INSURANCE SALES ANALYSIS

BUSSINESS REPORT

PGPDSBA 2022 BATCH
BY AKASH JHA

Abstract: -

The main objective of this capstone project is to develop an Insurance Sales (Agent Bonus Predictor) model that can help Company to predict the Agent Bonus according to Customer No of Policies and Insurance Sum assured.

Keywords: *Missing data, Outliers, Capping Technique, Central Tendency, Multi collinearity, Clustering- KMeans, Feature Selection, Scaling, Model Building, Model interpretation, Model tuning, Recommendations, Business Insights.*

Table of Contents:

List of Figures:

PROJECT OBJECTIVE:

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.

SECTION 1 A: INTRODUCTION, PROBLEM, OBJECTIVES, SCOPE, DATA SOURCES, METHODOLOGY........

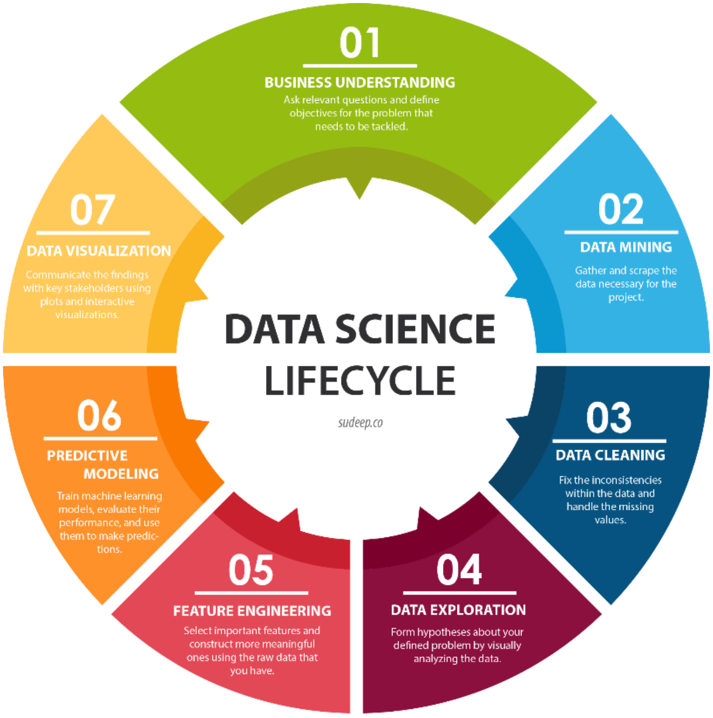
1.1)Introduction:

Big Data Analytics is receiving high growing demand in analysis of Insurance Sales in recent past, this is due to high beneficial outcomes in terms of customer behaviour analysis, sales trends & Churn rate etc.

Big Data Analytics has arrived as means of more precise method of predicting the customers need better, insurance sales trend, customer risk assessment, customers churn rate and causes.

How to use Data science to solve business problems?

Fig 1.1 Data Analytics Life Cycle


DATA SCIENCE LIFECYCLE
sudeep.co
01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

Above fig (1.1) shows life cycle of any raw data transferring to become data model, first we understand business problem, secondly, we collect data from different sources necessary for model building followed by cleaning data to reduce inconsistency in data and handle missing values, Data visualization is done for getting clear about which variables are relevant for model.

Out of large set of Variables, once we are clear with which variables, we can move towards evaluation model and after applying different algorithms on data we conclude business recommendations.

1.2) THE 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.

Company wants to understand what effects of variables are on Agent Bonus, which will motivate them to increase benefits for high performing agents and create programs or give some perks to low performing agents for increasing their efficiency.

Check causes e.g. Does no of complaints directly related to no of calls or not? Does number of policies is directly related to Agent Bonus, Reasons for high churn rate.

1.3) OBJECTIVES OF THE STUDY:

Objective of this study are to:
1) Analyses the Agent Bonus.
2)Identify cause of less or more Bonus.
3) Identify the impact of customer attributes on insurance sales.
4)Build the model which can predict the Agent Bonus
5) Fine tune model.
6) Identify and interpret the best model.
7) Give Business insights for defined problem.

1.4) SCOPE:
The Scope of this study is limited to the data set provided by Insurance Company and using the models mentioned in the objectives.

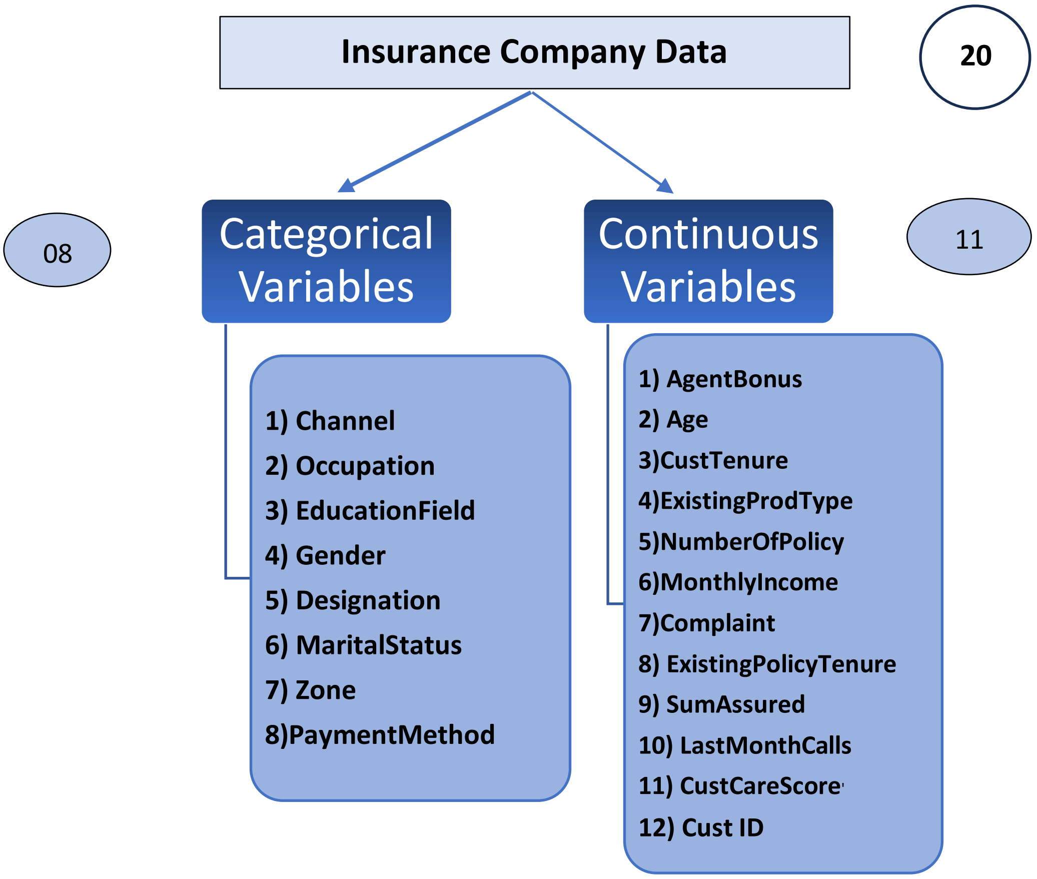
1.5) DATA SOURCE:

The given Dataset contains data given from insurance company with contains information of customers who are holding their policies, which consists of variables like education, marital status etc.

Data that is gathered had 4520 rows/records having 20 attributes / variables. The data contained both Quantitative: Numerical Variables that have are measured on a numeric or quantitative scale and Qualitative variable, also called a categorical variable, are variables that are not numerical.

The Data provided has been collected of customers on policies basis. Data is represented in following diagram for better understanding.

Fig 1.2 Data Report



Insurance Company Data

20

08

Categorical Variables

Continuous Variables

11

1) Channel
2) Occupation
3) EducationField
4) Gender
5) Designation
6) MaritalStatus
7) Zone
8)PaymentMethod

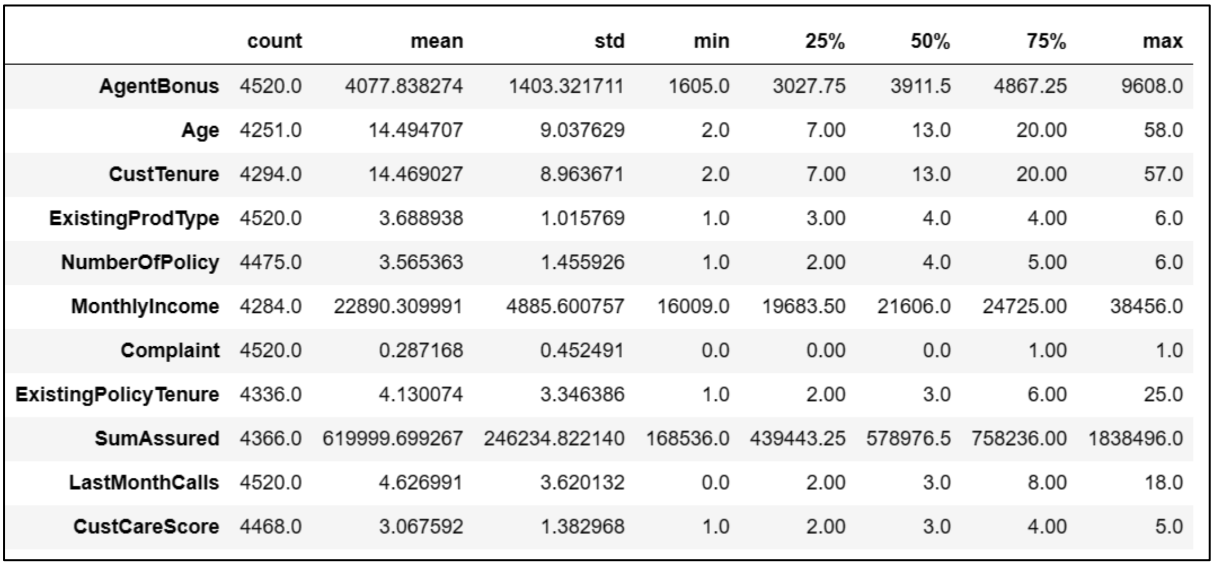
1) AgentBonus
2) Age
3)CustTenure
4)ExistingProdType
5)NumberOfPolicy
6)MonthlyIncome
7)Complaint
8) ExistingPolicyTenure
9) SumAssured
10) LastMonthCalls
11) CustCareScore
12) Cust ID

<u>From Above Figure we can see Data Report</u>

Data Dictionary:

Variable	Discerption
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 policies 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

Descriptive Data:



	count	mean	std	min	25%	50%	75%	max
AgentBonus	4520.0	4077.838274	1403.321711	1605.0	3027.75	3911.5	4867.25	9608.0
Age	4251.0	14.494707	9.037629	2.0	7.00	13.0	20.00	58.0
CustTenure	4294.0	14.469027	8.963671	2.0	7.00	13.0	20.00	57.0
ExistingProdType	4520.0	3.688938	1.015769	1.0	3.00	4.0	4.00	6.0
NumberOfPolicy	4475.0	3.565363	1.455926	1.0	2.00	4.0	5.00	6.0
MonthlyIncome	4284.0	22890.309991	4885.600757	16009.0	19683.50	21606.0	24725.00	38456.0
Complaint	4520.0	0.287168	0.452491	0.0	0.00	0.0	1.00	1.0
ExistingPolicyTenure	4336.0	4.130074	3.346386	1.0	2.00	3.0	6.00	25.0
SumAssured	4366.0	619999.699267	246234.822140	168536.0	439443.25	578976.5	758236.00	1838496.0
LastMonthCalls	4520.0	4.626991	3.620132	0.0	2.00	3.0	8.00	18.0
CustCareScore	4468.0	3.067592	1.382968	1.0	2.00	3.0	4.00	5.0

Fig 1.3 Descriptive Table of Data set

Inference from Descriptive Table:

- Std Deviation of Agent Bonus is very high compared to mean, which conveys agents are falling in different bunch of clusters from high to low.
- Mean and median for Agent Bonus is almost same, there is symmetry in distribution curve with some extreme outliers.
- Median age is 13, which indicates high % of customers are children and dependent.
- There can be possibilities parents of child are holding 2 or more policies.
- Mean of no of policies is 4, customers are having 3 or more policies.
- On an average customer are receiving 4 to 5 calls monthly from agents, it must be interesting to see what effect no of calls have on churn rate.
- Customer care has rating of avg 3 to 4, which is good but there is scope of improvement.

1.6) METHODOLOGY

The approach that was used to resolve the regression problem in the case study was by using machine learning and prediction modelling techniques like Logistic Regression, Naïve Bayes, KNN, CART, Random Forest, Ensembling Models and Python as the software tool.

SECTION 2: EXPLORATORY DATA ANALYSIS INCLUDING DATA PREPARATION, CLEANING, AND IMPUTATION

The below exploratory data analysis was conducted with the data set

2.1. Variable Identification
2.2. Univariate and Bivariate Analysis
2.3. Missing Value Treatment
2.4. Outlier Treatment
2.5. Check for Multi collinearity.
2.6. Data preparation- Feature Scaling, Balancing, Clustering
2.7. Feature Exploration

2.1 VARIABLE IDENTIFICATION:

This data set has 20 Variables/objects/columns and 4520 observations/rows, preliminary study was conducted to understand the variables.

- The variable Cust ID has unique customer id, which can't be summarized, hence all values of this variable will not add much value to generate any inference. Hence, this variable was removed.
- Age variable has highest % of missing value, need to impute missing value with simple imputer method or any other methods can be used.
- Martial Status variable has two unique value unmarried and single with same meaning, need to merge them.
- Designation variable has same messy data problem, must be organised before proceeding.
- Outliers are present in many numeric variables; further study was conducted, and treatment of outliers were performed.

2.2 UNIVARIATE AND BIVARIATE ANALYSIS:

What is Univariate Analysis:
Uni means One – It is method of picking one variable and analysing the data observations pertaining to that variable using descriptive statistics methods like Histograms, Density plots, Box plots to understand data patterns and distribution of the data.

