CAPSTONE PROJECT

Life Insurance

Name:	**Irene Asha Moses**
Batch:	**G-2 DSBA Aug B 20**
Date:	**11-Sep-2021**

INDEX

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION OF THE BUSINESS PROBLEM:

a) 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.

b) Purpose

- ➢ The purpose of this project to draw out valuable insights and make predictions from the given data.
- ➢ The predictions will be made on the agent data, and these will later be classified into two groups (i.e. high performing and low performing agents).
- ➢ Depending on the classification, the company will need to either upskill the agents or provide engagement activity.

c) Business/social opportunity

The project may provide some valuable information on the factors that are important to focus on for other Life Insurance studies as well.

2. EDA AND BUSINESS IMPLICATION:

a) Key points regarding the dataset

The dataset appears to be from country India, as in the Data Description, there is a variable called 'Zone' which tells us which zone in India, the customers are from.

The column/attribute information is given below

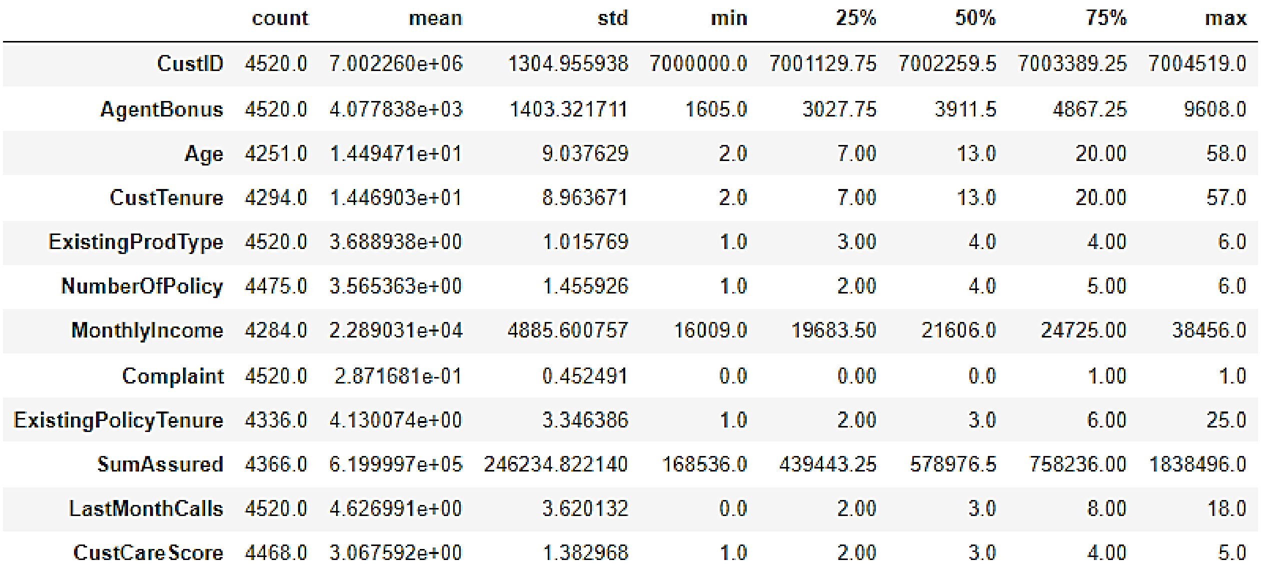
VARIABLES	DESCRIPTION
CUSTID	Unique customer ID
AGENTBONUS	Bonus amount given to each agent in last month
AGE	Age of customer
CUSTTENURE	Tenure of customer in organization
CHANNEL	Channel through which acquisition of customer is done
OCCUPATION	Occupation of customer
EDUCATIONFIELD	Field of education 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 2.1 – Variable information

- There are totally 4520 observations and 20 columns.
- There are totally 8 object type data and 12 numeric data (5 being discrete and 7 being continuous)
- There are some null values in the data set and these will be imputed or replaced with the required value later.

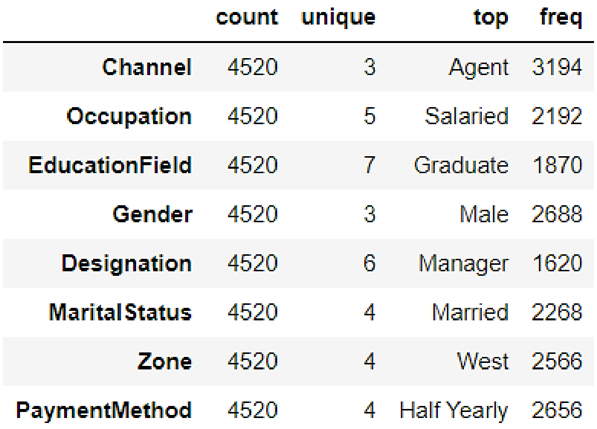
- The **Descriptive Statistics** of the numeric data is taken.



	count	mean	std	min	25%	50%	75%	max
CustID	4520.0	7.002260e+06	1304.955938	7000000.0	7001129.75	7002259.5	7003389.25	7004519.0
AgentBonus	4520.0	4.077838e+03	1403.321711	1605.0	3027.75	3911.5	4867.25	9608.0
Age	4251.0	1.449471e+01	9.037629	2.0	7.00	13.0	20.00	58.0
CustTenure	4294.0	1.446903e+01	8.963671	2.0	7.00	13.0	20.00	57.0
ExistingProdType	4520.0	3.688938e+00	1.015769	1.0	3.00	4.0	4.00	6.0
NumberOfPolicy	4475.0	3.565363e+00	1.455926	1.0	2.00	4.0	5.00	6.0
MonthlyIncome	4284.0	2.289031e+04	4885.600757	16009.0	19683.50	21606.0	24725.00	38456.0
Complaint	4520.0	2.871681e-01	0.452491	0.0	0.00	0.0	1.00	1.0
ExistingPolicyTenure	4336.0	4.130074e+00	3.346386	1.0	2.00	3.0	6.00	25.0
SumAssured	4366.0	6.199997e+05	246234.822140	168536.0	439443.25	578976.5	758236.00	1838496.0
LastMonthCalls	4520.0	4.626991e+00	3.620132	0.0	2.00	3.0	8.00	18.0
CustCareScore	4468.0	3.067592e+00	1.382968	1.0	2.00	3.0	4.00	5.0

Figure 2.1 – Numeric data Descriptive Statistics

- And below is the descriptive statistics for object/string type data:



	count	unique	top	freq
Channel	4520	3	Agent	3194
Occupation	4520	5	Salaried	2192
EducationField	4520	7	Graduate	1870
Gender	4520	3	Male	2688
Designation	4520	6	Manager	1620
MaritalStatus	4520	4	Married	2268
Zone	4520	4	West	2566
PaymentMethod	4520	4	Half Yearly	2656

Figure 2.2 – Object data Descriptive Statistics

Observations:

- Around 50% of the records have the Agent Bonus around 3911.5, with the minimum limit as 1605.0 and maximum limit as 9608.0.
- A majority of the customers' age (75%) is 20 years, with the maximum being 58 years. The minimum customer age is 2 years, which is strange.
- Most of the records (75%) have the Customer Tenure as 20 years.
- Around 50% of customers have monthly income as 21606.0.
- The maximum Existing Policy Tenure is 25 years.
- The most frequently used Channel for customer acquisition is 'Agent'
- A large proportion of the customers are Salaried.
- A lot of the records have 'West' Zone.
- There appear to be many categories for 'Gender' and 'Married' which is strange. This will be analysed in the later sections.
- There are no duplicate records.
- Renaming of variables will not be done.

b) Univariate analysis

The below analysis is done for the variables ***that are relevant*** for the final machine learning model.

