LI_BFSI_01- LIFE INSURANCE SALES – FINAL REPORT

Prepared by ARUN ASISH

Batch: G5

I. List of Contents

II. List of Tables

III. List of Figures

1) Introduction

Then main objective of the problem is to build machine learning model to predict the agent bonus to get insight on who are performing well and underperforming, so that we could take necessary actions to provide more bonus for the good performing agents and take necessary upskilling measures for the underperforming agents. We are provided with the data that contains the details of the customers of the insurance company.

The main objective of the project for the insurance company is to improve their sales that would lead to the growth of the company. So, based on the performance analysis of the agents we could find the good performing agents and provide the necessary incentive to improve their performance further. Also, for the underperforming agents could take necessary actions to upskill them. Which would result in the agents being more skillful and perform better. This would improve the sales of the company.

We are provided with the data of customer with age, gender, customer tenure, channel through which acquisition is done, payment method, occupation, income, sum assured, number of policies, etc., which would help in predicting the performance of Agents by their bonus.

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

Table 1 Data Dictionary

2) EDA and Business Implication

Univariate Analysis:

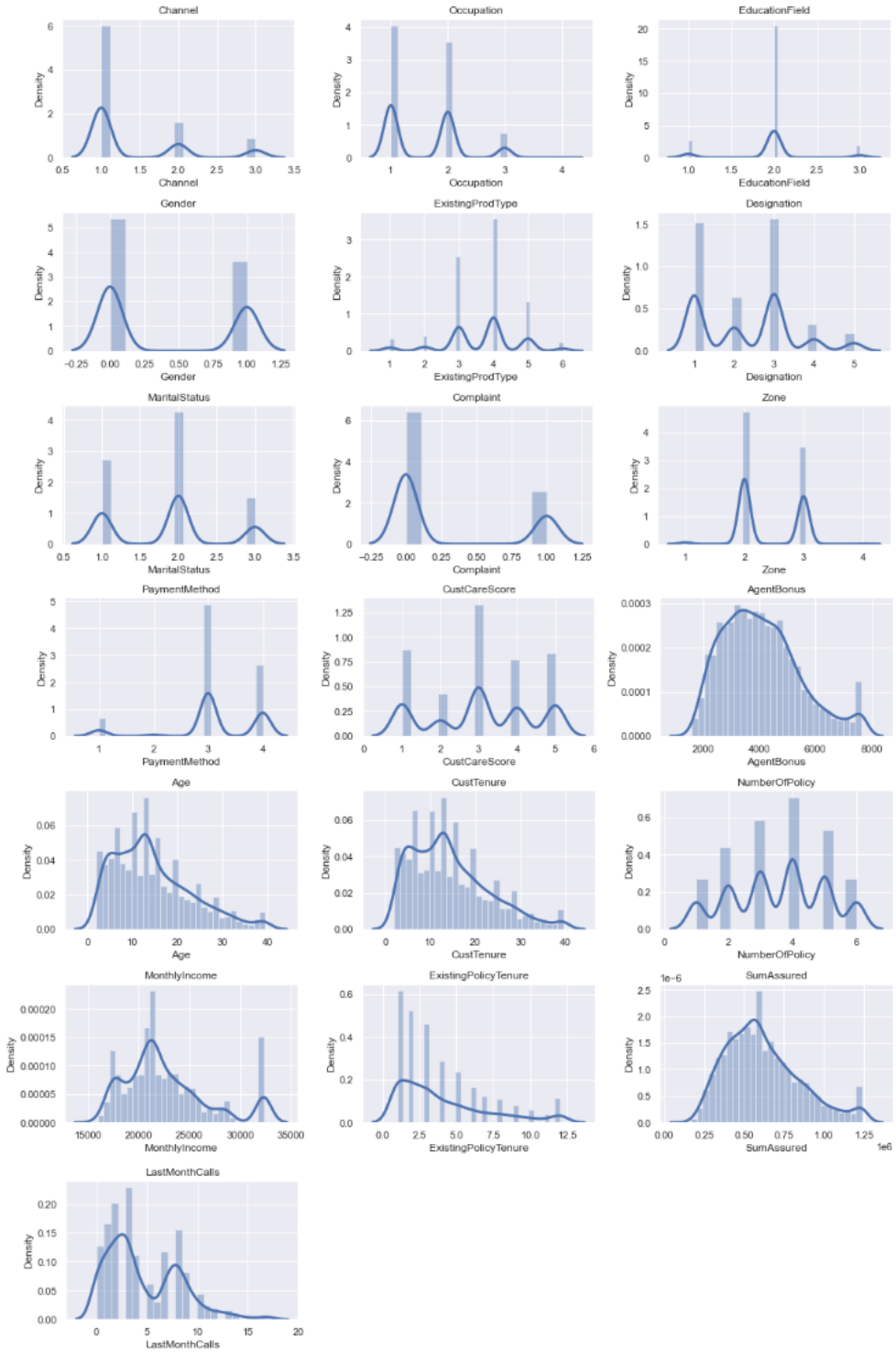

Figure 1 Distribution plot for Univariate analysis

Channel	1.421920
Occupation	0.614589
EducationField	-0.224439
Gender	0.385870
ExistingProdType	-0.401100
Designation	0.402365
MaritalStatus	0.195879
Complaint	0.941129
Zone	0.112471
PaymentMethod	-1.208737
CustCareScore	-0.138120
AgentBonus	0.630068
Age	0.819749
CustTenure	0.794531
NumberOfPolicy	-0.108161
MonthlyIncome	0.951884
ExistingPolicyTenure	1.138628
SumAssured	0.704322
LastMonthCalls	0.790936

Table 2 Skewness in data

Above graph shows that there is skewness in the data. And the above table contains the value of skewness for the variables in the data, among the key variables

- Agent bonus, Cust tenure, Monthly Income, existing policy tenure, sum assured are right skewed
- Number of policy, customer care score, payment method are left skewed

Bivariate Analysis:

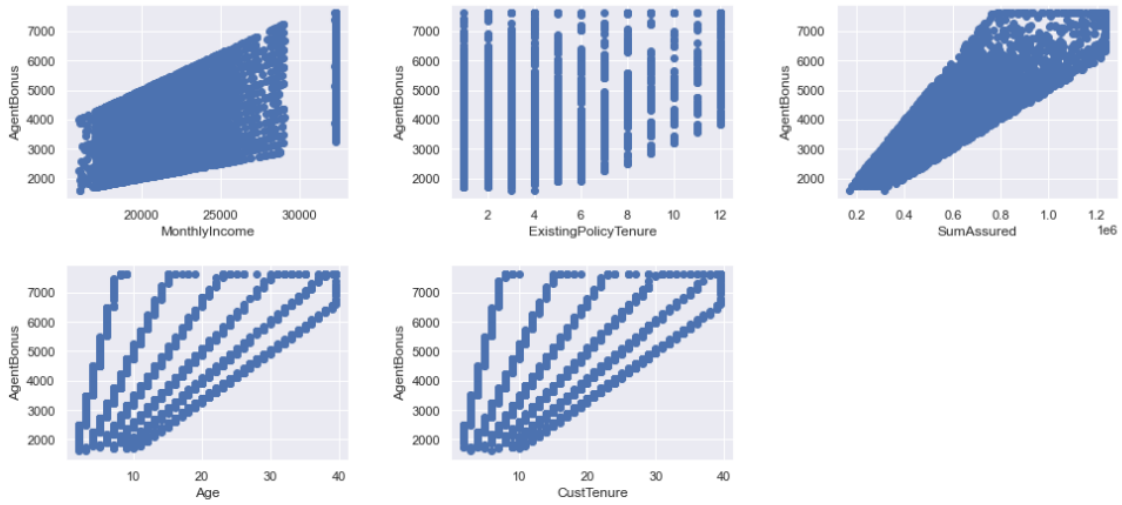

Figure 2 Bivariate analysis for corelated variables

From the above graphs we are able to observe that there is a positive collinearity between the agent bonus and the below listed fields