What is Bivariate Analysis:
Bi-Variate analysis as the name indicates involves simultaneous analysis of two variables for the purpose of determining empirical relationship between them. Bi variate analysis helps to test the hypothesis of association i.e. It explores the concept of relationship between two variables in terms of a. Whether there exists an association and strength of this association b. Or whether there are differences between two variables and significance of the difference.

In this data study there are Numerical and Categorical variables. The dependent variable is a Numerical variable, and the independent variables are both Numeric and Categorical.

Bonus Distribution:

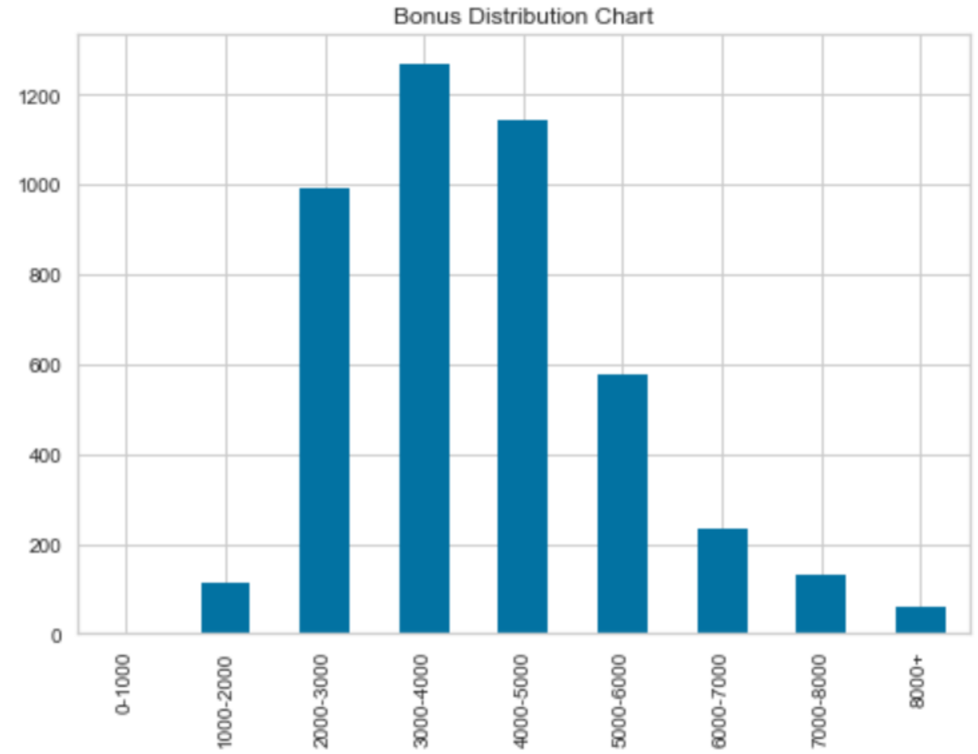


Fig 1.3 Agent Bonus Distribution

Inference:

- Agent with Bonus between 3000 to 4000 are highest in no followed by 4000-5000
- Agent with high bonus is very less in nos.
- We can see Agent with mid-lower bonuses falls in major pieces of cake.

Boxplot of Data:

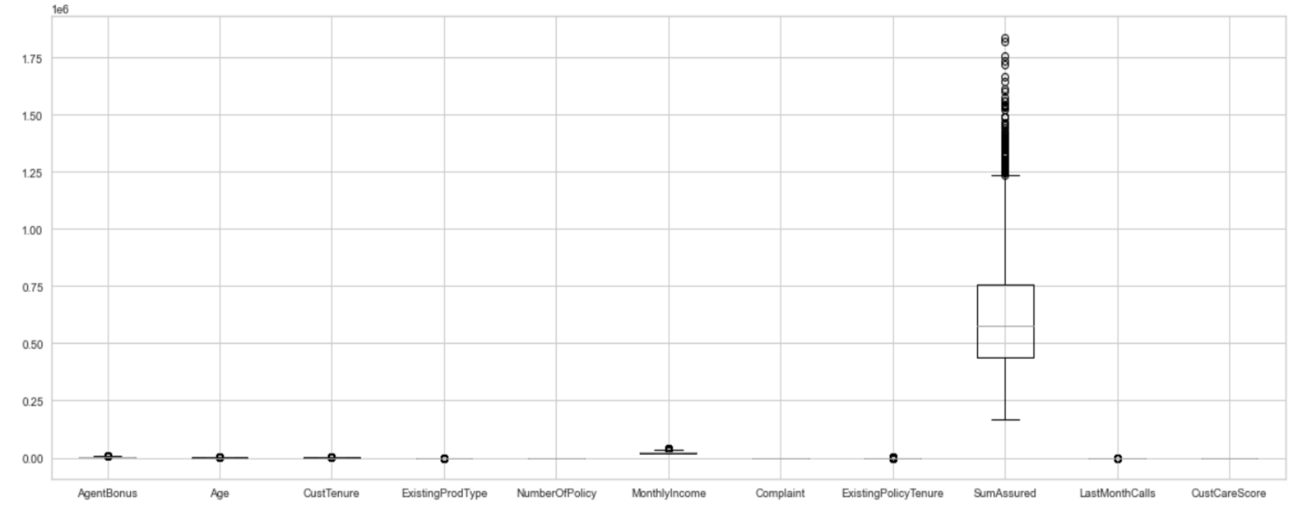
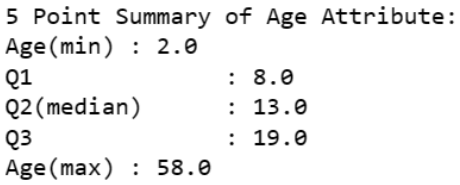


Fig 1.4 Boxplot of Data

We can see max numerical attributes have outliers.

Univariate Analysis with 5-point descriptive Analysis:

Age:



```
5 Point Summary of Age Attribute:
Age(min) : 2.0
Q1              : 8.0
Q2(median)      : 13.0
Q3              : 19.0
Age(max) : 58.0
```

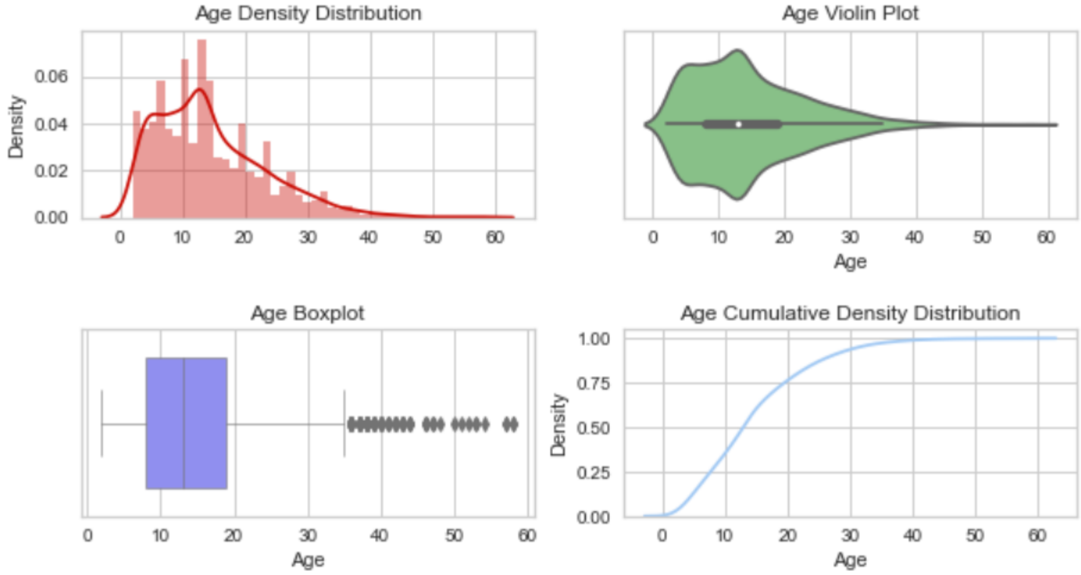


Fig 1.5 Age Distribution

- The Age of the insured approximately follow uniform distribution with Mean of 14 and Median of 13.0, and with lowest age being 2 and highest being 57.

Customer Tenure (Policy Tenure):

```
5 Point Summary of Custtenure Attribute:
Custtenure(min) : 2.0
Q1              : 8.0
Q2(median)      : 13.0
Q3              : 19.0
Custtenure(max) : 57.0
```

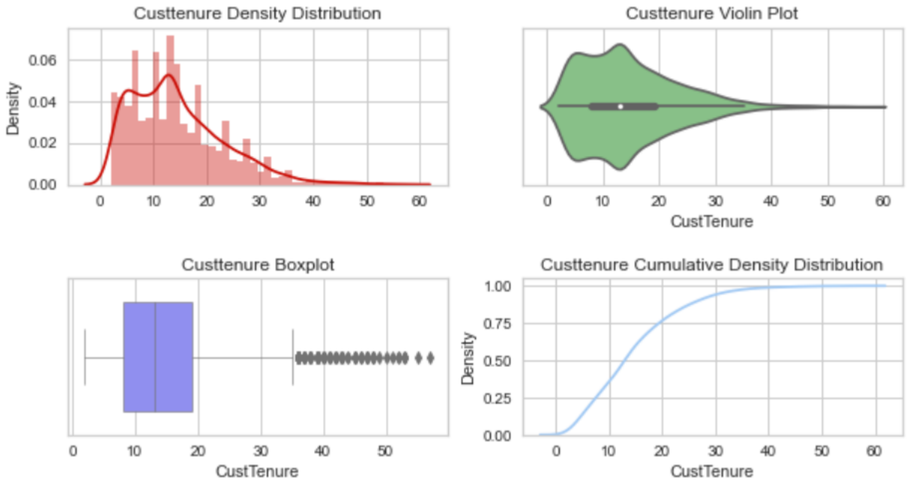


Fig 1.6 Customer Tenure

- Tenure for Customer of insurance approximately follow uniform distribution with Mean of 14 and Median of 13.0, and with lowest age being 2 and highest being 57.
- There are multiple outlier values in the Age distribution in the data.

Monthly Income of Customer:

```
5 Point Summary of Monthlyincome Attribute:
Monthlyincome(min) : 16009.0
Q1                 : 19858.0
Q2(median)         : 21606.0
Q3                 : 24531.75
Monthlyincome(max) : 38456.0
```

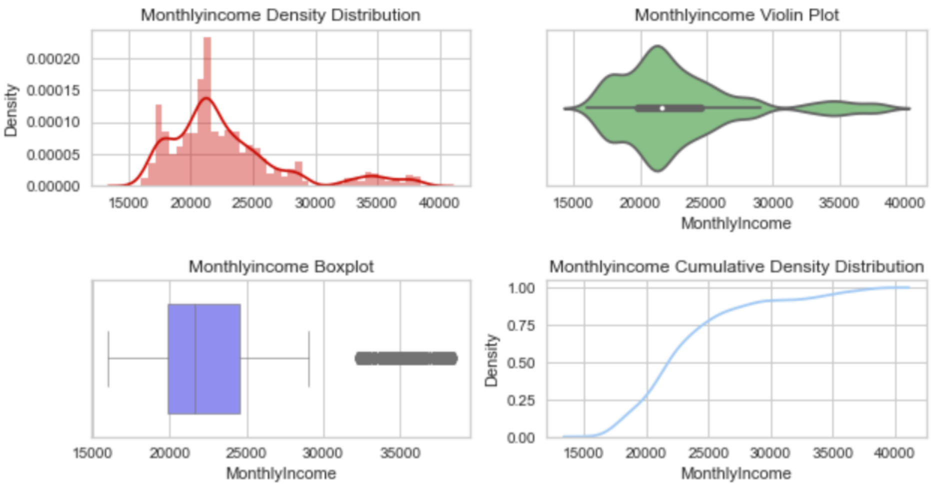
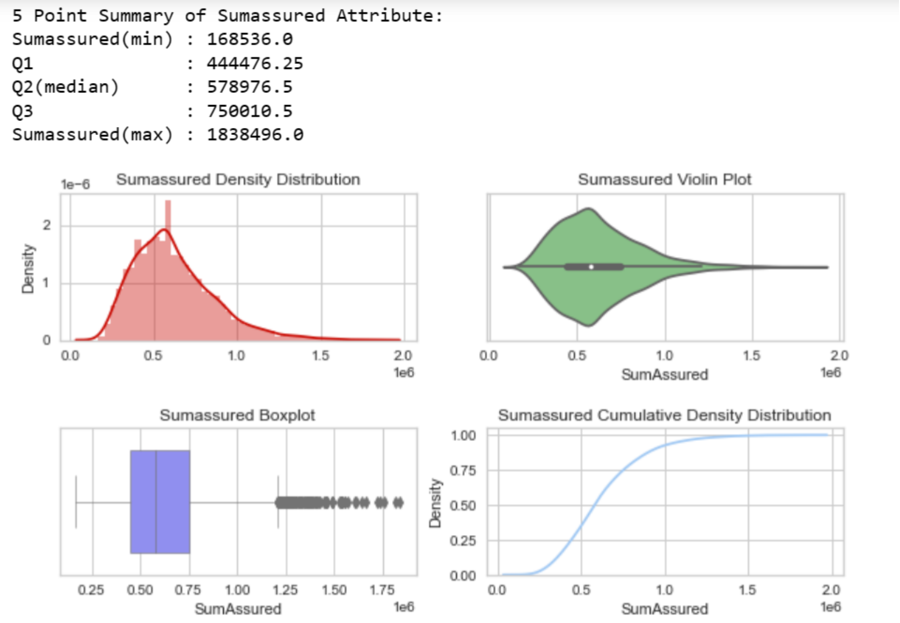


Fig 1.7 Monthly Income

Observations:

- The Monthly Income distribution of the Insured is skewed to the left (median < mean) with a Mean of 22823.22 and Median of 21606.06. The lowest charged amount is 16009.9 and the highest charged amount is 38456.6.
- Out of a total of 4520 data points, there are many outlier values in the distribution of charges, all in the higher side. The highest charges paid is 38456.6.