Univariate analysis for Continuous variables:

I. AGENT BONUS

- The box plot shows that there are outliers (values that are too small or too large), where the values are almost above 7500.
- The Distribution plots shows the spread of the data, and here, it is observed that the distribution is a bit right skewed (i.e. there are a large number of data points having large values)

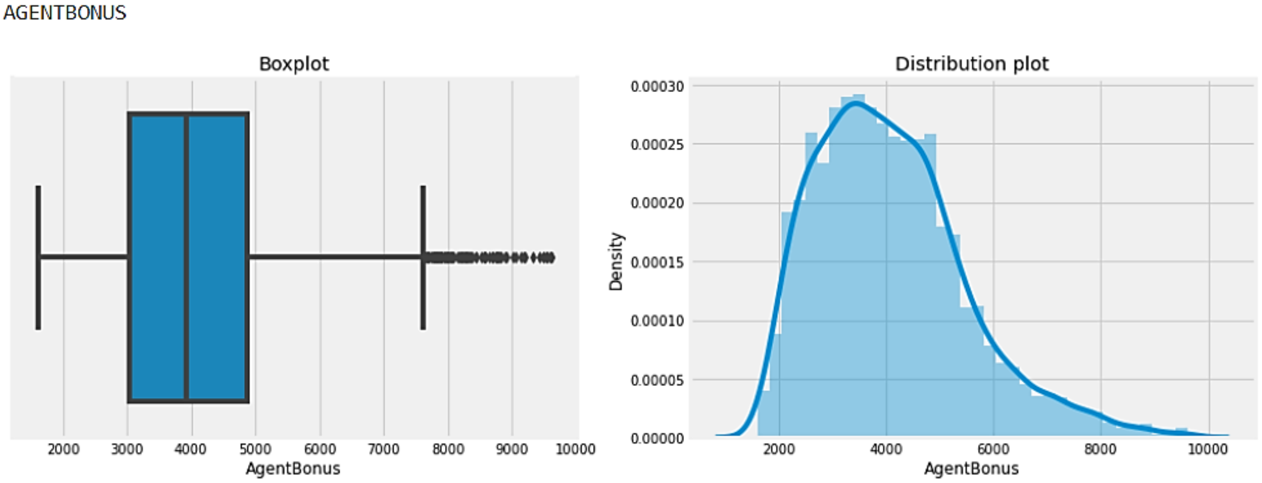


Figure 2.3– Boxplot & Distribution plots for AgentBonus

II. AGE

- The box plot shows that there are outliers where the values are almost above 40.
- The Distribution plots shows that the distribution is a bit right skewed.

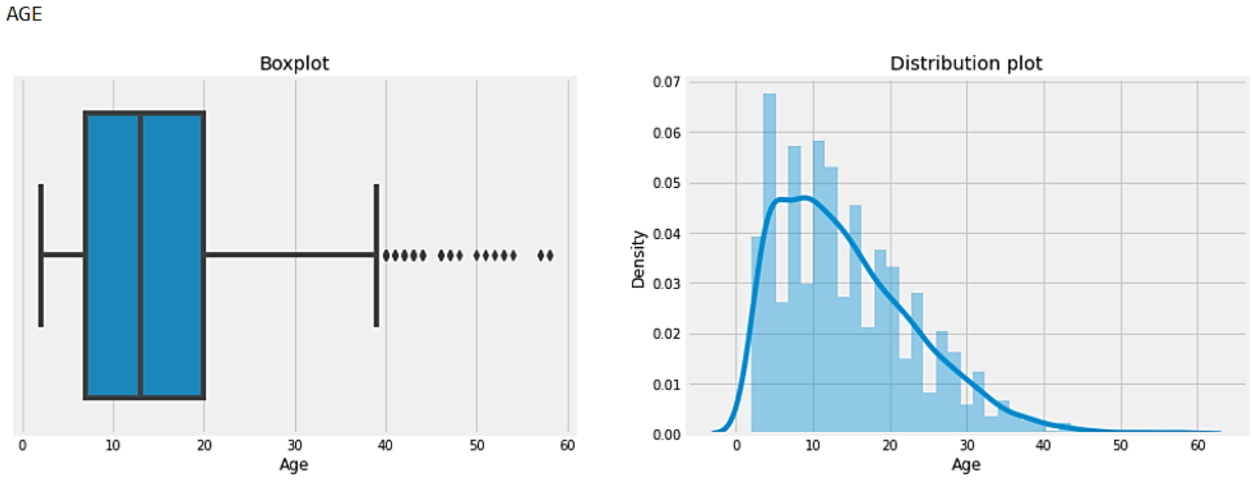


Figure 2.4– Boxplot & Distribution plots for Age

III. <u>**CUSTTENURE**</u>

- The highest customer tenure appears to be around 4 or 5 years.
- The box plot shows that there are outliers where the values are almost above 40.
- The Distribution plots shows that the distribution is a bit right skewed.

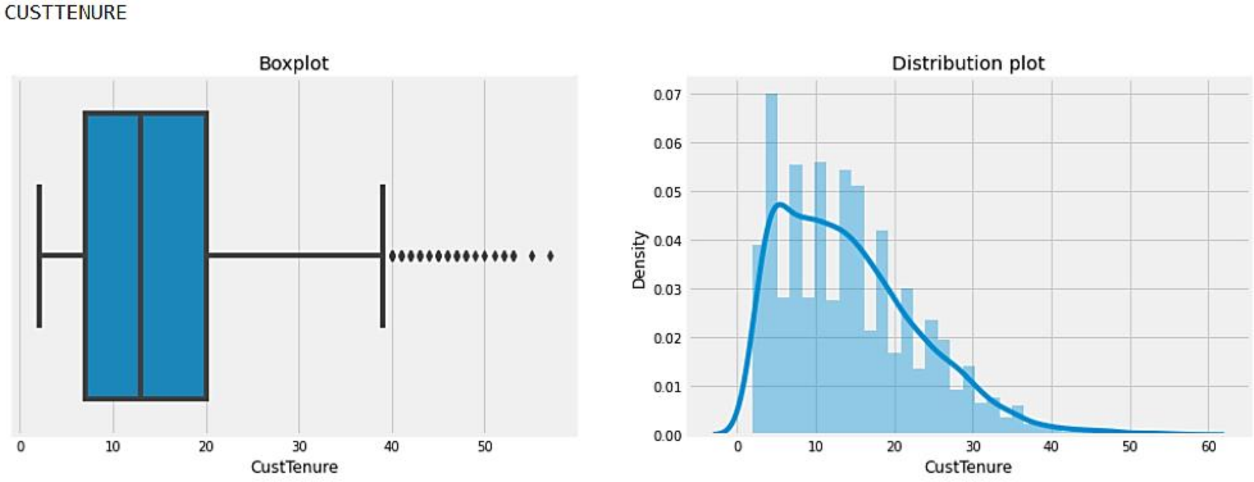


Figure 2.5– Boxplot & Distribution plots for CustTenure

IV. <u>**MONTHLYINCOME**</u>

- The box plot shows that there are outliers where the values are almost above 32000.
- The Distribution plots shows that the distribution is a bit uneven and right skewed.
- There appear to be 2 peaks

MONTHLYINCOME

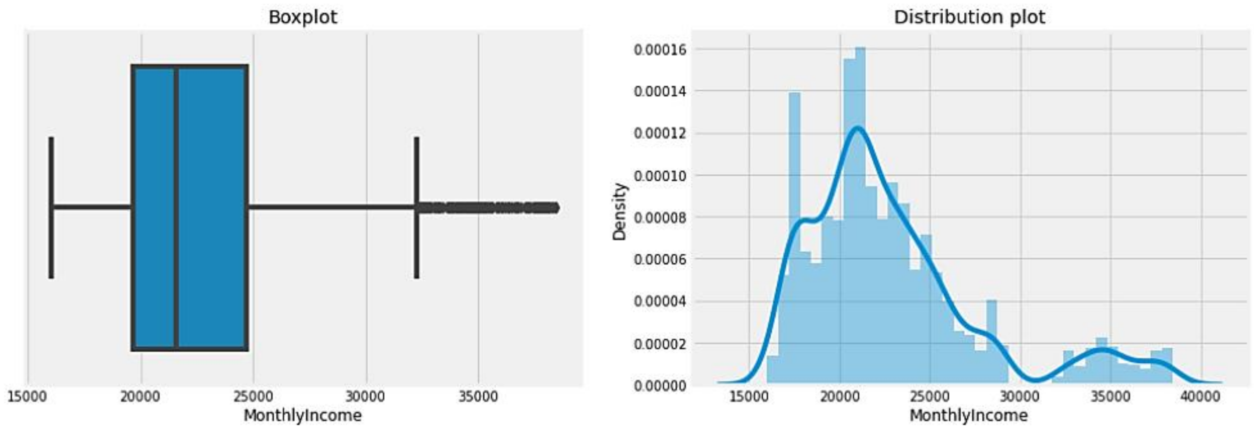


Figure 2.6 – Boxplot & Distribution plots for MonthlyIncome

V. SUMASSURED

- The box plot shows that there are some outliers with very high values.
- The Distribution plots shows that the distribution is right skewed.