- Monthly Income
- Existing policy Tenure
- SumAssured
- Age
- CustTenure

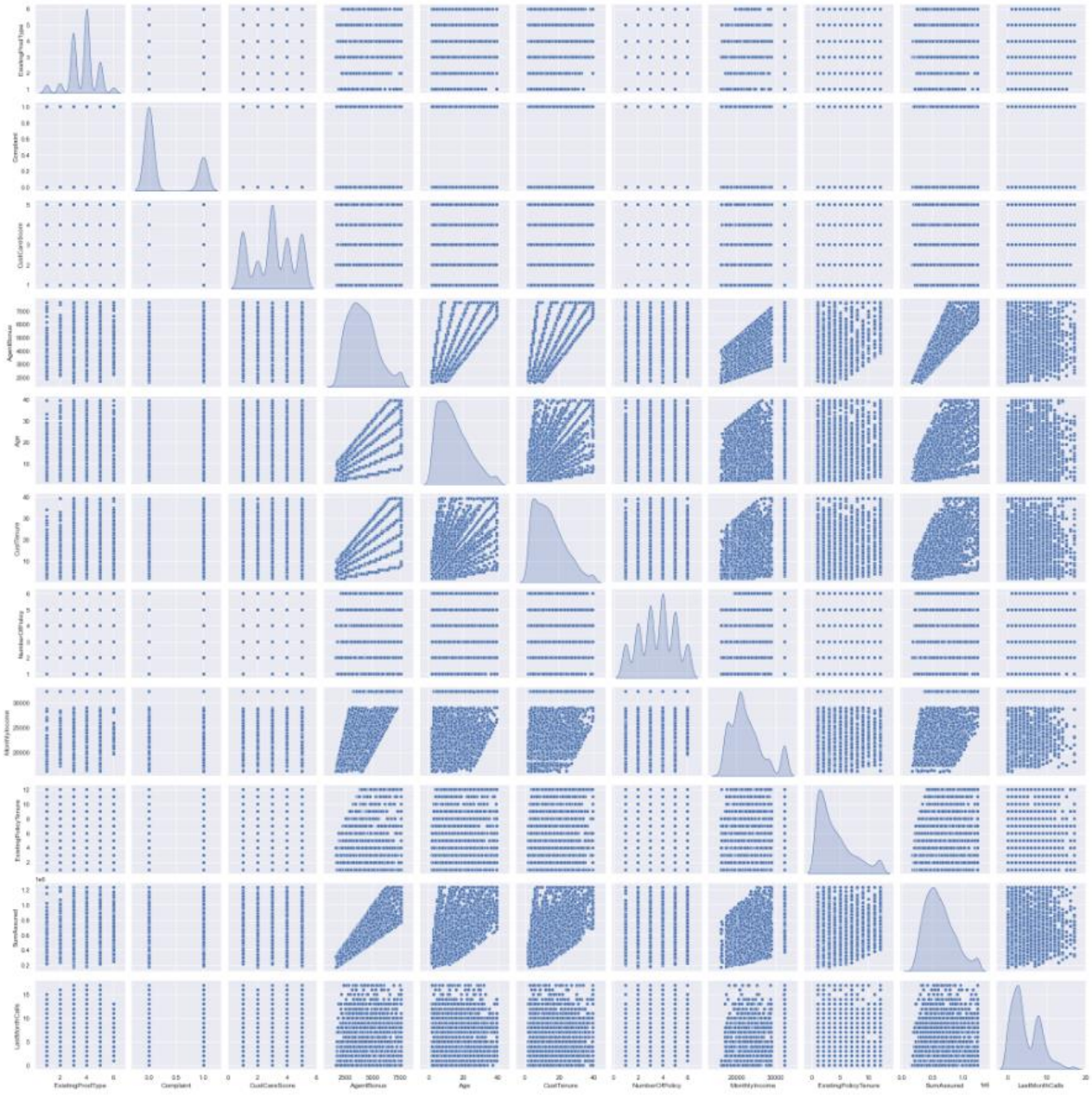

Figure 3 Pairplot for bivariate analysis

From the above pair plot we observe that there is less collinearity between the data. Some of the variables do have collinearity. To further analyze the data, we check for multicollinearity.

Multivariate Analysis:

