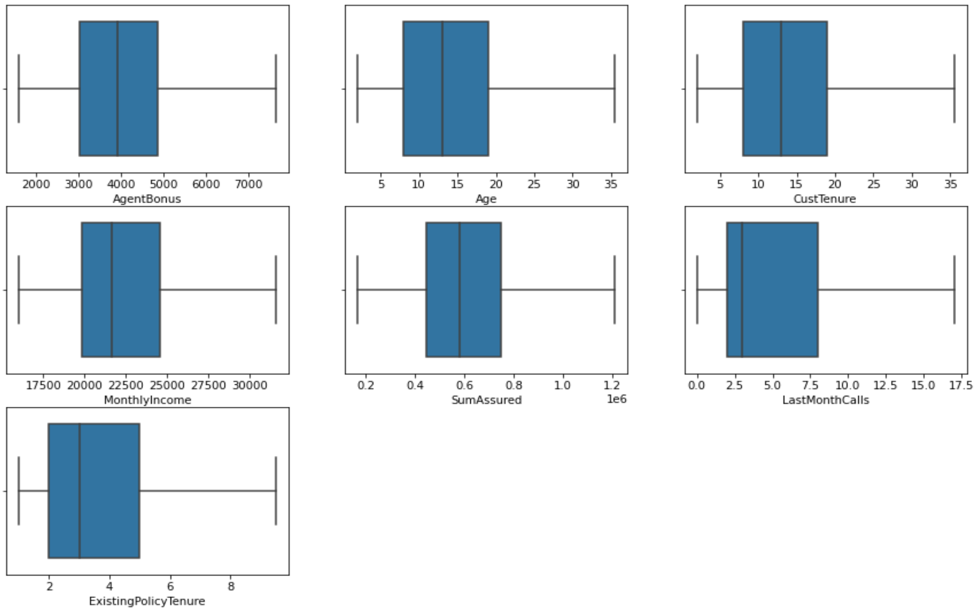

Figure 16: Box Plot after Outlier Treatment

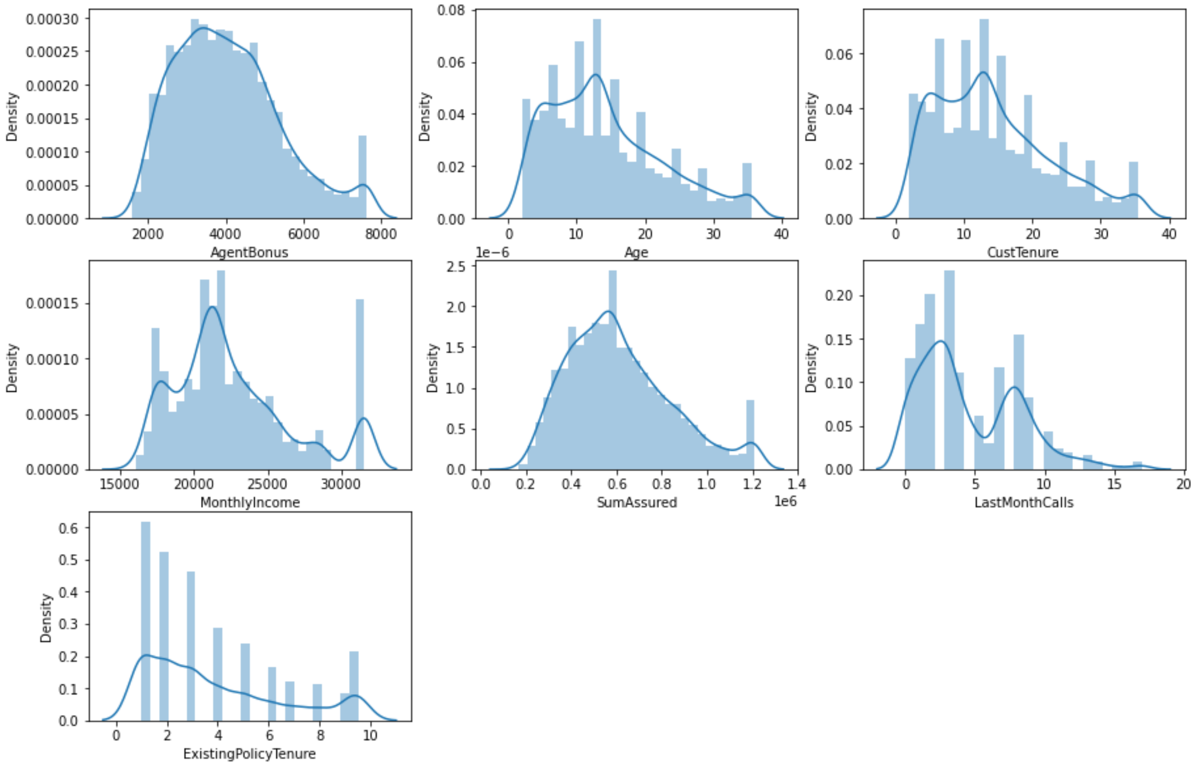

Figure 17: Histogram After Outlier Treatment

After Outlier Treatment also there is no improvement in distribution. Each of the variables are Right Skewed as seen earlier.

4. Business insights from EDA

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

We can see from Figure 4 and Figure 5 some of the categorical variables are unbalanced such as Zone, Occupation, Education Field etc. But our target variable is Agent Bonus which is continuous so we don't require to balance data.

b) Any business insights using clustering (if applicable)

Clustering is done by keeping number of clusters as 3 and by k means technique. Number of clusters is decided by taking look at the graph shown below.