SUMASSURED

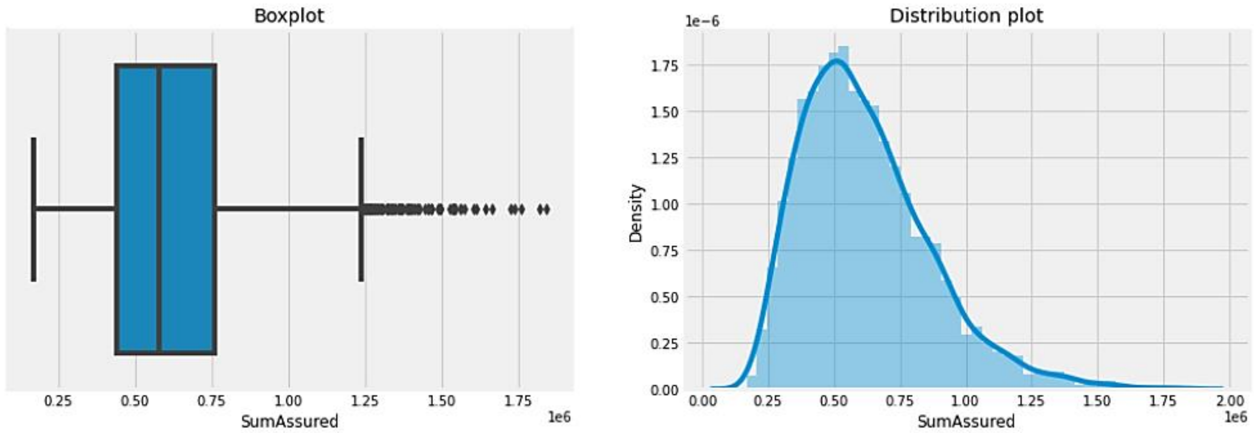


Figure 2.7– Boxplot & Distribution plots for SumAssured

Univariate analysis for Discrete variables:

VI. EXISTINGPOLICYTENURE

- Most of the records (990) have a policy tenure of only 1 year
- Few records have a policy tenure of over 5 years.

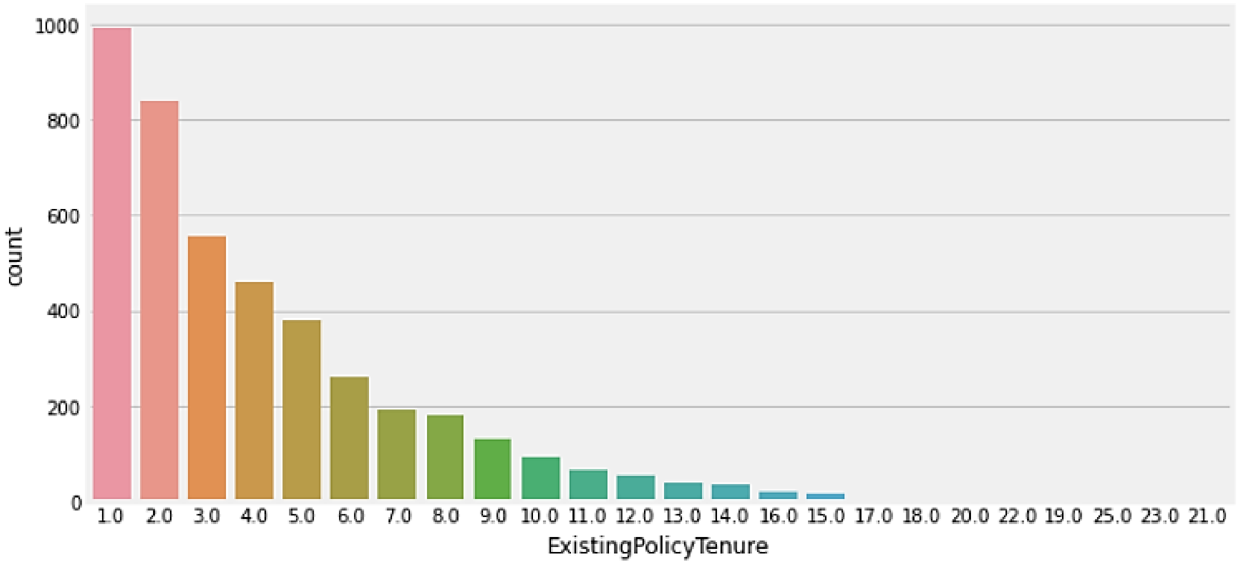


Figure 2.8– Countplot for ExistingPolicyTenure

Univariate analysis for Object/Categorical variables:

VII. DESIGNATION

- A majority of the customers are Managers (1620) and Executives (1535).
- There is a redundant word 'Exe' for Executive.

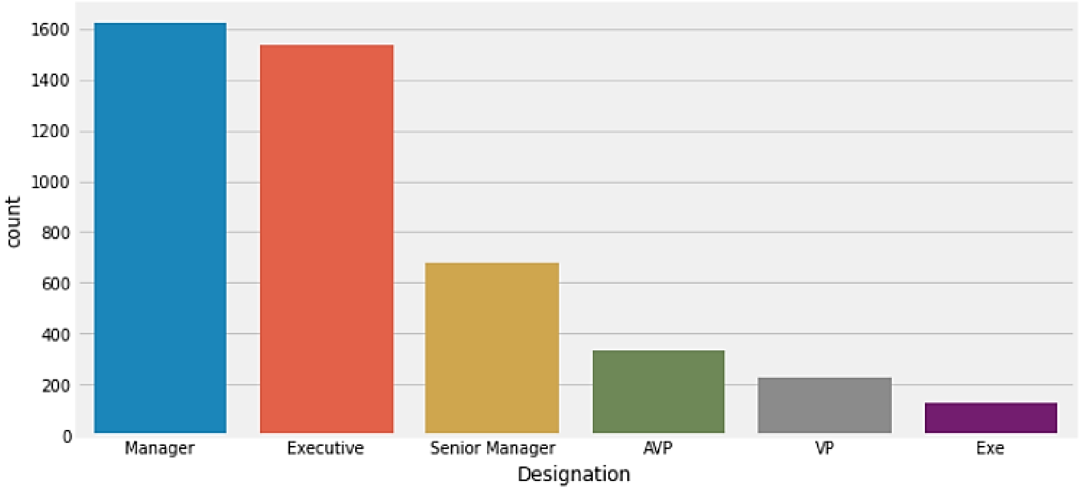
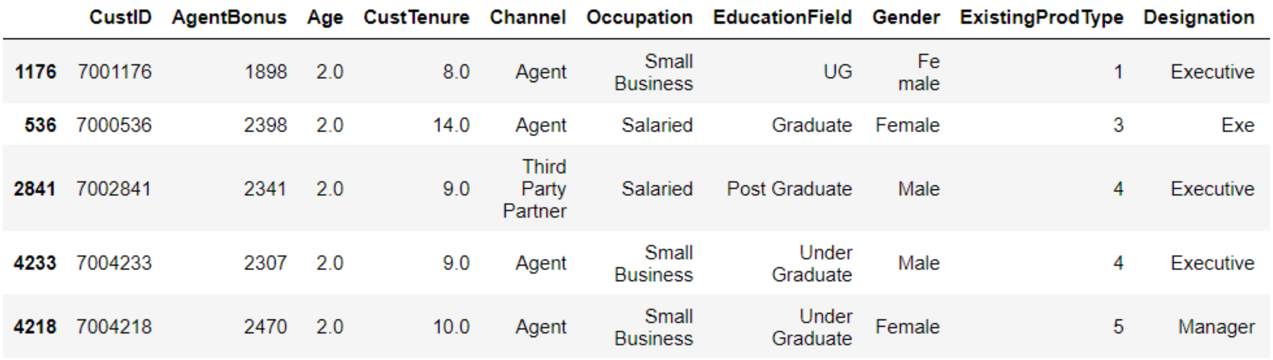


Figure 2.9– Countplot for Designation

UNIVARIATE ANALYSIS FOR 'AGE'

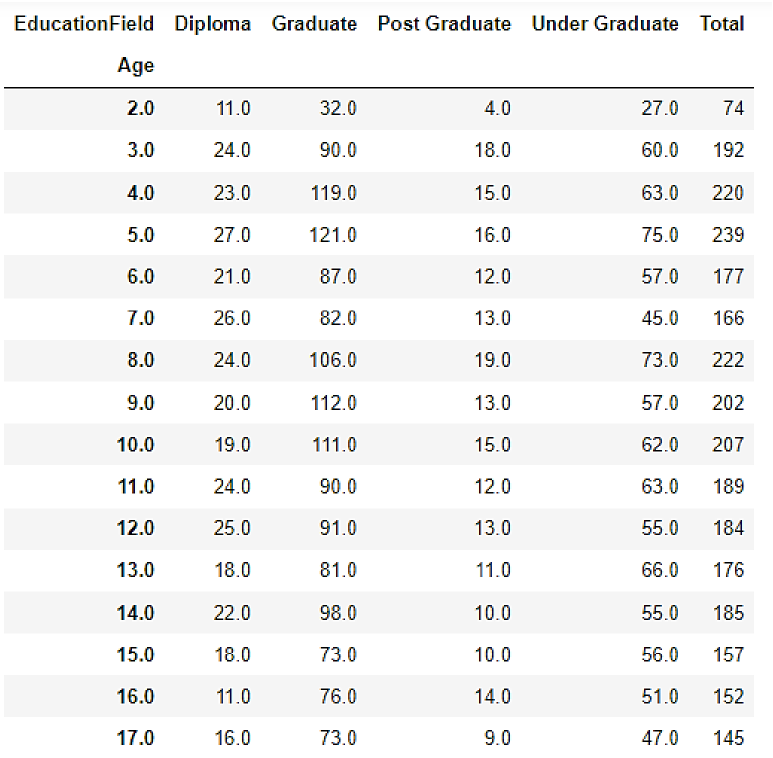
- On checking the dataset, it is observed that there are records with ages as less as 2, 3, 4..etc.
- These records have high values for Customer Tenure and high Education levels also (like Graduate, Diploma).