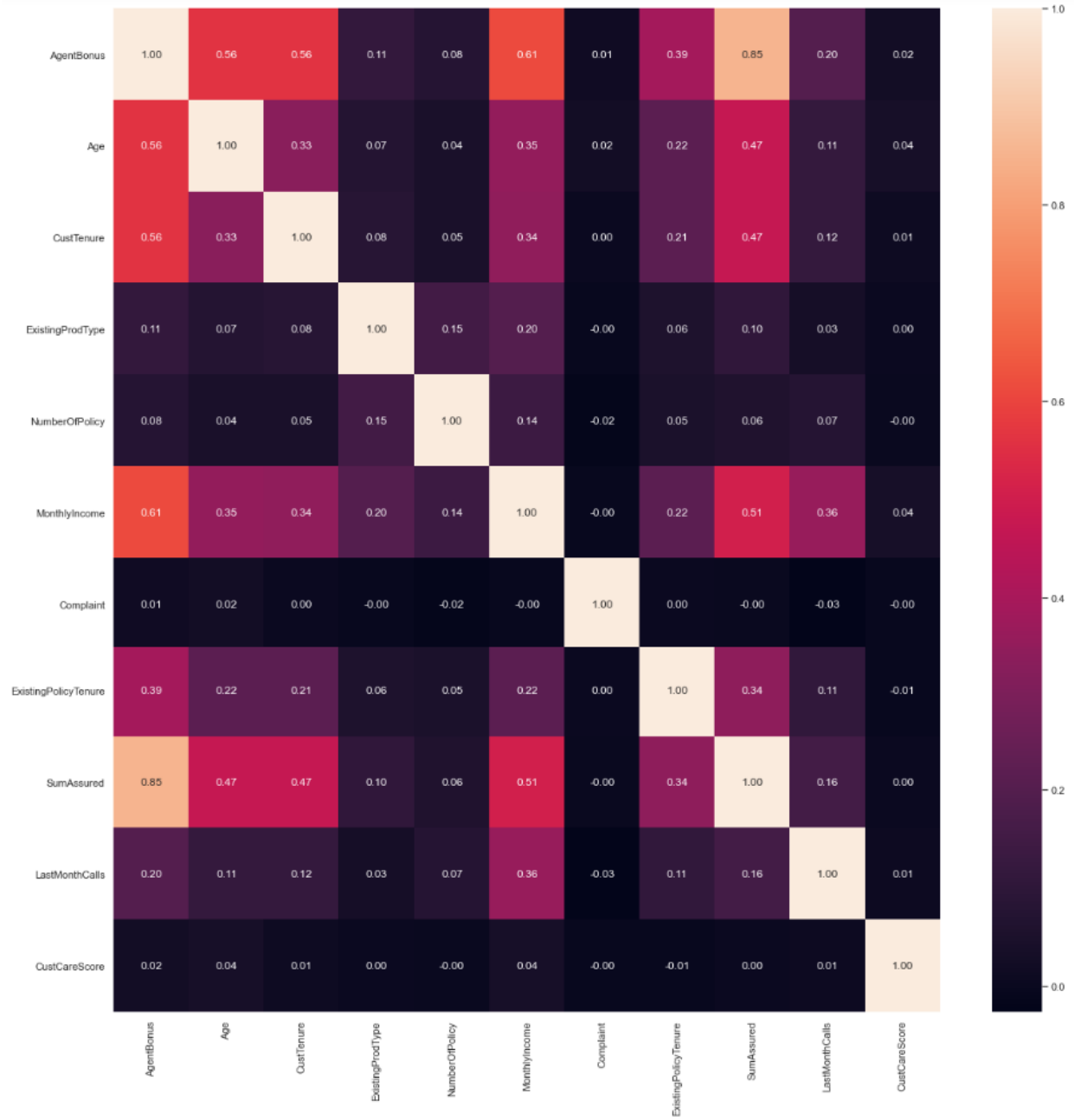

Figure 4 Heatmap for multicolinearity

From the above heatmap, Age, Customer Tenue, Monthly Income, Sum Assured plays vital role in agent bonus. Higher the values of these variable higher is the agent bonus.

Variables such as Number of policy, Existing product type shows sign of decreasing the customer bonus. Where more the number of policy, and higher the product type is, less is the agent bonus for those customers.

EDA for Business Implication:

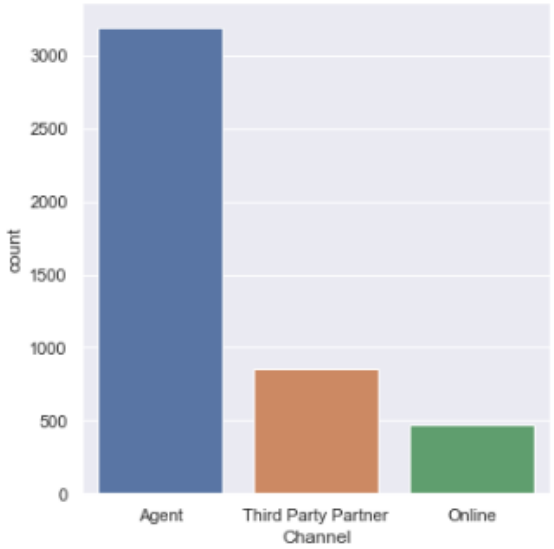

Figure 5 Count plot for Channel and Agent Bonus

From the above graph we observe that agents make higher level of bonus and are more in number.

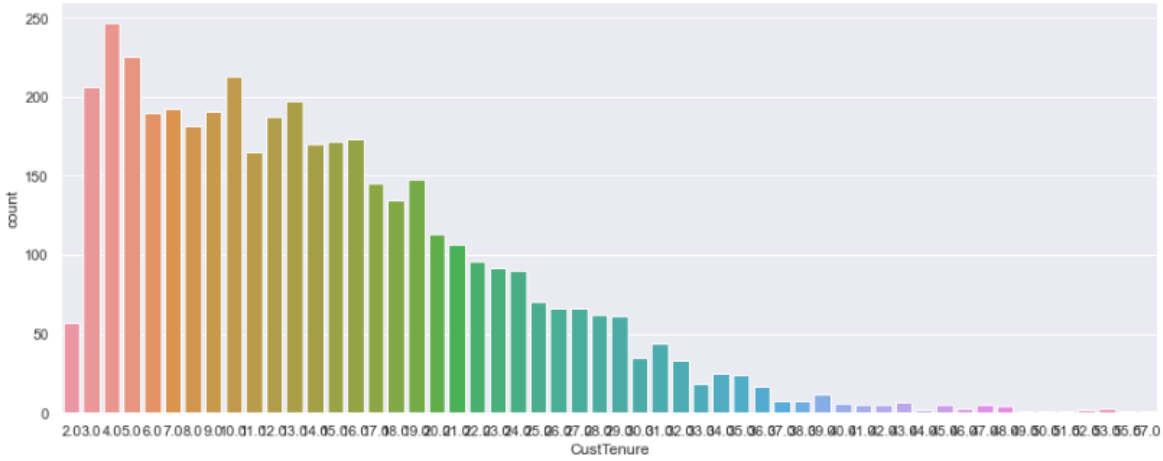

Figure 6 Count plot for Customer Tenure

Above chart has the data of customer tenure in years, where majority of the customers are 3 to 13 years tenured. And few customers have very high number of years tenure.

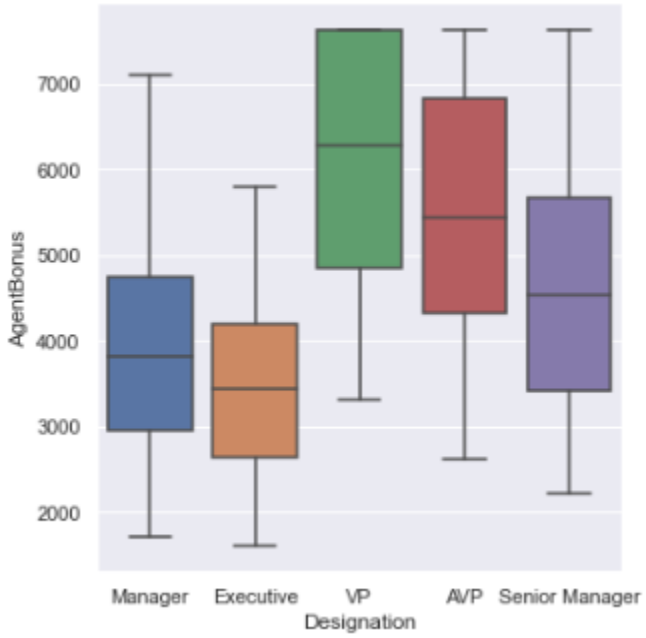

Figure 7 Boxplot for Designation and Agent Bonus

Above plot shows us that customers in designation as VP contribute more towards Agent bonus, followed by AVP and Senior managers and at last comes Manager and Executive.

