Capstone Project

Life Insurance Data

Business Report

Project Notes-2

Student Name: Vivek Bhatia

PGP-DSBA Online Jan_C 2022

Date: 24/12/2022

Table of Contents

1.Model building and Interpretation

a) Defining problem statement

The project notes 2 is an extension of project notes 1. The project notes 1 was completed and data was exported as Sales_Project_notes_2. We will continue from there and start our next phase of model building.

b) Need of the study/project

Companies often find themselves in complex situations of correctly analysing the performance factor for employees and the roadmap to upskill or reskill or reward them based on their performance metric. It is of paramount importance for the companies to recognize individual talent and contributions in order to ensure an unbiased approach towards their career progress and also to ensure right person for the right job philosophy.

By initiating this project company expects to benefit in knowing the exact performance metrics of each individual employee, their business needs, measures which can be taken to optimize the performance levels for low to medium performing employees and also to understand the working model of best performers to set a guideline for other peers.

Last but not the least this project will help company to prevent deliver the right amount of incentive for the right person thereby minimizing churn rates and retaining the in-house talent which will in turn lead to better performance and productivity and also reduce hiring and churning costs.

c) Understanding business/social opportunity

Understanding the business and the growth path is very crucial for the companies and it helps them forecast the future and also helps in making business related plans and projects to undertake in the next financial year or devise long term strategic objectives. By understanding the performance metrics of the agent company will be in a better position to mitigate risks associated with taking bigger projects based on their employees' approach and the bracket in which majority of the employees fall. Vision and Mission of the company and the objectives of CEO, CFO and CXO needs to be followed at grassroot level for any company to sustain and grow and combined effort of each and every induvial in the company plays a very significant role in achieving those targets. In turn for employees, it will yield better dividends and growth opportunities which in turn will drive them to deliver their best.

2)Data Report

a) Understanding how data was collected in terms of time, frequency and methodology

Importing the dataset and comparing the data with the last exported data for correctness and proceeding ahead with model building. The sample of the data set imported as a part of continuation of Project Notes 1 is as below.

	Age	CustTenure	Channel	Occupation	EducationField	Gender	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome	ExistingPolicyTenure	Zone	I
0	22.0	4.0	Agent	Salaried	Graduate	Female	Manager	2.0	Single	20993.0	2.0	North	
1	11.0	2.0	Third Party Partner	Salaried	Graduate	Male	Manager	4.0	Divorced	20130.0	3.0	North	
2	26.0	4.0	Agent	Free Lancer	Post Graduate	Male	Executive	3.0	Unmarried	17090.0	2.0	North	
3	11.0	2.0	Third Party Partner	Salaried	Graduate	Female	Executive	3.0	Divorced	17909.0	2.0	West	
4	6.0	4.0	Agent	Small Business	Under Graduate	Male	Executive	4.0	Divorced	18468.0	4.0	West	

The Sales data consisted of 20 different variable and was collected based on last month's bonus given to agents along with customer ID, Age of customer, tenure of customer and the channel through which the customer was acquired like agent, third party or online. The channel plays a very crucial role in determining the future business model as it will determine how many customers are acquired successfully via agents, online or third-party contract and a revenue model can be built across the thrse verticals.

After comparing the exported and imported data we can now proceed with model building as the data is good to go.

b) Understanding of attributes

The various attributes of the data are shown as below with their columns and unique categories after encoding.

```
Columns is : Channel
['Agent', 'Third Party Partner', 'Online']
Categories (3, object): ['Agent', 'Online', 'Third Party Partner']
[0 2 1]


Columns is : Occupation
['Salaried', 'Free Lancer', 'Small Business', 'Large Business']
Categories (4, object): ['Free Lancer', 'Large Business', 'Salaried', 'Small Business']
[2 0 3 1]


Columns is : EducationField
['Graduate', 'Post Graduate', 'Under Graduate', 'Engineer', 'Diploma', 'MBA']
Categories (6, object): ['Diploma', 'Engineer', 'Graduate', 'MBA', 'Post Graduate', 'Under Graduate']
[2 4 5 1 0 3]


Columns is : Gender
['Female', 'Male']
Categories (2, object): ['Female', 'Male']
[0 1]


Columns is : Designation
['Manager', 'Executive', 'VP', 'AVP', 'Senior Manager']
Categories (5, object): ['AVP', 'Executive', 'Manager', 'Senior Manager', 'VP']
[2 1 4 0 3]


Columns is : MaritalStatus
['Single', 'Divorced', 'Unmarried', 'Married']
Categories (4, object): ['Divorced', 'Married', 'Single', 'Unmarried']
[2 0 3 1]


Columns is : Zone
['North', 'West', 'East', 'South']
Categories (4, object): ['East', 'North', 'South', 'West']
[1 3 0 2]


Columns is : PaymentMethod
['Half Yearly', 'Yearly', 'Quarterly', 'Monthly']
Categories (4, object): ['Half Yearly', 'Monthly', 'Quarterly', 'Yearly']
[0 3 2 1]
```

3) Model building and interpretation

We will create a new variable X and drop the target variable “agent_bonus_per_policy” from it and we will put the dropped target variable in y.

The sample data of X and y is as below:

	Age	CustTenure	Channel	Occupation	EducationField	Gender	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome	ExistingPolicyTenure	Zone
0	22.0	4.0	0	2	2	0	2	2.0	2	20993.0	2.0	1
1	11.0	2.0	2	2	2	1	2	4.0	0	20130.0	3.0	1
2	26.0	4.0	0	0	4	1	1	3.0	3	17090.0	2.0	1
3	11.0	2.0	2	2	2	0	1	3.0	0	17909.0	2.0	3
4	6.0	4.0	0	3	5	1	1	4.0	0	18468.0	4.0	3

	Agent_bonus_Per_Policy
0	2204.500000
1	553.500000
2	1424.333333
3	597.000000
4	738.750000

We have 4520 rows of data and we need to split the data into train and test for model building. We need to find a good ratio for our split so that it can be justified and more accurate results can be derived from the models.
Let us analyse the different ratio and then choose the optimal ratio based on the size of test and train.

```
If split percentage of  data is 60/40 then train_data is 2712.0 and test data is 1808.0
If split percentage of  data is 65/35 then train_data is 2938.0 and test data is 1582.0
If split percentage of  data is 70/30 then train_data is 3164.0 and test data is 1356.0
If split percentage of  data is 75/25 then train_data is 3390.0 and test data is 1130.0
If split percentage of  data is 67/33 then train_data is 3028.4 and test data is 1491.6000000000001
```

As seen from the above results, we can split the data into 70:30 or 67:33.

If we split the data into 67:33 ratio then train data will have 3028 rows and test data will have 1491 rows.

I choose to go with 70:30 ratio as this looks better with train data of 3164 and test data of 1356. The shape of train and test data is shown below:

```
Shape for X_train is  (3164, 14)
Shape for X_test is  (1356, 14)
Shape for y_train is  (3164, 1)
Shape for y_test is  (1356, 1)
```

4)Linear Regression

a) What is Linear Regression?

Linear regression is a statistical method for modelling relationships between a dependent variable with a given set of independent variables.
In its simplest terms if the two variables are linearly related, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).
Multiple linear regression attempts to model the relationship between two or more features and a response by fitting a linear equation to the observed data.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$$
$$or$$
$$y_i = h(x_i) + \varepsilon_i \Rightarrow \varepsilon_i = y_i - h(x_i)$$

To build any Linear Regression Model we need to follow the following basic steps:

- Importing the libraries
- Reading the data set
- Exploring the data
- Cleaning the data
- Training the model
- Analysing and exploring the results

Now let us build our first model-SKLearn Linear model which is a basic model to begin with for linear regression.

5) Model 1: SKLearn Linear Model

with scikit-linear learn's regression approach, we will encounter the following fundamental concepts:

- Best Fit - The straight line in a plot that minimizes the divergence between related dispersed data points
- Coefficient - Also known as a parameter, is the factor that is multiplied by a variable. A coefficient in linear regression represents changes in a Response Variable
- Coefficient of Determination - It is the correlation coefficient. In a regression, this term is used to define the precision or degree of fit
- Correlation - the measurable intensity and degree of association between two variables, often known as the 'degree of correlation.' The values range from -1.0 to 1.0
- Dependent Feature - A variable represented as y in the slope equation y=ax+b. Also referred to as an Output or a Response
- Estimated Regression Line - the straight line that best fits a set of randomly distributed data points
- Independent Feature - a variable represented by the letter x in the slope equation y=ax+b. Also referred to as an Input or a predictor
- Intercept - It is the point at where the slope intersects the Y-axis, indicated by the letter b in the slope equation y=ax+b
- Least Squares - a method for calculating the best fit to data by minimizing the sum of the squares of the discrepancies between observed and estimated values
- Mean - an average of a group of numbers; nevertheless, in linear regression, Mean is represented by a linear function
- OLS (Ordinary Least Squares Regression) - sometimes known as Linear Regression.
- Residual - the vertical distance between a data point and the regression line
- Regression - is an assessment of a variable's predicted change in relation to changes in other variables
- Regression Model - The optimum formula for approximating a regression
- Response Variables - This category covers both the Predicted Response (the value predicted by the regression) and the Actual Response (the actual value of the data point)
- Slope - the steepness of a regression line. The linear relationship between two variables may be defined using slope and intercept: y=ax+b
- Simple linear regression - A linear regression with a single independent variable

After fitting the X_train and y_train into the Linear Regression Model we can get coefficients as shown below:

```
The coefficient for Age is 4.644165148928456
The coefficient for CustTenure is 4.106940177690477
The coefficient for Channel is -9.56561449923948
The coefficient for Occupation is 4.981901467505096
The coefficient for EducationField is -5.08030222761581
The coefficient for Gender is 6.321647575599218
The coefficient for Designation is 6.233434458852404
The coefficient for NumberOfPolicy is -133.62353676894256
The coefficient for MaritalStatus is 13.296456642725069
The coefficient for MonthlyIncome is 0.00955711142090671
The coefficient for ExistingPolicyTenure is 3.2879491753102488
The coefficient for Zone is -5.086378954059552
The coefficient for PaymentMethod is -1.2028217445142255
The coefficient for Sum_assured_Per_Policy is 0.004891189080896628
```

The coefficient means that for a unit change in the coefficient value the target variable will change by X units. Let us plot our coefficients in a more elaborate manner to understand the real meaning.

```
The meaning of coefficents is explained in a better way below:
For a unit change in Age agent bonus per policy will change by  4.644165148928456
The meaning of coefficents is explained in a better way below:
For a unit change in CustTenure agent bonus per policy will change by  4.106940177690477
The meaning of coefficents is explained in a better way below:
For a unit change in Channel agent bonus per policy will change by  -9.56561449923948
The meaning of coefficents is explained in a better way below:
For a unit change in Occupation agent bonus per policy will change by  4.981901467505096
The meaning of coefficents is explained in a better way below:
For a unit change in EducationField agent bonus per policy will change by  -5.08030222761581
The meaning of coefficents is explained in a better way below:
For a unit change in Gender agent bonus per policy will change by  6.321647575599218
The meaning of coefficents is explained in a better way below:
For a unit change in Designation agent bonus per policy will change by  6.233434458852404
The meaning of coefficents is explained in a better way below:
For a unit change in NumberOfPolicy agent bonus per policy will change by  -133.62353676894256
The meaning of coefficents is explained in a better way below:
For a unit change in MaritalStatus agent bonus per policy will change by  13.296456642725069
The meaning of coefficents is explained in a better way below:
For a unit change in MonthlyIncome agent bonus per policy will change by  0.00955711142090671
The meaning of coefficents is explained in a better way below:
For a unit change in ExistingPolicyTenure agent bonus per policy will change by  3.2879491753102488
The meaning of coefficents is explained in a better way below:
For a unit change in Zone agent bonus per policy will change by  -5.086378954059552
The meaning of coefficents is explained in a better way below:
For a unit change in PaymentMethod agent bonus per policy will change by  -1.2028217445142255
The meaning of coefficents is explained in a better way below:
For a unit change in Sum_assured_Per_Policy agent bonus per policy will change by  0.004891189080896628
```

The equation for the above data set can be written as follows:

```
Age * 4.644165148928456 + CustTenure * 4.106940177690477 + Channel * -9.56561449923948 + Occupation * 4.981901467505096 + Educa
tionField * -5.08030222761581 + Gender * 6.321647575599218 + Designation * 6.233434458852404 + NumberOfPolicy * -133.6235367689
4256 + MaritalStatus * 13.296456642725069 + MonthlyIncome * 0.00955711142090671 + ExistingPolicyTenure * 3.2879491753102488 + Z
one * -5.086378954059552 + PaymentMethod * -1.2028217445142255 + Sum_assured_Per_Policy * 0.004891189080896628 + 476.5055999170
405
```

We also need to understand the intercept which is the last value seen in the above equation.

```
The intercept for the model is 476.5055999170405
Intercept is the value of target variable , when all values of dependent variables  are Zero.
When all the target variables are 0, the agent bonus per policy still stands 476.5055999170405
```