- ➢ These are very strange and hence, they are imputed with the appropriate values. In this case study, the appropriate age groups are defined for different Education levels, and then a random number is selected from those age groups, with which the erratic age value can be imputed.



	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation
1176	7001176	1898	2.0	8.0	Agent	Small Business	UG	Fe male	1	Executive
536	7000536	2398	2.0	14.0	Agent	Salaried	Graduate	Female	3	Exe
2841	7002841	2341	2.0	9.0	Third Party Partner	Salaried	Post Graduate	Male	4	Executive
4233	7004233	2307	2.0	9.0	Agent	Small Business	Under Graduate	Male	4	Executive
4218	7004218	2470	2.0	10.0	Agent	Small Business	Under Graduate	Female	5	Manager

Figure 2.10– Data overview for erratic age values



EducationField	Diploma	Graduate	Post Graduate	Under Graduate	Total
Age					
2.0	11.0	32.0	4.0	27.0	74
3.0	24.0	90.0	18.0	60.0	192
4.0	23.0	119.0	15.0	63.0	220
5.0	27.0	121.0	16.0	75.0	239
6.0	21.0	87.0	12.0	57.0	177
7.0	26.0	82.0	13.0	45.0	166
8.0	24.0	106.0	19.0	73.0	222
9.0	20.0	112.0	13.0	57.0	202
10.0	19.0	111.0	15.0	62.0	207
11.0	24.0	90.0	12.0	63.0	189
12.0	25.0	91.0	13.0	55.0	184
13.0	18.0	81.0	11.0	66.0	176
14.0	22.0	98.0	10.0	55.0	185
15.0	18.0	73.0	10.0	56.0	157
16.0	11.0	76.0	14.0	51.0	152
17.0	16.0	73.0	9.0	47.0	145

Figure 2.11– Age data as per EducationField

Actions done:

- ✓ The ages are imputed as per Education level:
  - ➢ If the age is less than 18 and Education level is ‘Under Graduate’, replace it with any number between 18 to 21.
  - ➢ If the age is less than 22 and Education level is ‘Graduate’, replace it with any number between 22 to 25.

- If the age is less than 21 and Education level is 'Post Graduate', replace it with any number between 20 to 28.
- If the age is less than 15 and Education level is 'Diploma', replace it with any number between 14 to 19.

The distribution of Age now looks like it's right-skewed, and most of the records appear to have ages between 20 to 30 years.

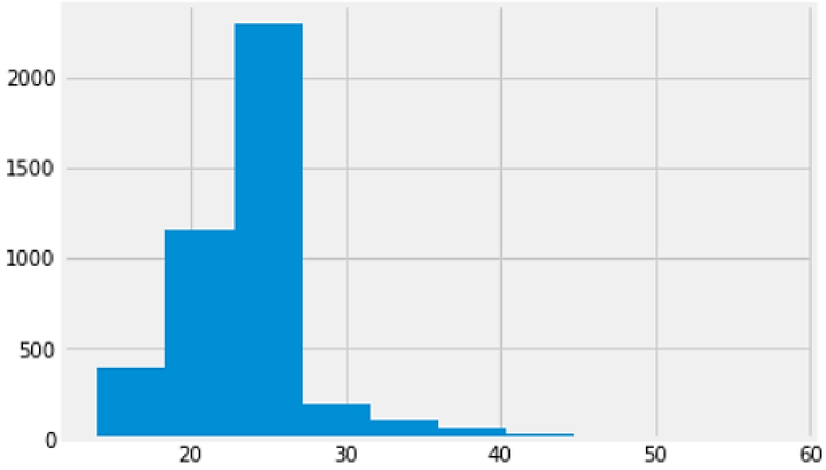


Figure 2.12– Age data Distribution after Imputing

UNIVARIATE ANALYSIS FOR 'AGENTBONUS'

On a benchmark can be selected after predicting the Agent Bonus by checking the Histogram. For example, in the current dataset, below is the Agent Bonus histogram.

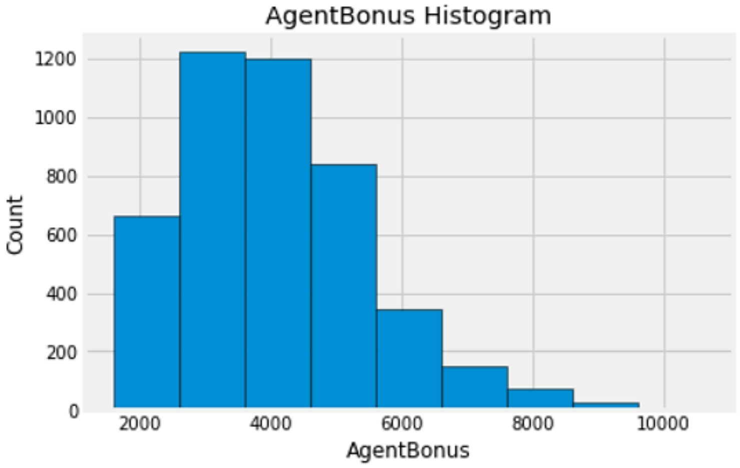


Figure 2.13– Agent Bonus Histogram

From this histogram, a benchmark of **5605** can be chosen i.e. any agent having a bonus above this value can be considered a high-performing agent. Based on this benchmark, around 594 agents (i.e 13.14% of the dataset) are high-performing. The benchmark taken after the predictions may vary a little from the above benchmark.

Fixing certain Attribute values:

Before going further with the Bivariate Analysis, certain attribute values are adjusted as some of them have synonymous words and spelling mistakes.

The below actions are done before proceeding further:

- ✓ **Occupation** – The value 'Laarge Business' is replaced with 'Large Business'
- ✓ **EducationField** –
  - ▪ The value 'UG' is replaced with 'Under Graduate' and 'Engineer' is replaced with 'Graduate' (since Engineer is a subtype of graduate and a person can say that they are an "Engineering Graduate").
  - ▪ The value 'MBA' is replaced with 'Post Graduate', since MBA is a PG degree.
- ✓ **Gender** – The value 'Fe male' is replaced with 'Female'
- ✓ **Designation** – The value 'Exe' is replaced with 'Executive'.
- ✓ **MaritalStatus** - The value 'Unmarried' is replaced with 'Single'.

c) Bivariate analysis

- ➢ The pairplot displays the relationships for all combination of the given variables. The pairplot is made based on all the important variables except the 'CustID' (since that is just the customer ID and is not important for analysis/modelling).
- ➢ The bivariate analysis that is displayed below was done for some important variables –
  - ✓ Age
  - ✓ SumAssured
  - ✓ MonthlyIncome
  - ✓ CustTenure
  - ✓ ExistingPolicyTenure