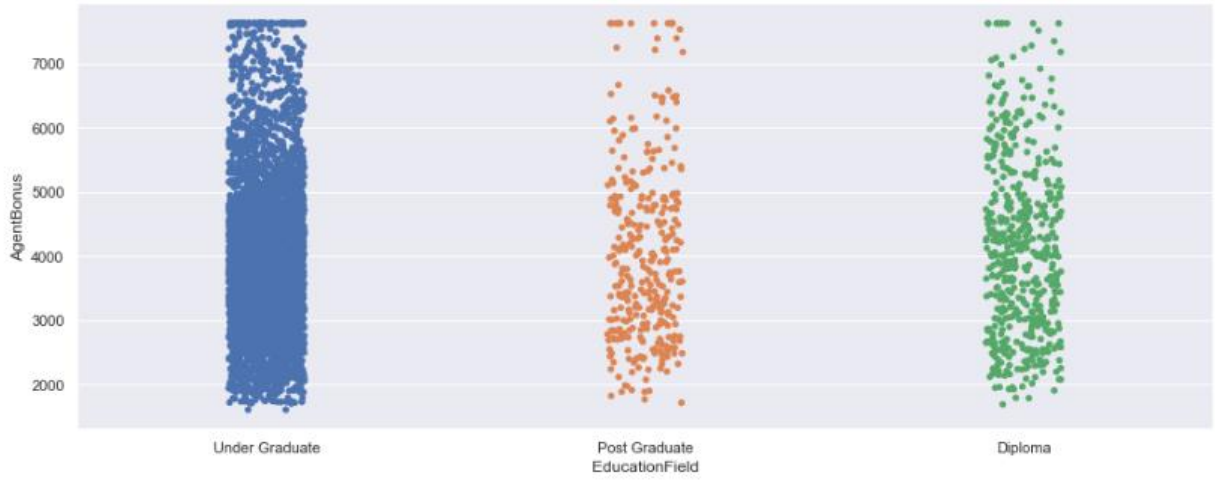

Figure 8 Strip plot for EducationField and Agent Bonus

From the above strip plot, we observe that Under graduates are more in number and contributes more to agent bonus. Diploma customers comes next and post graduates are very less in number as well as contribution to bonus.

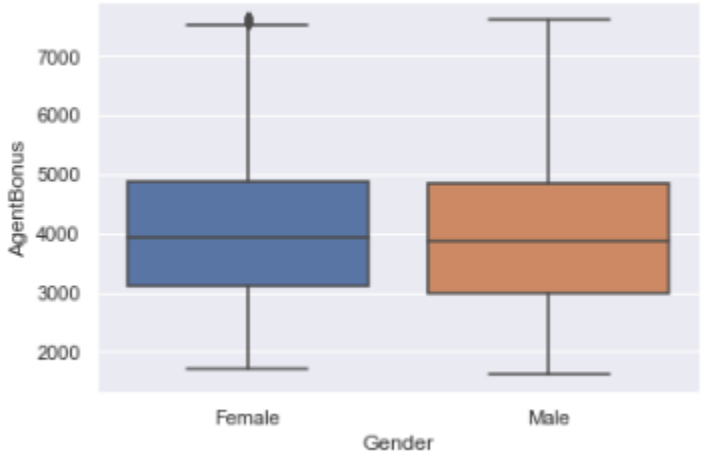

Figure 9 Box plot for Gender and Agent Bonus

From the above boxplot, we are able to observe that Male and Female customers are almost equal in number and contribution towards the agent bonus.

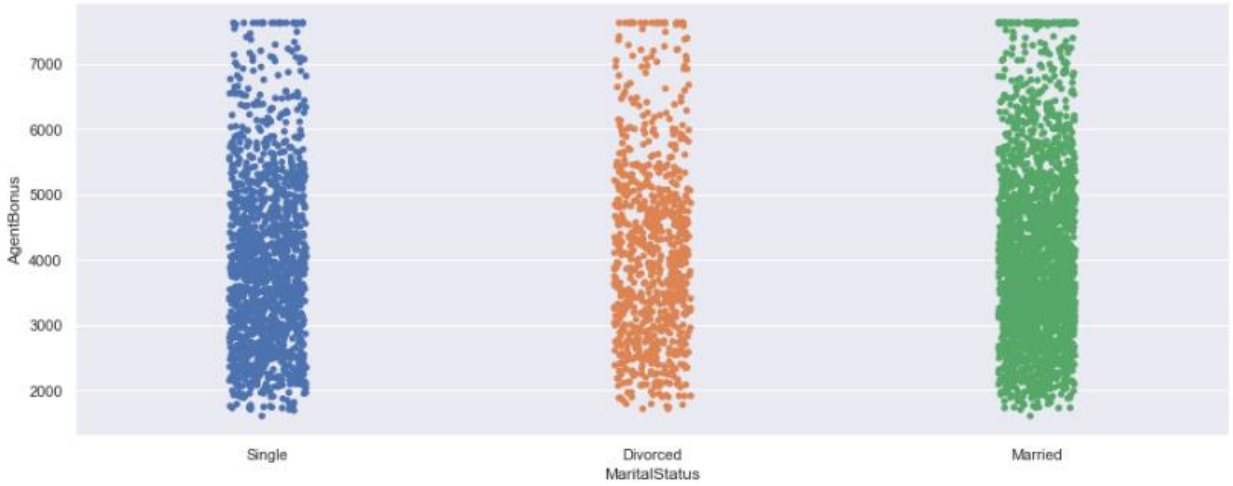

Figure 10 Strip plot for MaritalStatus and Agent Bonus

From the above strip plot, based on the marital status Married customers are more in number and contribute more for Agent bonus, next comes Divorced in terms of number and contribution for bonus. And at last comes Single customers.

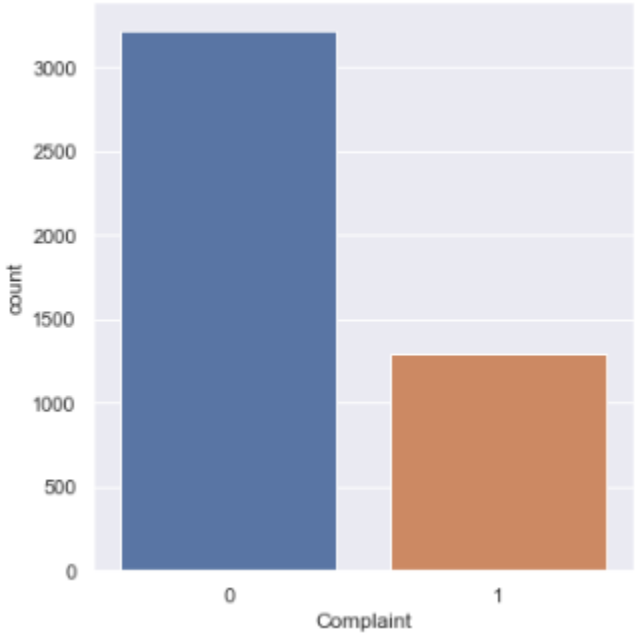

Figure 11 Count plot for Complaint

From the above count plot, we are able to see that one third of the customers has registered complaint in last one month which is high and necessary action should be taken on that.

