Life Insurance Sales

Project Notes – 2

Ujjwal Kumar

Post-Graduation (DSBA)

February 02, 2023

Table of Contents:

1. Introduction …… 03

a) Model Building and Interpretation

- Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purpose…… 03
- Test your predictive model against the test set using various appropriate performance metrics…….. 08
- Interpretation of the model(s)…… 08

b) Model Tuning

- Ensemble modelling, wherever applicable…… 09
- Any other model tuning measures (if applicable)…… 10
- Interpretation of the most optimum model and its implication on the business…….. 11

Life Insurance Sales

Introduction:

Insurance is nothing but one of the vital products for both business and human life. It not only provides necessary financial support in case of uncertainties but also safeguards against unpredictable events. It gives necessary cover and peace of mind against any catastrophic events which are not even in control of human being. Basically, Insurance is a financial safety net, helping us and our loved ones recover after something bad happens such as fire, theft, lawsuit or car accident etc. which is a legal contract between us and our insurance provider.

Model Buildings and Interpretations

- Since we know that the given problems statement is continuous, however the variables involved are continuous in nature, So probably seems regression is better for this problem.
- But we also have seen that there are some categorical variables are also present in the data set. Since the regression model uses only numerical variables so we have to convert those categorical variables into the numerical form.
- We also see that some of the categorical variables have more than two categories, so we apply One-Hot Encoding which means that it converts each categorical level within the category features into columns and makes it a binary feed.
  All in all it means that wherever the value holds 'true' it will give a value of 1, wherever the value is not available for a particular observation, it will give a value of zero.

Output after Encoding

	AgentBonus	Age	CustTenure	ExistingProdType	NumberOfPolicy	MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured	LastMonthCalls
0	4409.0	22.0	4.0	3.0	2.0	20993.0	1.0	2.0	806761.000000	5.0
1	2214.0	11.0	2.0	4.0	4.0	20130.0	0.0	3.0	294502.000000	7.0
2	4273.0	26.0	4.0	4.0	3.0	17090.0	1.0	2.0	619999.699267	0.0
3	1791.0	11.0	4.0	3.0	3.0	17909.0	1.0	2.0	268635.000000	0.0
4	2955.0	6.0	4.0	3.0	4.0	18468.0	0.0	4.0	366405.000000	2.0

Designation_VP	MaritalStatus_Married	MaritalStatus_Single	MaritalStatus_Unmarried	Zone_North	Zone_South	Zone_West	PaymentMethod_Monthly	PaymentMeth
0	0	1	0	1	0	0	0	
0	0	0	0	1	0	0	0	
0	0	0	1	1	0	0	0	
0	0	0	0	0	0	1	0	
0	0	0	0	0	0	1	0	

Since we are using the same data that we have used in Problem Statement 1, so we don't need to have EDA as we have already treated the biasedness that it had like null values, Outliers, so we could directly proceed for our Model Exercise.

Model Building : In model building the first step that we need to do is to split the data into Train and Test with their respective ratios. And here we have split the data into 75:25 ratio.

Data Shape after Train and Test split

```
Train data (3390, 38)
Test Data (1130, 38)
```

Insights of R Square and Root Mean Square Error (RMSE)

The value of R Square on Train data is 0.809 and RMSE on Train data is 596

The value of R Square on Test data is 0.781 and RMSE on Test data is 623

Using Linear Regression Model

Now the first iteration towards building Linear Regression model is that we used all the independent variables that the dataset carries which is given below:-

```
AgentBonus
Age
CustTenure
ExistingProdType
NumberOfPolicy
MonthlyIncome
Complaint
ExistingPolicyTenure
SumAssured
LastMonthCalls
CustCareScore
Channel_Online
Channel_Third Party Partner
Occupation_Laarge Business
Occupation_Large Business
Occupation_Salaried
Occupation_Small Business
EducationField_Engineer
EducationField_Graduate
EducationField_MBA
EducationField_Post Graduate
EducationField_UG
EducationField_Under Graduate
Gender_Female
Gender_Male
Designation_Exe
Designation_Executive
Designation_Manager
Designation_Senior Manager
Designation_VP
MaritalStatus_Married
MaritalStatus_Single
MaritalStatus_Unmarried
Zone_North
Zone_South
Zone_West
PaymentMethod_Monthly
PaymentMethod_Quarterly
PaymentMethod_Yearly
```