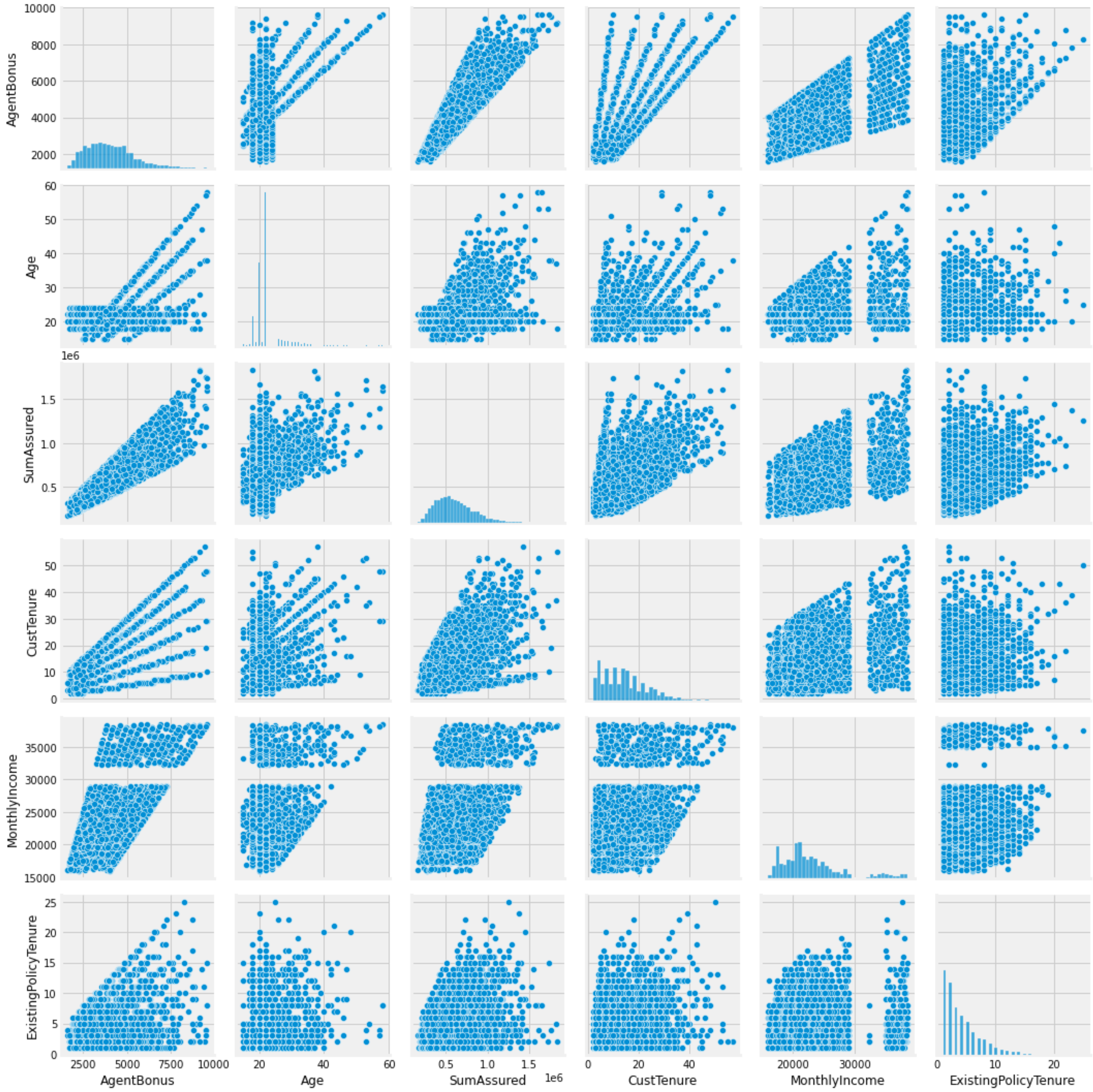


Figure 2.14– Pairplot for important attributes

And below is a notable observation between object variable Designation and AgentBonus

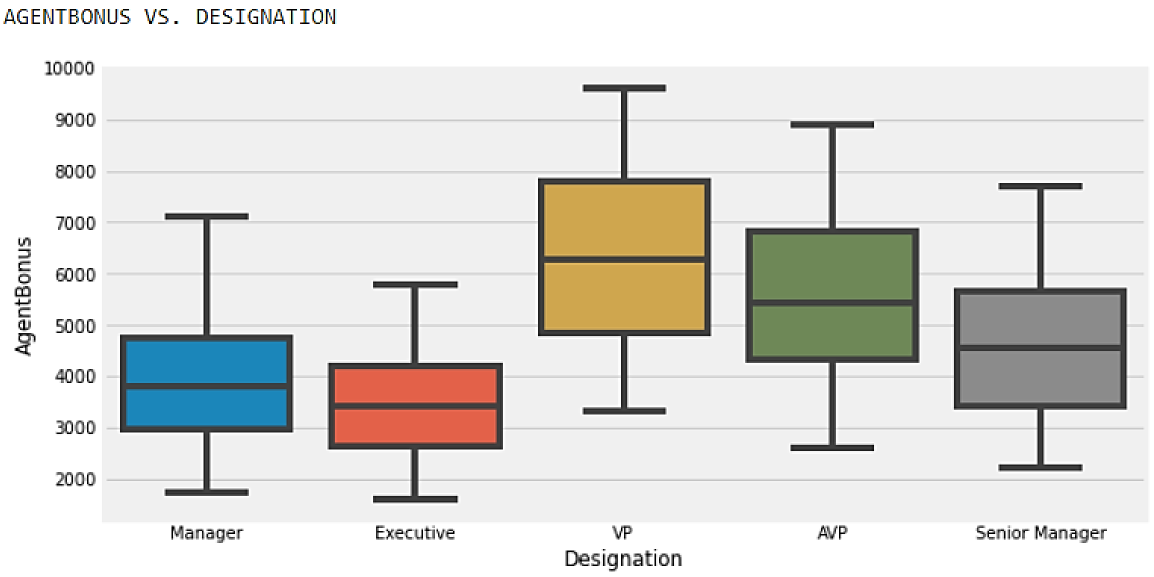


Figure 2.15– AgentBonus vs. Designation Boxplot

Below are the inferences

VARIABLES	OBSERVATIONS
Age & AgentBonus	There does not appear to be a very strong linear relationship between these two variables.
CustTenure & AgentBonus	There is a positive linear relationship
ExistingPolicyTenure & AgentBonus	There is a positive linear relationship
SumAssured & AgentBonus	There is a very strong positive linear relationship
Designation & AgentBonus	The Agent Bonus is high for customers with designation 'VP'.

Table 2.2– Bivariate analysis observations

3. DATA CLEANING AND PREPROCESSING:

a) Removal of unwanted variables

The column 'CustID' is dropped as it is just the customer ID and is not an important feature.

b) Encoding String values

There are two types of categorical data in the given dataset. They are Nominal variables and Ordinal variables, and they will be encoded differently.

Ordinal variables (Having rank/order):

- **EducationField:** Post Graduate > Graduate > Under Graduate > Diploma
- **Designation:** VP > AVP > Senior Manager > Manager > Executive
- **Occupation:** Large Business > Small Business > Salaried > Freelancer

Nominal variables (Having no rank/order):

- The variables 'Channel', 'Gender','MaritalStatus','Zone','PaymentMethod' are encoded using the ‘get_dummies’ method of the ‘pandas’ library.

c) Missing value Treatment

The below table shows the percentage of null values for each variable.

Variables	Null_%
Age	5.95
MonthlyIncome	5.22
CustTenure	5
ExistingPolicyTenure	4.07
SumAssured	3.41
CustCareScore	1.15
NumberOfPolicy	1

Table 3.1– Variables having null values

The percentage of null values is less than 6%, so the missing values will be replaced with the corresponding column statistic like median, mean, or they will be imputed.

So, in this case:

- The missing values for Age will be replaced with the Median (middle-most value in terms or ascending/descending order of the ages) as per each education level.
- The CustCareScore is also an ordinal variable which is already in numbers. The missing values will be replaced with the Mode (most recurring value). The same will also be done for NumberOfPolicy column and other discrete columns.
- The remaining continuous numeric columns will be imputed using the KNNImputer method with the ‘neighbours’ parameter as 10. This means that the method will replace the missing value with the Average calculated from the surrounding 10 data points.

d) Outlier Treatment

Below are the percentage of outliers for the independent variables.

Variables	Outliers_%
Channel_Third_Party_Partner	18.98
Age	11.24
Channel_Online	10.35
MonthlyIncome	8.38
PaymentMethod_Monthly	7.83
ExistingProdType	6.77
ExistingPolicyTenure	4.14
AgentBonus	2.21
SumAssured	2.21
PaymentMethod_Quarterly	1.68
CustTenure	1.42

LastMonthCalls	0.27
Zone_South	0.13

Table 3.2– Outlier% for features

From the above screenshot, we can see that most of the variables are categorical (i.e. those having 0's and 1's..etc). Hence, only the continuous variables - 'MonthlyIncome' and 'SumAssured' will be treated.