Sumassured of Policy:

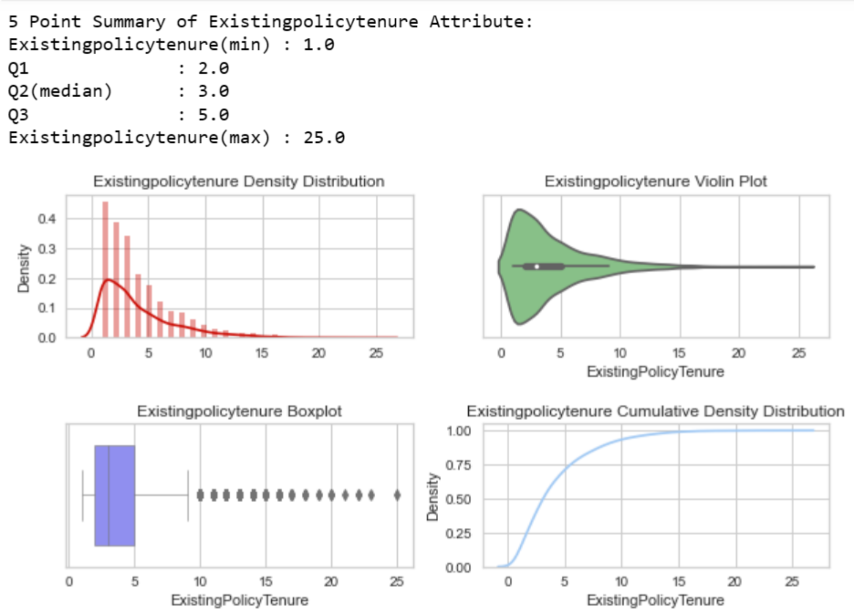


```
5 Point Summary of Sumassured Attribute:
Sumassured(min) : 168536.0
Q1              : 444476.25
Q2(median)      : 578976.5
Q3              : 750010.5
Sumassured(max) : 1838496.0
```

Fig 1.8 Sum Assured Distribution

- Charges doesn’t follow normal distribution; data is left skewed mean> median.
- By Violin plot we can see, major chunk of customer sum assured are between 5 lacs to 7 lacs, then we have outliers with higher sum assured with high tenure policy.
- Max sum assured customer is 18 lacs with Agent bonus 9192 and 55 total tenure.

Existing policy Tenure: (Fig 1.9)



```
5 Point Summary of Existingpolicytenure Attribute:
Existingpolicytenure(min) : 1.0
Q1              : 2.0
Q2(median)      : 3.0
Q3              : 5.0
Existingpolicytenure(max) : 25.0
```

Existingpolicytenure Density Distribution
Density
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
Existingpolicytenure Violin Plot
0
5
10
15
20
25
ExistingPolicyTenure
Existingpolicytenure Boxplot
0
5
10
15
20
25
ExistingPolicyTenure
Existingpolicytenure Cumulative Density Distribution
Density
0.00
0.25
0.50
0.75
1.00
0
5
10
15
20
25
ExistingPolicyTenure

- Current policy tenure of customer mostly in this data lies between 2 to 5 years, with max tenure 25 years.
- Existing Policy Tenure have outlier with extreme values.

Categorical Variables:

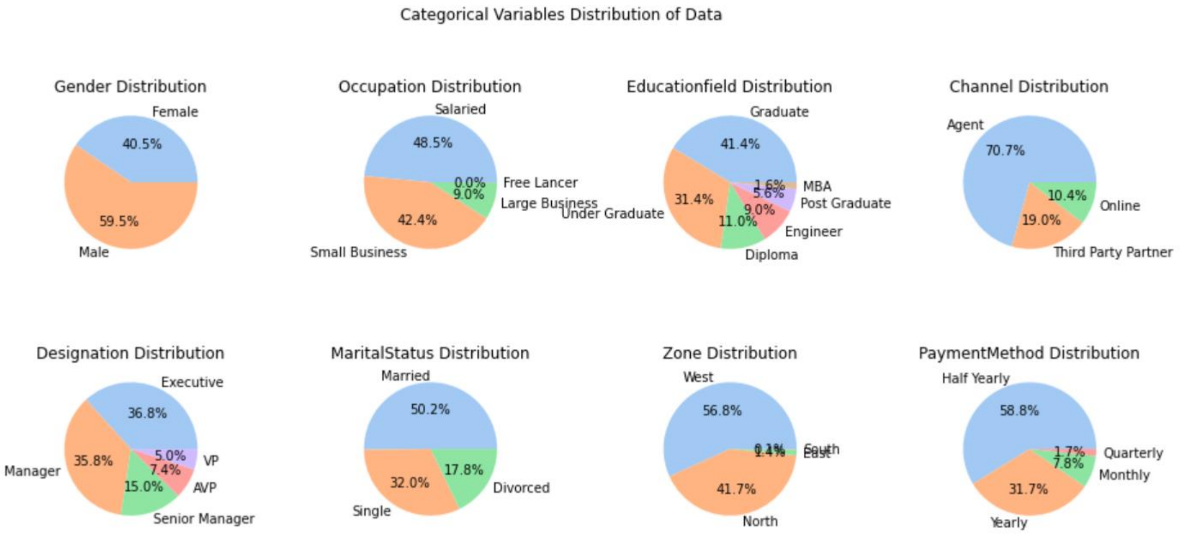


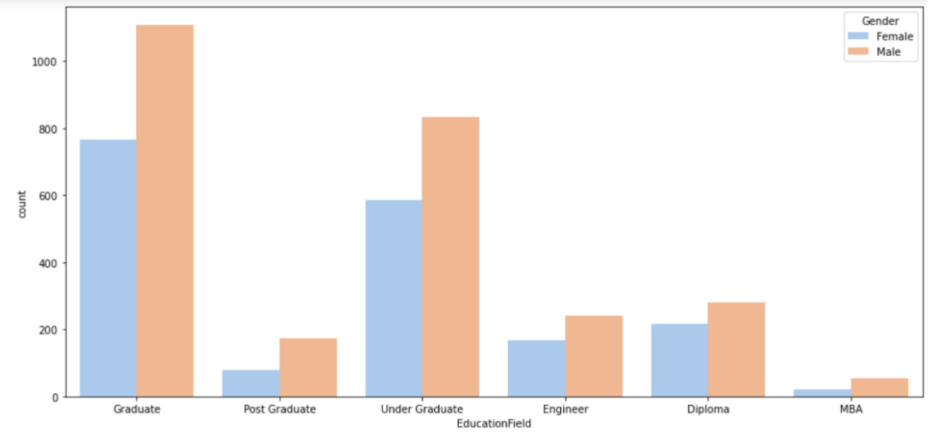
Fig 1.10. Categorical Variables Counts

Observation:

- There are more males' customers compared to female.
- Salaried Customers are highest followed by small business, large business are very less in counts, i.e. reason can be sum assured category amount is between 5 to 15 lacs, Large business must be opting for high sum assured policy.
- Max Insured customers are graduate and Undergraduate, more people with lesser education opt for life insurance policy.
- Agent Channel is bringing maximum customer, normally in insurance terms these agents are known as street agents who have high connection and network in local market.
- Max count of customers is married compared to single and divorced.
- We can see data from south and east zone is not much collected compared to West and North.
- Maximum customers are paying half yearly premium followed by yearly, Quarterly paying method is lowest.

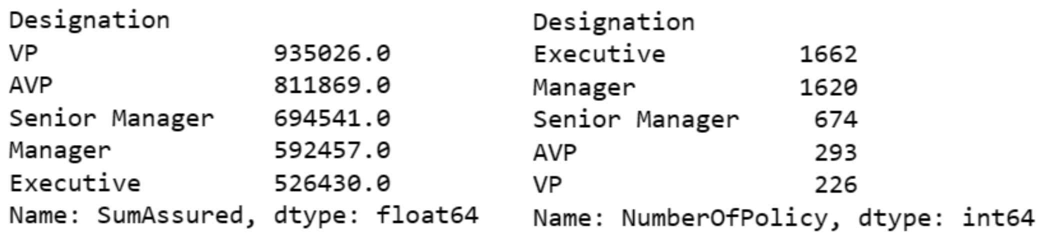
Bivariate Analysis for Numerical Variables:

Agent Bonus vs Educational (Fig 1.11)



Inference:

- Graduate and undergraduate males' customer are generating more bonuses compare to other education.
- Data seems to be skewed, as MBA and post graduate data is very less.



```
Designation
VP                  935026.0
AVP                 811869.0
Senior Manager      694541.0
Manager             592457.0
Executive           526430.0
Name: SumAssured, dtype: float64
```

```
Designation
Executive         1662
Manager           1620
Senior Manager     674
AVP                293
VP                 226
Name: NumberOfPolicy, dtype: int64
```

Fig 1.12 Designation vs Sum assured vs No of Policy

Inference:

- VP category has highest avg sum assured but low no of policy, reason can be high amt policies category.
- Executive category has highest no of policies but lowest sum assured , max executive category falls between 5 lacs to 7 lacs assured policy.
- Manager profile customers can be good grab for upgrading them for high insurance term.

Age Vs Agent Bonus:

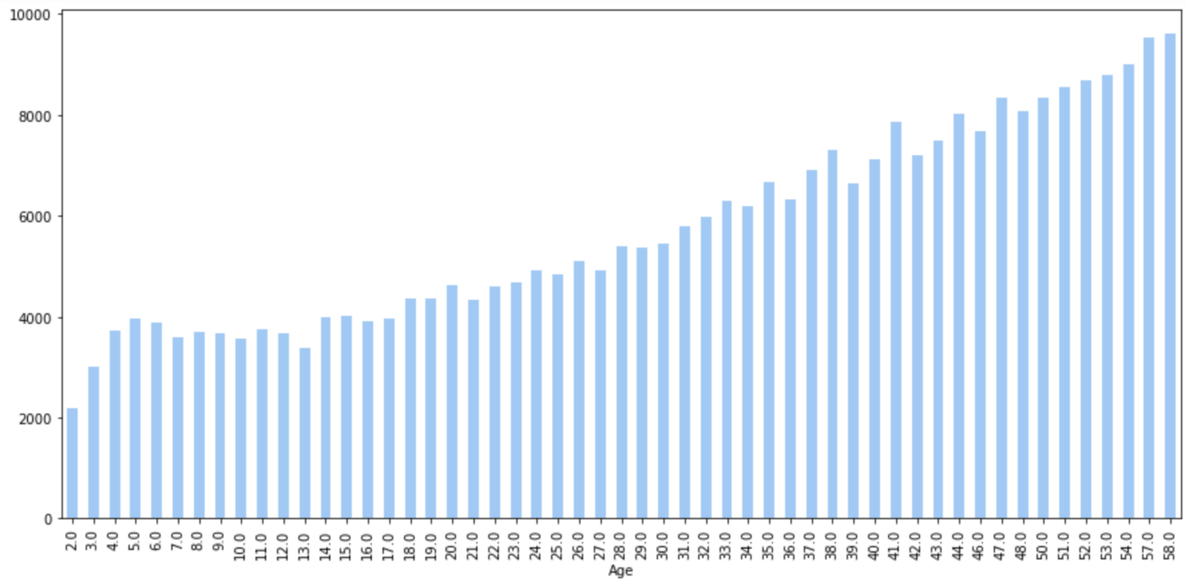


Fig 1.13 Age vs Agent Bonus

- As Age grows Agent Bonus goes up, shows positive trend.

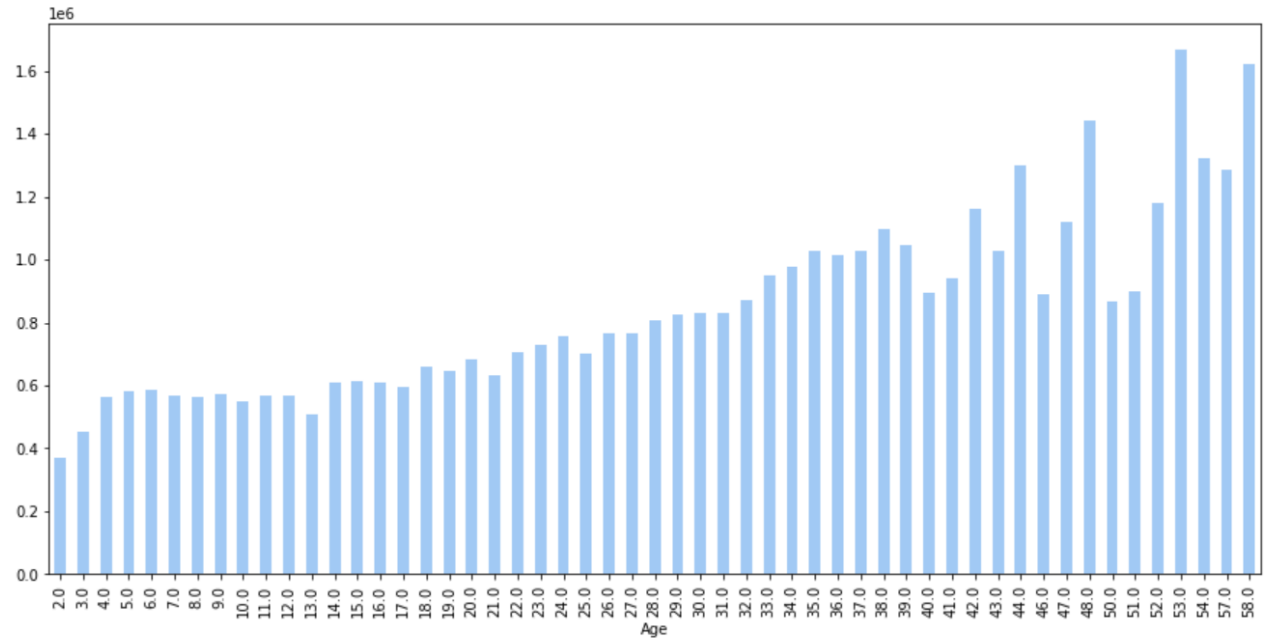


Fig 1.14 Age vs Sum assured

- As age grows sumassured goes up, but still we have some exception in upper cap of age where sumassured are less .