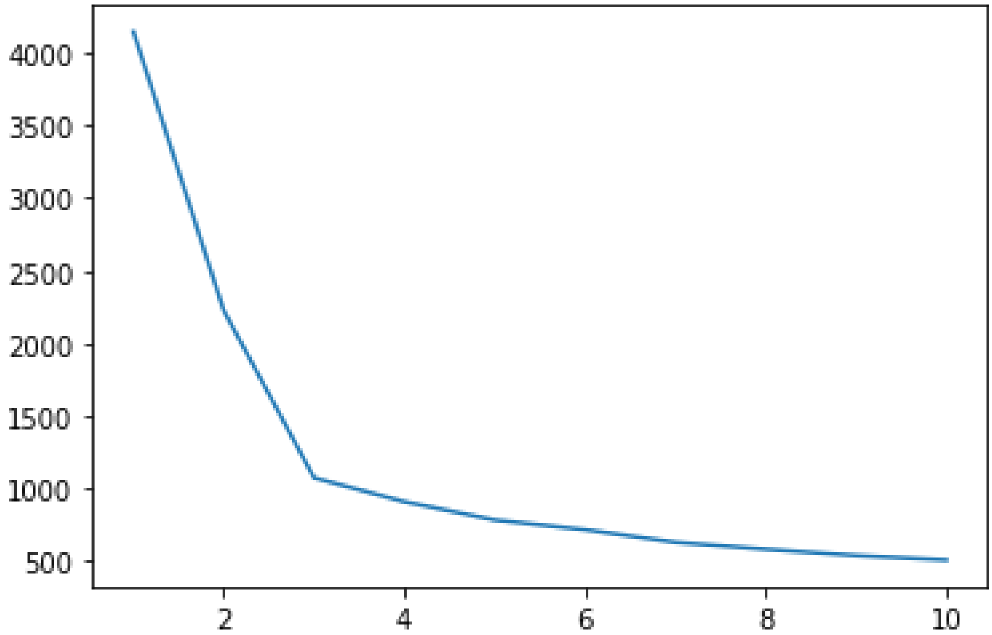

Figure 18: Number of Cluster

From figure above after 3 the difference in inertia is decreasing in linear fashion thus, we select number of clusters as 3.

```
Channel = ['Third Party Partner' 'Agent' 'Online']
Occupation = ['Salaried' 'Small Business' 'Large Business']
EducationField = ['Graduate' 'Under Graduate' 'Engineer' 'Diploma' 'Post Graduate' 'MBA']
Gender = ['Male' 'Female']
Designation = ['Manager' 'Executive' 'VP' 'Senior Manager' 'AVP']
MaritalStatus = ['Divorced' 'Single' 'Unmarried' 'Married']
Zone = ['North' 'West' 'East' 'South']
PaymentMethod = ['Yearly' 'Half Yearly' 'Quarterly' 'Monthly']
AgentBonus Max= 5146.0 Min= 1605.0
Age Max= 30.0 Min= 2.0
CustTenure Max= 30.0 Min= 2.0
NumberOfPolicy Max= 6.0 Min= 1.0
MonthlyIncome Max= 31542.375 Min= 16009.0
ExistingPolicyTenure Max= 9.5 Min= 1.0
SumAssured Max= 524195.0 Min= 168536.0
LastMonthCalls Max= 17.0 Min= 0.0
CustCareScore Max= 5.0 Min= 1.0
```

Figure 19: Cluster0

```
Channel = ['Agent' 'Online' 'Third Party Partner']
Occupation = ['Salaried' 'Free Lancer' 'Small Business' 'Large Business']
EducationField = ['Graduate' 'Post Graduate' 'Under Graduate' 'Engineer' 'Diploma' 'MBA']
Gender = ['Female' 'Male']
Designation = ['Manager' 'Executive' 'Senior Manager' 'VP' 'AVP']
MaritalStatus = ['Single' 'Unmarried' 'Married' 'Divorced']
Zone = ['North' 'West' 'East' 'South']
PaymentMethod = ['Half Yearly' 'Yearly' 'Quarterly' 'Monthly']
AgentBonus Max= 7626.5 Min= 1688.0
Age Max= 35.5 Min= 2.0
CustTenure Max= 35.5 Min= 2.0
NumberOfPolicy Max= 6.0 Min= 1.0
MonthlyIncome Max= 31542.375 Min= 16009.0
ExistingPolicyTenure Max= 9.5 Min= 1.0
SumAssured Max= 817741.0 Min= 524272.0
LastMonthCalls Max= 17.0 Min= 0.0
CustCareScore Max= 5.0 Min= 1.0
```

Figure 20: Cluster1

```
Channel = ['Online' 'Agent' 'Third Party Partner']
Occupation = ['Small Business' 'Salaried' 'Large Business']
EducationField = ['Under Graduate' 'Diploma' 'Graduate' 'Post Graduate' 'Engineer' 'MBA']
Gender = ['Male' 'Female']
Designation = ['AVP' 'Manager' 'Senior Manager' 'VP' 'Executive']
MaritalStatus = ['Divorced' 'Married' 'Single' 'Unmarried']
Zone = ['North' 'West' 'East']
PaymentMethod = ['Yearly' 'Half Yearly' 'Quarterly' 'Monthly']
AgentBonus Max= 7626.5 Min= 4157.0
Age Max= 35.5 Min= 4.0
CustTenure Max= 35.5 Min= 4.0
NumberOfPolicy Max= 6.0 Min= 1.0
MonthlyIncome Max= 31542.375 Min= 17322.0
ExistingPolicyTenure Max= 9.5 Min= 1.0
SumAssured Max= 1208311.875 Min= 818818.0
LastMonthCalls Max= 17.0 Min= 0.0
CustCareScore Max= 5.0 Min= 1.0
```

Figure 21: Cluster2

There are not major differences in Clusters. Freelancers are added in Cluster 1. Sum assured in all clusters are different Cluster 0 offer less sum and Cluster 1 offer moderate sum and Cluster 2 offer high sum.

c) Any other business insights

- The company have more clients in North and West region. There are less Clients in south and east region. Company should target in these regions.
- Freelancers are very less as they do not have regular income. Freelancers should be offered with yearly plans and half yearly plans.
- Company have very less complaints shows that agents are doing well.
- MBA students are very less in no. maybe they are investing in other investment or not happy with this company.
- Monthly income is directly proportional to sum assured.