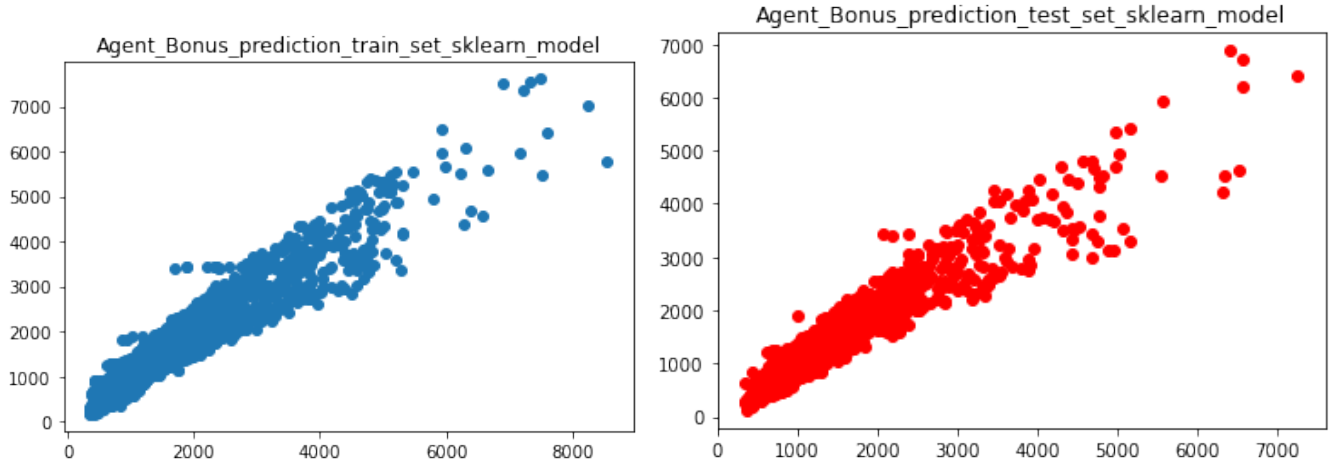
Coefficient of determination:
The coefficient of determination, denoted as R^2, tells you which amount of variation in y can be explained by the dependence on $\mathbf{x}$, using the particular regression model. A larger R^2 indicates a better fit and means that the model can better explain the variation of the output with different inputs. The larger the value of R^2 the less the unexplained variance is.
The value R^2 = 1 corresponds to SSR = 0. That's the perfect fit, since the values of predicted and actual responses fit completely to each other.
R-Squared: R2 is a statistic that will give some information about the goodness of fit of a model. It ranges from 0 to 1. Example: if the value of R-squared is 0.75 so it explains 75% of variance is explained by the model and 25% of variance cannot be explained.
In our model the coefficient of determination for train and test set is shown below:

```
The coefficient of determination R^2 of the prediction on Train set 0.9146462113091353
The coefficient of determination R^2 of the prediction on Test set 0.8996861106251208
```

For train set R^2 is 91.5% and it means that 91.5 % of variance is explained by the model and almost 8.5% of variance cannot be explained.
For train set R^2 is 91.5% and it means that 90 % of variance is explained by the model and almost 10% of variance cannot be explained.
One way to assess how well a regression model fits a dataset is to calculate the root mean square error, which is a metric that tells us the average distance between the predicted values from the model and the actual values in the dataset.
The lower the RMSE, the better a given model is able to "fit" a dataset.
The formula to find the root mean square error, often abbreviated RMSE, is as follows:
RMSE = Square root of (sum(Pi – Oi)**2 / n)
where:
Σ is a fancy symbol that means "sum" Pi is the predicted value for the ith observation in the dataset Oi is the observed value for the ith observation in the dataset n is the sample size
This means that the RMSE represents the square root of the variance of the residuals.
suppose we want to build a regression model to predict the exam score of students and we want to find the best possible model among several potential models.
Suppose we fit three different regression models and find their corresponding RMSE values:
RMSE of Model 1: 14.5 RMSE of Model 2: 16.7 RMSE of Model 3: 9.8
Then, Model 3 has the lowest RMSE, which tells us that it's able to fit the dataset the best out of the three potential models.
we also have other metrics like Mean Absolute deviation (MAD), Mean absolute percentage error (MAPE), Mean Squared Error (MSE) and Explained variance Score (EVS) to check our model evaluation. Which one to choose as the best metric depends on the business applications and we need to decide the best metric accordingly and justify the reason for choosing that metric against other metrics.

```
The Root Mean Square Error (RMSE) of the model is for training set is 303.7871800527234
The Root Mean Square Error (RMSE) of the model is for testing set is 311.7156165736742
```

Plotting the prediction train and test data:



From the scatter plot we can see that agent bonus per policy is mostly concentrated up to 4000 units (currency like rupees or dollars etc). Few policies give very high bonus up to 7000.

Let us also evaluate all other error metrics because depending on the business applications we can use different metric to determine the goodness of the model.

```
The train set MAE is: 208.99706704239236
The train set MSE is: 92286.65076438579
The train set RMSE is: 303.7871800527234
The train set MAPE is 0.15617807763321093
The train set EVS is 0.9146462113091353
```

```
The test set MAE is: 213.37624527983854
The test set MSE is: 97166.62561590585
The test set RMSE is: 311.7156165736742
The test set MAPE is 0.15733016829447444
The test set EVS is 0.8997497474660561
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.78718	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.15733	0.914646	0.89975

From the above results we can conclude that there is 15% of variance in terms of mean absolute percentage error for agent bonus per policy and in terms of Explained Variance score we are able to explain 91.4% of variance in train data and 89.9 or 90% of variance in test data. The unexplained variance is due to the other variables which we have dropped before building the model.

Here we can see RMSE, MAE, MSE, MAPE & EVS. Brief explanation about each of this metric is as follows:

1. RMSE (Root Mean Squared Error): It measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy).
   RMSE can be expressed as :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

where N is the number of data points, y(i) is the ith measurement, and ŷ(i) is its corresponding prediction.

Lower the value of RMSE, better the model fit.

2. MAE (Mean Absolute Error): It is a measure of errors between paired observations expressing the same phenomenon.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

3. MSE (Mean Squared Error): It measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate. In machine learning, specifically empirical risk minimization, MSE may refer to the empirical risk (the average loss on an observed data set), as an estimate of the true MSE (the true risk: the average loss on the actual population distribution). It is also known as MSD (Mean Squared Deviation)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

4. MAPE (Mean Absolute Percentage Error): It is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula:

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

   It is also known as MAPD (Mean Absolute Percentage Deviation). It is used as a loss function for regression problems and in model evaluation, because of its very intuitive interpretation in terms of relative error.
5. EVS (Explained Variance Score): It The explained variance score explains the dispersion of errors of a given dataset, and the formula is written as follows:

$$explained\ variance(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

6.) Model 2: Stats Model-OLS Regressor

In OLS method, we have to choose the values of and such that, the total sum of squares of the difference between the calculated and observed values of y, is minimised.

$$S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_1 x_1 - b_0)^2 = \sum_{i=1}^{n}(\hat{\epsilon}_i)^2 = min$$

....

Agent_bonus_Per_Policy is a function of all the dependent variables above and see the reults. We will slowly drop the insignificant features and optimize our policy to get the best results.

a) Stats Model 1:

f_1='Agent_bonus_Per_Policy~Age+CustTenure+Channel+Occupation+EducationField+Gender+Designation+NumberOfPolicy+MaritalStatus+MonthlyIncome+ExistingPolicyTenure+Zone+PaymentMethod+Sum_assured_Per_Policy'

Based on the formula created above (f_1) we will now create lm1 and analyse the parameters as follows:

```
Intercept               476.505600
Age                       4.644165
CustTenure                4.106940
Channel                  -9.565614
Occupation                4.981901
EducationField           -5.080302
Gender                    6.321648
Designation               6.233434
NumberOfPolicy         -133.623537
MaritalStatus            13.296457
MonthlyIncome             0.009557
ExistingPolicyTenure      3.287949
Zone                     -5.086379
PaymentMethod            -1.202822
Sum_assured_Per_Policy    0.004891
dtype: float64
```

The summary of the OLS lm1 model is shown as follows:

OLS Regression Results

Dep. Variable:	Agent_bonus_Per_Policy	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.914
Method:	Least Squares	F-statistic:	2410.
Date:	Sat, 24 Dec 2022	Prob (F-statistic):	0.00
Time:	17:31:20	Log-Likelihood:	-22576.
No. Observations:	3164	AIC:	4.518e+04
Df Residuals:	3149	BIC:	4.527e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	476.5056	43.316	11.001	0.000	391.575	561.437
Age	4.6442	0.680	6.825	0.000	3.310	5.978
CustTenure	4.1069	0.693	5.929	0.000	2.749	5.465
Channel	-9.5656	6.855	-1.395	0.163	-23.006	3.875
Occupation	4.9819	9.758	0.511	0.610	-14.152	24.115
EducationField	-5.0803	3.542	-1.434	0.152	-12.026	1.865
Gender	6.3216	11.021	0.574	0.566	-15.287	27.930
Designation	6.2334	6.207	1.004	0.315	-5.937	18.404
NumberOfPolicy	-133.6235	5.700	-23.443	0.000	-144.799	-122.448
MaritalStatus	13.2965	7.115	1.869	0.062	-0.654	27.246
MonthlyIncome	0.0096	0.001	6.441	0.000	0.007	0.012
ExistingPolicyTenure	3.2879	1.200	2.740	0.006	0.935	5.641
Zone	-5.0864	5.363	-0.948	0.343	-15.602	5.429
PaymentMethod	-1.2028	3.979	-0.302	0.762	-9.005	6.599
Sum_assured_Per_Policy	0.0049	5.03e-05	97.243	0.000	0.005	0.005

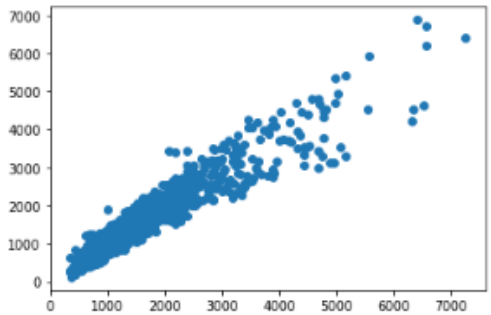
Omnibus:	1079.113	Durbin-Watson:	1.933
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9744.005
Skew:	1.357	Prob(JB):	0.00
Kurtosis:	11.157	Cond. No.	2.25e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.25e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

The OLS result summary is a very detailed summary where we can see the dependent variable, model, method, date, time and important results like R-Squared, Adj R-Squared-statistics, Probability of F-Statistics, AIC,BIC etc to name a few.

The equation can be shown as below:

```
Test data root mean squared error value for lm1 model is : 311.71561657367016
Train data root mean squared error value for lm1 model : 303.7871800527231
Equation for model is :
(476.51) * Intercept + (4.64) * Age + (4.11) * CustTenure + (-9.57) * Channel + (4.98) * Occupation + (-5.08) * EducationField
+ (6.32) * Gender + (6.23) * Designation + (-133.62) * NumberOfPolicy + (13.3) * MaritalStatus + (0.01) * MonthlyIncome + (3.2
9) * ExistingPolicyTenure + (-5.09) * Zone + (-1.2) * PaymentMethod + (0.0) * Sum_assured_Per_Policy +
```



A more comprehensive results consisting of all the error metrics is shown below for lm1 model:

```
The MAE for Train data lm1 model is: 208.9970670423894
The MSE for Train data lm1 model is: 92286.65076438579
The RMSE for Train data lm1 model is 303.7871800527234
The MAPE for Train data lm1 model is 0.15617807763320574
The EVS for Train data lm1 model is 0.9146462113091353
The MAE for Test data lm1 model is: 213.37624527983428
The MSE for Test data lm1 model is: 97166.62561590347
The RMSE for Test data lm1 model is 311.71561657367033
The MAPE for Test data lm1 model is 0.1573301682944708
The EVS for Test data lm1 model is 0.8997497474660554
```

The results can be combined and compared with the SKLearn model result and we see that there is no difference between the two models.

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.78718	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.15733	0.914646	0.89975
Stats Model 1	303.78718	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.15733	0.914646	0.89975

From the above results we can conclude that there is no difference in Linear Regression SKLearn model and Stats Model 1.

b) Stats Model 2:

As seen from the OLS Regression results a lot of features or variables have P value more than 0.05.

In the OLS Model result we get many results like R-squared, Adj. R-squared, F-statistic, Prob (F-statistic), Log-Likelihood, AIC and BIC.

According to the Model, Columns with P value Less than 0.05 should be consider for the Model.

Null hypothesis says that there is No relation between independent variable and Target (Dependent variable).

When P value is Greater than 0.05 then we do not have enough evidence in the data to reject the Null hypothesis and we do not reject H0, which means we accept the NULL Hypothesis and there is no relation between Independent and target variables.

The idea behind this analysis is to eliminate the multi collinearity problem.

Looking at the P Values for all the fields, we can Come to this conclusion, following fields can be ignored for the Model prediction:

	coef	std err	t	P>|t|
Channel	-9.5656	6.855	-1.395	0.163
Occupation	4.9819	9.758	0.511	0.610
EducationField	-5.0803	3.542	-1.434	0.152
Gender	6.3216	11.021	0.574	0.566
Designation	6.2334	6.207	1.004	0.315
MaritalStatus	13.2965	7.115	1.869	0.062
Zone	-5.0864	5.363	-0.948	0.343
PaymentMethod	-1.2028	3.979	-0.302	0.762

As we have lot of variables, we can calculate the VIF and see which one can be eliminated from the next model.

Calculate VIF

Calculate VIF for each Field and based on VIF value we decide, which column can be eliminated

VIF (variance inflation factor) is used for checking multicollinearity in regression Model. its Values can range between 1 to Infinite.

How to interpret a given VIF value?

Consider the following linear regression model:

$Y = \beta0 + \beta1 \times X1 + \beta2 \times X2 + \beta3 \times X3 + \varepsilon$

For each of the independent variables X1, X2 and X3 we can calculate the variance inflation factor (VIF) in order to determine if we have a multicollinearity problem.

Here's the formula for calculating the VIF for X1:

VIF1 = 1 / (1 - R**2)

VIF (variance inflating factor) formula for the first variable in the model R2 in this formula is the coefficient of determination from the linear regression model which has:

X1 as dependent variable X2 and X3 as independent variables In other words, R2 comes from the following linear regression model:

$X1 = \beta0 + \beta1 \times X2 + \beta2 \times X3 + \varepsilon$

And because R2 is a number between 0 and 1:

When R2 is close to 1 (i.e., X2 and X3 are highly predictive of X1): the VIF will be very large When R2 is close to 0 (i.e., X2 and X3 are not related to X1): the VIF will be close to 1 Therefore the range of VIF is between 1 and infinity.