Summary of Linear Model 1

Dep. Variable:	AgentBonus	R-squared:	0.801
Model:	OLS	Adj. R-squared:	0.799
Method:	Least Squares	F-statistic:	410.4
Date:	Sun, 29 Jan 2023	Prob (F-statistic):	0.00
Time:	10:56:59	Log-Likelihood:	-26546.
No. Observations:	3390	AIC:	5.316e+04
Df Residuals:	3356	BIC:	5.337e+04
Df Model:	33		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-308.8624	210.137	-1.470	0.142	-720.871	103.146
Age	21.5843	1.411	15.294	0.000	18.817	24.351
CustTenure	22.7989	1.408	16.189	0.000	20.038	25.560
ExistingProdType	-74.0961	23.404	-3.166	0.002	-119.983	-28.209
NumberOfPolicy	0.0983	7.655	0.013	0.990	-14.911	15.108
MonthlyIncome	0.0722	0.005	14.648	0.000	0.063	0.082
Complaint	29.5719	23.490	1.259	0.208	-16.485	75.629
ExistingPolicyTenure	38.2599	3.748	10.208	0.000	30.911	45.608
SumAssured	0.0035	6.01e-05	58.759	0.000	0.003	0.004
LastMonthCalls	0.6478	3.147	0.206	0.837	-5.522	6.817
CustCareScore	8.6291	7.749	1.114	0.266	-6.564	23.822
Channel_Online	24.9877	35.035	0.713	0.476	-43.704	93.679
Channel_Third_Party_Partner	-3.2896	27.360	-0.120	0.904	-56.933	50.353
Occupation_Large_Business	-27.6162	74.344	-0.371	0.710	-173.380	118.148
Occupation_Salaried	-0.4077	147.813	-0.003	0.998	-290.220	289.405
Occupation_Small_Business	-0.4595	148.313	-0.003	0.998	-291.252	290.333
EducationField_Engineer	-17.6585	139.264	-0.127	0.899	-290.710	255.393
EducationField_MBA	-127.4832	90.602	-1.407	0.160	-305.125	50.159
EducationField_Post_Graduate	12.8045	49.439	0.259	0.796	-84.128	109.737
EducationField_Under_Graduate	-33.5090	32.425	-1.033	0.301	-97.084	30.066
Gender_Male	15.1673	21.646	0.701	0.484	-27.273	57.608
Designation_Executive	105.4200	46.669	2.259	0.024	13.917	196.923
Designation_Manager	-70.6496	40.914	-1.727	0.084	-150.868	9.569
Designation_Senior_Manager	-5.7249	43.342	-0.132	0.895	-90.704	79.255
Designation_VP	47.1871	64.562	0.731	0.465	-79.398	173.772
MaritalStatus_Married	-52.9494	29.153	-1.816	0.069	-110.109	4.211
MaritalStatus_Single	11.4256	32.325	0.353	0.724	-51.953	74.804
MaritalStatus_Unmarried	-137.8457	60.636	-2.273	0.023	-256.734	-18.957
Zone_North	49.1826	93.357	0.527	0.598	-133.860	232.225
Zone_South	201.2722	289.706	0.695	0.487	-366.746	769.291
Zone_West	42.9594	92.886	0.462	0.644	-139.158	225.077
PaymentMethod_Monthly	-49.9428	57.238	-0.873	0.383	-162.167	62.282
PaymentMethod_Quarterly	-9.2841	86.238	-0.108	0.914	-178.368	159.800
PaymentMethod_Yearly	44.1151	34.186	1.290	0.197	-22.913	111.143

Omnibus:	136.383	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	156.022
Skew:	0.479	Prob(JB):	1.32e-34

- We know that we have the RMSE value is 608.92 and here the variation in R Square and Adjusted R Square is not that significant.
- So, in the second Iteration we are going to consider only those independent variables, whose P value is less than 0.05 and Hence we are going to drop all redundant variables or we can just ignore those variables to reduce the multicollinearity levels which is also the reason behind the change in values.