5. Model Building and Interpretation

a) Train Test Split

In this problem Agent Bonus is our target variable and rest variables are independent variable. The data is split in the ratio 70:30. 70 percent of data is train set and 30 percent of data is test set. There are 3164 and 1356 rows in train and test respectively. Also, One hot encoding is done for the string categorical variable.

b) Multi Collinearity check

The multi collinearity is check by using Variance Inflation Factor as one of the assumptions for regression is that data is not correlated with other variable and the formula is $vif = \frac{1}{1-R^2}$. Below image shows VIF values for different variables.

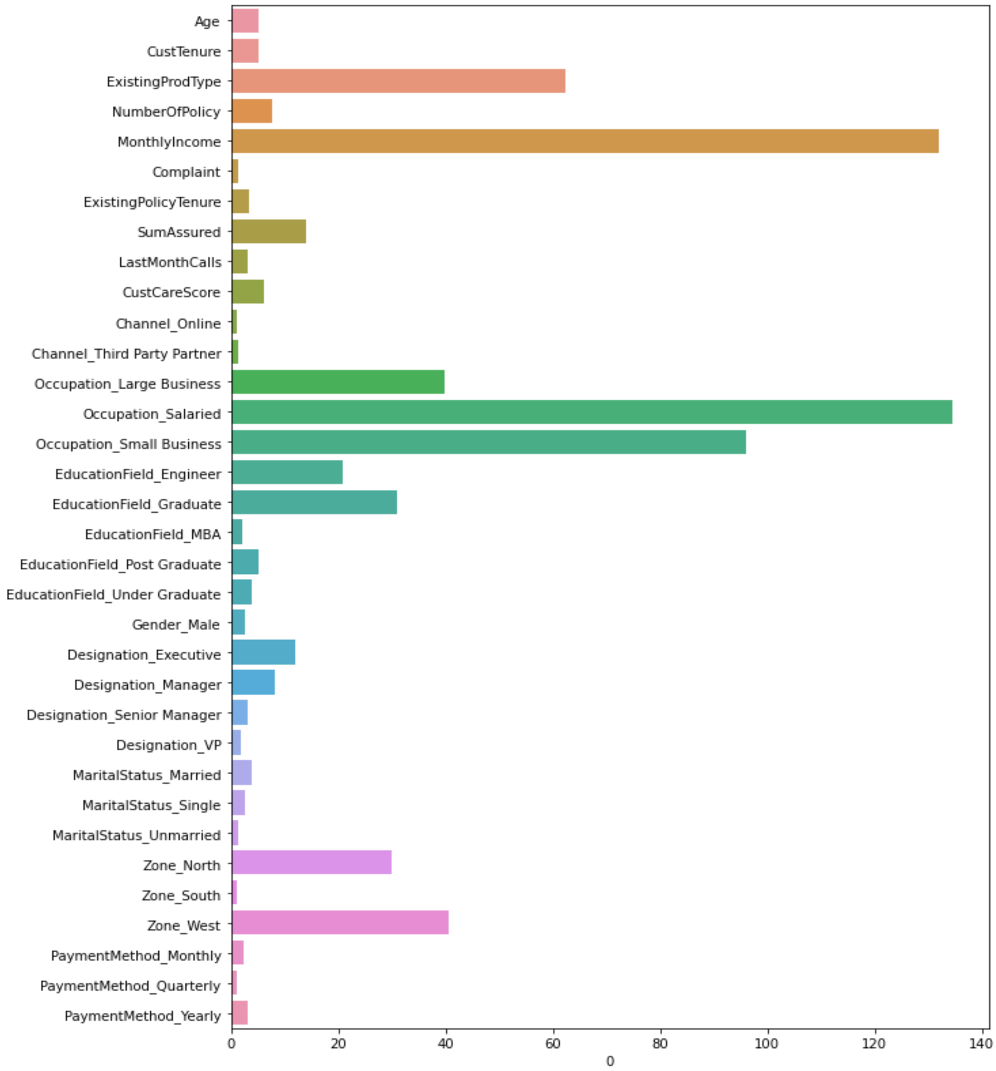

Figure 22: VIF Values of Variables

With the help of Variance Influence Factor multi collinearity is checked within variables. Columns such as ExistingProdType, Occupation_Salaried, MonthlyIncome, Zone_West, Occupation_Small Business, Occupation_Large Business,EducationField_Graduate, Zone_North are dropped as these are having high VIF as shown in image

c) Model Building.

1. Linear Regression Using Ordinary Least Square Method:

As the targets variable is continuous this is regression problem. First model is built using Linear regression with Ordinary Least Square Regression method. The basic formula for Linear Regression is y=mx+C, where C is intercept and x is variable which is multiplied by its coefficient m. In this method best fit line is drawn such that squared error from mean is lowest.