```
VIF for  CustTenure --> 3.760517160627696
VIF for  Channel --> 1.3518790536589753
VIF for  EducationField --> 3.014882945985352
VIF for  Gender --> 2.2386508131987
VIF for  Designation --> 3.655528588522457
VIF for  MaritalStatus --> 2.939762984445114
VIF for  ExistingPolicyTenure --> 2.176065528485394
VIF for  Zone --> 4.308308097008905
VIF for  PaymentMethod --> 1.5610572051537284
VIF for  Sum_assured_Per_Policy --> 2.8428058987570117
```

We have reduced the features from 14 to 10 using VIF.

f_2='Agent_bonus_Per_Policy~CustTenure+Channel+EducationField+Gender+Designation+MaritalStatus+ExistingPolicyTenure+Zone+PaymentMethod+Sum_assured_Per_Policy'

lm2 intercept and coefficients:

```
Intercept               72.409029
CustTenure               2.813706
Channel                 -6.576378
EducationField          -3.976172
Gender                  15.394656
Designation             16.468421
MaritalStatus           13.097822
ExistingPolicyTenure     3.994369
Zone                     1.388434
PaymentMethod           -2.181941
Sum_assured_Per_Policy   0.005782
dtype: float64
```

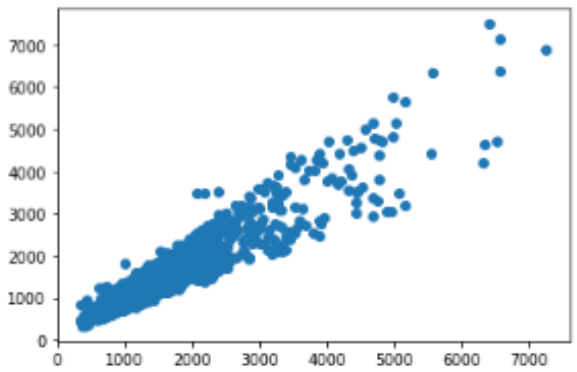
The OLS result summary fro lm2 model is shown below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Agent_bonus_Per_Policy   R-squared:                       0.899
Model:                               OLS   Adj. R-squared:                  0.899
Method:                    Least Squares   F-statistic:                     2811.
Date:                   Sat, 24 Dec 2022   Prob (F-statistic):               0.00
Time:                           17:31:22   Log-Likelihood:                -22840.
No. Observations:                   3164   AIC:                         4.570e+04
Df Residuals:                       3153   BIC:                         4.577e+04
Df Model:                             10
Covariance Type:               nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                   72.4090     25.110      2.884      0.004      23.175     121.643
CustTenure                   2.8137      0.707      3.977      0.000       1.427       4.201
Channel                     -6.5764      7.444     -0.883      0.377     -21.173       8.020
EducationField              -3.9762      3.370     -1.180      0.238     -10.583       2.631
Gender                      15.3947     11.960      1.287      0.198      -8.056      38.845
Designation                 16.4684      6.141      2.682      0.007       4.429      28.508
MaritalStatus               13.0978      7.698      1.701      0.089      -1.996      28.192
ExistingPolicyTenure         3.9944      1.218      3.279      0.001       1.606       6.383
Zone                         1.3884      5.818      0.239      0.811     -10.019      12.795
PaymentMethod               -2.1819      4.321     -0.505      0.614     -10.654       6.290
Sum_assured_Per_Policy       0.0058   3.61e-05    160.268      0.000       0.006       0.006
==============================================================================
Omnibus:                     1022.160   Durbin-Watson:                   1.950
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6733.279
Skew:                           1.370   Prob(JB):                         0.00
Kurtosis:                       9.601   Cond. No.                     1.22e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.22e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The equation of lm2 model is shown as follows along with the scatter plot.

```
test data root mean square error value is : 342.90517343060617
Train data root mean square error value is : 330.2453460384369
Equation for model is :
(72.41) * Intercept + (2.81) * CustTenure + (-6.58) * Channel + (-3.98) * EducationField + (15.39) * Gender + (16.47) * Designa
tion + (13.1) * MaritalStatus + (3.99) * ExistingPolicyTenure + (1.39) * Zone + (-2.18) * PaymentMethod + (0.01) * Sum_assured_
Per_Policy +
```



The other error metrics are shown as below:

```
The MAE for Train data lm2 model is: 226.66298925297846
The MSE for Train data lm2 model is: 109061.98858004669
The RMSE for Train data lm2 model is 330.24534603843654
The MAPE for Train data lm2 model is 0.1644402680303528
The EVS for Train data lm2 model is 0.8991310893789726
The MAE for Test data lm2 model is: 235.43640166127037
The MSE for Test data lm2 model is: 117583.95796547395
The RMSE for Test data lm2 model is 342.90517343060594
The MAPE for Test data lm2 model is 0.1691621989121059
The EVS for Test data lm2 model is 0.8787718085513729
```

The results are combined and represented as follows:

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772

From the results we can conclude that there is not much difference between Linear Regression SKLearn model, Stats Model 1 and Stats Model 2. Although the intercept has reduced from 474 to 72 meaning when all the features are zero the agent bonus per policy now is 72 units.

From the OLS Regression results of lm2 summary we can see that P value are still high for some features. We need to drop them one by one so that we can see the effect of reducing multi collinearity. If we drop them all together, we risk losing the vital data. The features are as follows in descending order of their P values

	coef	std err	t	P>|t|
Zone	1.3884	5.818	0.239	0.811
PaymentMethod	-2.1819	4.321	-0.505	0.614
Channel	-6.5764	7.444	-0.883	0.377
EducationField	-3.9762	3.370	-1.180	0.238
Gender	15.3947	11.960	1.287	0.198
MaritalStatus	13.0978	7.698	1.701	0.089
Designation	16.4684	6.141	2.682	0.007

As zone has the highest P value we will drop it and see how the model improves.

c) Stats Model 3:

f_3='Agent_bonus_Per_Policy~CustTenure+Channel+EducationField+Gender+Designation+MaritalStatus+Existing PolicyTenure+PaymentMethod+Sum_assured_Per_Policy'

The coefficient, intercept and the OLS model summary for lm3 is shown as below:

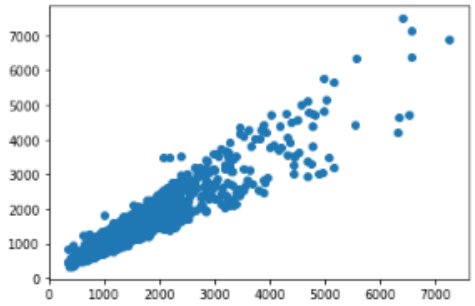
```
Intercept                75.069631
CustTenure                2.813798
Channel                  -6.575002
EducationField           -3.978598
Gender                   15.467191
Designation              16.472657
MaritalStatus            13.189378
ExistingPolicyTenure      4.001438
PaymentMethod            -2.149817
Sum_assured_Per_Policy    0.005782
dtype: float64
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Agent_bonus_Per_Policy   R-squared:                       0.899
Model:                                OLS   Adj. R-squared:                  0.899
Method:                     Least Squares   F-statistic:                     3124.
Date:                    Sat, 24 Dec 2022   Prob (F-statistic):               0.00
Time:                            17:31:23   Log-Likelihood:                -22840.
No. Observations:                    3164   AIC:                         4.570e+04
Df Residuals:                        3154   BIC:                         4.576e+04
Df Model:                               9
Covariance Type:                nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                75.0696     22.496      3.337      0.001      30.961     119.178
CustTenure                2.8138      0.707      3.978      0.000       1.427       4.201
Channel                  -6.5750      7.443     -0.883      0.377     -21.169       8.019
EducationField           -3.9786      3.369     -1.181      0.238     -10.584       2.627
Gender                   15.4672     11.955      1.294      0.196      -7.972      38.907
Designation              16.4727      6.140      2.683      0.007       4.435      28.511
MaritalStatus            13.1894      7.688      1.716      0.086      -1.884      28.262
ExistingPolicyTenure      4.0014      1.218      3.286      0.001       1.614       6.389
PaymentMethod            -2.1498      4.318     -0.498      0.619     -10.616       6.317
Sum_assured_Per_Policy    0.0058   3.61e-05    160.351      0.000       0.006       0.006
==============================================================================
Omnibus:                     1021.489   Durbin-Watson:                   1.950
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6725.349
Skew:                           1.369   Prob(JB):                         0.00
Kurtosis:                       9.597   Cond. No.                     1.11e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.11e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The equation, scatter plot and other error metrics for lm3 are shown as below:

```
Test data root mean square error value is : 342.88779489904476
Train data root mean square error value is : 330.24832877131877
Equation for model is :
(75.07) * Intercept + (2.81) * CustTenure + (-6.58) * Channel + (-3.98) * EducationField + (15.47) * Gender + (16.47) * Designa
tion + (13.19) * MaritalStatus + (4.0) * ExistingPolicyTenure + (-2.15) * PaymentMethod + (0.01) * Sum_assured_Per_Policy +
```



```
The MAE for Train data lm3 model is: 226.69784713357828
The MSE for Train data lm3 model is: 109063.95865624904
The RMSE for Train data lm3 model is 330.24832877131877
The MAPE for Train data lm3 model is 0.16449989145887225
The EVS for Train data lm3 model is 0.8991292673010611
The MAE for Test data lm3 model is: 235.43965099058485
The MSE for Test data lm3 model is: 117572.03989072937
The RMSE for Test data lm3 model is 342.88779489904476
The MAPE for Test data lm3 model is 0.16916244159325244
The EVS for Test data lm3 model is 0.8787854542540674
```

The combined result is shown below:

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785

As seen, The RMSE and other metrics have increased for lm3 model.

From the OLS Regression results of lm3 summary we can see that P value are still high for some features. We need to drop them one by one so that we can see the effect of reducing multi collinearity. If we drop them all together, we risk losing the vital data. The features are as follows in descending order of their P values

	coef	std err	t	P>|t|	[0.025	0.975]
PaymentMethod	-2.1498	4.318	-0.498	0.619	-10.616	6.317
Channel	-6.5750	7.443	-0.883	0.377	-21.169	8.019
EducationField	-3.9786	3.369	-1.181	0.238	-10.584	2.627
Gender	15.4672	11.955	1.294	0.196	-7.972	38.907
MaritalStatus	13.1894	7.688	1.716	0.086	-1.884	28.262
Designation	16.4727	6.140	2.683	0.007	4.435	28.511

We will now drop PaymentMethod and check the result.

d) Stats Model 4:

f_4='Agent_bonus_Per_Policy~CustTenure+Channel+EducationField+Gender+Designation+MaritalStatus+Existing PolicyTenure+Sum_assured_Per_Policy'

The coefficient, intercept and the OLS model summary for lm4 is shown as below:

```
Intercept                72.819220
CustTenure                2.819517
Channel                  -6.521566
EducationField           -3.976237
Gender                   15.404716
Designation              16.420704
MaritalStatus            13.100247
ExistingPolicyTenure      4.016928
Sum_assured_Per_Policy    0.005782
dtype: float64
```

OLS Regression Results

Dep. Variable:	Agent_bonus_Per_Policy	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.899
Method:	Least Squares	F-statistic:	3515.
Date:	Sat, 24 Dec 2022	Prob (F-statistic):	0.00
Time:	17:31:23	Log-Likelihood:	-22840.
No. Observations:	3164	AIC:	4.570e+04
Df Residuals:	3155	BIC:	4.575e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	72.8192	22.035	3.305	0.001	29.615	116.023
CustTenure	2.8195	0.707	3.987	0.000	1.433	4.206
Channel	-6.5216	7.442	-0.876	0.381	-21.112	8.069
EducationField	-3.9762	3.369	-1.180	0.238	-10.581	2.629
Gender	15.4047	11.952	1.289	0.198	-8.031	38.840
Designation	16.4207	6.138	2.675	0.008	4.386	28.455
MaritalStatus	13.1002	7.685	1.705	0.088	-1.967	28.167
ExistingPolicyTenure	4.0169	1.217	3.300	0.001	1.631	6.403
Sum_assured_Per_Policy	0.0058	3.61e-05	160.376	0.000	0.006	0.006

Omnibus:	1022.079	Durbin-Watson:	1.949
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6736.972
Skew:	1.369	Prob(JB):	0.00
Kurtosis:	9.603	Cond. No.	1.09e+06

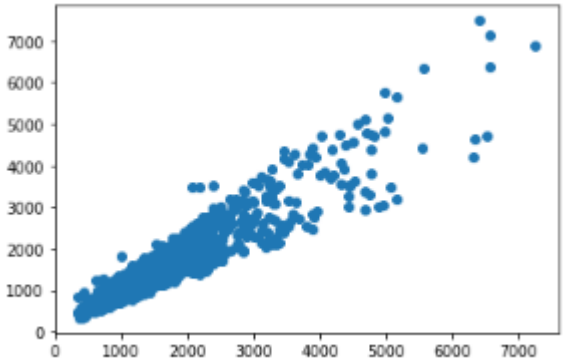
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.09e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