Summary of Linear Model 2

OLS Regression Results

Dep. Variable:	AgentBonus	R-squared:	0.801
Model:	OLS	Adj. R-squared:	0.799
Method:	Least Squares	F-statistic:	541.4
Date:	Sun, 29 Jan 2023	Prob (F-statistic):	0.00
Time:	10:57:03	Log-Likelihood:	-26550.
No. Observations:	3390	AIC:	5.315e+04
Df Residuals:	3364	BIC:	5.331e+04
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-134.5120	94.870	-1.418	0.156	-320.522	51.498
Age	21.4809	1.409	15.243	0.000	18.718	24.244
CustTenure	22.6278	1.406	16.095	0.000	19.871	25.384
ExistingProdType	-60.1828	21.281	-2.828	0.005	-101.907	-18.459
NumberOfPolicy	0.6519	7.626	0.085	0.932	-14.300	15.604
MonthlyIncome	0.0676	0.004	18.866	0.000	0.061	0.075
Complaint	29.6797	23.455	1.265	0.206	-16.308	75.668
ExistingPolicyTenure	38.7012	3.738	10.352	0.000	31.371	46.031
SumAssured	0.0035	5.96e-05	59.012	0.000	0.003	0.004
LastMonthCalls	0.0369	3.121	0.012	0.991	-6.081	6.155
CustCareScore	9.0957	7.719	1.178	0.239	-6.038	24.230
Channel_Online	22.2588	34.481	0.646	0.519	-45.346	89.864
EducationField_Engineer	-21.6990	37.380	-0.580	0.562	-94.990	51.592
EducationField_MBA	-118.1478	89.715	-1.317	0.188	-294.050	57.754
EducationField_Post_Graduate	20.2667	47.923	0.423	0.672	-73.695	114.228
Gender_Male	18.3735	21.570	0.852	0.394	-23.919	60.666
Designation_Manager	-147.1686	23.883	-6.162	0.000	-193.996	-100.342
Designation_Senior_Manager	-68.0573	33.280	-2.045	0.041	-133.308	-2.807
MaritalStatus_Married	-51.5621	29.122	-1.771	0.077	-108.661	5.537
MaritalStatus_Single	16.7452	32.164	0.521	0.603	-46.318	79.808
MaritalStatus_Unmarried	-150.3305	60.216	-2.497	0.013	-268.394	-32.267
Zone_South	156.9224	274.897	0.571	0.568	-382.060	695.905
Zone_West	-4.8346	21.371	-0.226	0.821	-46.737	37.068
PaymentMethod_Monthly	-26.6094	54.455	-0.489	0.625	-133.379	80.160
PaymentMethod_Quarterly	4.4188	85.496	0.052	0.959	-163.211	172.048
PaymentMethod_Yearly	30.6159	32.555	0.940	0.347	-33.215	94.446

Omnibus:	123.398	Durbin-Watson:	1.995
Prob(Omnibus):	0.000	Jarque-Bera (JB):	140.124
Skew:	0.451	Prob(JB):	3.74e-31
Kurtosis:	3.422	Cond. No.	1.72e+07

Variance Inflation Factor (VIF):

Moving ahead we are going to look for VIF which is (Variance Inflation Factor) but before that we have to understand what VIF actually is. Basically A variance Inflation Factor is a measure of the amount of multicollinearity in regression analysis. The Variance Inflation Factor is nothing but a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

The higher the value of Variance Inflation Factor is, the higher the correlation between the variables. Below is the pictorial view of VIF .