Dep. Variable:	AgentBonus	**R-squared:**	0.803
Model:	OLS	**Adj. R-squared:**	0.802
Method:	Least Squares	**F-statistic:**	493.2
Date:	Sun, 08 Jan 2023	**Prob (F-statistic):**	0
Time:	19:45:25	**Log-Likelihood:**	-24719
No. Observations:	3164	**AIC:**	4.95E+04
Df Residuals:	3137	**BIC:**	4.97E+04
Df Model:	26		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1429.0266	76.993	18.561	0	1278.065	1579.988
Age	23.4785	1.462	16.054	0	20.611	26.346
CustTenure	24.1339	1.476	16.351	0	21.24	27.028
NumberOfPolicy	20.2702	7.463	2.716	0.007	5.638	34.903
Complaint	39.1494	23.68	1.653	0.098	-7.281	85.58
ExistingPolicyTenure	36.2995	4.189	8.665	0	28.086	44.513
SumAssured	0.0036	6.06E-05	59.529	0	0.003	0.004
LastMonthCalls	-3.1881	3.215	-0.992	0.321	-9.491	3.115
CustCareScore	11.8437	7.794	1.52	0.129	-3.438	27.125
Channel_Online	19.3786	37.131	0.522	0.602	-53.425	92.182
Channel_Third_Party_Partner	-15.5464	27.278	-0.57	0.569	-69.031	37.938
EducationField_Engineer	-33.9267	37.746	-0.899	0.369	-107.936	40.083
EducationField_MBA	-17.117	89.829	-0.191	0.849	-193.247	159.013
EducationField_Post_Graduate	-46.7724	48.041	-0.974	0.33	-140.968	47.423
EducationField_Under_Graduate	-20.0675	24.401	-0.822	0.411	-67.911	27.776
Gender_Male	2.3956	21.953	0.109	0.913	-40.648	45.44
Designation_Executive	-692.6192	47.101	-14.705	0	-784.971	-600.268
Designation_Manager	-600.8232	44.932	-13.372	0	-688.922	-512.724
Designation_Senior_Manager	-377.4341	48.739	-7.744	0	-472.997	-281.871
Designation_VP	149.2648	61.743	2.418	0.016	28.204	270.326
MaritalStatus_Married	-7.5739	29.899	-0.253	0.8	-66.197	51.049
MaritalStatus_Single	71.2407	32.699	2.179	0.029	7.127	135.354
MaritalStatus_Unmarried	-205.7284	61.729	-3.333	0.001	-326.761	-84.696
Zone_South	-118.2118	301.516	-0.392	0.695	-709.401	472.977
PaymentMethod_Monthly	50.8461	41.164	1.235	0.217	-29.865	131.557

PaymentMethod_Quarterly	118.6395	84.424	1.405	0.16	-46.892	284.171
PaymentMethod_Yearly	-18.2888	23.539	-0.777	0.437	-64.442	27.864

Omnibus:	125.184	**Durbin-Watson:**	2.061
Prob(Omnibus):	0	**Jarque-Bera (JB):**	140.179
Skew:	0.492	**Prob(JB):**	3.64E-31
Kurtosis:	3.31	**Cond. No.**	1.84E+07

Table 8: Model 1 Summary

RMSE Train value: 598.04

RMSE Test value: 642.84

R^2 Train: 0.803

R^2 Test: 0.789

Interpretation: From above results we can see RMSE (Root mean squared error) of test is higher than train but their R Square is almost same. R Square indicates how much our data can predict target variable, 80% is not a bad value model can be a good predictor. From summary we can see that Agent bonus is highly positively depended on Designation_VP and highly negatively depended on Designation_Executive.

2. Linear Regression using Sklearn.

Second model is built using Sklearn library. The train data is fit into model. Sklearn Linear Regression also uses Least Squared Method but this is by machine.

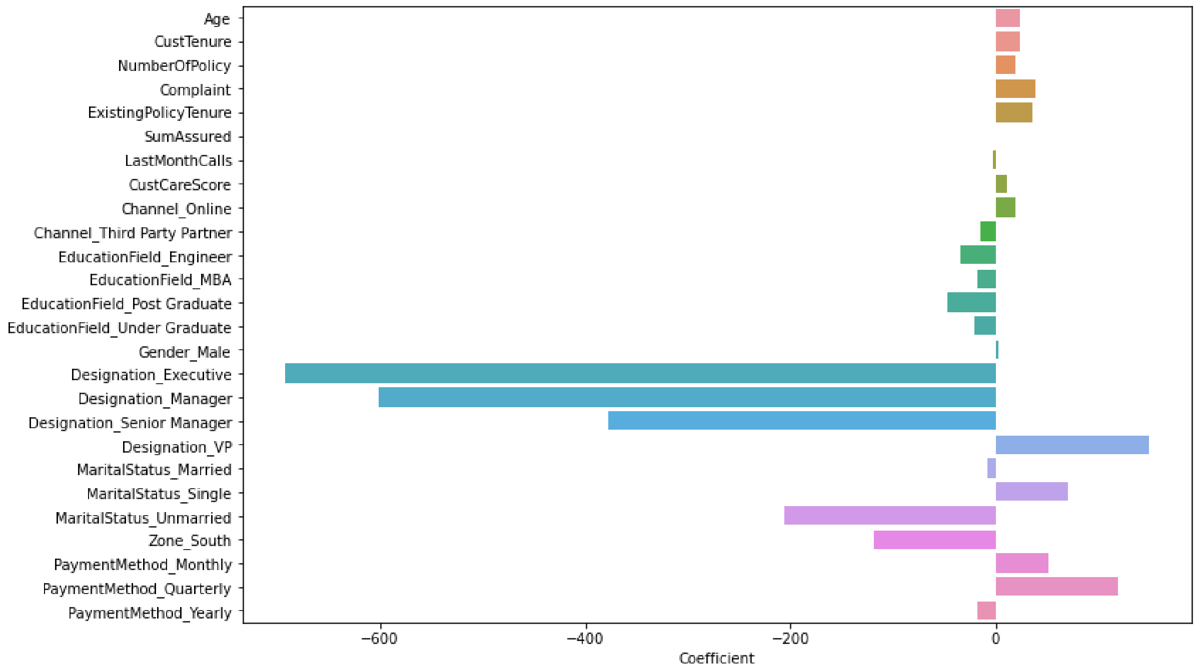

Figure 23: Coefficient of Variables

RMSE Train value: 598.04

RMSE Test value: 642.84

R^2 Train: 0.803

R^2 Test: 0.782

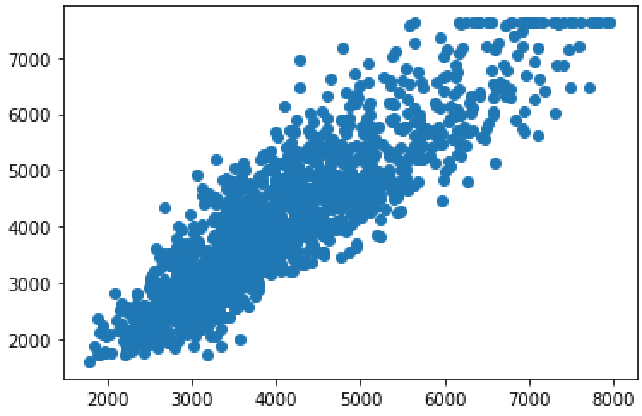

Figure 24: Predict vs Actual Linear Regression for Train

Interpretation: The RMSE values and R Squared values are same as OLS model. The depended values are also same as OLS model. Both train and test can predict 80% of Agent Bonus. This model is good to predict values. From above Scatter Plot we can see with increase in actual bonus predicted values is also increasing.