The intercept and coefficients, equation and scatter plot are shown below:

Intercept	72.819220
CustTenure	2.819517
Channel	-6.521566
EducationField	-3.976237
Gender	15.404716
Designation	16.420704
MaritalStatus	13.100247
ExistingPolicyTenure	4.016928
Sum_assured_Per_Policy	0.005782

dtype: float64

```
Test data root mean square error value is : 342.77851512783303
Train data root mean square error value is : 330.26130573549875
Equation for model is :
(72.82) * Intercept + (2.82) * CustTenure + (-6.52) * Channel + (-3.98) * EducationField + (15.4) * Gender + (16.42) * Designat
ion + (13.1) * MaritalStatus + (4.02) * ExistingPolicyTenure + (0.01) * Sum_assured_Per_Policy +
```



The error metrics and the combined result for lm4 model is shown as below:

```
The MAE for Train data lm4 model is: 226.71647599117708
The MSE for Train data lm4 model is: 109072.5300661167
The RMSE for Train data lm4 model is 330.2613057354989
The MAPE for Train data lm4 model is 0.16451943018180268
The EVS for Train data lm4 model is 0.8991213398023322
The MAE for Test data lm4 model is: 235.45440143828287
The MSE for Test data lm4 model is: 117497.11043324205
The RMSE for Test data lm4 model is 342.77851512783303
The MAPE for Test data lm4 model is 0.16925266292699
The EVS for Test data lm4 model is 0.878861629746393
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	109072.530066	117497.110433	0.164519	0.169253	0.899121	0.878862

From the OLS Regression results of lm4 summary we can see that P value are still high for some features. We need to drop them one by one so that we can see the effect of reducing multi collinearity. If we drop them all together, we risk losing the vital data. The features are as follows in descending order of their P values

	coef	std err	t	P>|t|	[0.025	0.975]
Channel	-6.5216	7.442	-0.876	0.381	-21.112	8.069
EducationField	-3.9762	3.369	-1.180	0.238	-10.581	2.629
Gender	15.4047	11.952	1.289	0.198	-8.031	38.840
MaritalStatus	13.1002	7.685	1.705	0.088	-1.967	28.167
Designation	16.4207	6.138	2.675	0.008	4.386	28.455

We will drop Channel and see the results in next model.

e) Stats Model 5:

f_5='Agent_bonus_Per_Policy~CustTenure+EducationField+Gender+Designation+MaritalStatus+ExistingPolicyTenure+Sum_assured_Per_Policy'

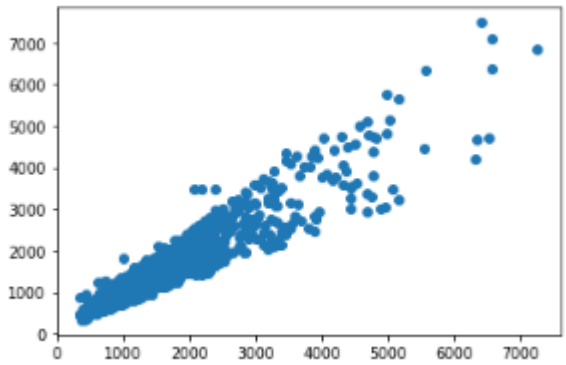
The OLS model summary for lm5 is shown as below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Agent_bonus_Per_Policy   R-squared:                       0.899
Model:                               OLS   Adj. R-squared:                  0.899
Method:                    Least Squares   F-statistic:                     4017.
Date:                   Sat, 24 Dec 2022   Prob (F-statistic):               0.00
Time:                           17:31:24   Log-Likelihood:                -22841.
No. Observations:                   3164   AIC:                         4.570e+04
Df Residuals:                       3156   BIC:                         4.575e+04
Df Model:                              7
Covariance Type:               nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                   69.8785     21.777      3.209      0.001      27.180     112.577
CustTenure                   2.8213      0.707      3.990      0.000       1.435       4.208
EducationField              -4.0560      3.367     -1.205      0.228     -10.658       2.546
Gender                      15.5997     11.950      1.305      0.192      -7.831      39.030
Designation                 16.3615      6.137      2.666      0.008       4.328      28.395
MaritalStatus               12.8985      7.681      1.679      0.093      -2.161      27.958
ExistingPolicyTenure         4.0504      1.216      3.330      0.001       1.665       6.435
Sum_assured_Per_Policy       0.0058   3.61e-05    160.392      0.000       0.006       0.006
==============================================================================
Omnibus:                     1025.136   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6771.505
Skew:                           1.373   Prob(JB):                         0.00
Kurtosis:                       9.620   Cond. No.                     1.08e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.08e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The intercept and coefficients, equation and scatter plot are shown below:

```
Test data root mean square error value is : 342.6067684600315
Train data root mean square error value is : 330.3015012802034
Equation for model is :
(69.88) * Intercept + (2.82) * CustTenure + (-4.06) * EducationField + (15.6) * Gender + (16.36) * Designation + (12.9) * Marit
alStatus + (4.05) * ExistingPolicyTenure + (0.01) * Sum_assured_Per_Policy +
```



The error metrics and the combined result for lm5 model is shown as below:

```
The MAE for Train data lm5 model is: 226.6201794928076
The MSE for Train data lm5 model is: 109099.08174795608
The RMSE for Train data lm5 model is 330.3015012802032
The MAPE for Train data lm5 model is 0.16441814035079433
The EVS for Train data lm5 model is 0.8990967827659425
The MAE for Test data lm5 model is: 235.3869914673677
The MSE for Test data lm5 model is: 117379.39779462545
The RMSE for Test data lm5 model is 342.6067684600312
The MAPE for Test data lm5 model is 0.16928797174593233
The EVS for Test data lm5 model is 0.8789832662622046
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	109072.530066	117497.110433	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	109099.081748	117379.397795	0.164418	0.169288	0.899097	0.878983

As seen, we don't see any significant improvement in stats model results even after dropping the params.

From the OLS Regression results of lm5 summary we can see that P value are still high for some features. We need to drop them one by one so that we can see the effect of reducing multi collinearity. If we drop them all together, we risk losing the vital data. The features are as follows in descending order of their P values

	coef	std err	t	P>|t|	[0.025	0.975]
EducationField	-4.0560	3.367	-1.205	0.228	-10.658	2.546
Gender	15.5997	11.950	1.305	0.192	-7.831	39.030
MaritalStatus	12.8985	7.681	1.679	0.093	-2.161	27.958
Designation	16.3615	6.137	2.666	0.008	4.328	28.395

We will drop EducationField and see the results in next model.

e) Stats Model 6:

f_6='Agent_bonus_Per_Policy~CustTenure+Gender+Designation+MaritalStatus+ExistingPolicyTenure+Sum_assured_Per_Policy'

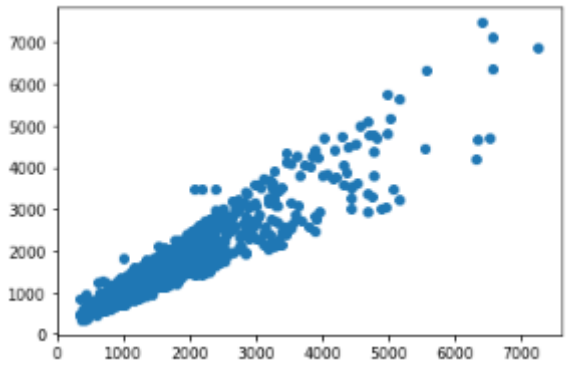
The OLS model summary for lm6 is shown as below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Agent_bonus_Per_Policy   R-squared:                       0.899
Model:                               OLS   Adj. R-squared:                  0.899
Method:                    Least Squares   F-statistic:                     4686.
Date:                   Sat, 24 Dec 2022   Prob (F-statistic):               0.00
Time:                           17:31:25   Log-Likelihood:                -22841.
No. Observations:                   3164   AIC:                         4.570e+04
Df Residuals:                       3157   BIC:                         4.574e+04
Df Model:                              6
Covariance Type:               nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                   58.4473     19.602      2.982      0.003      20.014      96.881
CustTenure                   2.8129      0.707      3.978      0.000       1.426       4.199
Gender                      15.2344     11.947      1.275      0.202      -8.190      38.659
Designation                 16.7137      6.131      2.726      0.006       4.693      28.734
MaritalStatus               12.8321      7.681      1.671      0.095      -2.228      27.893
ExistingPolicyTenure         4.0681      1.216      3.344      0.001       1.683       6.453
Sum_assured_Per_Policy       0.0058   3.61e-05    160.376      0.000       0.006       0.006
==============================================================================
Omnibus:                     1025.425   Durbin-Watson:                   1.948
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6775.048
Skew:                           1.374   Prob(JB):                         0.00
Kurtosis:                       9.621   Cond. No.                     9.89e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.89e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The intercept and coefficients, equation and scatter plot are shown below:

```
Test data root mean square error value is : 342.3681147969265
Train data root mean square error value is : 330.3015012802034
Equation for model is :
(58.45) * Intercept + (2.81) * CustTenure + (15.23) * Gender + (16.71) * Designation + (12.83) * MaritalStatus + (4.07) * Exist
ingPolicyTenure + (0.01) * Sum_assured_Per_Policy +
```



The error metrics and the combined result for lm6 model is shown as below:

```
The MAE for Train data lm6 model is: 226.6934391055961
The MSE for Train data lm6 model is: 109149.23696749142
The RMSE for Train data lm6 model is 330.3774159465072
The MAPE for Train data lm6 model is 0.16444770433918315
The EVS for Train data lm6 model is 0.8990503953634904
The MAE for Test data lm6 model is: 235.25029423607177
The MSE for Test data lm6 model is: 117215.92602960153
The RMSE for Test data lm6 model is 342.36811479692665
The MAPE for Test data lm6 model is 0.16917569531090462
The EVS for Test data lm6 model is 0.8791532396992957
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997087	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997087	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	109072.530066	117497.110433	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	109099.081748	117379.397795	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	109149.236967	117215.926030	0.164448	0.169176	0.899050	0.879153

From the OLS Regression results of lm6 summary we can see that P value are still high for some features. We need to drop them one by one so that we can see the effect of reducing multi collinearity. If we drop them all together, we risk losing the vital data. The features are as follows in descending order of their P values

	coef	std err	t	P>|t|	[0.025	0.975]
Gender	15.2344	11.947	1.275	0.202	-8.190	38.659
MaritalStatus	12.8321	7.681	1.671	0.095	-2.228	27.893
Designation	16.7137	6.131	2.726	0.006	4.693	28.734

We can remove Gender and check now based on P value.

f) Stats Model 7:

f_7='Agent_bonus_Per_Policy~CustTenure+Designation+MaritalStatus+ExistingPolicyTenure+Sum_assured_Per_Policy'

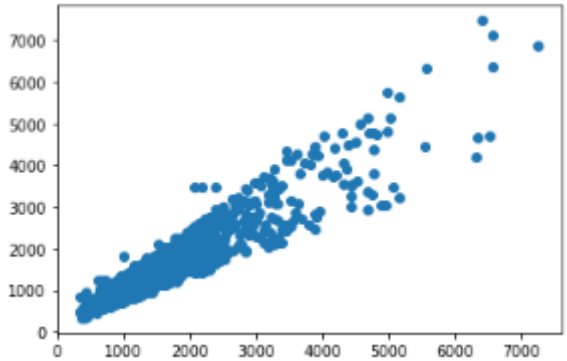
The OLS model summary for lm7 is shown as below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Agent_bonus_Per_Policy   R-squared:                       0.899
Model:                               OLS   Adj. R-squared:                  0.899
Method:                    Least Squares   F-statistic:                     5622.
Date:                   Sat, 24 Dec 2022   Prob (F-statistic):               0.00
Time:                           17:31:25   Log-Likelihood:                -22842.
No. Observations:                   3164   AIC:                         4.570e+04
Df Residuals:                       3158   BIC:                         4.573e+04
Df Model:                              5
Covariance Type:               nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                   67.8821     18.154      3.739      0.000      32.288     103.476
CustTenure                   2.8043      0.707      3.966      0.000       1.418       4.191
Designation                 16.7111      6.131      2.725      0.006       4.689      28.733
MaritalStatus               12.5966      7.680      1.640      0.101      -2.461      27.654
ExistingPolicyTenure         4.0174      1.216      3.304      0.001       1.633       6.401
Sum_assured_Per_Policy       0.0058   3.61e-05    160.390      0.000       0.006       0.006
==============================================================================
Omnibus:                     1025.150   Durbin-Watson:                   1.948
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6757.041
Skew:                           1.374   Prob(JB):                         0.00
Kurtosis:                       9.611   Cond. No.                     8.98e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.98e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The intercept and coefficients, equation and scatter plot are shown below:

```
Test data root mean square error value is : 342.11049231742624
Train data root mean square error value is : 330.4624879794362
Equation for model is :
(67.88) * Intercept + (2.8) * CustTenure + (16.71) * Designation + (12.6) * MaritalStatus + (4.02) * ExistingPolicyTenure + (0.
01) * Sum_assured_Per_Policy +
```



The error metrics and the combined result for lm7 model is shown as below:

```
The MAE for Train data lm7 model is: 226.84574221153798
The MSE for Train data lm7 model is: 109205.4559615589
The RMSE for Train data lm7 model is 330.462487979436
The MAPE for Train data lm7 model is 0.1646053342717496
The EVS for Train data lm7 model is 0.8989983997162294
The MAE for Test data lm7 model is: 234.97192212432788
The MSE for Test data lm7 model is: 117039.5889536717
The RMSE for Test data lm7 model is 342.1104923174262
The MAPE for Test data lm7 model is 0.16876242684882808
The EVS for Test data lm7 model is 0.8793478649283756
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	109072.530066	117497.110433	0.164519	0.169253	0.899121	0.878882
Stats Model 5	330.301501	342.606768	226.620179	235.386991	109099.081748	117379.397795	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	109149.236967	117215.926030	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	109205.455962	117039.588954	0.164605	0.168762	0.898998	0.879348

From the OLS Regression results of lm7 summary we can see that P value are still high for some features. We need to drop them one by one so that we can see the effect of reducing multi collinearity. If we drop them all together, we risk losing the vital data. The features are as follows in descending order of their P values

	coef	std err	t	P>|t|	[0.025	0.975]
MaritalStatus	12.5966	7.680	1.640	0.101	-2.461	27.654
Designation	16.7111	6.131	2.725	0.006	4.689	28.733

We will now drop MaritalStatus and check the P values

g) Stats Model 8:

f_8 = 'Agent_bonus_Per_Policy~CustTenure+Designation+ExistingPolicyTenure+Sum_assured_Per_Policy'

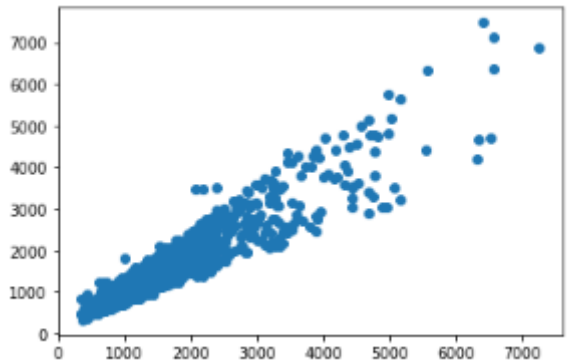
The OLS model summary for lm7 is shown as below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Agent_bonus_Per_Policy   R-squared:                       0.899
Model:                               OLS   Adj. R-squared:                  0.899
Method:                    Least Squares   F-statistic:                     7023.
Date:                   Sat, 24 Dec 2022   Prob (F-statistic):               0.00
Time:                           17:31:26   Log-Likelihood:                -22844.
No. Observations:                   3164   AIC:                         4.570e+04
Df Residuals:                       3159   BIC:                         4.573e+04
Df Model:                              4
Covariance Type:               nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                     84.0705     15.240      5.516      0.000      54.188     113.953
CustTenure                     2.7360      0.706      3.875      0.000       1.352       4.121
Designation                   16.1432      6.123      2.636      0.008       4.137      28.149
ExistingPolicyTenure           3.9385      1.215      3.241      0.001       1.556       6.321
Sum_assured_Per_Policy         0.0058    3.6e-05    160.899      0.000       0.006       0.006
==============================================================================
Omnibus:                     1028.648   Durbin-Watson:                   1.947
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6774.277
Skew:                           1.380   Prob(JB):                         0.00
Kurtosis:                       9.616   Cond. No.                     7.43e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.43e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The intercept and coefficients, equation and scatter plot are shown below:

```
Test data root mean square error value is : 342.0499544096864
Train data root mean square error value is : 330.60322459242826
Equation for model is :
(84.07) * Intercept + (2.74) * CustTenure + (16.14) * Designation + (3.94) * ExistingPolicyTenure + (0.01) * Sum_assured_Per_Po
licy +
```



The error metrics and the combined result for lm7 model is shown as below:

```
The MAE for Train data lm8 model is: 226.82411630302252
The MSE for Train data lm8 model is: 109298.49211091165
The RMSE for Train data lm8 model is 330.6032245924284
The MAPE for Train data lm8 model is 0.16438380152557963
The EVS for Train data lm8 model is 0.8989123527336301
The MAE for Test data lm8 model is: 234.77372287017386
The MSE for Test data lm8 model is: 116998.17131166876
The RMSE for Test data lm8 model is 342.04995440968673
The MAPE for Test data lm8 model is 0.16828681121074118
The EVS for Test data lm8 model is 0.8794001065422812
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	109072.530066	117497.110433	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	109099.081748	117379.397795	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	109149.236967	117215.926030	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	109205.455962	117039.588954	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	109298.492111	116998.171312	0.164384	0.168287	0.898912	0.879400

From the OLS Regression results of lm8 summary we can see that P value are still high for some features. We need to drop them one by one so that we can see the effect of reducing multi collinearity. If we drop them all together, we risk losing the vital data. The features are as follows in descending order of their P values

	coef	std err	t	P>|t|	[0.025	0.975]
Designation	16.1432	6.123	2.636	0.008	4.137	28.149

We will now finally drop Designation and check P value

h) Stats Model 9:

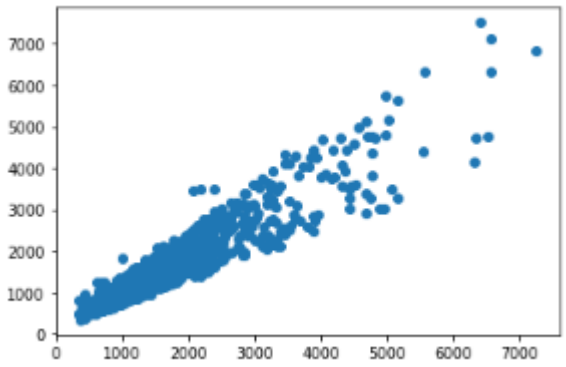
f_9 = 'Agent_bonus_Per_Policy~CustTenure+ExistingPolicyTenure+Sum_assured_Per_Policy'
The OLS model summary for lm7 is shown as below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Agent_bonus_Per_Policy   R-squared:                       0.899
Model:                               OLS   Adj. R-squared:                  0.899
Method:                    Least Squares   F-statistic:                     9344.
Date:                   Sat, 24 Dec 2022   Prob (F-statistic):               0.00
Time:                           17:31:27   Log-Likelihood:                -22847.
No. Observations:                   3164   AIC:                         4.570e+04
Df Residuals:                       3160   BIC:                         4.573e+04
Df Model:                              3
Covariance Type:               nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept        107.1357     12.491      8.577      0.000      82.645     131.627
CustTenure         2.9594      0.702      4.218      0.000       1.584       4.335
ExistingPolicyTenure       4.0234      1.216      3.309      0.001       1.639       6.408
Sum_assured_Per_Policy     0.0058   3.59e-05    161.195      0.000       0.006       0.006
==============================================================================
Omnibus:                     1032.033   Durbin-Watson:                   1.947
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6901.383
Skew:                           1.380   Prob(JB):                         0.00
Kurtosis:                       9.688   Cond. No.                     5.92e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.92e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The intercept and coefficients, equation and scatter plot are shown below:

```
Test data root mean square error value is : 341.5257294612619
Train data root mean square error value is : 330.9667203173249
Equation for model is :
(107.14) * Intercept + (2.96) * CustTenure + (4.02) * ExistingPolicyTenure + (0.01) * Sum_assured_Per_Policy +
```



The error metrics and the combined result for lm7 model is shown as below:

```
The MAE for Train data lm9 model is: 226.76699744512692
The MSE for Train data lm9 model is: 109538.96995760659
The RMSE for Train data lm9 model is 330.96672031732527
The MAPE for Train data lm9 model is 0.164434435840662
The EVS for Train data lm9 model is 0.8986899403355029
The MAE for Test data lm9 model is: 234.06036375397642
The MSE for Test data lm9 model is: 116639.82388404713
The RMSE for Test data lm9 model is 341.525729461262
The MAPE for Test data lm9 model is 0.16774397811459646
The EVS for Test data lm9 model is 0.8797840273969011
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	109072.530066	117497.110433	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	109099.081748	117379.397795	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	109149.236967	117215.926030	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	109205.455962	117039.588954	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	109298.492111	116998.171312	0.164384	0.168287	0.898912	0.879400
Stats Model 9	330.966720	341.525729	226.766997	234.060364	109538.969958	116639.823884	0.164434	0.167744	0.898690	0.879784

From the OLS Regression results of lm9 summary we can see that P value are below 0.05 for all features. However, there is no improvement in the RMSE, MAE, MSE, MAPE and EVS. As a matter of fact, the error metrics increased as we dropped more and more features. The final result of our stats model linear regression can be summarised as follows:

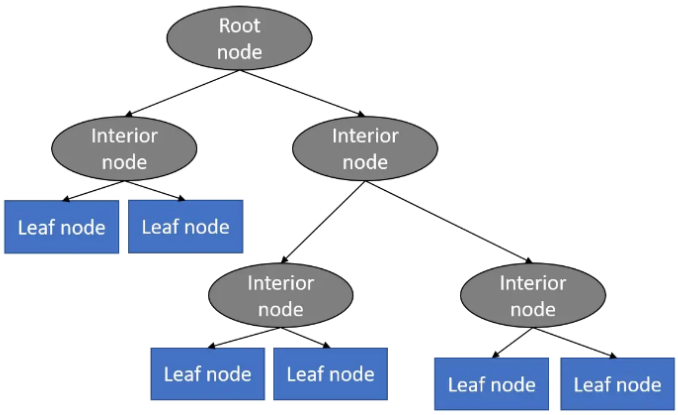
```
(107.14) * Intercept + (2.96) * CustTenure + (4.02) * ExistingPolicyTenure + (0.01) * Sum_assured_Per_Policy +
```

As, we didn’t see any significant improvement in our stats model we will move on to other models and evaluate the results.

7.Model 3: Decision Tree Regressor

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.



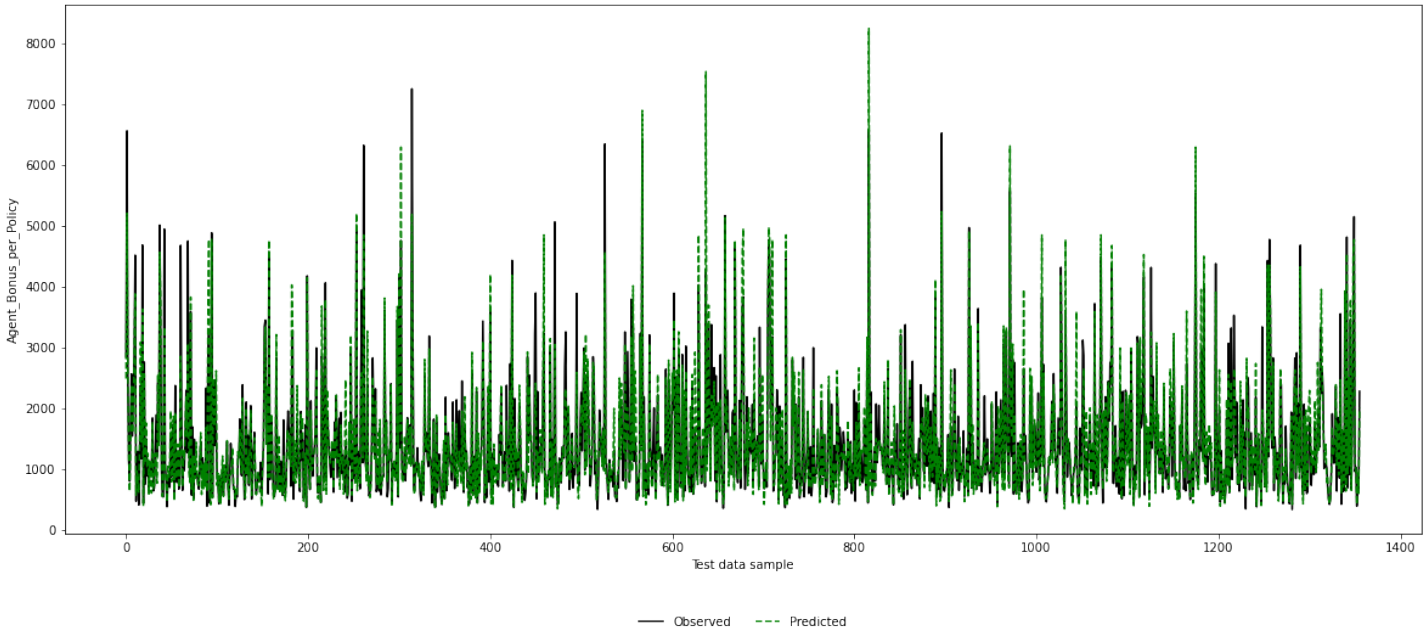
After training the model on to Decision Tree Regressor, we can now see the error metric and an overall combined result for better evaluation.

```
The MAE of train data DT model is: 0.0
The MSE of train data DT model is: 0.0
The MAPE of train data DT modelis 0.0
The EVS of train data DT model is 1.0
The RMSE of train data DT model is:0.00
The MAE for Test data DT model is: 235.78820058997053
The MSE for Test data DT model is: 138710.37781588003
The MAPE for Test data DT model is 0.16840953644501427
The EVS for Test data DT model is 0.8567971942205577
The RMSE for Test data DT model is:372.44
```

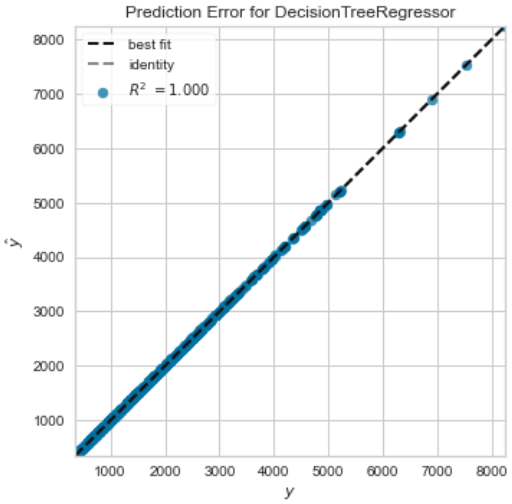
	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	109072.530066	117497.110433	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	109099.081748	117379.397795	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	109149.236967	117215.926030	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	109205.455962	117039.588954	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	109298.492111	116998.171312	0.164384	0.168287	0.898912	0.879400
Stats Model 9	330.966720	341.525729	226.766997	234.060364	109538.969958	116639.823884	0.164434	0.167744	0.898690	0.879784
DT Model	0.000000	372.438422	0.000000	235.788201	0.000000	138710.377816	0.000000	0.168410	1.000000	0.856797

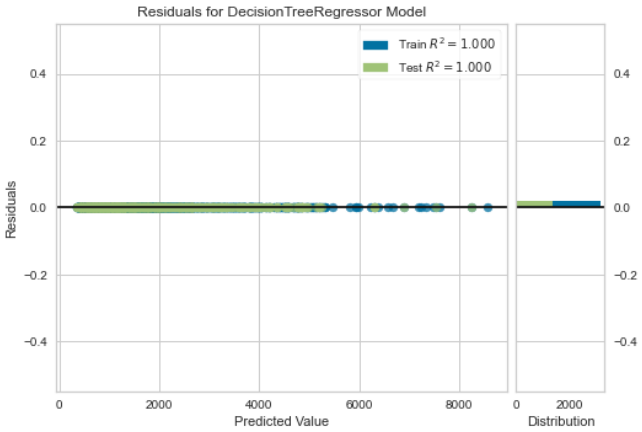
As we can see the train error metrics are all zero and this is normal because Decision Tree model tends to cover all the data points present in the training set and hence gives zero error.