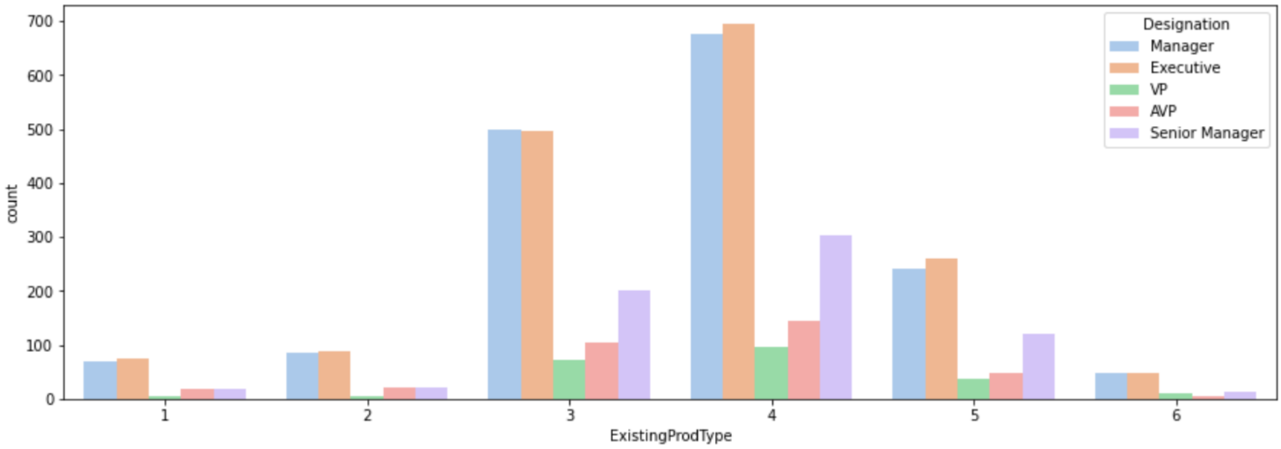


Fig 1.15 Product type vs Agent Bonus

- Product 2 and product 6 have highest bonus for agent.
- Product 3 contribution low in Agent Bonus.
- Agent Bonus are directly related to sum assured.

Pairplot

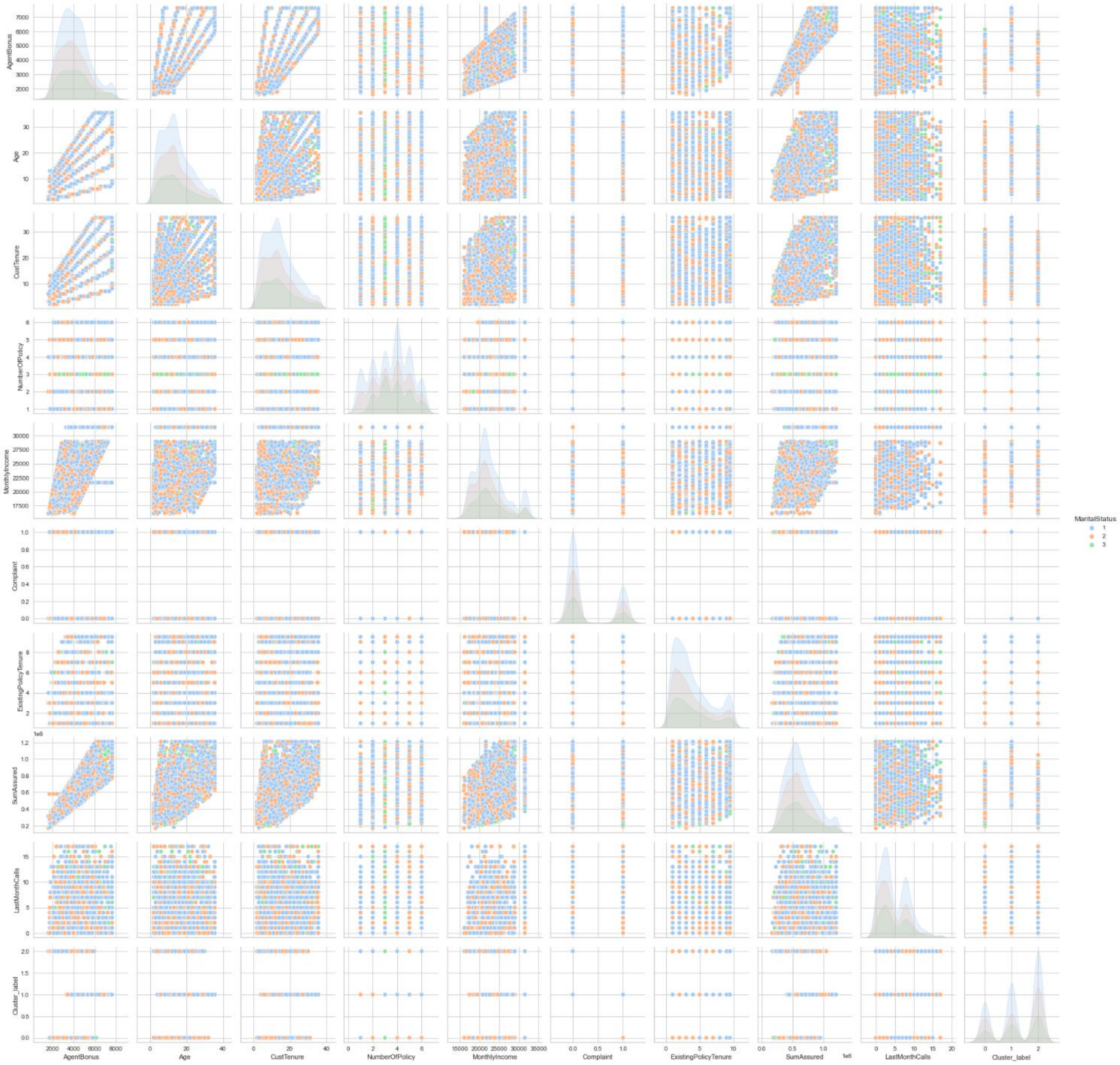


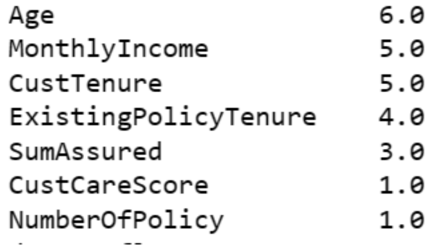
Fig 1.16 Pair plot of Data

Inference:

- We can see as monthly income is high agent bonus is also high.
- High Sum assured assure high Agent Bonus.
- Policy type 4 have high bonus.
- Last Month calls are scattered relation with Agent Bonus, can pull out more inference.
- We can see married customer has more complaints compared to non-married.
- Type 3 policy is more opted by Divorced category compare to any other , need to dig more details of type 3 product and agents can target if any special benefits are there.
- Married Customer has high tenure compared to non-married or divorced.

2.3 MISSING VALUE TREATMENT

There were few missing values identified in the data set, Age- 6%, Monthly income-5%, CustTenure-5%, due to missing values less we will impute missing values by Simple imputer method.



```
Age                      6.0
MonthlyIncome            5.0
CustTenure               5.0
ExistingPolicyTenure     4.0
SumAssured               3.0
CustCareScore            1.0
NumberOfPolicy           1.0
```

2.4 OUTLIER TREATMENT:

Outlier Identification was conducted for the numeric variables in the dataset and Boxplot review was conducted to identify outliers with the subset of variables identified in the previous step. Box plot shows there are outliers in most of the numeric variables.

Since the logistic regression models are sensitive to outliers, hence the outliers were treated by capping technique using capping upper and lower quartile.

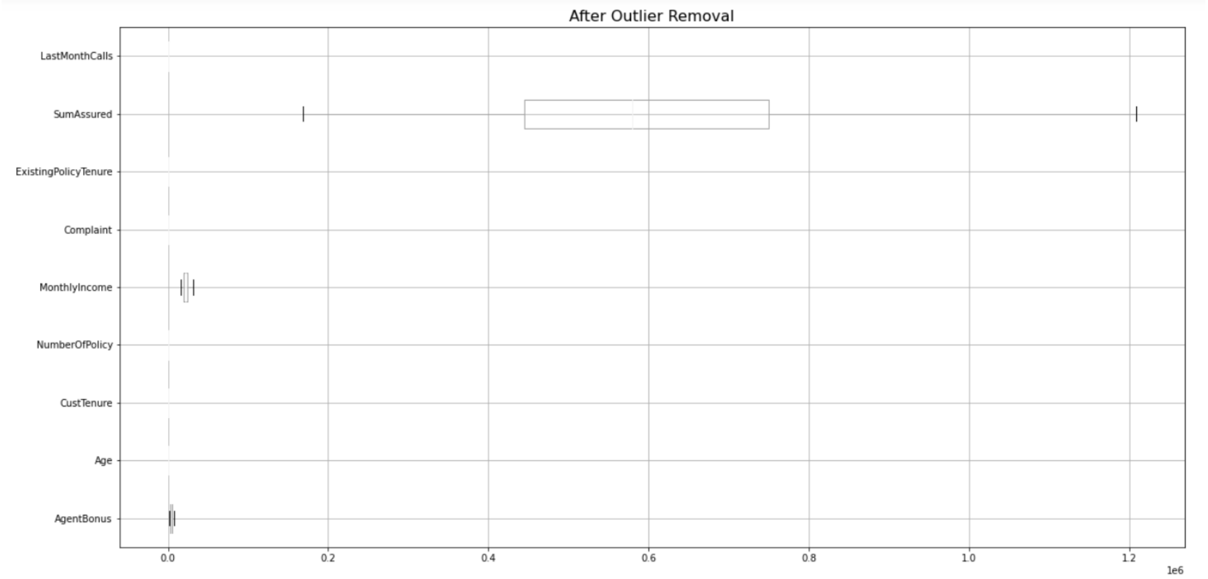


Fig 1.17 Boxplot After Outliers Treatment

2.5 CHECK FOR MULTICOLLINEARITY

Definition of Multicollinearity: -

Multicollinearity occurs when the independent variables of a regression model are correlated and if the degree of collinearity between independent variables is high, it becomes difficult to estimate the relationship between each independent variable and the dependent variable and the overall precision of the estimated coefficients.

Disadvantages of Multicollinearity: -

For Regression Multicollinearity is a problem because a. If two independent variables contain essentially same information to a large extent, one may become insignificant (or) may become significant b. Unstable estimates as it tends to increase the variances of regression coefficients.

Heatmap:

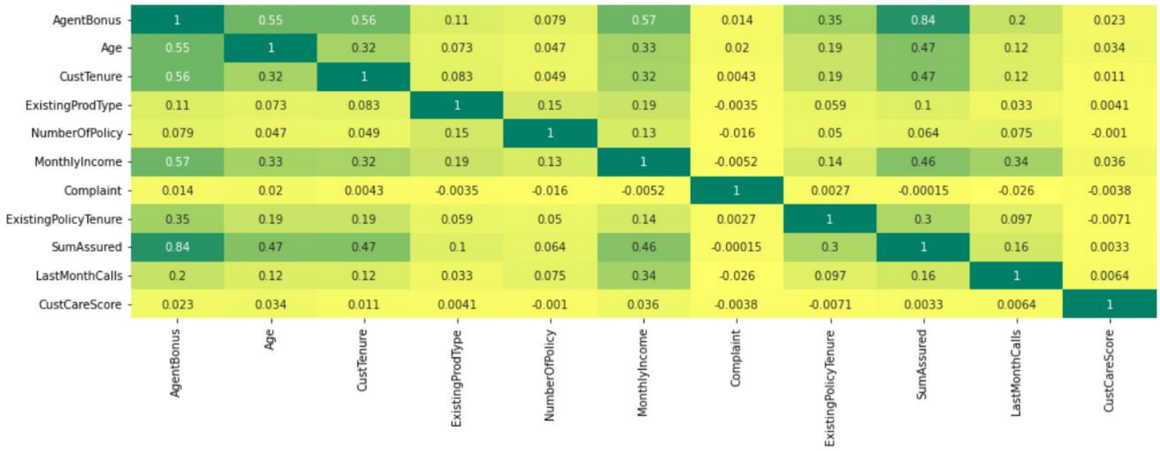


Fig 1.18 Heatmap for checking collinearity in Data

Observations:

- Highly correlated columns with target variables Agent Bonus- Sum assured, Monthly income, Cust Tenure, Age, Existing Tenure.
- Customer Tenure is highly Correlated with Sum assured, high tenure means high insurance value.
- Sum assured have high correlation with age.
- Monthly income has direct relation with last month calls, Agents try to target high income group for more policy sales.
- We can remove multi collinearity in Data by Variance Inflation Factor Elimination method, by removing highly correlated variables.

For this data set clustering technique Kmeans was performed to reduce the dimension of correlated independent variables which is covered in next section- Data Preparation.

2.6 DATA PREPARATION – FEATURE SCALING, BALANCING AND CLUSTERING:

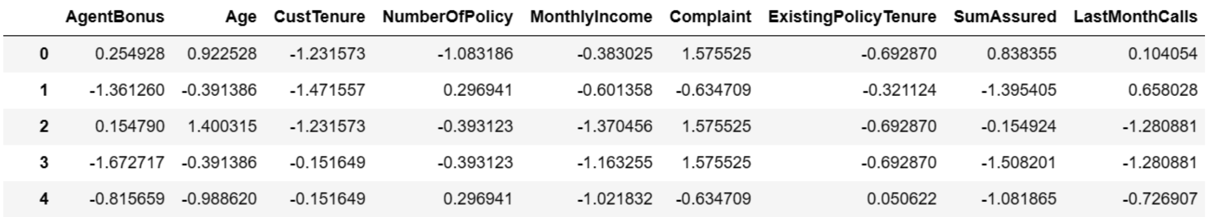
Feature Scaling: -

Why Feature Scaling Needed?