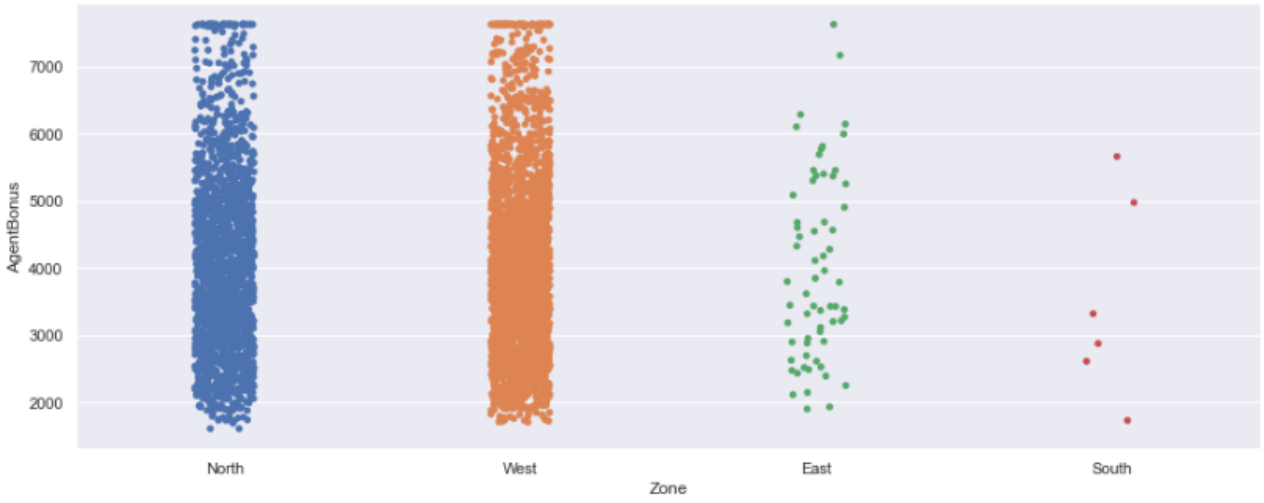

Figure 12 Strip plot for Zone and Agent Bonus

From the above strip plot, we observe that customers in north are high in number and contribute more towards bonus, followed by West in second, East with considerably very less in both and South with negligible number of customers.

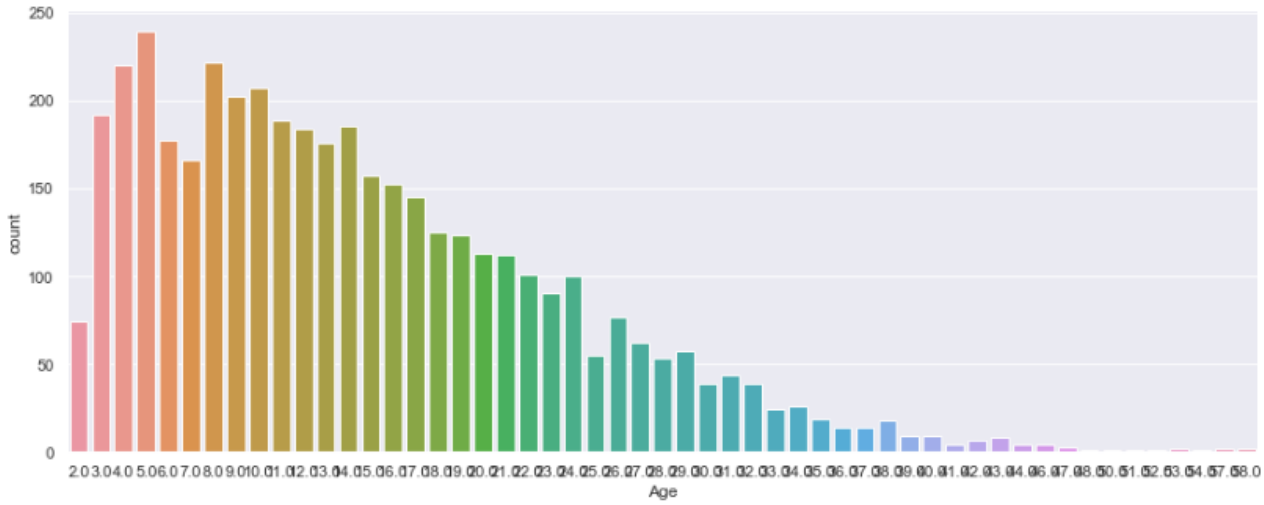

Figure 13 Count plot for Age

Above chart has the data of age of customers in years, where majority of the customers are in age group 3 to 5 and 9 to 15. And as age increases the count of customers decreases.

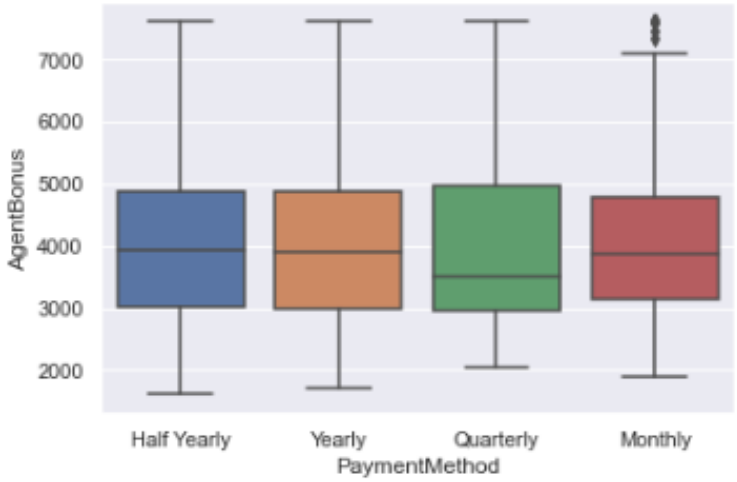

Figure 14 Box plot for PaymentMethod and Agent Bonus

Above boxplot shows Customers who do Half yearly payment and yearly payment contribute more towards agent bonus, with Quarterly contributing same level but has reduced recently and the ones who do monthly payment is considerably less and have some high value outliers.

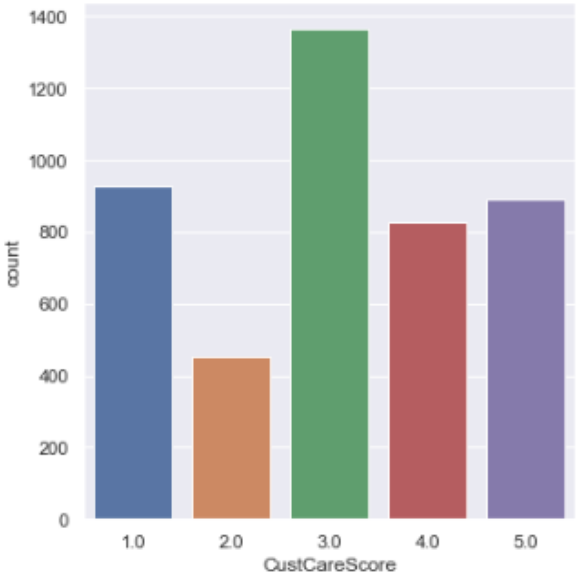

Figure 15 CustomerCareScore

From the above count plot, we observe that among the Customer care score given by the customers 3 is very high in number, with 1 and 2 having considerable number which forms the majority. 4 and 5 are comparatively less. Hence care has to be taken in improving the customer support.