The observed and predicted data can be visualized as per the following plot.



The prediction error and Residuals for DT Regressor Model can be seen as below:

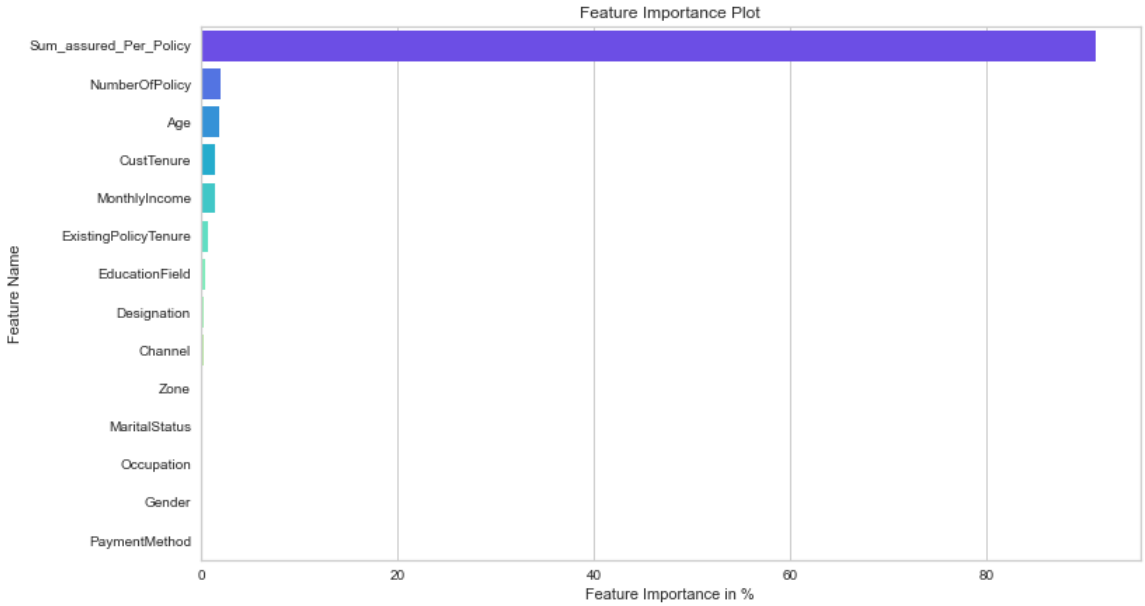




As seen from the above plot the prediction error and residual is nil and R^2 value is 1.0 which means there is no error.

The feature importance of a DT regressor model can be seen as below:

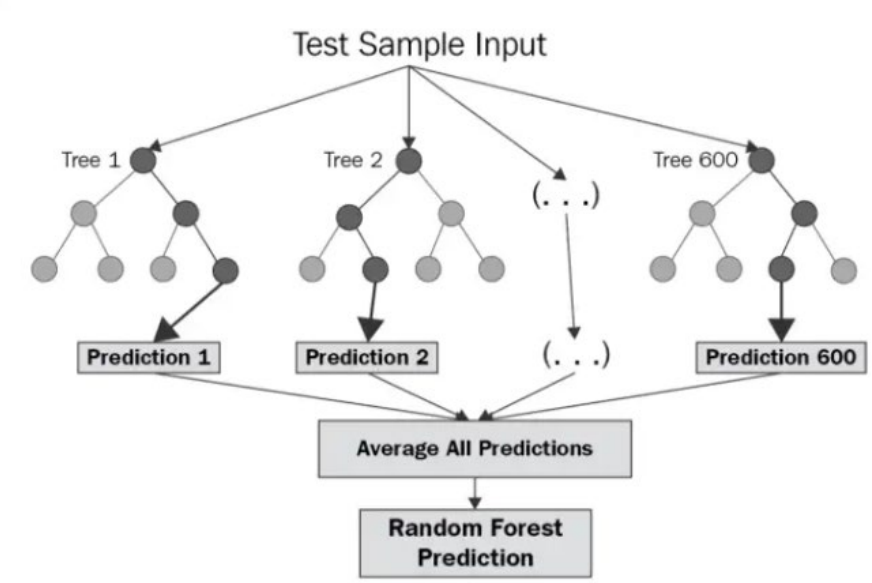
```
Sum_assured_Per_Policy    0.910520
NumberOfPolicy            0.019860
Age                       0.017973
CustTenure                0.014652
MonthlyIncome             0.014398
ExistingPolicyTenure      0.006560
EducationField            0.004192
Designation               0.002858
Channel                   0.002621
Zone                      0.001597
MaritalStatus             0.001435
Occupation                0.001339
Gender                    0.001191
PaymentMethod             0.000805
dtype: float64
```



From the above plots we can see that DT model is train error metrics of zero but it does have test error metrics. The prediction error and the residual error plot also shows that there is very minimal error and looking at the feature importance plot we can see that agent bonus per policy is significantly dependent on sum assured which impacts the target variable almost 91% while all other features have a significance of less than 2%.

8.Model 4: Random Forest Regressor

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

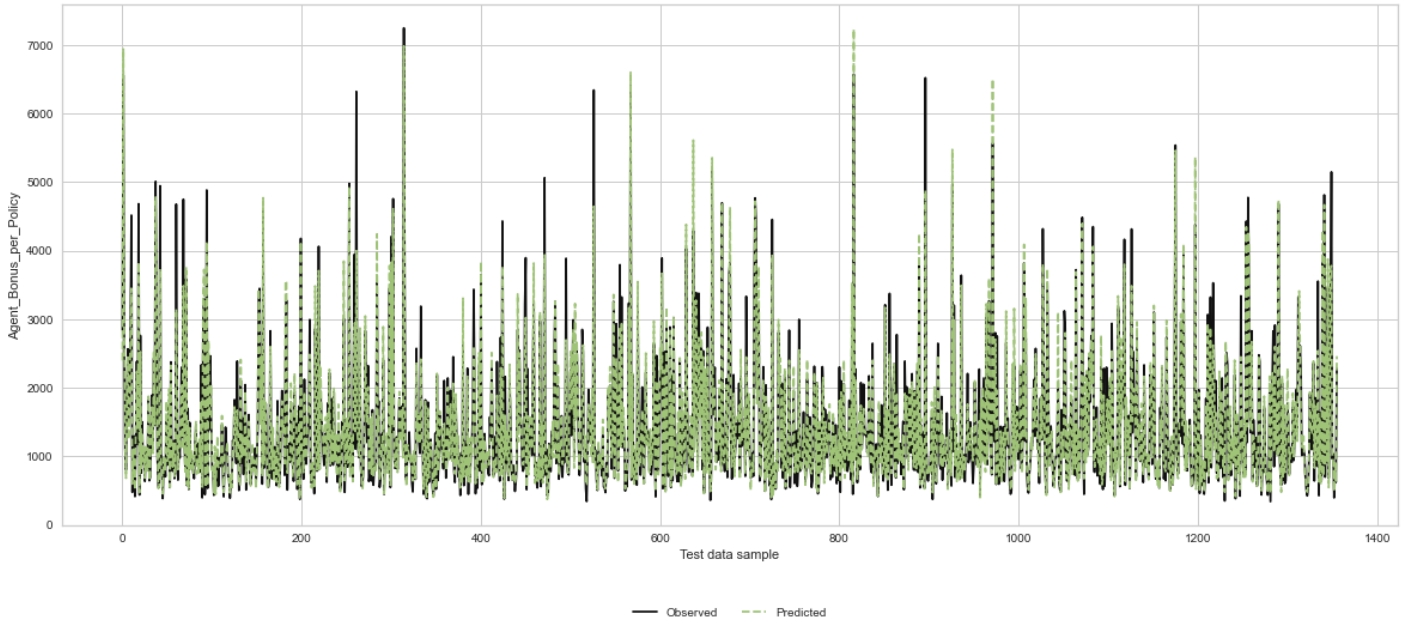


After training the model on to Random Forest Regressor with number of estimators as 30 and random state as 30, we can now see the error metric and an overall combined result for better evaluation.

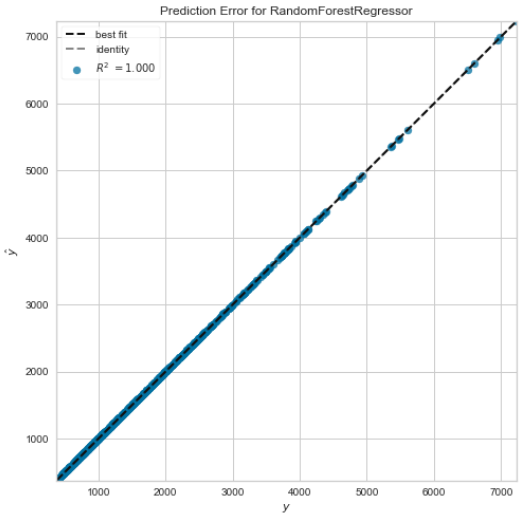
```
The MAE of train data RF model is: 66.05605509903076
The MSE of train data RF model is: 10784.346290070485
The MAPE of train data RF modelis 0.04772033093621116
The EVS of train data RF model is 0.9900261210618498
The RMSE of train data RF model is:103.85
The MAE for Test data RF model is: 178.60636553588986
The MSE for Test data RF model is: 74328.66337661126
The MAPE for Test data RF model is 0.12766478957305577
The EVS for Test data RF model is 0.9232642228832235
The RMSE for Test data RF model is:272.63
```

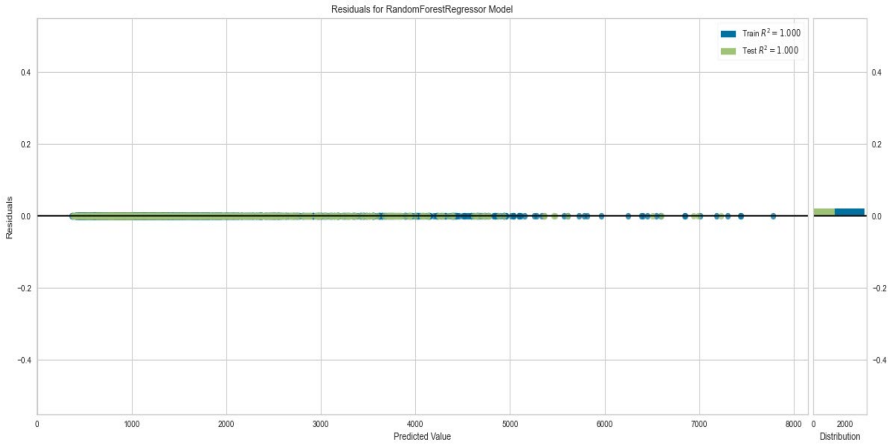
	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	92286.650764	97166.625616	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	109061.988580	117583.957965	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	109063.958656	117572.039891	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	109072.530066	117497.110433	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	109099.081748	117379.397795	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	109149.236967	117215.926030	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	109205.455962	117039.588954	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	109298.492111	116998.171312	0.164384	0.168287	0.898912	0.879400
Stats Model 9	330.966720	341.525729	226.766997	234.060364	109538.969958	116639.823884	0.164434	0.167744	0.898690	0.879784
DT Model	0.000000	372.438422	0.000000	235.788201	0.000000	138710.377816	0.000000	0.168410	1.000000	0.856797
RF Model	103.847707	272.632836	66.056055	178.606366	10784.346290	74328.663377	0.047720	0.127665	0.990026	0.923264

The observed and predicted data can be visualized as per the following plot.