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

Most Common methods for feature scaling are:
1) Standard Scaler method.
2) Z score Method.

We have used Zscore method for Scaling Data.



	AgentBonus	Age	CustTenure	NumberOfPolicy	MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured	LastMonthCalls
0	0.254928	0.922528	-1.231573	-1.083186	-0.383025	1.575525	-0.692870	0.838355	0.104054
1	-1.361260	-0.391386	-1.471557	0.296941	-0.601358	-0.634709	-0.321124	-1.395405	0.658028
2	0.154790	1.400315	-1.231573	-0.393123	-1.370456	1.575525	-0.692870	-0.154924	-1.280881
3	-1.672717	-0.391386	-0.151649	-0.393123	-1.163255	1.575525	-0.692870	-1.508201	-1.280881
4	-0.815659	-0.988620	-0.151649	0.296941	-1.021832	-0.634709	0.050622	-1.081865	-0.726907

Scaled Numeric Data

Business insights from EDA

Clustering: By Kmeans Method.

Clustering is the task of dividing the population (or) data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

For e.g. If Grocery store owner want to categorize their customer into yearly purchase capability, he/she can segregate customer in 10 groups from high to loyal by this yearly bills amt and can know which items are high selling and low selling.

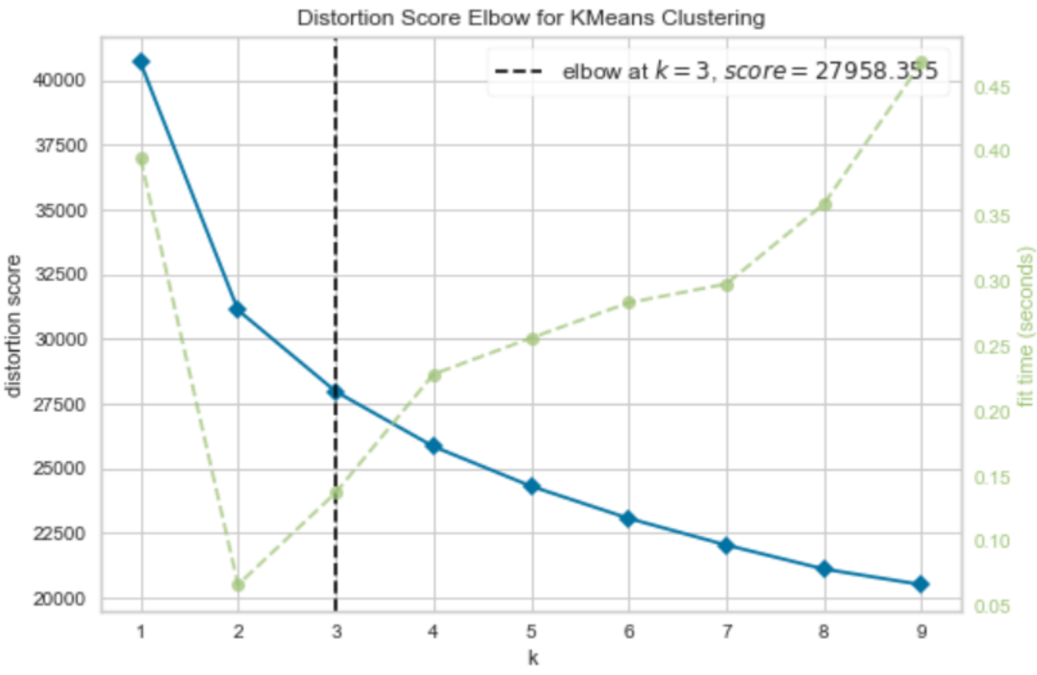
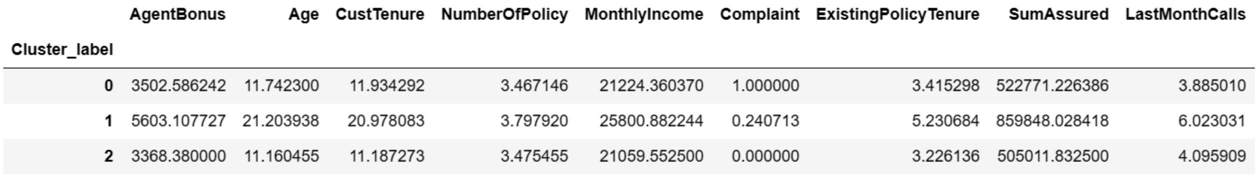


Fig 1.19 Elbow Graph for deciding No of clusters, here n_clusters=3



	AgentBonus	Age	CustTenure	NumberOfPolicy	MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured	LastMonthCalls
Cluster_label									
0	3502.586242	11.742300	11.934292	3.467146	21224.360370	1.000000	3.415298	522771.226386	3.885010
1	5603.107727	21.203938	20.978083	3.797920	25800.882244	0.240713	5.230684	859848.028418	6.023031
2	3368.380000	11.160455	11.187273	3.475455	21059.552500	0.000000	3.226136	505011.832500	4.095909

2.7 VARIABLE TRANSFORMATION

When a categorical variable has more than two categories, it can be represented by a set of dummy variables, with one variable for each category for the algorithm to function. We identified Character variables which had more than 2 categories, we have transformed categorical variables by mapping.

Data Balancing:

What is Balanced and Imbalanced Datasets?

Balanced dataset: Let us take simple example in a dataset we have positive and negative values. If the positive values and equal to negative values, then we can say the dataset is balanced.

Imbalanced dataset: In the same example if there is very high difference between positive and negative values then the data set is imbalance data set.

We can see in our data target variable Agent Bonus seems balanced and there is some discrepancy in category variable like zone with imbalanced in data.

Insights:

- **Cluster 1 shows Agent Bonus > Age > No of Policy > Complaint < Existing Policy Tenure > Sum Assured > Last month Calls compared to other two clusters.**
- **Cluster 0 has high complaint ratio and least calls, agents need to focus more on calling these customers for dissatisfaction.**
- **Cluster 2 has maximum amt of customer and most of them are in younger age group customer with low sum assured and low monthly income.**
- **Agents must focus more on this Cluster 2 customer to upgrade them in Cluster 1, with some lucrative offers.**
- **No of policy doesn't not play much impact on customer satisfaction and dissatisfaction.**
- **Making more calls regarding new products will help customers to get more chooses to purchase.**

Section 2:

ALL MODEL DEVELOPMENT INCLUDING TESTING OF ASSUMPTIONS AND PERFORMANCE EVALUATION METRICS.

The objective of the model development is to build appropriate prediction model on the train data and apply the predicted train model on the test data to find its robustness in maintaining the correctness of the prediction.

The predictive models that were built for this case study are using Linear Regression, Lasso Regression, Decision Tree Regressor of predictive model techniques. Ensemble methods like Random Forest Regressor were also used to create models, post the model development interpretation of the model outputs, necessary modifications like tuning the parameters were done to find the optimal model outputs.

Overfitting it impact & Sample Split Purpose:

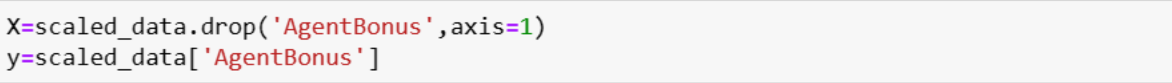
In statistical machine learning techniques, there is problem of data overfitting i.e., Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This problem occurs when the model is too complex. The problem of overfitting can be avoided by spitting the data in to Training and Test data.

This report covers the model build and evaluation that were performed with Train and Test data.
This report is covering the model build and evaluation in below sequence: -

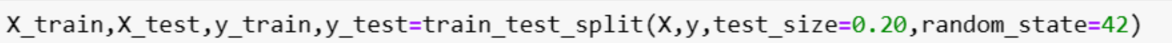
2.3) Applying Linear Regression, Model Evaluation & Result Interpretation.
2.4) Applying Lasso Regression, Model Evaluation & Result Interpretation.
2.5) Applying Polynomial Regression, Model Evaluation & Result Interpretation.
2.6) Applying Decision Tree Regression, Model Tuning, Model Evaluation & Result Interpretation.
2.7) Applying Random Forest Regression, Model Tuning, Model Evaluation & Result Interpretation.
2.8) Applying XG Booster Regression, Model Tuning, Model Evaluation & Result Interpretation.

2.2) Splitting Data: (80/20 Ratio)

- **Train_Test_Spilt**: It's function from sklearn. model_selection module. It is used to randomly split data in training and testing data two subset. It ensures randomization in split data.
- **X and y**: These variables represent input features(X) and target variable(y). They are dataset you are obtained after drooping Target variable and separating it from input feature.
- **Test_size**: In this case test size is 20%, which means training to testing ratio is 80/20.
- **X_train, X_test, y_train, y_test**: This Variables stores dataset after splitting, X_train, y_train represent training data, X_test , y_test represent testing data.
- Separation of X and y done when preparing data for machine learning task, specifically for learning Algorithm.
- Independent (X) Variables will be used for making predictions and have predictive power.
- Target variable (y) is variable which we want to predict, we had dropped cluster column from EDA part and prepared data for modelling.
- Before Splitting Data, we need to scale Data (We are scaling only X variables and keeping target variable as it is).
- For scaling standard scaler method is used.
- Spilt Data into 80/20, train/test.



```
X=scaled_data.drop('AgentBonus',axis=1)
y=scaled_data['AgentBonus']
```



```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.20,random_state=42)
```

- **Random State**: When splitting a dataset into training and testing sets, the random state parameter is used to control the random shuffling of the data before the split. you ensure that every time you run the code with the same value, you'll get the same split, allowing you to reproduce the exact same training and testing datasets.

2.3) Applying Linear Regression, Model Evaluation & Result Interpretation.

Linear regression is a basic and widely used statistical technique for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the dependent variable (target) and the independent variables (features). The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted values and the actual values of the target variable.

Single Variable Linear Regression:

Y= b0 + b1*x

Multi Variable Linear Regression:

Y=b0 + b1*x1 + b2*x2+b3*x3+…+bn*xn

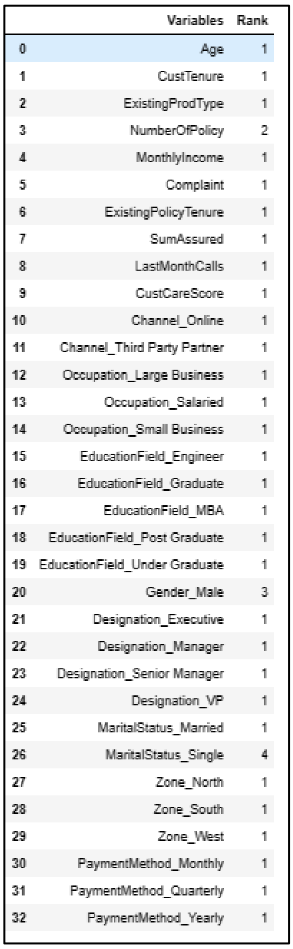
Where:

- Y is target variable (Prediction continuous variable).
- B0 is y-intercept (the value of y when all input features are zero).
- B1, B2,…,Bn are slopes (coefficients) of respective input features.

- X1, X2,..,Xn are the input features values.

For making model fit more efficient we are using First recursive elimination method to eliminate unwanted features.

- **Recursive Feature Elimination (RFE)** is a feature selection technique used in machine learning to identify the most important features for a given predictive model. It is an iterative method that works by recursively fitting the model and removing the least significant feature(s) at each step until a desired number of features is reached.



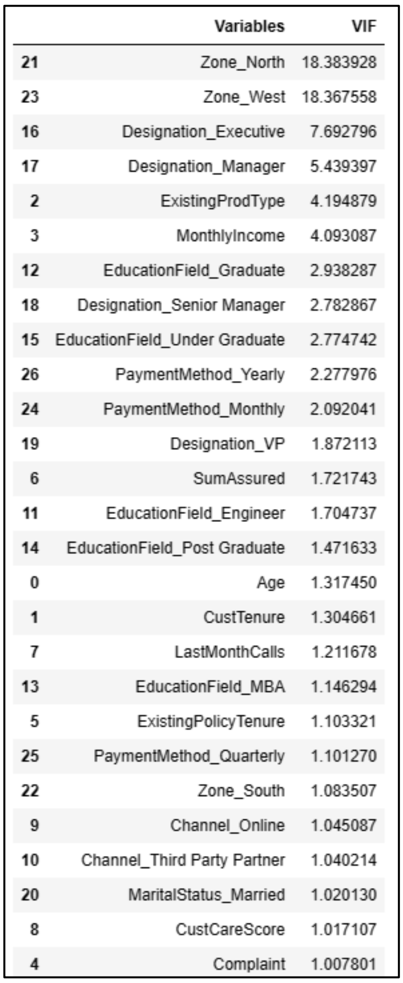
	Variables	Rank
0	Age	1
1	CustTenure	1
2	ExistingProdType	1
3	NumberOfPolicy	2
4	MonthlyIncome	1
5	Complaint	1
6	ExistingPolicyTenure	1
7	SumAssured	1
8	LastMonthCalls	1
9	CustCareScore	1
10	Channel_Online	1
11	Channel_Third Party Partner	1
12	Occupation_Large Business	1
13	Occupation_Salaried	1
14	Occupation_Small Business	1
15	EducationField_Engineer	1
16	EducationField_Graduate	1
17	EducationField_MBA	1
18	EducationField_Post Graduate	1
19	EducationField_Under Graduate	1
20	Gender_Male	3
21	Designation_Executive	1
22	Designation_Manager	1
23	Designation_Senior Manager	1
24	Designation_VP	1
25	MaritalStatus_Married	1
26	MaritalStatus_Single	4
27	Zone_North	1
28	Zone_South	1
29	Zone_West	1
30	PaymentMethod_Monthly	1
31	PaymentMethod_Quarterly	1
32	PaymentMethod_Yearly	1

Fig 2.1) RFE Features Ranking for Eliminating Unwanted features.