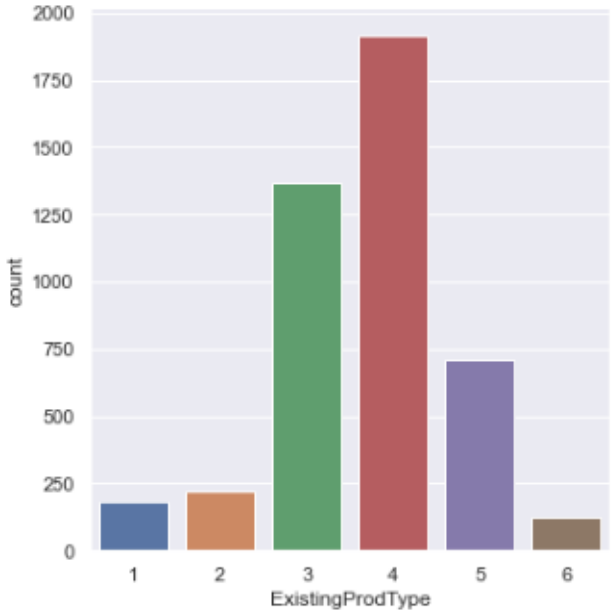

Figure 16 Count plot for ExistingProductType and Agent Bonus

Form the above boxplot, customers having product type 6 contribute more towards agent bonus, next followed by 5, 4, 2, 3 and 1 respectively.

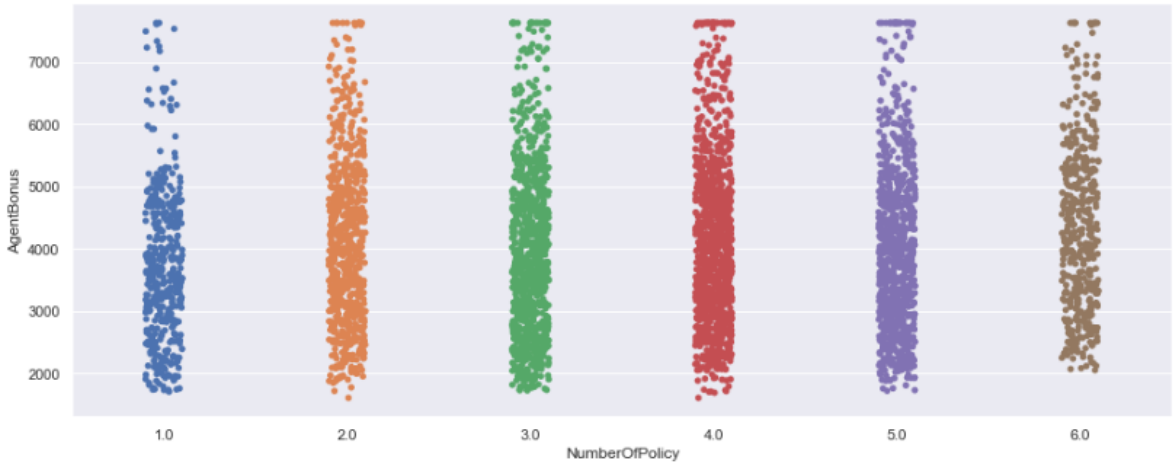

Figure 17 Strip plot for NumberOfPolicy and Agent Bonus

From the above plot, in the number of policies customers have, most customers have 4 policies and contribute more towards bonus, next comes customers with 3 policies, third comes customers with 5 policies followed customers with 2, 6 and 1 policy respectively.

3) Data Cleaning and Pre-processing

Missing value treatment:

As we saw from the above table there are a few fields missing is some of the columns. Hence, we need to fill values for those missing fields. Here we could use the imputation process and we use the median value to fill missing values as these variables are continuous variables.

#	Column	Non-Null Count	Dtype
0	Channel	4520 non-null	object
1	Occupation	4520 non-null	object
2	EducationField	4520 non-null	object
3	Gender	4520 non-null	object
4	ExistingProdType	4520 non-null	int64
5	Designation	4520 non-null	object
6	MaritalStatus	4520 non-null	object
7	Complaint	4520 non-null	int64
8	Zone	4520 non-null	object
9	PaymentMethod	4520 non-null	object
10	CustCareScore	4520 non-null	float64
11	AgentBonus	4520 non-null	float64
12	Age	4520 non-null	float64
13	CustTenure	4520 non-null	float64
14	NumberOfPolicy	4520 non-null	float64
15	MonthlyIncome	4520 non-null	float64
16	ExistingPolicyTenure	4520 non-null	float64
17	SumAssured	4520 non-null	float64
18	LastMonthCalls	4520 non-null	float64

Table 3 Row and column analysis after missing value treatment

Above table contains the variable information and we could observe that there are no missing values in it after treatment.

Outlier value treatment:

In the below graphs we are able to observe that some of the numeric variables are having outliers. Some of these would play a crucial role in model building. Hence, we need to do outlier treatment for those variables.

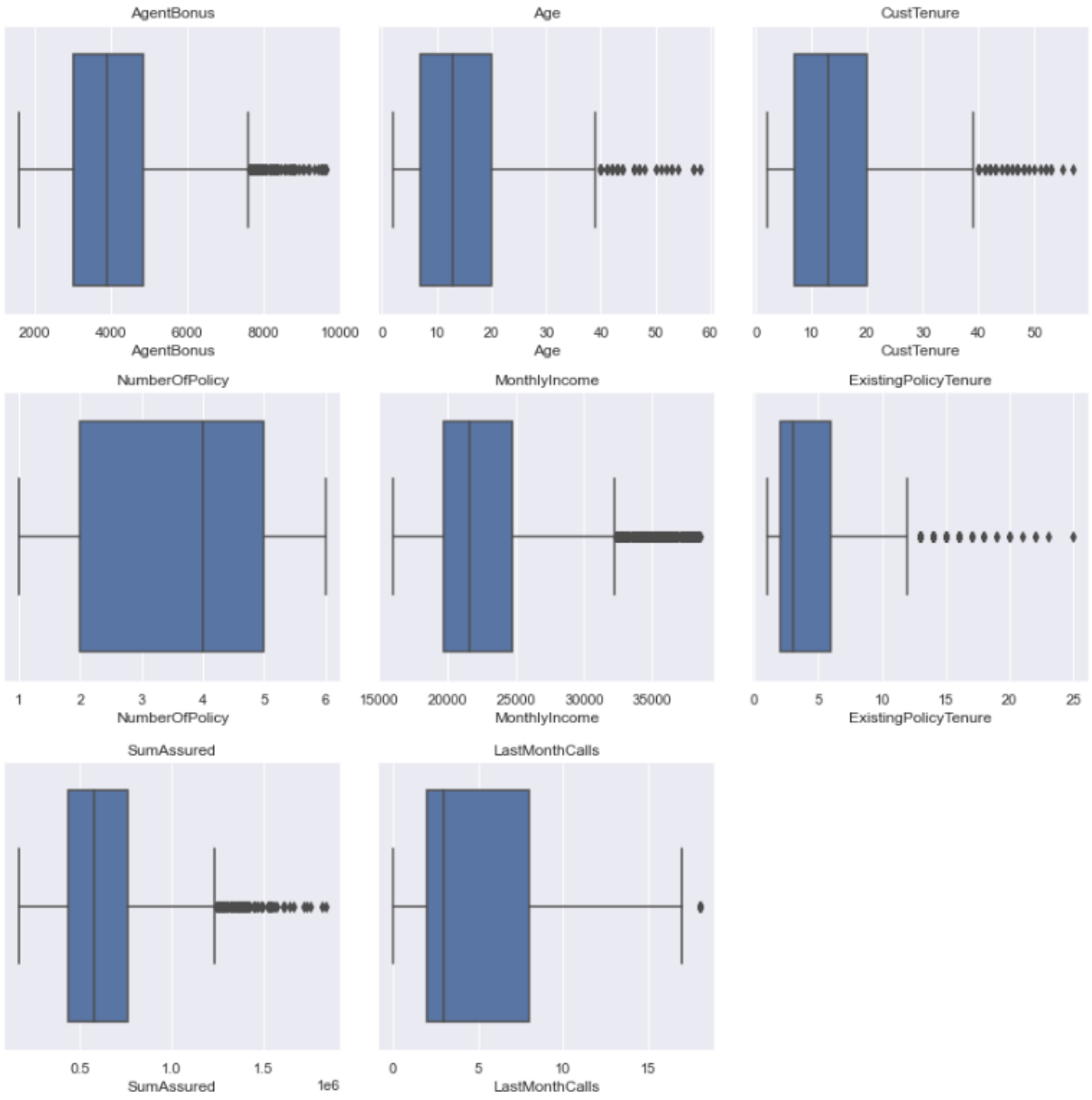

Figure 18 Boxplot before outlier treatement

For outlier treatment we do capping and flooring of the outliers. For that we use the first quartile (25th percentile) for lower values and third quartile (75th percentile) for the higher values.