3. Random Forest Model.

Random Forest is a Machine Learning Technique where model is built by no. of different Discission tree and result with maximum votes are considered as predicted value for classification model and average of all results are taken for regression model. With the help of Bootstrap and aggregation predicted values are found.

Random Forest is present in Sklearn library. Train data is fiit into Random Forest with n_estimator as 350

RMSE Train value: 189.68

RMSE Test value: 545.94

R^2 Train: 0.980

R^2 Test: 0.843

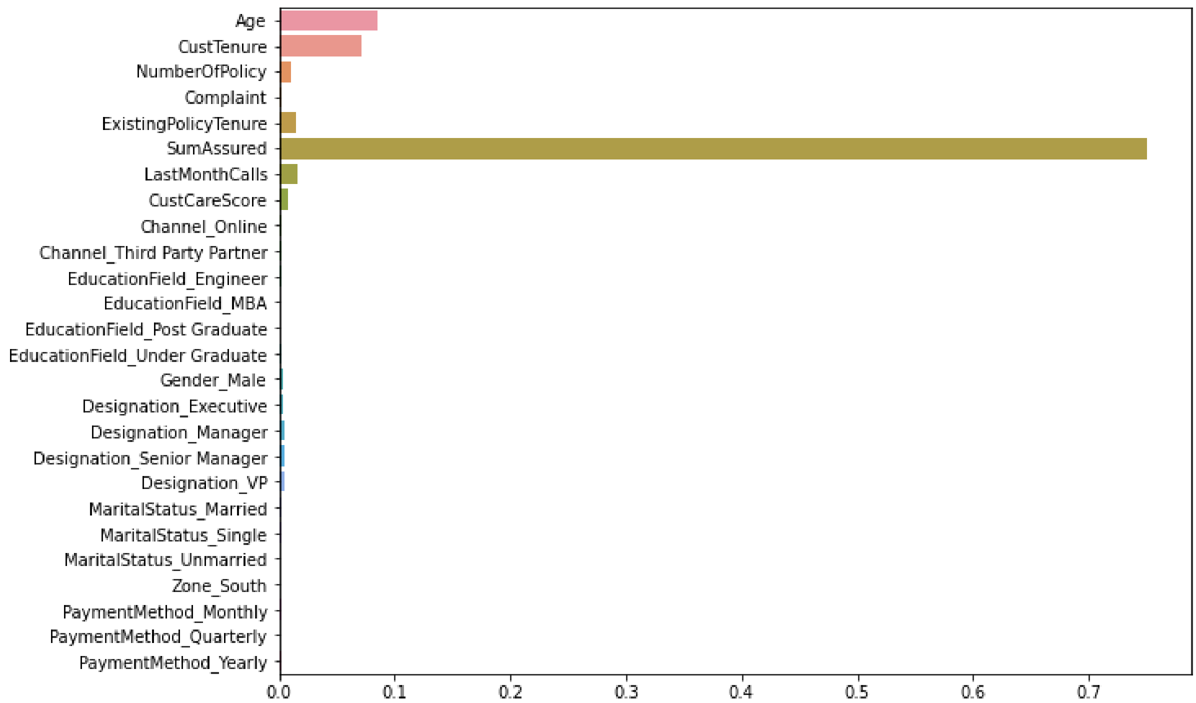

Figure 25: Feature Importance Random Forest

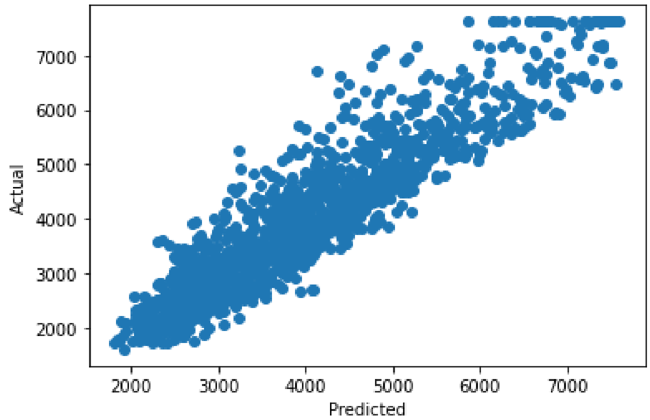

Figure 26: Actual vs Predicted Test Random Forest

Interpretation: From above data we can see that model can predict well in train as r squared values 98% but in test it can predict 84% hence the difference is huge and model is not good predictor for unknown data compared to train. RMSE score for both train and test are different RMSE is lower for train and higher for test which again shows model has not done well for unknown data. This is overfit model. From figure 25 we can see that Sum Assured is important factor for predicting Agent Bonus.

4. Lasso and Ridge model

Lasso and Ridge is Regularization technique on Linear Model. Penally λ is added to Residual value. Lasso is fit in train data using λ values as 2.3

RMSE Train: 599.14

RMSE Test: 642.38

R Square Train: 0.803

R Square Test: 0.783

	Coefficient
Age	23.827659
CustTenure	24.433032
NumberOfPolicy	19.115546
Complaint	27.01607
ExistingPolicyTenure	35.62071
SumAssured	0.00364
LastMonthCalls	-1.334978
CustCareScore	11.207713
Channel_Online	0
Channel_Third Party Partner	-2.859142
EducationField_Engineer	-3.47484
EducationField_MBA	0
EducationField_Post Graduate	0
EducationField_Under Graduate	0
Gender_Male	0
Designation_Executive	-618.32804
Designation_Manager	-523.77354
Designation_Senior Manager	-303.53722
Designation_VP	153.653353
MaritalStatus_Married	0
MaritalStatus_Single	65.019354
MaritalStatus_Unmarried	-146.71963
Zone_South	0
PaymentMethod_Monthly	17.047504
PaymentMethod_Quarterly	0
PaymentMethod_Yearly	-15.603244

Table 9: Coefficient of Variables

From above table we can see that some of the coefficient is zero which suggest those variables are not significant and can be dropped. Designation_VP is highly positively correlated to Agent Bonus and Designation Executive is highly negatively correlated to Agent Bonus.