Actions done:

- ✓ The outliers are treated by Replacing very low values with the 25th quantile and very high values with the 75th quantile.

e) Multicollinearity Check

The heatmap is taken for all the attributes.

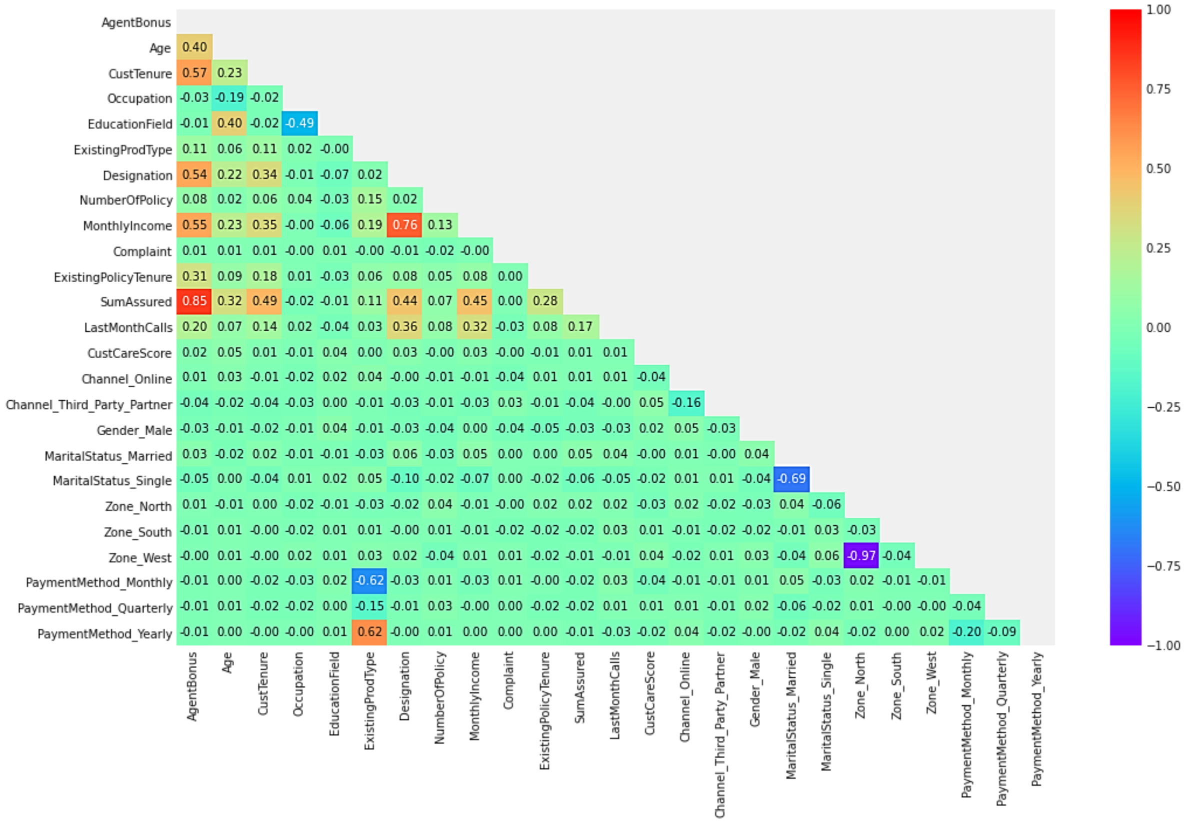


Figure 3.1– Heatmap

Observations:

1) AgentBonus is highly correlated with the below variables:

- ✓ Age
- ✓ CustTenure
- ✓ Designation
- ✓ MonthlyIncome
- ✓ SumAssured

2) There is also some multicollinearity between some independent variables i.e. changes in one variable affects the other variables, and this may lead to some incorrect predictions in the machine learning model.

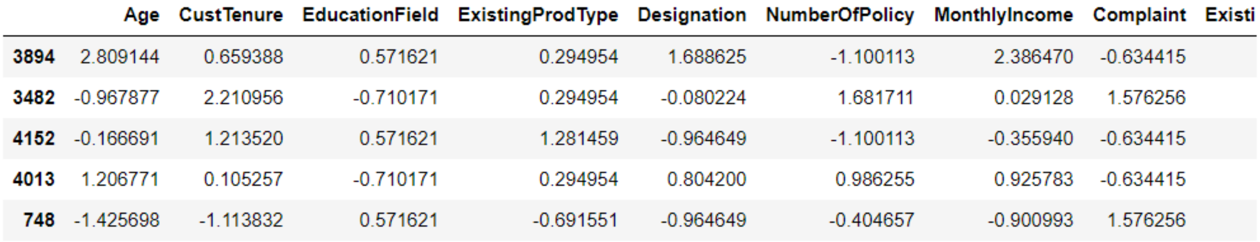
f) Data Split

Two dataframes are created, where one contains all the attributes except the target variable 'AgentBonus', and this set is named 'X'. The other dataframe only contains the target variable 'AgentBonus', and this set is named 'y'.

The data from X and y is split into train and test sets, in the 70:30 ratio i.e. 70% of the data from X and y are for training, and the remaining 30% are for testing.

g) Scaling Predictors

Since distance calculation is involved in some regression models, any differences in value ranges in the predictors will impact the learning process. Hence, Standard Scaling is done.



	Age	CustTenure	EducationField	ExistingProdType	Designation	NumberOfPolicy	MonthlyIncome	Complaint	Existi
3894	2.809144	0.659388	0.571621	0.294954	1.688625	-1.100113	2.386470	-0.634415	
3482	-0.967877	2.210956	-0.710171	0.294954	-0.080224	1.681711	0.029128	1.576256	
4152	-0.166691	1.213520	0.571621	1.281459	-0.964649	-1.100113	-0.355940	-0.634415	
4013	1.206771	0.105257	-0.710171	0.294954	0.804200	0.986255	0.925783	-0.634415	
748	-1.425698	-1.113832	0.571621	-0.691551	-0.964649	-0.404657	-0.900993	1.576256	

Figure 3.2–Scaled dataset

h) Variable transformation

The variable transformation will not be done and the scaled train sets will be used for further modelling.

i) Addition of new variables

There will be no addition of new variables.

4) MODEL BUILDING:

a) Build various models

The various supervised machine learning models used are as follows:

i. Linear Regression
ii. Random Forest Regressor
iii. Gradient Boosting Regressor
iv. K-Neighbours Regressor

So, two ensemble models are used.

First, the models are built without any optimization, just to check the general performance of each model.

i. LINEAR REGRESSION:

- It is a supervised machine learning model which is used only if the relationship between the target and the predictors is linear in nature.
- The linear regression model is fit to the train set and then made to predict on the test set.

The below performance metrics are used for this model as well as the other models

i. ***Root Mean Squared (RMSE)***
ii. ***R-Squared Error***
iii. ***Adjusted R-Squared***
iv. ***Mean Absolute Error (MAE)***

Model	R2_ train	R2_ test	Adj_R2_ train	Adj_R2_ test	RMSE_train	RMSE_test	MAE_ train	MAE_ test
Linear Regression	81.07	81.11	80.96	80.77	606.27	618.17	484.08	489.02

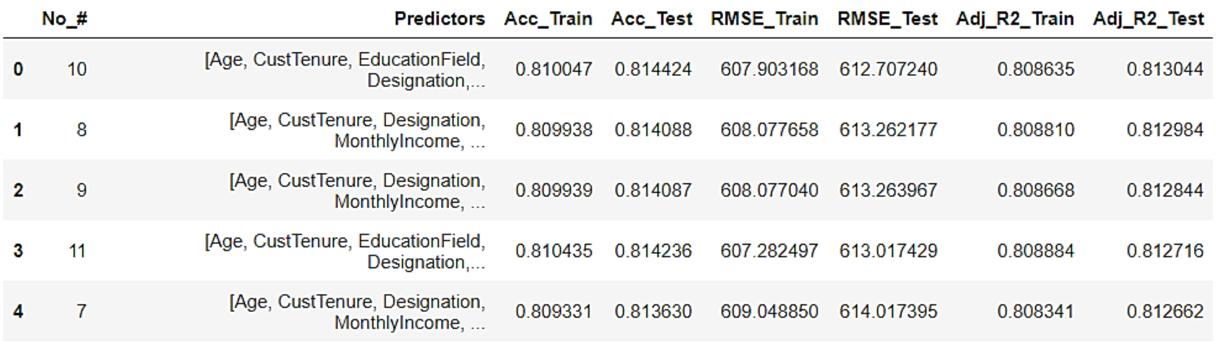
Table 4.1–Linear Regression base model metrics

The above metrics show that the basic Linear Regression model is relatively good, however, it is observed that the RSME of the test set is a bit higher than that of the train set, hence, **Feature Selection** will be done in order to leave out any un-important variables, and to check of the Adjusted R-Squared score can be increased.