And after outlier treatment we get the below graphs. We are able to see that now there are no outliers.

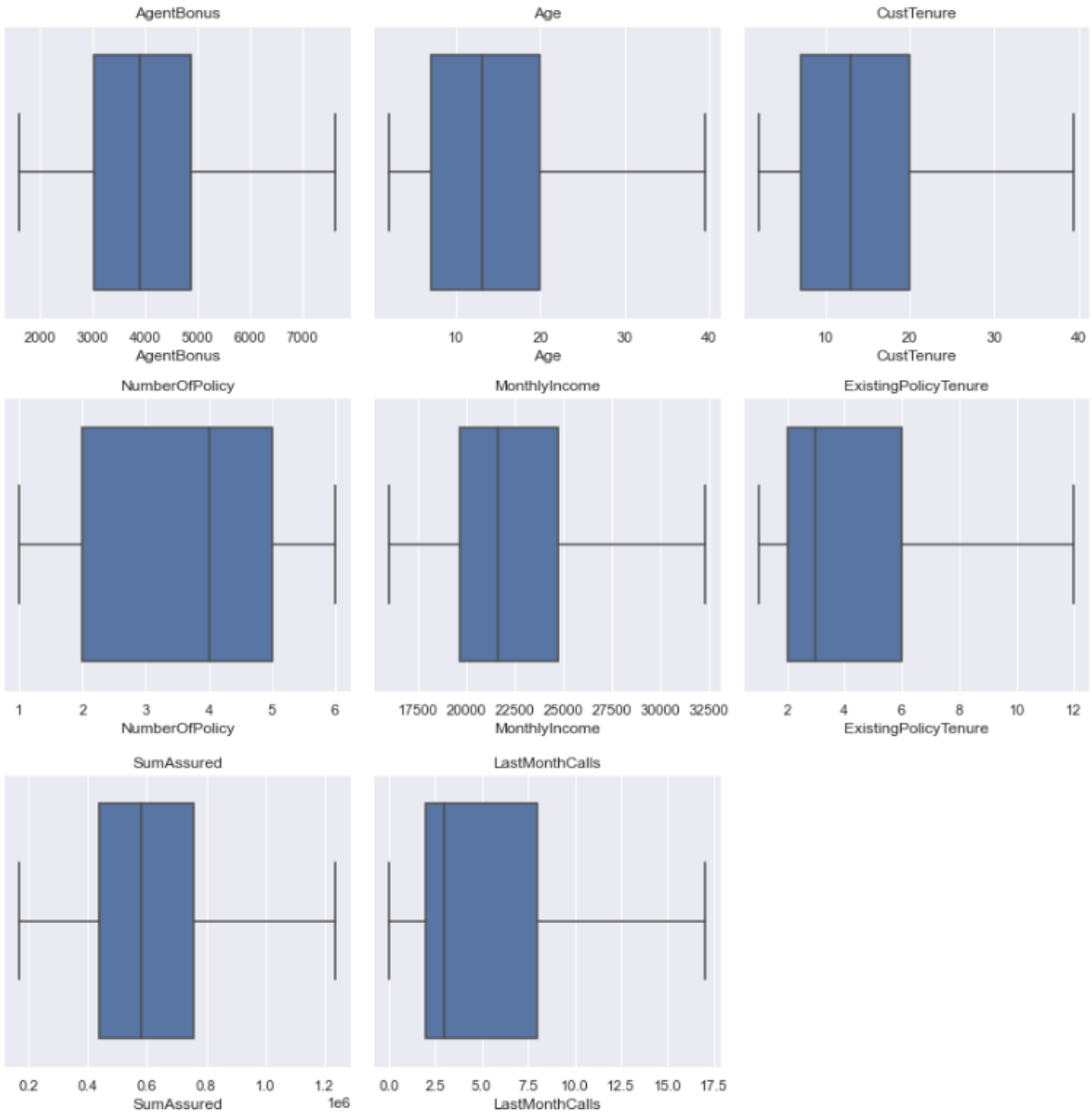

Figure 19 Boxplot after outlier treatement

Variable Transformation:

We do have categorical variables in the data hence we need to encode them into numeric variables so that it would be easy for model building, as some of the models could process only continuous variables.

	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	MaritalStatus	Complaint	Zone	PaymentMethod	CustCareScore	AgentBonus
0	1	1	2	1	3	1	1	1	3	3	2.0	4409.0
1	2	1	2	0	4	1	3	0	3	4	3.0	2214.0
2	1	4	3	0	4	3	1	1	3	4	3.0	4273.0
3	2	1	2	1	3	3	3	1	2	3	5.0	1791.0
4	1	2	2	0	3	3	3	0	2	3	5.0	2955.0

Table 4 Data set after variable transformation

#	Column	Non-Null Count	Dtype
0	Channel	4520 non-null	int64
1	Occupation	4520 non-null	int64
2	EducationField	4520 non-null	int64
3	Gender	4520 non-null	int64
4	ExistingProdType	4520 non-null	int64
5	Designation	4520 non-null	int64
6	MaritalStatus	4520 non-null	int64
7	Complaint	4520 non-null	int64
8	Zone	4520 non-null	int64
9	PaymentMethod	4520 non-null	int64
10	CustCareScore	4520 non-null	float64
11	AgentBonus	4520 non-null	float64
12	Age	4520 non-null	float64
13	CustTenure	4520 non-null	float64
14	NumberOfPolicy	4520 non-null	float64
15	MonthlyIncome	4520 non-null	float64
16	ExistingPolicyTenure	4520 non-null	float64
17	SumAssured	4520 non-null	float64
18	LastMonthCalls	4520 non-null	float64

Table 5 Row and column analysis after variable transformation

From the above two tables we observe that the categorical variables are encoded into numerical variables, which will be suitable for model building.

Removal and addition of new variables:

From the dataset we observe that that field CustID cannot be user in modeling, hence it is dropped.