Ridge is fit into train data by using λ values as 5

RMSE Train: 598.27

RMSE Test: 643.08

R Square Train: 0.803

R Square Test: 0.782

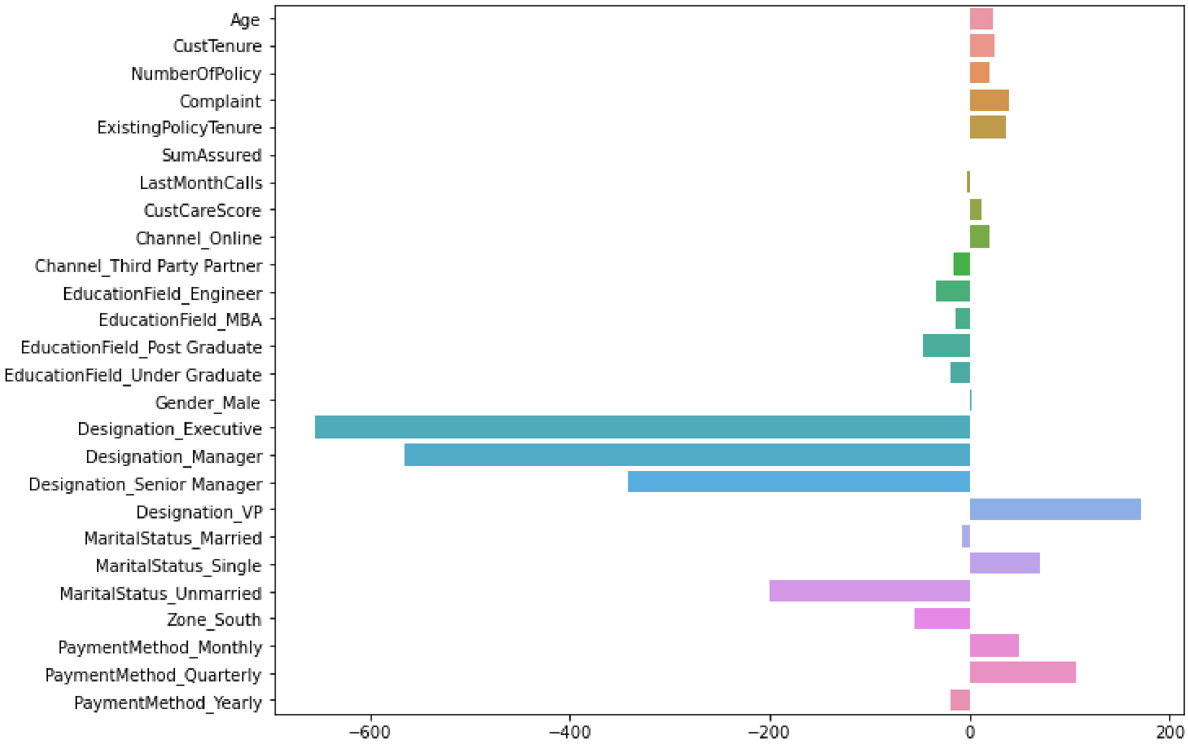

Figure 27: Ridge Coefficient

From above image we can see that Designation Executive is highly negatively correlated to Agent Bonus and highly positive to Designation Vp.

Interpretations: Lasso and Ridge both show equal performance not much difference is seen from Linear Regression model after regularization. But from lasso model we can see important predictors for agent bonus.

6. Ensemble Model

a) Bagging Model

Bagging is technique where random rows are collected to build different model is fit and the predicted values with highest votes is taken as final value in case of classification problem and for regression mean of all predicted values are taken as final predicted values.

In this problem bagging is on random subset of data using decision tree regressor with 250 as no. of estimators. The average of all predicted is taken as final value.

RMSE Train value: 190.49

RMSE Test value: 545.22

R^2 Train: 0.98

R^2 Test: 0.84

Interpretation: From above values we can that model is actually performing like Random Forest as done before. Model fits well for train and lacks same accuracy when shown unknown data.

b) Boosting Model

1. Adaptive Boosting Regressor

Adaptive Boosting model is built from Sklearn Library.

RMSE Train value: 17.42

RMSE Test value: 553.38

R^2 Train: 0.99

R^2 Test: 0.83

2. Gradient Boosting Regressor

Gradient Boosting Regression model is built form Sklearn Library.