The prediction error and Residuals for DT Regressor Model can be seen as below:

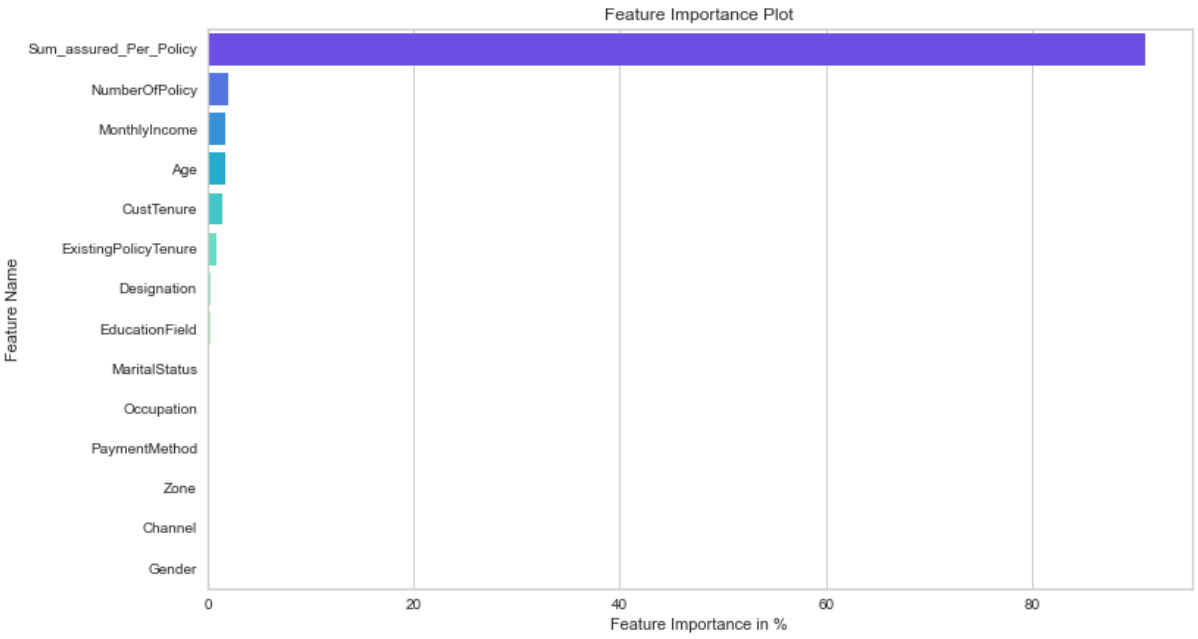




As seen from the above plot the prediction error and residual is nil and R^2 value is 1.0 which means there is no error.

The feature importance of a RF regressor model can be seen as below:

```
Sum_assured_Per_Policy    0.909266
NumberOfPolicy            0.019961
MonthlyIncome             0.017461
Age                       0.017131
CustTenure                0.014817
ExistingPolicyTenure      0.008367
Designation               0.002548
EducationField            0.002300
MaritalStatus             0.001663
Occupation                0.001471
PaymentMethod             0.001436
Zone                      0.001236
Channel                   0.001230
Gender                    0.001113
dtype: float64
```



From the above plots we can see that RF model the following error metrics:

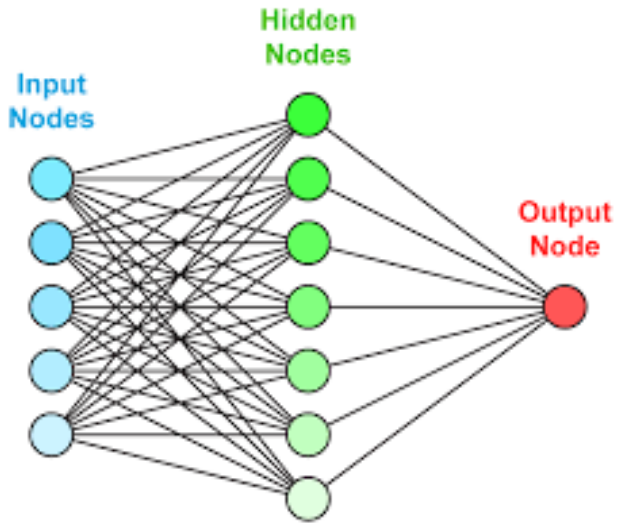
Train RMSE : 103.847707
Test RMSE : 272.632836
Train MAE : 66.056055
Test MAE : 178.606366
Train MSE : 10784.346290
Test MSE : 74328.663377
Train MAPE : 0.047720
Test MAPE : 0.127665
Train EVS : 0.990026
Test EVS : 0.923264

The prediction error and the residual error plot also shows that there is very minimal error and looking at the feature importance plot we can see that agent bonus per policy is significantly dependent on sum assured which impacts the target variable almost 90% while all other features have a significance of less than 2%

RF model is by far the best regression model in terms of its RMSE value and MAPE value. We can also consider EVS score and confirm that it is the best model so far with a highest EVS test score of 92.3%

9.Model 5 : Artificial Neural Network Regressor

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes and improve continuously. Thus, artificial neural networks attempt to solve complicated problems, like summarizing documents or recognizing faces, with greater accuracy.



We cannot build ANN model without scaling of data. Hence, we need to do scaling first. After scaling we can fit the train and test data and find out the best grid parameters and best grid annunciators shown as below:

```
{'activation': 'logistic',
 'hidden_layer_sizes': 100,
 'max_iter': 250,
 'solver': 'sgd',
 'tol': 0.1}
```

```
MLPRegressor(activation='logistic', hidden_layer_sizes=100, max_iter=250,
             random_state=1, solver='sgd', tol=0.1)
```

The results are shown as below for the ANN Regressor model:

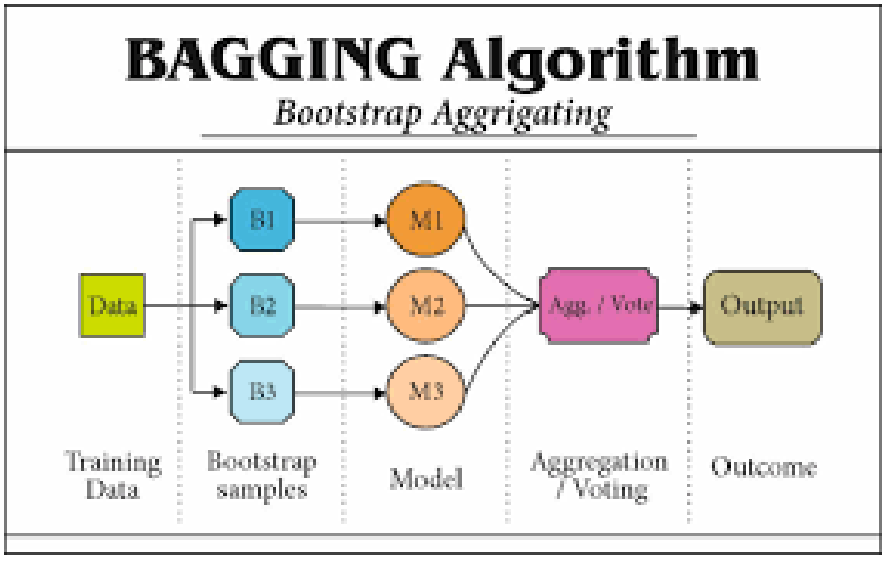
```
The MAE of Train data for ANN Regressor  is: 5338.11874733934
The MSE of Train data for ANN Regressor is: 29561248.54824166
The MAPE of Train data for ANN Regressor is 5.5930581961842405
The EVS of Train data for ANN Regressor is -0.029718853547272106
The RMSE of Train data for ANN Regressor is:5437.03
The MAE of Test data for ANN Regressor is: 5358.526912267968
The MSE of Test data for ANN Regressor is: 29708039.241327155
The MAPE of Test data for ANN Regressor is 5.633304412658901
The EVS of Test data for ANN Regressor is -0.034118430913623 32
The RMSE of Test data for ANN Regressor is:5450.51
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	1.090620e+05	1.175840e+05	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	1.090640e+05	1.175720e+05	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	1.090725e+05	1.174971e+05	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	1.090991e+05	1.173794e+05	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	1.091492e+05	1.172159e+05	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	1.092055e+05	1.170396e+05	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	1.092985e+05	1.169982e+05	0.164384	0.168287	0.898912	0.879400
Stats Model 9	330.966720	341.525729	226.766997	234.060364	1.095390e+05	1.166398e+05	0.164434	0.167744	0.898690	0.879784
DT Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797
RF Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
ANN Model	5437.025708	5450.508164	5338.118747	5358.526912	2.956125e+07	2.970804e+07	5.593058	5.633304	-0.029719	-0.034118

As seen ANN Model has performed very poorly on the data with highest RMSE and other errors so far.

10.Model 6: Bagging regressor

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once. After several data samples are generated, these weak models are then trained independently, and depending on the type of task—regression or classification, for example—the average or majority of those predictions yield a more accurate estimate.



After training the model on a Bagging Regressor at a random state of 1 we can predict the errors as follows:

```
The MAE of Train data for Bagging Regressor  is: 70.18071533923303
The MSE of Train data for Bagging Regressor is: 12984.21555630531
The MAPE of Train data for Bagging Regressor is 0.05027091654714961
The EVS of Train data for Bagging Regressor is 0.9879930438286032
The RMSE of Train data for Bagging Regressor is:113.95
The MAE of Test data for Bagging Regressor is: 187.17977753195674
The MSE of Test data for Bagging Regressor is: 83313.32376233407
The MAPE of Test data for Bagging Regressor is 0.13316826159510783
The EVS of Test data for Bagging Regressor is 0.9139888112128083
The RMSE of Test data for Bagging Regressor is:288.64
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	9.228865e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	9.228865e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	1.090620e+05	1.175840e+05	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	1.090640e+05	1.175720e+05	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	1.090725e+05	1.174971e+05	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	1.090991e+05	1.173794e+05	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	1.091492e+05	1.172159e+05	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	1.092055e+05	1.170396e+05	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	1.092985e+05	1.169982e+05	0.164384	0.168287	0.898912	0.879400
Stats Model 9	330.966720	341.525729	226.766997	234.060364	1.095390e+05	1.166398e+05	0.164434	0.167744	0.898690	0.879784
DT Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797
RF Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
ANN Model	5437.025708	5450.508164	5338.118747	5358.526912	2.956125e+07	2.970804e+07	5.593058	5.633304	-0.029719	-0.034118
Bagging Model	113.948302	288.640475	70.180715	187.179778	1.298422e+04	8.331332e+04	0.050271	0.133168	0.987993	0.913989

The bagging Regressor has done a much better job after and has the second best RMSE, MAPE values after RF Regressor model.

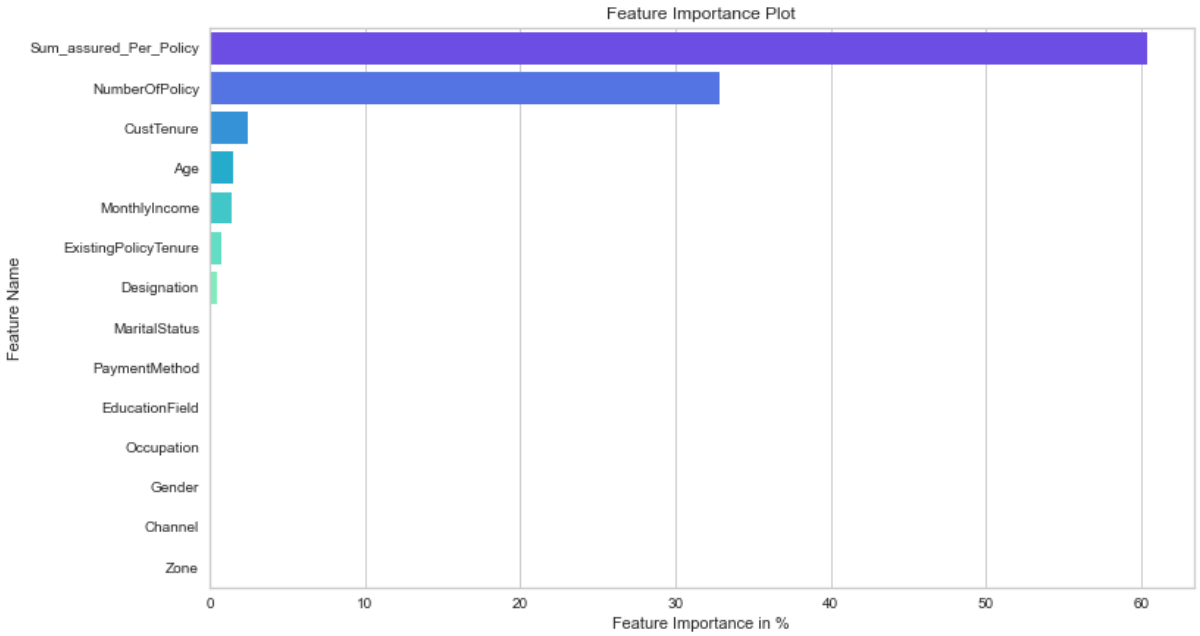
11.Model 7: Random Forest Regressor with hyper Tuned Parameters

In this model we will try to tune RF Regressor model with hyper parameters and see if there is any improvement in the results.