From RFE we can eliminate No of Policy, Male Gender & Single Martial Status feature before model fitting.

Variance Inflation Factor for Feature Elimination:

Eliminating more input features by VIF method, Variables whose VIF value comes more then 5, need to be eliminated from model, to get more efficient model with less multicollinearity between two or more features.



	Variables	VIF
21	Zone_North	18.383928
23	Zone_West	18.367558
16	Designation_Executive	7.692796
17	Designation_Manager	5.439397
2	ExistingProdType	4.194879
3	MonthlyIncome	4.093087
12	EducationField_Graduate	2.938287
18	Designation_Senior Manager	2.782867
15	EducationField_Under Graduate	2.774742
26	PaymentMethod_Yearly	2.277976
24	PaymentMethod_Monthly	2.092041
19	Designation_VP	1.872113
6	SumAssured	1.721743
11	EducationField_Engineer	1.704737
14	EducationField_Post Graduate	1.471633
0	Age	1.317450
1	CustTenure	1.304661
7	LastMonthCalls	1.211678
13	EducationField_MBA	1.146294
5	ExistingPolicyTenure	1.103321
25	PaymentMethod_Quarterly	1.101270
22	Zone_South	1.083507
9	Channel_Online	1.045087
10	Channel_Third Party Partner	1.040214
20	MaritalStatus_Married	1.020130
8	CustCareScore	1.017107
4	Complaint	1.007801

Fig 2.2) VIF method for Feature Elimination.

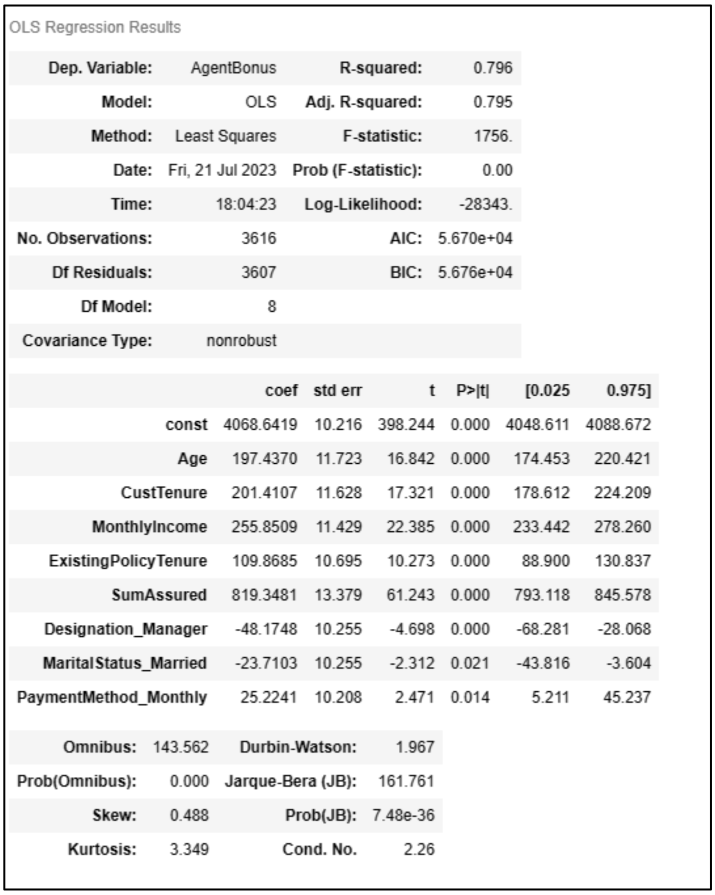
Model Building:

We are using Stats model for creating best possible model and eliminating unwanted features with iterative process.

The most common method to estimate these coefficients is the Ordinary Least Squares (OLS) method, which minimizes the sum of squared differences between the actual and predicted values.

Interpreting the summary of an Ordinary Least Squares (OLS) regression model is essential to understand the results of the linear regression analysis.

OLS Model Interpretation:



```
OLS Regression Results
   Dep. Variable:        AgentBonus          R-squared:      0.796
           Model:               OLS     Adj. R-squared:      0.795
          Method:     Least Squares        F-statistic:      1756.
            Date:  Fri, 21 Jul 2023  Prob (F-statistic):      0.00
            Time:          18:04:23     Log-Likelihood:    -28343.
No. Observations:              3616                AIC:  5.670e+04
    Df Residuals:              3607                BIC:  5.676e+04
        Df Model:                 8
 Covariance Type:         nonrobust

                             coef  std err        t  P>|t|    [0.025    0.975]
               const    4068.6419   10.216  398.244  0.000  4048.611  4088.672
                 Age     197.4370   11.723   16.842  0.000   174.453   220.421
          CustTenure     201.4107   11.628   17.321  0.000   178.612   224.209
       MonthlyIncome     255.8509   11.429   22.385  0.000   233.442   278.260
 ExistingPolicyTenure    109.8685   10.695   10.273  0.000    88.900   130.837
          SumAssured     819.3481   13.379   61.243  0.000   793.118   845.578
  Designation_Manager    -48.1748   10.255   -4.698  0.000   -68.281   -28.068
 MaritalStatus_Married   -23.7103   10.255   -2.312  0.021   -43.816    -3.604
 PaymentMethod_Monthly    25.2241   10.208    2.471  0.014     5.211    45.237

      Omnibus:  143.562     Durbin-Watson:     1.967
Prob(Omnibus):    0.000  Jarque-Bera (JB):   161.761
         Skew:    0.488          Prob(JB):  7.48e-36
     Kurtosis:    3.349          Cond. No.      2.26
```

Fig 2.3 OLS model

Let's try to go through some key elements in OLS summary and interpret them.

- **Dep. Variable:** This indicates dependent variable (target) of linear regression analysis (Agent Bonus).
- **Model:** Provides type of model (Ols, Ordinary least Square).
- **Method:** Estimation method used (i.e., Least Squares)
- **No of Observation:** No of data point of Training Dataset. (3616)
- **DF Residuals:** Degrees of freedom of the residuals (3616- 8).
- **DF_Model:** Degrees of freedom of Model (No of Parameter -1)
- **R_Squared:** The coefficient of variance measures the proportion of variance in the dependent variables that is predictable from independent variables. It ranges from 0 to1. 1 indicates perfect fit (our model has around 0.80)
- **Adjusted_R Squared:** The adjusted R-squared adjusts the R-squared value for the number of predictors in the model, penalizing models with too many predictors.
- **F-statistic:** The F-statistic tests the overall significance of the model. It compares the fit of the current model with a model that has no predictors (null model). Higher F-statistic values indicate a better fit of the model. Our F statistics have high value (1756)

- **Prob (F-statistic):** The p-value associated with the F-statistic. A small p-value (typically < 0.05) indicates that the model's overall fit is statistically significant. We have created model with features having p value lower than 0.05.
- **coef:** This table provides information about the estimated coefficients of the independent variables (features). Each coefficient represents the estimated effect of the corresponding independent variable on the dependent variable.

Analysis of Coefficients and their Signs on Target Variable:

1) Features which will have positive or increasing effect on Agent Bonus when their values increase is:

Sum Assured, Age, Cust Tenure, Monthly Income, Existing Policy Tenure, Payment Mode-Monthly.

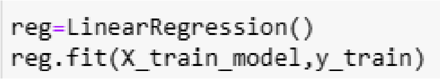
2) Features which will have decreasing effect on Agent Bonus when their values increase is:

Manager Designation, Married.

Linear Regression Equation form this OLS Model:

Y(Agent Bonus)= 4086 +(197.43 * Age) + (201.41 * Cust Tenure) + (255.80 * Monthly Income) + (109.8 * Existing Policy Tenure) + (819 * Sum assured) + (-48.18 * Manager Designation) +(-23.17 * Married) + (25.49 * Monthly payment mode).

Evaluate Linear Regression Model by Fitting in Scikit Learn.



```
reg=LinearRegression()
reg.fit(X_train_model,y_train)
```

Fitting Linear regression model by scikit learn linear model after dropping constant.
After fitting model on training and test Data and predicting value on Testing Set, Let's see the plot.

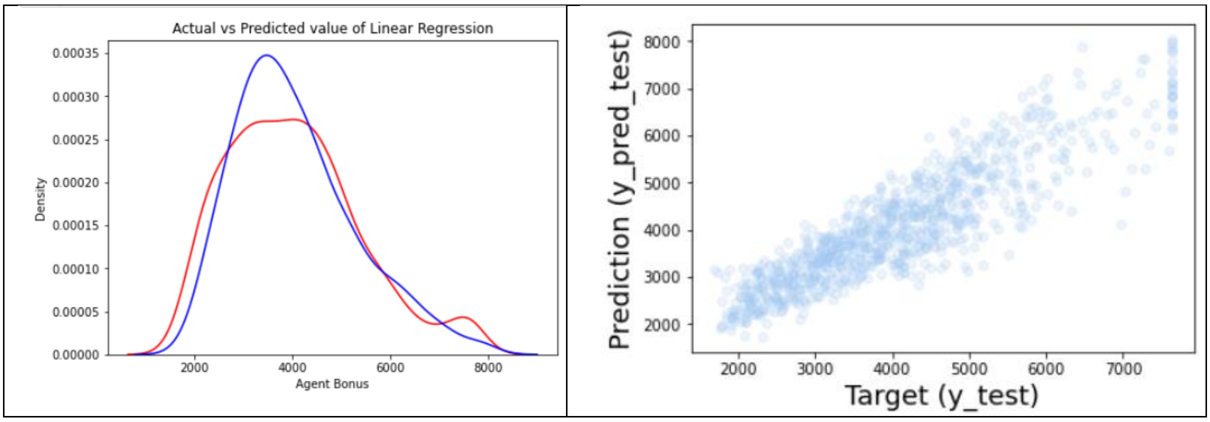


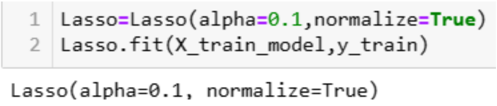
Fig 2.4 Linear Regression Actual vs Predicted Plot

Inference: There is discrepancy in Actual and predicted value, our model is prediction Agent bonus very high compared to actual upper high value.

2.4) Applying Lasso Regression, Model Evaluation & Result Interpretation.

Lasso Regression, short for "Least Absolute Shrinkage and Selection Operator," is a linear regression technique that adds a penalty term to the linear regression cost function to perform both feature selection and regularization. It is particularly useful when dealing with high-dimensional datasets where some features may be irrelevant or less important for predicting the target variable.

Fitting lasso regression in model.



```
Lasso=Lasso(alpha=0.1,normalize=True)
Lasso.fit(X_train_model,y_train)
```

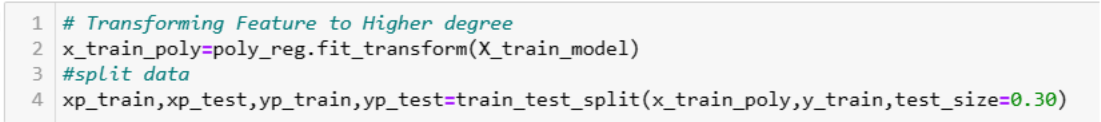
```
Lasso(alpha=0.1, normalize=True)
```

<u>Accuracy of Training & Testing both – 0.79</u>

2.4) Applying Polynomial Regression, Model Evaluation & Result Interpretation.

Polynomial Regression is a form of linear regression where the relationship between the independent variable(s) and the dependent variable is modelled as an nth-degree polynomial. It is a powerful technique that allows us to capture non-linear relationships between variables by fitting a polynomial curve to the data.

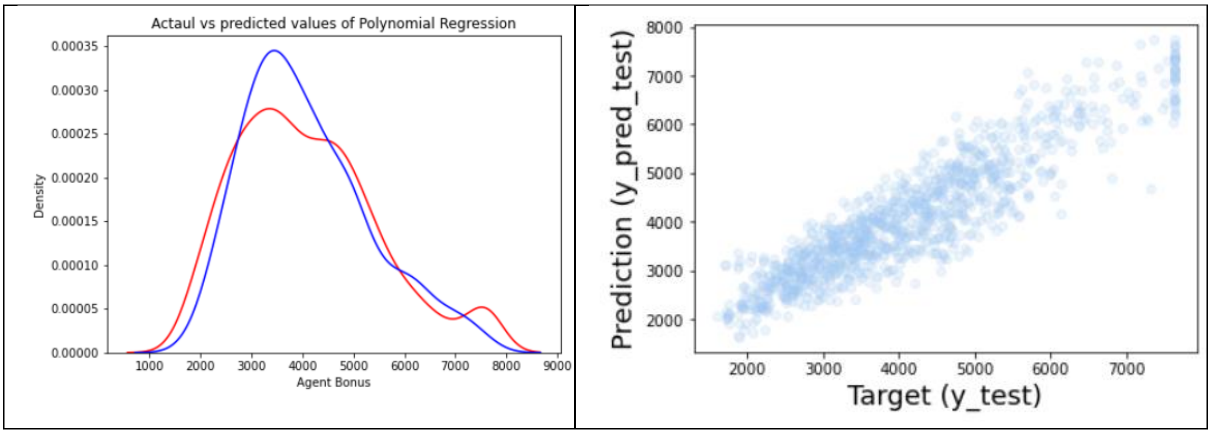
Transforming Features to higher degree and fit in new training data.



```
# Transforming Feature to Higher degree
x_train_poly=poly_reg.fit_transform(X_train_model)
#split data
xp_train,xp_test,yp_train,yp_test=train_test_split(x_train_poly,y_train,test_size=0.30)
```

Fitting Polynomial regression model.