RMSE Train value: 484.47

RMSE Test value: 572.03

R^2 Train: 0.87

R^2 Test: 0.83

Image Below shows performance of model for Train and test

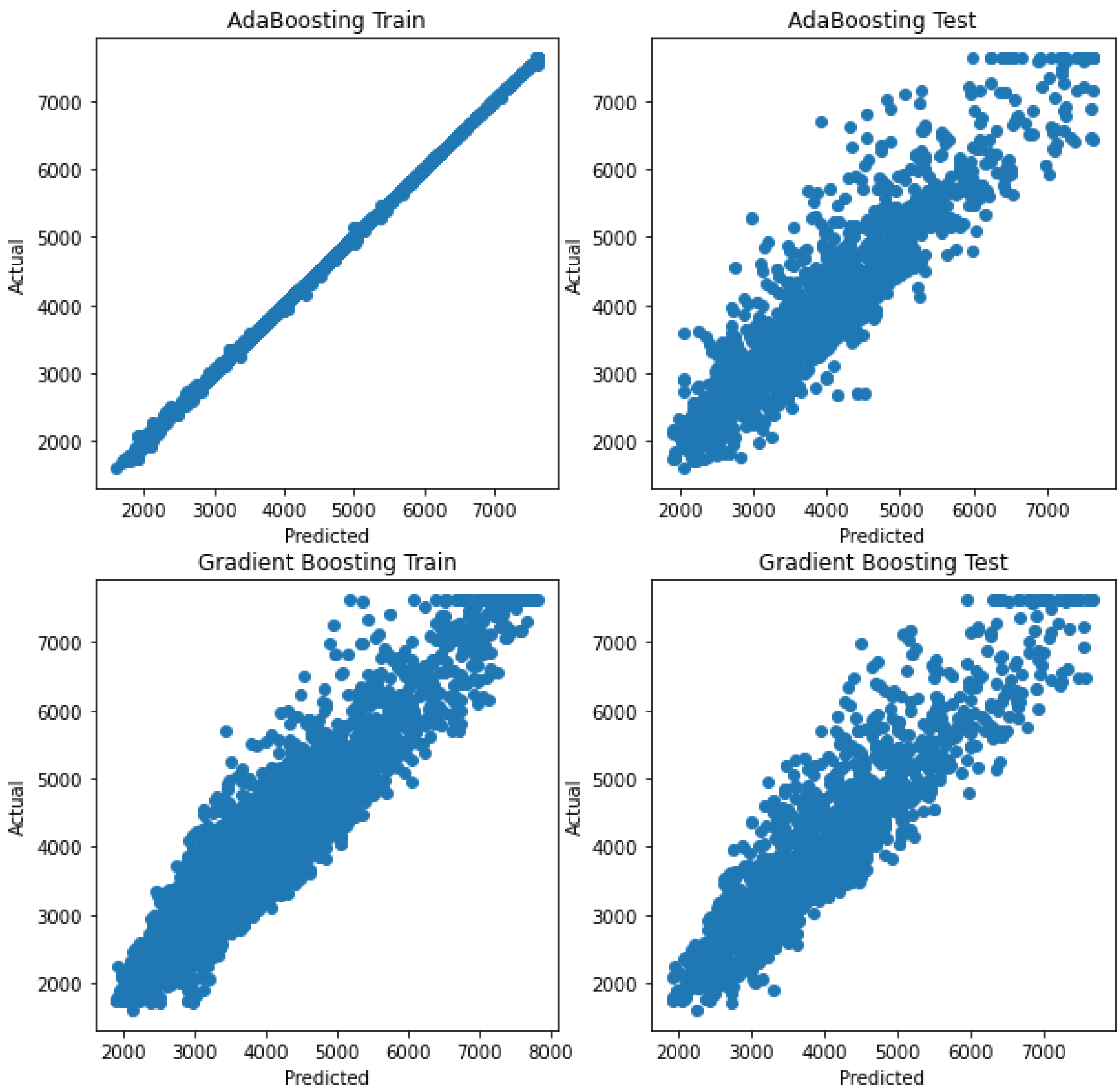

Figure 28: Ensemble Performance

Interpretation: From Boosting Model we can see that Ada Boosting works well for train data but not accurate for test data by looking at R square values. Gradient Boosting has done better for train and test as R square is good for both train and test and RMSE seems good. From Scatter Plot we can see that Gradient Boosting for train and test almost looks same. This model is good for prediction.

c) Model Tuning

1. Grid Search on Random Forest

Model tuning is done by using GRID Search and parameters were used on Random Forest model. Best parameters are No. of estimators= 200, min sample leaf = 1 and min sample split = 6. Model is fit into train data and results are shown below.

RMSE Train value: 293.03

RMSE Test value: 544.90

R^2 Train: 0.97

R^2 Test: 0.84

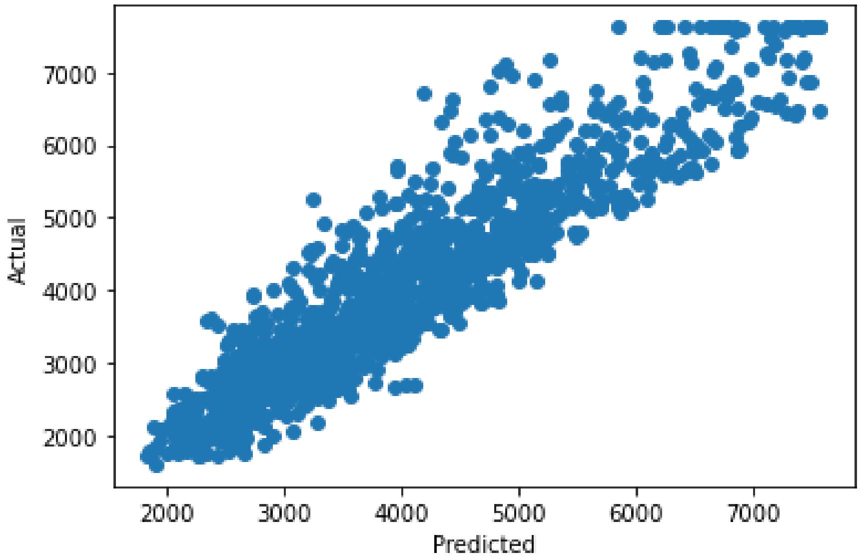

Figure 29: Model Performance After Tuning

Interpretation: After tuning there no improvement in model Random Forest is doing good for train but not so accurate for unknown data.

2. Model from Lasso Insight

From Lasso model we have seen that, coefficient for columns Channel_Online, EducationField_MBA, EducationField_Post Graduate,EducationField_Under Graduate, Gender_Male, MaritalStatus_Married,Zone_South, PaymentMethod_Quarterly is zero hence building Least Square Model without this columns. Data is fit in to train model.