FEATURE SELECTION USING RFE (Recursive Feature Elimination):

- This algorithm is selected to iterate through all the features and to one by one, eliminate the features that do not contribute a huge importance to the model.

- The number of features to be selected are given from 5 to 20, and the Linear Regression model is used in the RFE and fit to the train set.
- The same metrics as mentioned above are calculated and the main focal point here, is the Adjusted R-Squared.



	No_#	Predictors	Acc_Train	Acc_Test	RMSE_Train	RMSE_Test	Adj_R2_Train	Adj_R2_Test
0	10	[Age, CustTenure, EducationField, Designation,...	0.810047	0.814424	607.903168	612.707240	0.808635	0.813044
1	8	[Age, CustTenure, Designation, MonthlyIncome, ...	0.809938	0.814088	608.077658	613.262177	0.808810	0.812984
2	9	[Age, CustTenure, Designation, MonthlyIncome, ...	0.809939	0.814087	608.077040	613.263967	0.808668	0.812844
3	11	[Age, CustTenure, EducationField, Designation,...	0.810435	0.814236	607.282497	613.017429	0.808884	0.812716
4	7	[Age, CustTenure, Designation, MonthlyIncome, ...	0.809331	0.813630	609.048850	614.017395	0.808341	0.812662

Figure 4.1–Feature selection

- It is observed that the highest Adjusted R-Squared score belongs to the model having the below 10 features:

- ✓ 'Age', 'CustTenure', 'EducationField', 'Designation', 'MonthlyIncome' 'ExistingPolicyTenure', 'SumAssured', 'Occupation_Large_Business', 'Occupation_Salaried', 'Occupation_Small_Business'

The above 10 features will be used for the remaining regression models.

ii. RANDOM FOREST REGRESSOR:

- ➢ It is a supervised machine learning model that uses the Ensemble method, where the mean of all the outputs generated by each tree is taken as the final prediction.
- ➢ Below is the feature importance according to this model.

Variables	Imp_%
SumAssured	76.07
CustTenure	7.52
MonthlyIncome	7.23
Age	3.93
ExistingPolicyTenure	2.33
Designation	2.25
EducationField	0.67

Table 4.2–Random Forest Feature importance

iii. GRADIENT BOOSTING REGRESSOR:

- ➢ It is a supervised machine learning model that operates in a sequential manner. It combines multiple weak models to construct a strong model which produces highly accurate predictions.
- ➢ The Gradient Boosting Regressor is built and then fitted to the train set.

iv. KNEIGHBORS REGRESSOR:

- ➢ It is a supervised machine learning model that operates based on “feature similarity”, where it takes a data point and find other similar data points, and finally takes the mean of all the surrounding data points to make a prediction.
- ➢ Initially, the number of neighbors are taken from 1 to 21 (all odd values are taken in-between this range) and the misclassification scores are plotted.

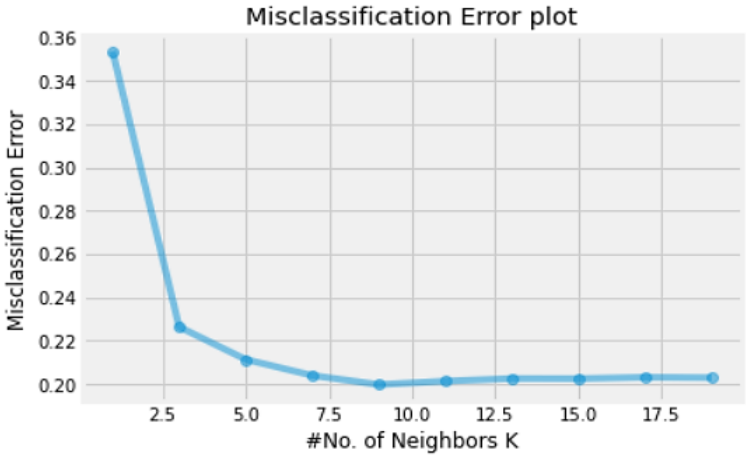


Figure 4.2 –KNeighbors Misclassification error

- The optimal number of neighbors appear to be 9, as it gives the lowest error.
- The KNeighbors Regressor is built and then fitted to the train set.

b) Testing predictive model against the test set using various appropriate performance metrics

The predictions are made for the 4 models which were built, and below is the comparison.

Metrics	Linear_Regression	Random Forest	Gradient Boosting	KNeighbors_Regressor
R2_train	81.11	97.93	87.13	85.08
R2_test	81.11	85.05	85.00	80.61
Adj_R2_train	80.96	97.92	87.11	85.05
Adj_R2_test	80.77	84.97	84.92	80.71
RMSE_train	606.27	200.82	500.24	538.68
RMSE_test	618.17	549.92	550.93	624.60
MAE_train	484.08	150.47	391.37	412.26
MAE_test	489.03	414.32	424.44	471.87

Table 4.3 –Performance metrics of base models

c) Interpretation of the models

Below are the inferences of the **base models** from the above comparison table:

- It appears that overall, the Linear Regression and the Gradient Boosting models have done a relatively good job especially regarding the R-Square Score and Adjusted R-Square score.

- The Random Forest Regressor model has overfit the train set a bit i.e. it is able to perform well on the train set, but not as well on the test set.

- In order to check if we can get better performance, Hyperparameter tuning will be done to build optimized models.

d) Hyperparameter Tuning

The Hyperparameter tuning is done using the RandomizedSearchCV method. It tries random combination of parameter values to get an efficient model. Below are the optimized models:

MODEL	OPTIMIZED PARAMETERS
Linear Regression	None
Random Forest Regressor	oob_score=False, n_jobs=4, n_estimators=200, min_samples_split=20, min_samples_leaf=10, max_depth=15
Gradient Boosting Regressor	tol=0.0001, n_estimators=50, min_samples_split=80, min_samples_leaf=30, max_depth=10
Kneighbors Regressor	n_neighbors=11, metric='minkowski', n_jobs=2

Table 4.4 – Optimized model parameters

5) MODEL VALIDATION:

a) Performance metrics

The Model Comparison is shown below for the optimized models.

Model	Linear_Regression	Random Forest	Gradient Boosting	KNeighbors_Regressor
R2_train	80.95	89.15	90.67	84.46
R2_test	81.24	84.69	85.75	80.88
Adj_R2_train	80.91	89.12	90.65	84.42
Adj_R2_test	81.15	84.61	85.68	80.78
RMSE_train	608.71	459.51	426.08	549.87
RMSE_test	615.99	556.45	536.88	621.92
MAE_train	485.56	351.81	325.64	420.37
MAE_test	487.01	427.97	406.02	469.80

Table 5.1–Performance metrics of optimized models

- The Gradient Boosting model has done better than the other models across all the performance metrics
- There is almost 5-6% difference between the R-squared and Adjusted R-Squared scores of the train and test sets. So, now, the Gradient Boosting model alone will be optimized further.
- RFE is again done, this time using the Gradient Boosting model as an estimator and below are the important features:
    - *'Age', 'CustTenure', 'Designation', 'MonthlyIncome', 'ExistingPolicyTenure', 'SumAssured'*
- The final optimized parameters of the Gradient Boosting model are as follows:

MODEL	OPTIMIZED PARAMETERS
Gradient Boosting Regressor	tol=0.0001, n_estimators=50, min_samples_split=30, min_samples_leaf=30, max_depth=5

Table 5.2–Gradient Boosting final optimized parameters

- Below are the performance metrics:

	R2	ADJ_R2	RMSE	MAE
TRAIN	87.51	87.49	492.90	381.10
TEST	85.41	85.36	543.20	415.00

Table 5.3–Gradient Boosting optimized model metrics

b) Interpretation of the optimized models

1. The Gradient Boosting Regressor model outperforms all the other models, especially in the aspect of R-Square and Adjusted R- Square scores. The test scores are almost close to those of the train sets.

   - The difference between the RMSE and MAE of the train and test sets are also very less compared to the other models.

2. The Linear Regression model has the lowest of R-Square and Adjusted R- Squares scores out of all the models, and high RMSE scores.
3. The Random Forest Regressor and KNeighbors Regressor models appear to have done well in terms of R-Square and Adjusted R- Squares scores, however, the RSME scores are relatively high, meaning there is a high difference between the actual and predicted data.