<u>After fitting model on training and test Data and predicting value on Testing Set, Let's see the plot.</u>



<u>Fig 2.5 Polynomial Regression Actual vs Predicted Plot</u>

<u>Inference: There is discrepancy in Actual and predicted value, our model is prediction Agent bonus very high compared to actual upper high value.</u>

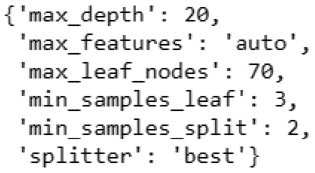
2.5) Applying Decision Tree Regression, Model Evaluation & Result Interpretation.

The decision tree regressor makes predictions using a tree-like structure, where each internal node represents a feature, each branch represents a decision based on the feature value, and each leaf node contains the predicted value.

Hyperparameter Tuning to increase accuracy with optimal parameter values.

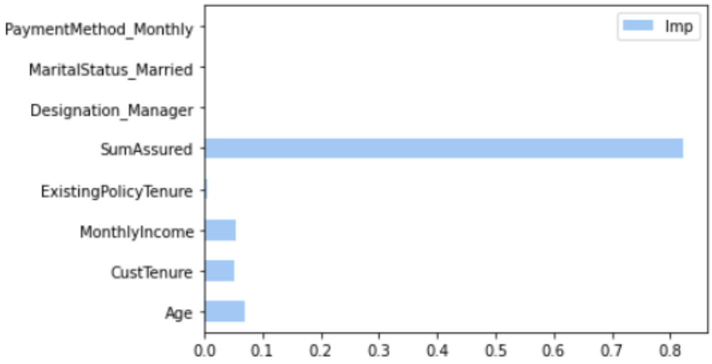
Hyperparameter tuning, also known as hyperparameter optimization, is the process of finding the best set of hyperparameters for a machine learning algorithm.

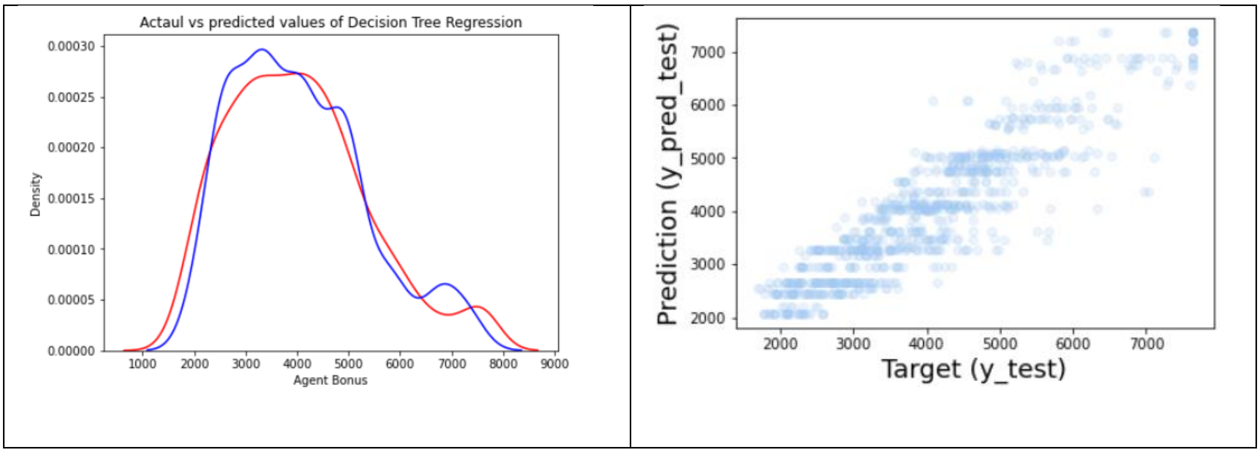
<u>Best parameters</u>:



```
{'max_depth': 20,
 'max_features': 'auto',
 'max_leaf_nodes': 70,
 'min_samples_leaf': 3,
 'min_samples_split': 2,
 'splitter': 'best'}
```

After fitting Decision tree model , following is importance of Features.





<u>Fig 2.6 Decision Tree Regression Actual vs Predicted Plot</u>

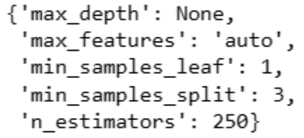
Inference:

X_train and y_train. Then, we use the trained model to predict the target variable. The predicted values are obtained using regressor. Predict(X_test) and stored in the prediction's variable.
We can conclude we got better prediction graph compared to linear regression.
After Hyper Tuning Model accuracy rose to 0.83 compare to 0.73.

2.5) Applying Random Forest Regression, Model Evaluation & Result Interpretation.

Random Forest Regressor is an ensemble learning technique used for regression tasks. It is an extension of the decision tree algorithm and combines multiple decision trees to make more accurate predictions.
Fitting Random Forest Model.
For increasing accuracy and R-Score we had done hyper parameter tuning.

Best Parameters:



```
{'max_depth': None,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 3,
 'n_estimators': 250}
```

Fitting random Forest regressor model with Best parameters.

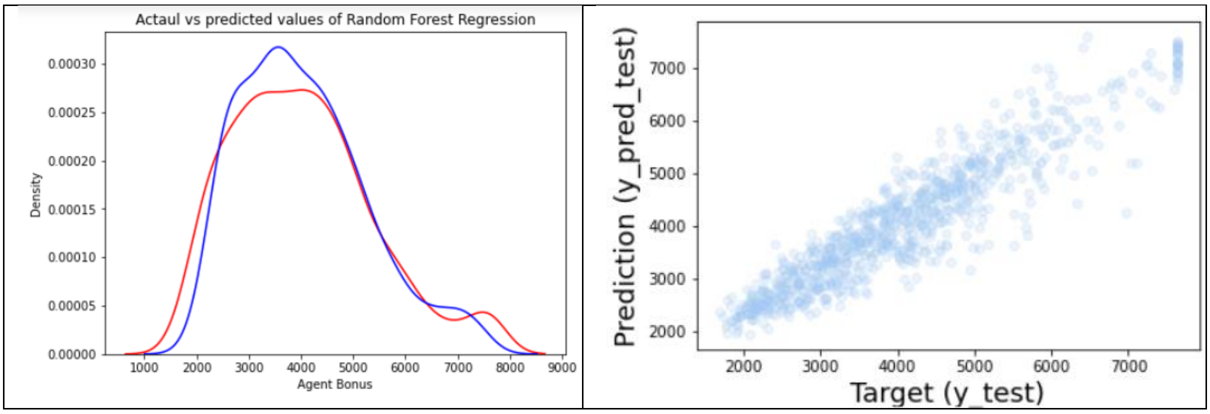


Fig 2.7 Random Forest Regression Actual vs Predicted Plot

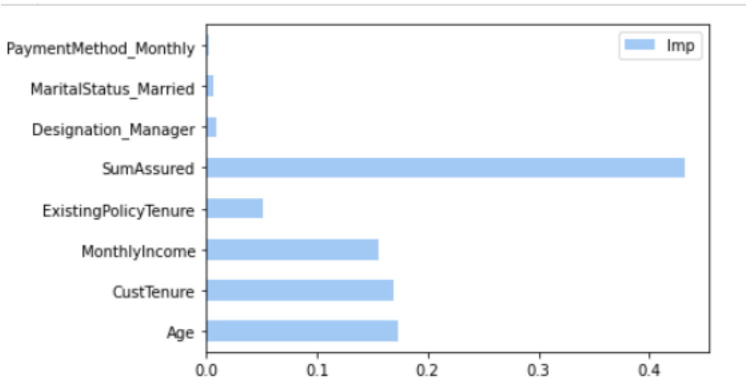


Fig 2.8 Random Forest Regression Feature Importance.

Inference:

Sum assured is most influential and important feature followed by Cust Tenure and Age.

By comparing the performance of the model on the training data and a validation set, we can identify if the model is not overfitted (model has same performance on training data and performance on unseen data).

2.6) Applying XG Booster Regression, Model Evaluation & Result Interpretation.

In XGBoost Regression, the algorithm is used for regression tasks, where the goal is to predict continuous numerical values for the target variable.

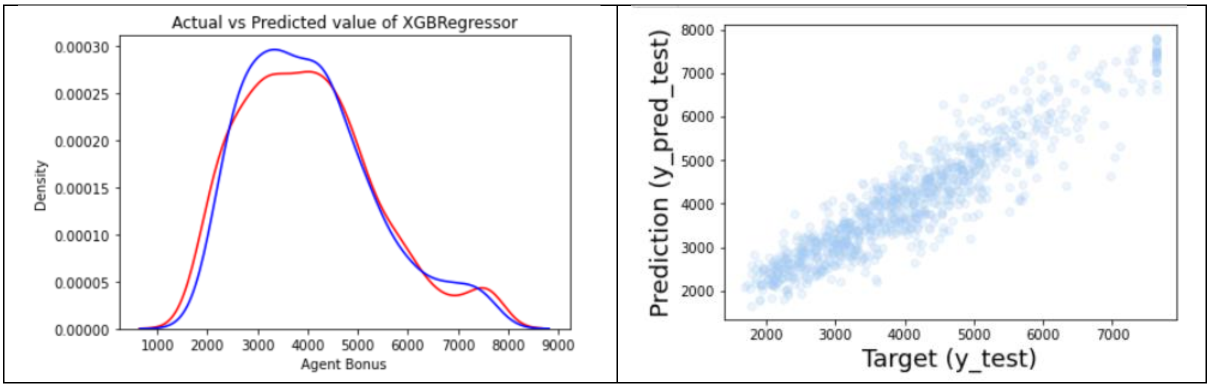


Fig 2.9 XG Boost Regression Actual vs Predicted Plot

Inference:
XG Boost Model is performing much better on unseen data compare to other models. Testing Accuracy for model is approx.. 0.86, still we have much scope to improve model by optimizing model.

Comparison of basis of Metrics values:

Regression Model Type	Test R2 Score	Test MSE	Test MAPE	Test MAE	Test RMSE	Train: Test Split
Linear Regression	0.79	378826.83	12%	484.62	615.48	80/20
Lasso Regression	0.79	377540.29	0.12	484.36	614.44	80/20
Decision Tree Regressor	0.82	499401.02	13%	492.89	706.68	80/20
Random Forest Regressor	0.86	250927.46	9%	376.07	500.92	80/20
XG Boost Regressor	0.85	268854.75	10%	397.3	518.51	80/20

Fig 2.10 Model Evaluation on metric Values

Inference:

- Mean Squared Error (MSE): MSE is one of the most widely used regression metrics. It measures the average squared difference between predicted values and actual target values. A lower MSE indicates better model performance.
  Lowest out of all models: Random Forest – 25092.46
- Root Mean Squared Error (RMSE): RMSE is the square root of the MSE. It has the same units as the target variable and is often used to interpret the error in a more understandable scale.
  Lowest out of all models: Random Forest – 500.92
- Mean Absolute Error (MAE): MAE measures the average absolute difference between predicted values and actual target values. It is less sensitive to outliers compared to MSE.
  Lowest out of all models: Random Forest – 376.09
- R-squared (R2): R-squared, also known as the coefficient of determination, indicates the proportion of variance in the target variable explained by the regression model. It ranges from 0 to 1, where 1 indicates a perfect fit.
  Highest out of all models: Random Forest – 0.86
- Mean Absolute Percentage Error (MAPE): MAPE is like MPE but takes the absolute percentage differences, making it more robust to outliers.
  Lowest % out of all models: Random Forest – 9%

Residual Plot Comparison for Training Data of all Model:

A residual plot is a graphical representation used to visualize the differences between the predicted values and the actual target values (residuals) in a regression model.

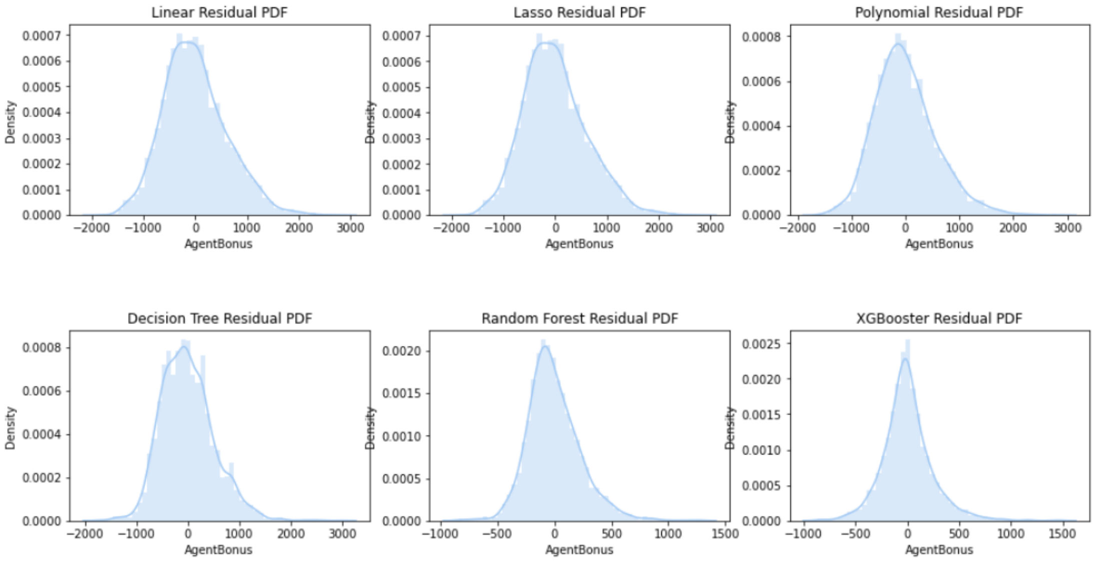


Fig 2.11 Residual PDF Plot Comparison

Inference: Random Forest Regression and XG Boost Regression residual plot with a random scatter of points around the zero line (y = 0) indicates that the model's predictions are unbiased and that the model captures the underlying relationships in the data reasonably well. This is a positive sign that the model is a good fit for the data.

Let’s Dig Deeper and try models with only top 5 important input and compare how model performs.