Dep. Variable:	AgentBonus	**R-squared:**	0.803
Model:	OLS	**Adj. R-squared:**	0.802
Method:	Least Squares	**F-statistic:**	713
Date:	Sun, 08 Jan 2023	**Prob (F-statistic):**	0
Time:	22:23:07	**Log-Likelihood:**	-24721
No. Observations:	3164	**AIC:**	4.95E+04
Df Residuals:	3145	**BIC:**	4.96E+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1419.3797	71.859	19.752	0	1278.484	1560.275
Age	23.5766	1.459	16.156	0	20.715	26.438
CustTenure	24.0978	1.474	16.347	0	21.207	26.988
NumberOfPolicy	20.5316	7.419	2.767	0.006	5.985	35.078
Complaint	38.7787	23.631	1.641	0.101	-7.555	85.112
ExistingPolicyTenure	36.4451	4.172	8.736	0	28.266	44.625
SumAssured	0.0036	6.06E-05	59.6	0	0.003	0.004
LastMonthCalls	-3.0781	3.208	-0.959	0.337	-9.368	3.212
CustCareScore	11.9612	7.775	1.538	0.124	-3.284	27.206
Channel_Third_Party_Partner	-16.8667	26.883	-0.627	0.53	-69.577	35.843
EducationField_Engineer	-25.7409	36.419	-0.707	0.48	-97.149	45.667
Designation_Executive	-694.4254	46.941	-14.793	0	-786.464	-602.386
Designation_Manager	-600.9653	44.83	-13.406	0	-688.863	-513.067
Designation_Senior_Manager	-380.9358	48.607	-7.837	0	-476.241	-285.631
Designation_VP	145.1237	61.43	2.362	0.018	24.677	265.571
MaritalStatus_Single	76.3065	23.956	3.185	0.001	29.335	123.278
MaritalStatus_Unmarried	-202.6813	57.375	-3.533	0	-315.178	-90.185
PaymentMethod_Monthly	45.7637	40.924	1.118	0.264	-34.477	126.004
PaymentMethod_Yearly	-21.6539	23.386	-0.926	0.355	-67.508	24.2

Omnibus:	124.615	**Durbin-Watson:**	2.063
Prob(Omnibus):	0	**Jarque-Bera (JB):**	139.486
Skew:	0.49	**Prob(JB):**	5.14E-31
Kurtosis:	3.31	**Cond. No.**	6.18E+06

Table 10: Summary Tuned Model

RMSE Train value: 598.42

RMSE Test value: 633.93

R^2 Train: 0.803

R^2 Test: 0.788

Interpretation: From above values we can see that values are not much different than Linear Regression model. This model is good for prediction. Model is properly fit in train and test.

7. Interpretations

Below table shows performance for each model by comparing their RMSE and R squared values for both train and test.

	RMSE		R Square	
Method	**Train**	**Test**	**Train**	**Test**
OLS Method	598.04	642.84	80.30%	78.90%
Linear Regression (Sklearn)	598.04	642.84	80.30%	78.20%
Random Forest	189.68	545.94	98.01%	84.30%
Lasso	599.14	642.38	80.30%	78.30%
Ridge	598.27	643.08	80.30%	78.20%
Bagging	190.49	545.22	98.01%	84.31%
AdaBoost	17.42	553.38	99.98%	83.00%
Gradient Boosting	484.47	572.03	87.08%	82.79%
Grid Search	293.03	544.9	97.00%	84.35%
Model from Lasso Insight	598.42	633.93	80.30%	78.80%

Table 11: Model Performance

a) Insights and Analysis

- From above table we can see that Random Forest, Bagging, Ada Boost, Grid Search is Overfitting as they performing well for train and no so well for test.
- Gradient boost model is performing well for Train and Test but difference between RMSE values for train and test is higher that Linear Regression models.
- Other models are Linear Regression models two models are regularized with Lasso and Ridge techniques and one model is built by dropping insignificant variables.
- OLS model is good model for prediction as most of work is done manually and other models has used machine for building model.

So, for predicting Bonus for the Agent OLS model is taken and below is the equation of Linear Regression generated from OLS.

(1429.03) * Intercept + (23.48) * Age + (24.13) * CustTenure + (20.27) * NumberOfPolicy + (39.15) * Complaint + (36.3) * ExistingPolicyTenure + (0.0) * SumAssured + (-3.19) * LastMonthCalls + (11.84) * CustCareScore + (19.38) * Channel_Online + (-15.55) * Channel_Third_Party_Partner + (-33.93) * EducationField_Engineer + (-17.12) * EducationField_MBA + (-46.77) * EducationField_Post_Graduate + (-20.07) * EducationField_Under_Graduate + (2.4) * Gender_Male + (-692.62) * Designation_Executive + (-600.82) * Designation_Manager + (-377.43) * Designation_Senior_Manager + (149.26) * Designation_VP + (-7.57) * MaritalStatus_Married + (71.24) * MaritalStatus_Single + (-205.73) * MaritalStatus_Unmarried + (-118.21) * Zone_South + (50.85) * PaymentMethod_Monthly + (118.64) * PaymentMethod_Quarterly + (-18.29) * PaymentMethod_Yearly

b) Recommendation

- Agent whose Bonus are less than 4000 should be considered as under performers.
- Upskill programs for such Agents is recommended.
- Underperformers should target for VPs of company also they should be trained to interact with their customers.
- People prefer to pay quarterly so underperformer should request their clients to pay quarterly thus decreases load form pocket and customers can be there for long time.
- Underperformers are recommended to target Large Business and should consider Age of clients; higher age can give more bonus.
- Company should request their performing Agents to target west and east to spread their sale throughout.
- Underperformer should take care of North and south region until they perform.

5. Appendix

a) List of Figures

b) List of Tables

c) Bibliography

- https://economictimes.indiatimes.com/industry/banking/finance/insure/life-insurance-penetration-in-india-reaches-3-2-close-to-global-averages-benori-knowledge/articleshow/93793635.cms?from=mdr
- https://iiflinsurance.com/insurance-companies/life-insurance-companies-in-india
- https://www.bankrate.com/insurance/life-insurance/life-insurance-statistics/
- https://www.basunivesh.com/latest-irda-claim-settlement-ratio-2023/