From the heatmap we observe that there is multicollinearity between the data. On further analysis of multicollinearity based on the Variance Inflation Factor value in the below table, “Existing Product Type” and “Payment Method” is having VIF value of 3.49345 and 3.28916 respectively. This is high

compared to the other feature's values, which means the features are having multicollinearity, hence it can be removed from the data.

```
Channel ---> 1.0102323148236587
Occupation ---> 1.1095743923197483
EducationField ---> 1.1315479717212364
Gender ---> 1.0132694998900091
ExistingProdType ---> 3.4934540441427484
Designation ---> 1.1146984494585517
MaritalStatus ---> 1.0158668053074633
Complaint ---> 1.0037637159037436
Zone ---> 1.0101676816821261
PaymentMethod ---> 3.2891630295369665
CustCareScore ---> 1.0074422266945788
Age ---> 1.317244683153126
CustTenure ---> 1.3174145452682358
NumberOfPolicy ---> 1.0932350468771412
MonthlyIncome ---> 1.623371587555275
ExistingPolicyTenure ---> 1.116241929566706
SumAssured ---> 1.7128814236760648
LastMonthCalls ---> 1.160455005787569
```

Table 6 VIF value of variables

4) Model building

Below mentioned models were built with the final set of variables, after removing the multicollinear variables

- Linear Regression
- Ridge Regression
- Lasso Regression
- KNN
- ANN

Below table contains the metrics for all the models built. From the metrics we see that most models are likely suitable for the data.

	RMSE Train	RMSE Test	MAPE Train	MAPE Test	R2 Train	R2 Test
LinearRegression	0.44617	0.45737	1.80357	4.96002	0.80104	0.79055
Ridge	0.44617	0.45353	1.80347	5.24310	0.80104	0.79406
Lasso	0.47340	0.48352	1.69886	4.38479	0.77601	0.76592
KNN	0.47382	0.60148	1.82464	6.24221	0.77561	0.63777
ANN	0.41611	0.44052	1.89944	4.22962	0.82694	0.80570

Table 7 Evaluation metrics of models built

Below ensemble method models were built, as ensemble techniques will help in improving accuracy

- Random Forest
- Ada Boosting
- Gradient Boosting

	RMSE Train	RMSE Test	MAPE Train	MAPE Test	R2 Train	R2 Test
RandomForest	0.13899	0.38293	0.58924	2.56306	0.98069	0.85318
AdaBoosting	0.47131	0.49075	2.26774	5.29454	0.77798	0.75887
GradientBoosting	0.28071	0.37961	1.16028	2.99394	0.92124	0.85572

Table 8 Metrics of ensemble models

Above table contains the scores of models built using the ensemble techniques, we have Random Forest, Ada Boosting and Gradient Boosting models built.

We observe that Ada boosting is having very less value of R squared. Random Forest is having very high R square value. And Gradient boosting is having considerable good R squared, MAPE and RMSE value.

To have further more improved scores, we did hyperparameter tuning of the below models by Grid Search Cross Validation.

- KNN
- ANN
- Random Forest

Below table contains the improved scores of models KNN, ANN and Random Forest by tuning.

	RMSE Train	RMSE Test	MAPE Train	MAPE Test	R2 Train	R2 Test
KNN_GC	0.49380	0.58417	1.58372	4.68577	0.75628	0.65832
ANN_GC	0.41162	0.44195	1.75110	4.26077	0.83065	0.80443
RandomForest_GC	0.13899	0.38293	0.72438	2.63631	0.97164	0.85542

Table 9 Metrics of models after tuning

5) Model validation

	RMSE Train	RMSE Test	MAPE Train	MAPE Test	R2 Train	R2 Test
LinearRegression	0.446165	0.457369	1.803565	4.960022	0.801037	0.790553
Ridge	0.446165	0.453527	1.803467	5.243104	0.801037	0.794057
Lasso	0.473401	0.483515	1.698860	4.384791	0.776005	0.776005
KNN	0.473820	0.601484	1.824638	6.242215	0.775608	0.637765
ANN	0.416069	0.443265	1.888912	4.574432	0.826974	0.803271
RandomForest	0.138991	0.382935	0.589238	2.563062	0.980691	0.853178
AdaBoosting	0.471308	0.490747	2.267738	5.294542	0.777981	0.758867
GradientBoosting	0.280708	0.379609	1.160277	2.993942	0.921243	0.855718
KNN_GC	0.493805	0.584167	1.583721	4.685772	0.756280	0.658323
ANN_GC	0.411624	0.441952	1.751098	4.260769	0.830651	0.804435
RandomForest_GC	0.138991	0.382935	0.713459	2.600150	0.971793	0.855483

Table 10 Metrics of all models built

From the above table we observe that for KNN model the difference between the train and test values of MAPE and RMSE is high and the difference between test and train R square is also high compared to most of the models, for before and after tuning, hence it is not suitable for the problem.

For ANN model the difference between R squared and RMSE of train and test is less, however the difference between the test and train value of MAPE is high. Same is the case for ANN grid search model. So, this is not suitable for the problem.

In Random forest model and Random forest grid search model, we observe that the difference in test and train value of R square, RMSE and MAPE are high. So, it is not suitable for the problem.

Ada boosting, Linear Regression, Ridge Regression, Lasso Regression, model has good test and train R square value and RMSE value. However, the difference is the test and train MAPE is high. Hence it is also not suitable for the problem.

Random Forest gradient boosting model has good scores in terms of R squared, RMSE and MAPE. Also, the difference between the train and test value of RMSE and MAPE is very less. Hence, Random forest gradient boosting is the best model for this data.

6) Final interpretation / recommendation

Insights:

- From the different models built we could see that Rando grid search cross validation is the best model for the data.
- Agents are doing good job in terms of insurance sales compared to online sales and third-party sales.

- West and North zone is having higher sales respectively, however East has very less sales and South zone has negligible sales.
- Under graduates contribute more towards sales while Post graduate and diploma graduates contribute considerably less.
- On designation VP, AVP and Senior manage contributes more for Agent bonus, however Manager and executives contribute less.
- Product type 4 is having good sales, next comes product type 3 and 5 respectively. Product types 1, 2 and 6 are having very low sales.
- Younger customers are more in number, as age increases number of customer decreases.
- One third of the customers have filed complaint in last month which is a high number.
- Customers having less tenure are more in number, as customer tenure increases the number of customers decreases.

Recommendations:

- Agents should concentrate more on East and South region as company could get new customers and new sales that would eventually increase profit.
- More the income is more the agent bonus is, customer with higher income should be targeted, also higher the sum assured is higher the bonus is, so customer should be provided with high sum assured.
    - Agents who so far has been targeting customers with high income and providing higher sum assured policies to customers should be given higher bonus.
    - It is better to provide with specific plan or create new plans for customers with high income and also provide with specific plan or create new plans that gives high sum assured to customers.
- On designation Managers and Executives who are comparatively in lower designation contribute less to agent bonus which is an impact of the monthly income. Hence providing plans with less price or premium will help in increasing sales among them.
- Agents who worked on product type 4, 3 and 5 should be rewarded more bonus for good sales. Agents who worked on product type 1, 2 and 6 should be given training to improve sales.
- Age plays a major role in agent bonus, we observe as customers grow older the policy might not be beneficial for them, hence it is better to classify the age groups and provide plans for them.
- Increasing the customer service would reduce the complaints raised, which would eventually improve the credibility and sales.
- Customers should be provided benefits to continue their current policy which would improve customer tenure agent bonus and sales.
- Diploma graduates might have less income, so they must be provided with plans of less price. Post graduates should have more income that they should think that they could manage without insurance, so they should be provided with plans with more benefits irrespective of price.