Top five features: Age, Monthly Income, Sum assured, Existing policy tenure, Cust Tenure.

Regression Model Type	Test R2 Score	Train Accuracy	Test Accuracy	Test MSE	Test MAPE	Test MAE	Test RMSE	Train: Test Split
Decision Tree Regressor	0.73	0.99	0.73	496708.02	13%	493.89	704.68	80/20
Random Forest Regressor	0.863	0.98	0.86	253306.46	9%	376.01	503.92	80/20
XG Boost Regressor	0.86	0.96	0.85	253037.75	10%	385.3	518.51	80/20

Inference:

By fitting models on top 5 features still we are getting random forest regression as best model, but surprisingly XG Boost model accuracy and RMSE Score got better.

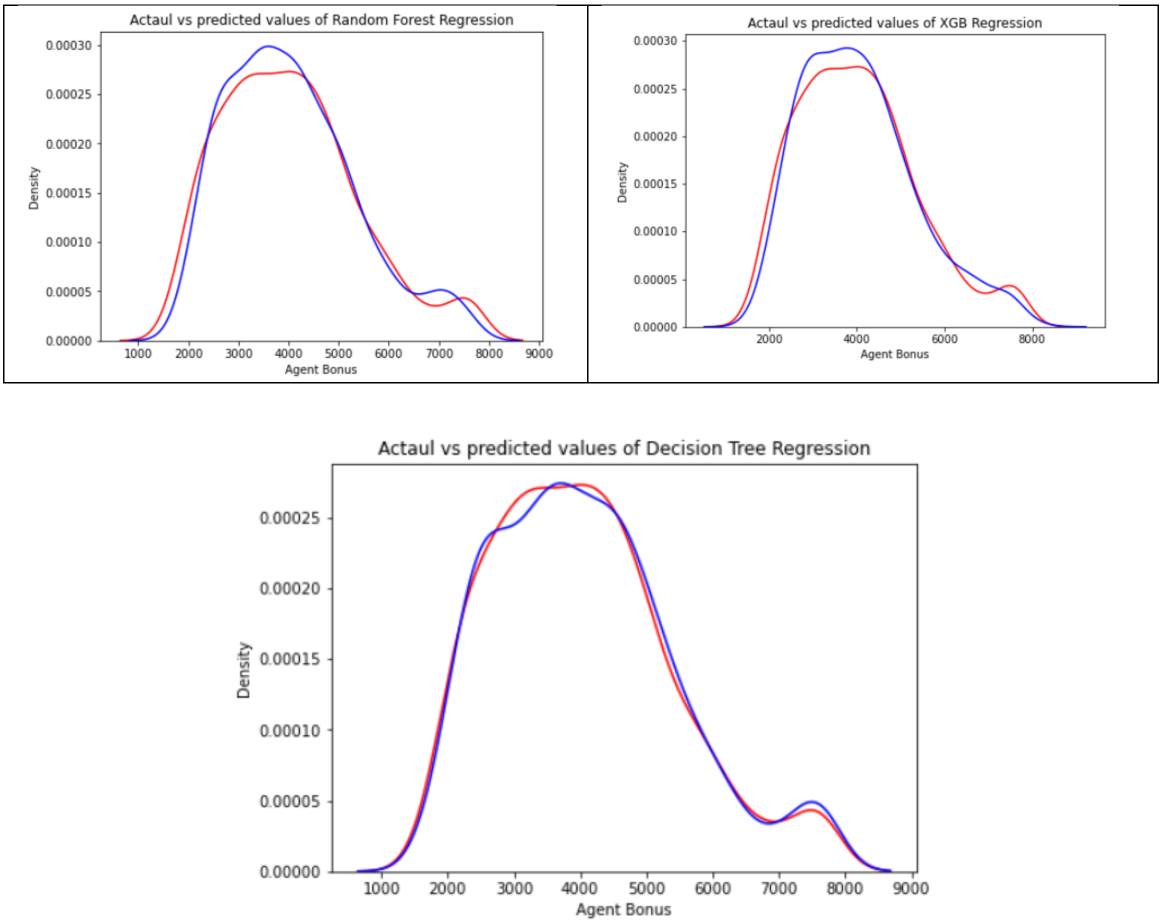


Fig 2.12 Actual vs Predicted Plot with Top 5 important features model.

# Comparison of Different Model R2 Score with 9 features

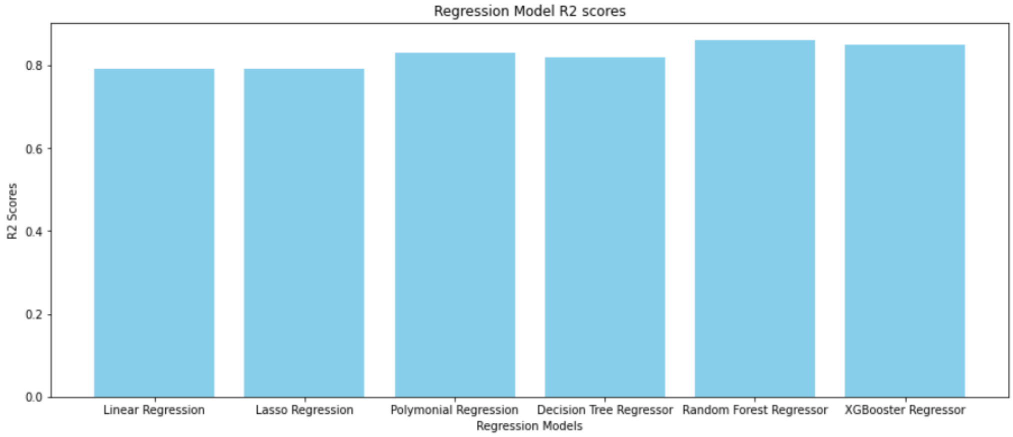


Fig 2.13 R2 Score Comparison of different models

# Comparison of Different Model RMSE with 9 features

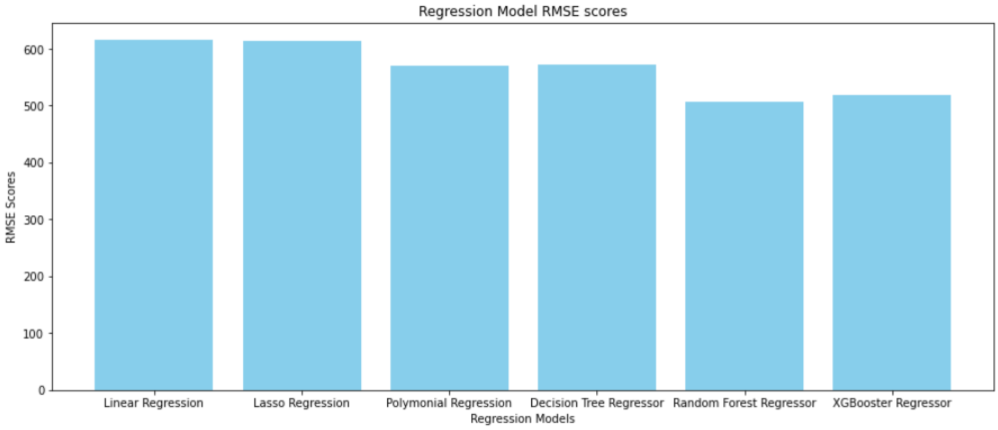


Fig 2.14 RMSE Score Comparison of different models

Conclusion:

- Our best model is the Random Forest Regressor & XG Boost Regression.
- The hyperparameters of our model are:
  n_estimator: 250
  max_depth: none
  min_samples_split: 3
  min_samples_leaf: 1
  max_features: auto
- The evaluation metrics test results of our model are:
  R 2: 0.86
  MSE: 250927.46
  RMSE: 500.9
- As the Agent Bonus increases, the RMSE also increases. We are aware that overfitting may be present, but our model still produces the best results on the test data. Even though we

are aware that we can achieve better results on this specific test set by using different hyperparameters, we have chosen to utilize the hyperparameters assigned during the grid search process. This decision was made to ensure that the model remains robust when exposed to different types of data.

- From the above models, we can see that XG Boost Regressor and Random Forest Regressor are giving the best results. But Random Forest Regressor is giving the best results with the least RMSE value. Therefore, I will use Random Forest Regressor to predict the Agent Bonus of Insurance Policy.
- Thus, from the overall analysis, we can conclude that the Bonus of Agent depends on their age, sum assured, Customer Tenure & Monthly Income.

SECTION 4: FINDINGS & INSIGHTS, DATA CONSTRAINTS & MODEL INTERPRETATION

The problem statement and the reason for embarking on this model development was predicting agent bonus for insurance company problems. The underlying problem associated with the data is there are low agent bonus which lead to demotivating agent experience, which affects the profitability- both Top Line and Bottom line, decrease in Sales, and hence prediction modelling study was conducted with the provided data set to.

- Predict Agent Bonus
- Improving Agent Performance

Insights from model:

- As our model suggest Monthly payment mode is most popular mode of payment, agent must present customer monthly payment model, as monthly payment value pictures as less amt compared to half yearly.
- Age is big factor in insurance policy, agent must educate customers of higher the age higher premium rates concept to customers.
- If Agent proposes customer with longer tenure policy with lucrative offer and benefits, Agent have high chances of getting good bonuses.
- Company must give goods perks to agents in form of festive incentives or extra target achieving incentive to attend agent loyalty, as insurance agent can be collaborated with multiple insurance companies.

Which variables are contributing to increase of agent Bonus?

- Sum Assured, Age, Cust Tenure, Monthly Income, Existing Policy Tenure, Payment Mode- Monthly.

Which variables are contributing to decrease of agent Bonus?

- Manager Designation, Married.

4.2 DATA CONSTRAINTS & MODEL INTERPRETATION:

The interpretation of the presented models and the data study that was conducted and presented above (Section 2 & 3), which the readers of this report to be aware of are listed below.

- The given data is mix of continuous and categorical variables.
- Many of the variables did not have impact on the Target variable- Agent Bonus, hence could be filtered at early stage.
- Few variables had missing values, but the proportion was high, hence variables was ignored.
- Many of the continuous variables had outliers, hence outlier treatment was necessary.
- The Independent variables were highly correlated amongst each other, hence the situation of multicollinearity which will affect the model existed.
- Variance inflation factor method used to eliminate multicollinear variables before modelling.
- Independent Categorical variables were also correlated, hence correlated categorical variables were dropped.
- The data had scale differences, hence scaling helped to standardise/normalise the data.
- Linear Regression it was essential to consider only the important uncorrelated independent variables.
- Lasso regression used to address the issue of multicollinearity and to encourage sparse solutions in linear regression models.
- KNN works best only for continuous variables with no outliers, hence only numeric independent variables considered for the model building.
- RANDOM Forest different mtry combination may yield different results.
- Decision Trees and Random Forest produced better results compared to Linear regression (or) Frequency Based Algorithms.
- Ensemble models - XGBoost works with matrices that contain all numeric variables. All categorical to be converted to dummies.
- Learning model XG Boost produced the best results and hence considered as best model based on the parametric evaluation.
- Random forest presented best model comparing all other models.

SECTION 5: CHALLENGES FACED DURING RESEARCH OF PROJECT AND TECHNIQUES USED TO OVERCOME THE CHALLENGES

Data Preparation:

The data had many outliers, many predictors were correlated amongst each other leading to situation of multi-collinearity and many predictors did not have predicting capability. Hence, 80% of time spent on cleaning and preparing the data to improve its quality i.e., to make it consistent, before utilising for analysis.

Getting The Right Data:

Quality is better than quantity is the call of the hour in this case. The business problem involves understanding the reason for low Agent Bonus, however with the given data set such reasons could not be identified, hence recommended (in recommendation section) for additional data on A. Product Flow B. Information Flow C. Revenue Flow.

Thus, to build an accurate model which works well with the business it is necessary to get the right data with the most meaningful features at the first instance. To overcome this data issue, would need to communicate with the business to get enough data and then use domain understanding to get rid of the irrelevant features. This is a backward elimination process but one which often comes handy in most occasions.

SECTION 6: RECOMMENDATIONS, CONCLUSIONS/APPLICATIONS.

The objective of this case study is to find “The best model which can predict Agent Bonus. Also, which variables are a significant predictor behind the decision. We developed prediction models by studying the data set provided using Linear Regression, lasso regression, Decision tree regressor, Random Forest regressor, we found Machine learning- XG Boost method to have provided the better model considering higher Accuracy, Sensitivity and Precision to identify the Agent Bonus.

<u>Possible methods to improve Agent’s performance.</u>

- ➢ **Training and Development**: Provide comprehensive and ongoing training programs for insurance agents to improve their product knowledge, sales techniques, and customer service skills. Training should cover various insurance products, market trends, objection handling, and communication skills.
- ➢ **Market Research and Analysis:** Provide agents with market research and analysis to help them understand customer preferences, identify potential leads, and tailor their sales strategies accordingly.
- ➢ **Digital Marketing and social media:** Train agents on digital marketing techniques and social media strategies to expand their reach and engage with potential clients online.
- ➢ **Identify Top Performer:** From bonus model we can identify which agents are achieving or exceeding targets, reward agents and motivate for future success.
- ➢ **Incentives Alignment**: Ensure the bonus structure are aligned with company strategy's goal and objectives. Consider adjusting bonus criteria according to products or targeting customer segments.

- **Continuous Evaluation:** Regularly review and refine the bonus model to ensure it remains aligned with changing business goals, market trends, and customer preferences.
- **Feedback Loop:** Establish a feedback pattern where agents can provide input on the bonus model's effectiveness and suggest potential improvements.
- **Low performing Boost:** Ensure the bonus structure are tailor in such a way, low performing agents are earning at least some incentives in form of vouchers. Evaluate reasons for low performance i.e., It can be specific region wrong product selling strategy.
- Company must give goods perks to agents in form of festive incentives or extra target achieving incentive to attend agent loyalty, as insurance agent can be collaborated with multiple insurance companies.

----End of Report-----