4. The predictions made by each model are plotted against the actual data.

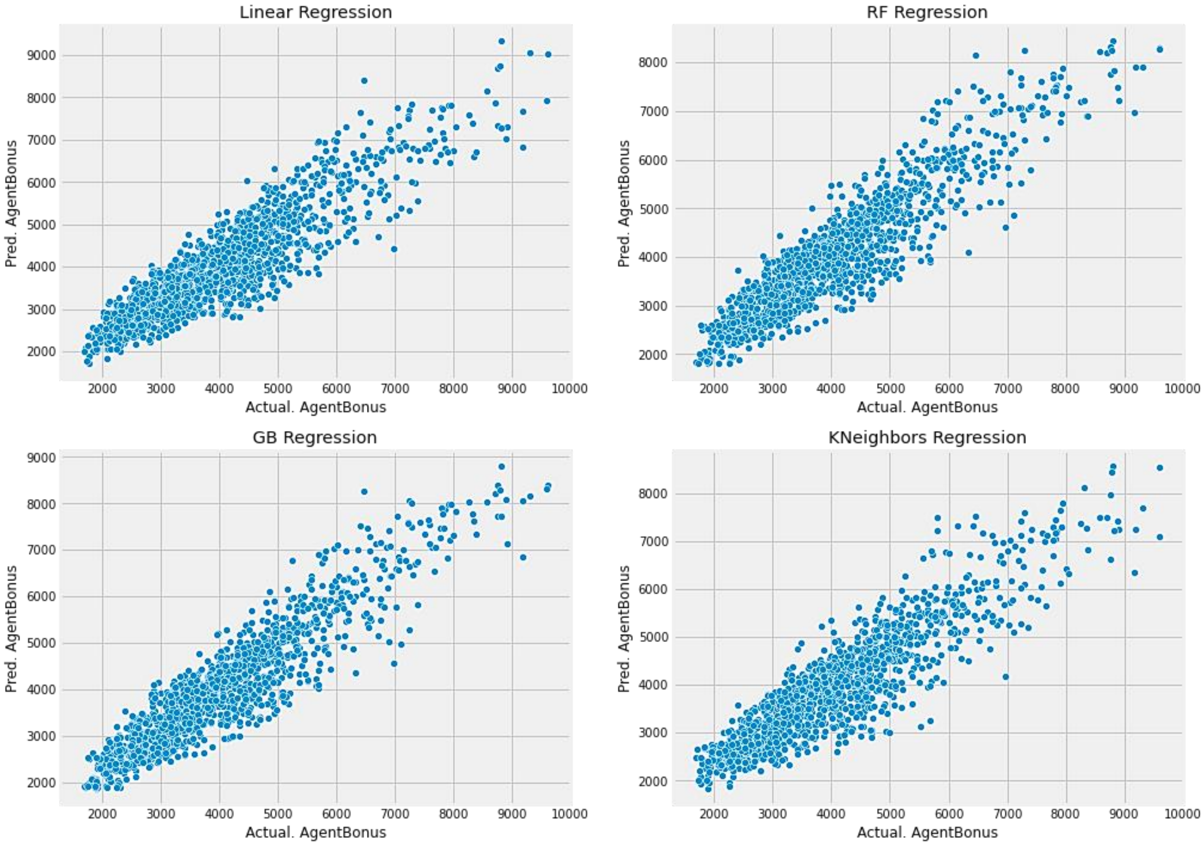


Figure 5.1–Predictions visualization for each model

From the above graph, it is observed that Linear Regressor and Gradient Boosting Regressor do relatively well, as not a majority of the predicted and actual points are too far apart.

Hence, as per, model validation metrics, especially Adjusted R-Square, the Gradient Boosting model is selected. It explains 85.36% of the variability in AgentBonus in the test set.

6) FINAL INTERPRETATIONS/RECOMMENDATIONS:

a) INSIGHTS/INTERPRETATIONS

1. A majority of the customers have Existing Policy Tenure above 10 years.
2. About 75% of the customers chose product type 4.
3. Most of the customer acquisitions have been done through ‘Agent’ channel.
4. A lot of the customers are graduates and under graduates
5. A lot of the customers (57%) belong to ‘West’ Zone.
6. Only 18-19% of the customers gave customer care scores of 4 and 5. Most gave a score of 3.
7. The important features as per Gradient Boosting model are:

Variables	Imp_%
SumAssured	80.96
CustTenure	5.84
MonthlyIncome	4.87
Designation	3.61
Age	3.56
ExistingPolicyTenure	1.16

Table 6.1–Gradient Boosting important features

8. All of the above-mentioned variables have a positive linear relationship with AgentBonus i.e. they cause an increase in AgentBonus. Hence, the company can focus more on these predictors in future when making predictions.

9. The scaled dataset is reverted to its original value range and a table is created with the important features along with the AgentBonus predictions. The below histogram of the predicted Agent Bonus depicts the count of agents having different bonus ranges.

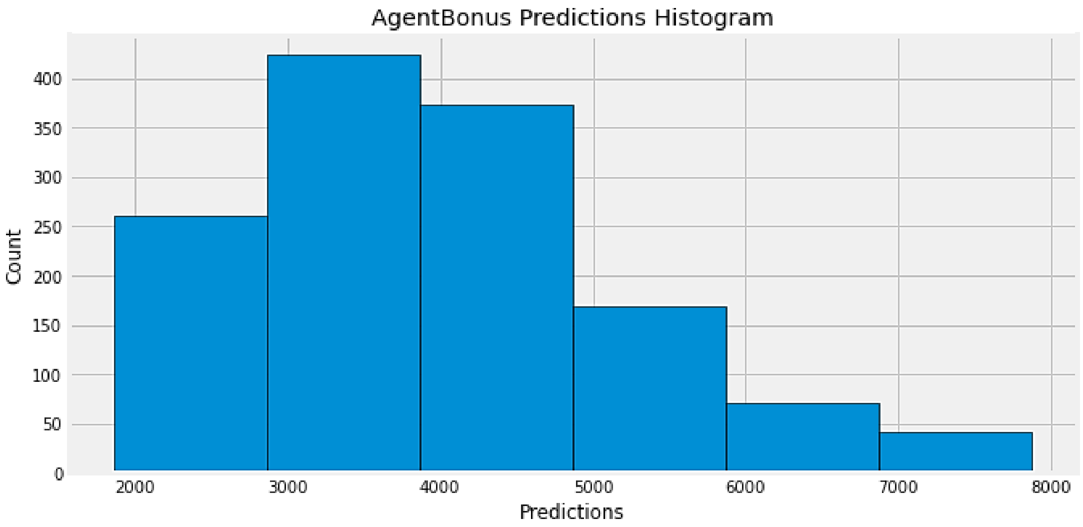


Figure 6.1–AgentBonus Predictions Histogram

10. The agents having a bonus of **4868.50** or above are considered as High-Performing agents. (*A different benchmark can also be taken by the company if required*).

    - According to this benchmark, about 77.95% of the agents are Low-performing and 22.05% are High-Performing (1057 agents and 299 agents respectively in terms of count)

11. The overall average agent bonus of each group i.e. Low and High performing are as follows:

Performance	Pred_AgentBonus
High-Performing	5993.18
Low-Performing	3510.46

Table 6.2–AgentBonus Mean of each class

b) RECOMMENDATIONS

1. The company can give more focus on customers who fall within the age range of 20 to 30 years.

2. Special benefits can be provided to customers who have Existing Policy Tenure above 10 years.

3. Since most of the customers choose a Product type of 4, then more policies catered to this type can be made.

4. In future, a large portion of the customer acquisitions can be done through 'Agent' channel.

5. The company can cater more policies to customers in the West zone.

6. A survey can be sent to customers who gave customer care scores of 3 or less in order to understand what they were dissatisfied with.

7. Certain policies can be specially designed for graduates and under graduates.

8. AgentBonus is highly correlated with the below variables, and are hence, important:

   - ✓ Age
   - ✓ CustTenure
   - ✓ Designation
   - ✓ MonthlyIncome
   - ✓ ExistingPolicyTenure
   - ✓ SumAssured

9. The Gradient Boosting model can be used for making the predictions (with following parameters):

   - *tol=0.0001, n_estimators=50, min_samples_split=30, min_samples_leaf=30, max_depth=5*

10. Hence, Gradient Boosting Regression is a good model as it explains 85.36% of the variability in the Agent Bonus.

11. Based on the predictions and selected benchmark (4868.50), 22.05% of the agents can be provided appropriate engagement activities, and the remaining 77.95% can be upskilled.

    The company can also choose a higher or lower benchmark if needed.