The best parameters are shown as follows along with the feature importance.

```
{'max_depth': 10,
 'max_features': 5,
 'min_samples_leaf': 30,
 'min_samples_split': 100,
 'n_estimators': 100}
```

	Imp
Sum_assured_Per_Policy	0.603552
NumberOfPolicy	0.328542
CustTenure	0.024523
Age	0.015215
MonthlyIncome	0.014059
ExistingPolicyTenure	0.008077
Designation	0.004508
MaritalStatus	0.000723
PaymentMethod	0.000248
EducationField	0.000162
Occupation	0.000109
Gender	0.000103
Channel	0.000100
Zone	0.000079



The results can be seen as follows:

```
The MAE for Train data of RF Tuned Model is: 66.05605509903076
The MSE for Train data of RF Tuned Model is: 10784.346290070485
The MAPE for Train data of RF Tuned Model is 0.04772033093621116
The EVS for Train data of RF Tuned Model is 0.9900261210618498
The RMSE for Train data of RF Tuned Model is:103.85
The MAE for Test data of RF Tuned Model is: 178.60636553588986
The MSE for Test data of RF Tuned Model is: 74328.66337661126
The MAPE for Test data of RF Tuned Model is 0.12766478957305577
The EVS for Test data of RF Tuned Model is 0.9232642228832235
The RMSE for Test data of RF Tuned Model is:272.63
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	1.090620e+05	1.175840e+05	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	1.090640e+05	1.175720e+05	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	1.090725e+05	1.174971e+05	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	1.090991e+05	1.173794e+05	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	1.091492e+05	1.172159e+05	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	1.092055e+05	1.170396e+05	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	1.092985e+05	1.169982e+05	0.164384	0.168287	0.898912	0.879400
Stats Model 9	330.966720	341.525729	226.766997	234.060364	1.095390e+05	1.166398e+05	0.164434	0.167744	0.898690	0.879784
DT Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797
RF Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
ANN Model	5437.025708	5450.508164	5338.118747	5358.526912	2.956125e+07	2.970804e+07	5.593058	5.633304	-0.029719	-0.034118
Bagging Model	113.948302	288.640475	70.180715	187.179778	1.298422e+04	8.331332e+04	0.050271	0.133168	0.987993	0.913989
Random Forest Hyper Tuned Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264

Random Forest Hyper Tuned Model and Random Forest model have given exactly similar results irrespective of hyper parameter tuning and have performed best so far with RMSE of 103.84 and Test MAPE of 92.3%.

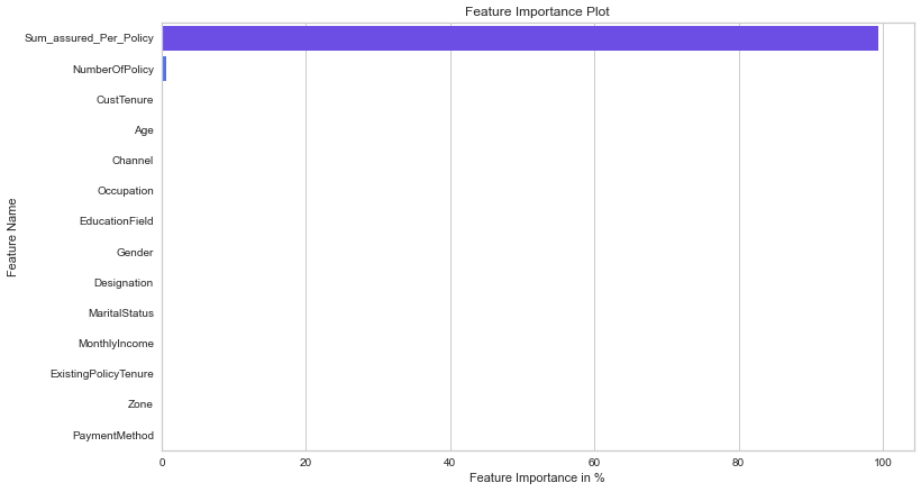
12.Model 8: Decision Tree Regressor with hyper Tuned Parameters

In this model we will try to tune DT Regressor model with hyper parameters and see if there is any improvement in the results.

The best parameters are shown as follows along with the feature importance.

```
{'criterion': 'squared_error', 'max_depth': 8, 'min_samples_leaf': 100, 'min_samples_split': 300}
```

	Imp
Sum_assured_Per_Policy	0.993201
NumberOfPolicy	0.006548
CustTenure	0.000251
Age	0.000000
Channel	0.000000
Occupation	0.000000
EducationField	0.000000
Gender	0.000000
Designation	0.000000
MaritalStatus	0.000000
MonthlyIncome	0.000000
ExistingPolicyTenure	0.000000
Zone	0.000000
PaymentMethod	0.000000



The final results are shown as below:

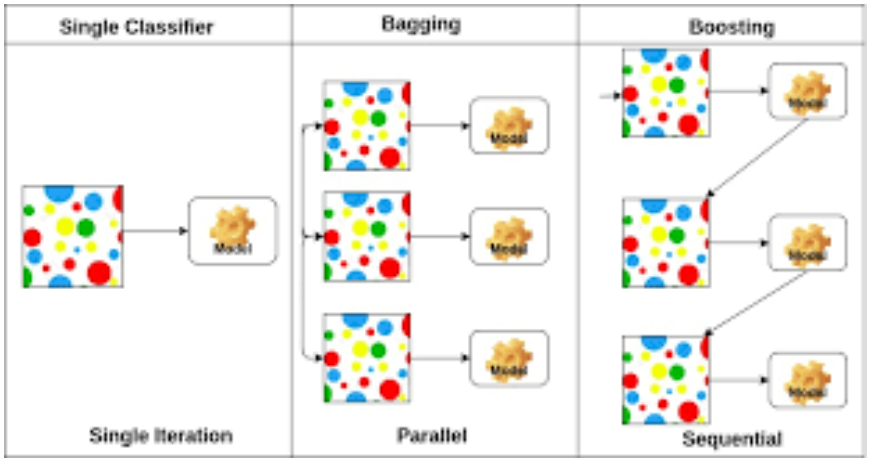
```
The MAE for Train data of DT Tuned Model is: 0.0
The MSE for Train data of DT Tuned Model is: 0.0
The MAPE for Train data of DT Tuned Model is 0.0
The EVS for Train data of DT Tuned Model is 1.0
The RMSE for Train data of DT Tuned Model is:0.00
The MAE for Test data of DT Tuned Model is: 235.78820058997053
The MSE for Test data of DT Tuned Model is: 138710.37781588003
The MAPE for Test data of DT Tuned Model is 0.16840953644501427
The EVS for Test data of DT Tuned Model is 0.8567971942205577
The RMSE for Test data of DT Tuned Model is:372.44
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	1.090620e+05	1.175840e+05	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	1.090640e+05	1.175720e+05	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	1.090725e+05	1.174971e+05	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	1.090991e+05	1.173794e+05	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	1.091492e+05	1.172159e+05	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	1.092055e+05	1.170396e+05	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	1.092985e+05	1.169982e+05	0.164384	0.168287	0.898912	0.879400
Stats Model 9	330.966720	341.525729	226.766997	234.060364	1.095390e+05	1.166398e+05	0.164434	0.167744	0.898690	0.879784
DT Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797
RF Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
ANN Model	5437.025708	5450.508164	5338.118747	5358.526912	2.956125e+07	2.970804e+07	5.593058	5.633304	-0.029719	-0.034118
Bagging Model	113.948302	288.640475	70.180715	187.179778	1.298422e+04	8.331332e+04	0.050271	0.133168	0.987993	0.913989
Random Forest Hyper Tuned Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
Decision Tree Hyper Tuned Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797

we can see similar result with Decision tree regressor model with hyper tuned paramenter as was seen with Decision tree regressor model.

13.Model 9: AdaBosst Regressor

AdaBoost is short for Adaptive Boosting and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”. AdaBoost uses an iterative approach to learn from the mistakes of weak classifiers, and turn them into strong ones.



After training the model on AdaBoost Regressor we can see the results and compare them.

```
The MAE of Train data for AdaBoost Regressor  is: 248.63571918180824
The MSE of Train data for AdaBoost Regressor is: 103462.07083501344
The MAPE of Train data for AdaBoost Regressor is 0.21248886873716744
The EVS of Train data for AdaBoost Regressor is 0.9105240611914388
The RMSE of Train data for AdaBoost Regressor is:321.66
The MAE of Test data for AdaBoost Regressor is: 265.5191111171686
The MSE of Test data for AdaBoost Regressor is: 123259.82142867075
The MAPE of Test data for AdaBoost Regressor is 0.2232065331894396
The EVS of Test data for AdaBoost Regressor is 0.8790883559204712
The RMSE of Test data for AdaBoost Regressor is:351.08
```

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 1	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 2	330.245346	342.905173	226.662989	235.436402	1.090620e+05	1.175840e+05	0.164440	0.169162	0.899131	0.878772
Stats Model 3	330.248329	342.887795	226.697847	235.439651	1.090640e+05	1.175720e+05	0.164500	0.169162	0.899129	0.878785
Stats Model 4	330.261306	342.778515	226.716476	235.454401	1.090725e+05	1.174971e+05	0.164519	0.169253	0.899121	0.878862
Stats Model 5	330.301501	342.606768	226.620179	235.386991	1.090991e+05	1.173794e+05	0.164418	0.169288	0.899097	0.878983
Stats Model 6	330.377416	342.368115	226.693439	235.250294	1.091492e+05	1.172159e+05	0.164448	0.169176	0.899050	0.879153
Stats Model 7	330.462488	342.110492	226.845742	234.971922	1.092055e+05	1.170396e+05	0.164605	0.168762	0.898998	0.879348
Stats Model 8	330.603225	342.049954	226.824116	234.773723	1.092985e+05	1.169982e+05	0.164384	0.168287	0.898912	0.879400
Stats Model 9	330.966720	341.525729	226.766997	234.060364	1.095390e+05	1.166398e+05	0.164434	0.167744	0.898690	0.879784
DT Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797
RF Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
ANN Model	5437.025708	5450.508164	5338.118747	5358.526912	2.956125e+07	2.970804e+07	5.593058	5.633304	-0.029719	-0.034118
Bagging Model	113.948302	288.640475	70.180715	187.179778	1.298422e+04	8.331332e+04	0.050271	0.133168	0.987993	0.913989
Random Forest Hyper Tuned Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
Decision Tree Hyper Tuned Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797
AdaBoost Regressor Model	321.655205	351.083781	248.635719	265.519111	1.034621e+05	1.232598e+05	0.212489	0.223207	0.910524	0.879088

The AdaBoost Regressor model has not performed as well as RF model with a RMSE of 351.08 and MAPE of 22.3%.

14.Comparing all these models and selecting the best one

Sorting by RMSE or MAPE gives us the following results

	Train RMSE	Test RMSE	Train MAE	Test MAE	Train MSE	Test MSE	Train MAPE	Test MAPE	Train EVS	Test EVS
Random Forest Hyper Tuned Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
RF Model	103.847707	272.632836	66.056055	178.606366	1.078435e+04	7.432866e+04	0.047720	0.127665	0.990026	0.923264
Bagging Model	113.948302	288.640475	70.180715	187.179778	1.298422e+04	8.331332e+04	0.050271	0.133168	0.987993	0.913989
Stats Model 1	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Linear Regression SKlearn	303.787180	311.715617	208.997067	213.376245	9.228665e+04	9.716663e+04	0.156178	0.157330	0.914646	0.899750
Stats Model 9	330.966720	341.525729	226.766997	234.060364	1.095390e+05	1.166398e+05	0.164434	0.167744	0.898690	0.879784
Stats Model 8	330.603225	342.049954	226.824116	234.773723	1.092985e+05	1.169982e+05	0.164384	0.168287	0.898912	0.879400
Stats Model 7	330.462488	342.110492	226.845742	234.971922	1.092055e+05	1.170396e+05	0.164605	0.168762	0.898998	0.879348
Stats Model 6	330.377416	342.368115	226.693439	235.250294	1.091492e+05	1.172159e+05	0.164448	0.169176	0.899050	0.879153
Stats Model 5	330.301501	342.606768	226.620179	235.386991	1.090991e+05	1.173794e+05	0.164418	0.169288	0.899097	0.878983
Stats Model 4	330.261306	342.778515	226.716476	235.454401	1.090725e+05	1.174971e+05	0.164519	0.169253	0.899121	0.878862
Stats Model 3	330.248329	342.887795	226.697847	235.439651	1.090640e+05	1.175720e+05	0.164500	0.169162	0.899129	0.878785
Stats Model 2	330.245346	342.905173	226.662989	235.436402	1.090620e+05	1.175840e+05	0.164440	0.169162	0.899131	0.878772
AdaBoost Regressor Model	321.655205	351.083781	248.635719	265.519111	1.034621e+05	1.232598e+05	0.212489	0.223207	0.910524	0.879088
Decision Tree Hyper Tuned Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797
DT Model	0.000000	372.438422	0.000000	235.788201	0.000000e+00	1.387104e+05	0.000000	0.168410	1.000000	0.856797
ANN Model	5437.025708	5450.508164	5338.118747	5358.526912	2.956125e+07	2.970804e+07	5.593058	5.633304	-0.029719	-0.034118

15.Final Concussions:

- Both in terms of RMSE and MAPE RF and Random Forest Hyper Tuned Model has performed best among all the 17 models created using different regression techniques. A point to note is RF model is exactly same and we can use RF model with or without hyper tuned parameters for the best results.
- Basis RMSE we can say that we can expect an average difference of 272 in agent bonus per policy as per RF model or Random Forest Hyper Tuned Model.
- Basis MAPE we can say that we can expect 12% deviation or variance in agent bonus per policy as per RF model or Random Forest Hyper Tuned Model.
- ANN model has performed the worst with RMSE of 5450.50 and MAPE of 563.33% which is too high.
- Bagging regressor model comes second with RMSE of 288.64 and MAPE of 13.31%.
- Rest all the models have a RMSE of between 300-400 and MAPE between 15% to 17%.
- For this data set RF model or Random Forest Hyper tuned Model is the best option and should be selected for production.

xxxxxxxxxxxxxxxxxxxxEnd of Project Notes2xxxxxxxxxxxxxxxxxxxx