VIF values (Before Variables dropped)

```
Age  VIF =  1.41
CustTenure  VIF =  1.38
ExistingProdType  VIF =  4.75
NumberOfPolicy  VIF =  1.12
MonthlyIncome  VIF =  5.24
Complaint  VIF =  1.01
ExistingPolicyTenure  VIF =  1.12
SumAssured  VIF =  1.76
LastMonthCalls  VIF =  1.2
CustCareScore  VIF =  1.03
Channel_Online  VIF =  1.05
Channel_Third_Party_Partner  VIF =  1.04
Occupation_Laarge Business  VIF =  62.39
Occupation_Large_Business  VIF =  101.63
Occupation_Salaried  VIF =  432.81
Occupation_Small_Business  VIF =  440.93
EducationField_Engineer  VIF =  18.07
EducationField_Graduate  VIF =  17.29
EducationField_MBA  VIF =  2.0
EducationField_Post_Graduate  VIF =  4.44
EducationField_UG  VIF =  1.57
EducationField_Under_Graduate  VIF =  2.58
Gender_Female  VIF =  4.77
Gender_Male  VIF =  4.54
Designation_Exe  VIF =  2.3
Designation_Executive  VIF =  8.62
Designation_Manager  VIF =  6.08
Designation_Senior_Manager  VIF =  2.82
Designation_VP  VIF =  1.84
MaritalStatus_Married  VIF =  1.92
MaritalStatus_Single  VIF =  1.89
MaritalStatus_Unmarried  VIF =  1.37
Zone_North  VIF =  19.24
Zone_South  VIF =  1.12
Zone_West  VIF =  19.21
PaymentMethod_Monthly  VIF =  2.22
PaymentMethod_Quarterly  VIF =  1.12
PaymentMethod_Yearly  VIF =  2.4
```

VIF values (After variables dropped)

```
Age  VIF =  1.4
CustTenure  VIF =  1.37
ExistingProdType  VIF =  3.73
NumberOfPolicy  VIF =  1.11
MonthlyIncome  VIF =  1.98
Complaint  VIF =  1.01
ExistingPolicyTenure  VIF =  1.11
SumAssured  VIF =  1.74
LastMonthCalls  VIF =  1.18
CustCareScore  VIF =  1.02
Channel_Online  VIF =  1.02
Occupation_Laarge Business  VIF =  1.58
EducationField_Engineer  VIF =  1.68
EducationField_Graduate  VIF =  1.26
EducationField_MBA  VIF =  1.04
EducationField_Post_Graduate  VIF =  1.09
EducationField_UG  VIF =  1.26
Gender_Female  VIF =  4.74
Gender_Male  VIF =  4.51
Designation_Exe  VIF =  1.19
Designation_Manager  VIF =  1.22
Designation_Senior_Manager  VIF =  1.29
MaritalStatus_Married  VIF =  1.92
MaritalStatus_Single  VIF =  1.88
MaritalStatus_Unmarried  VIF =  1.36
Zone_South  VIF =  1.01
Zone_West  VIF =  1.02
PaymentMethod_Monthly  VIF =  1.98
PaymentMethod_Quarterly  VIF =  1.1
PaymentMethod_Yearly  VIF =  2.11
```

We see in the above case that there are many variables which expresses multicollinearity, they also have VIF values which is more than 5, so we dropped it and highlighted again in the next column.

Comparing Linear Model Results

The RMSE of Linear Model 1 on Train data is 608 and RMSE of Linear Model 2 on Train Data is 609.

The RMSE of Linear Model 1 on Test data is 633 and RMSE of Linear Model 2 on Test Data is 632

- So, we have observed that there is no significant changes in R Square or Root Mean Square Values in both the iterations, so this may not be the ideal way to choose the best model.
- It is required for us to check for different models like Decision Tree, Random Forest or Artificial Neural Network with base parameters and then compare their results to choose the best model

Data Scaling:

First we need to understand what Data Scaling is and why do we need it. Scaling is related too the numeric features in the data.

We need Scaling because when we observe the numeric features in the data, the scale of the numeric features differs, and some of the algorithms are sensitive to this. They start giving higher weightage to the features that have higher values comparatively to the features who have smaller values.

Talking about the data that we have we had the following observations:-

- We have observed that the feature like 'Sum', 'Sum assured', are carrying higher weightage, so in order to make our decision based on them we have to normalize the data and bring them to common scale using the Data Scaling method.
- As discussed above it is true that Scaling does not impact the coefficient of attributes or its intercept values, however it is useful in reducing multicollinearity.

Different Models and their Scores (Base Parameter)

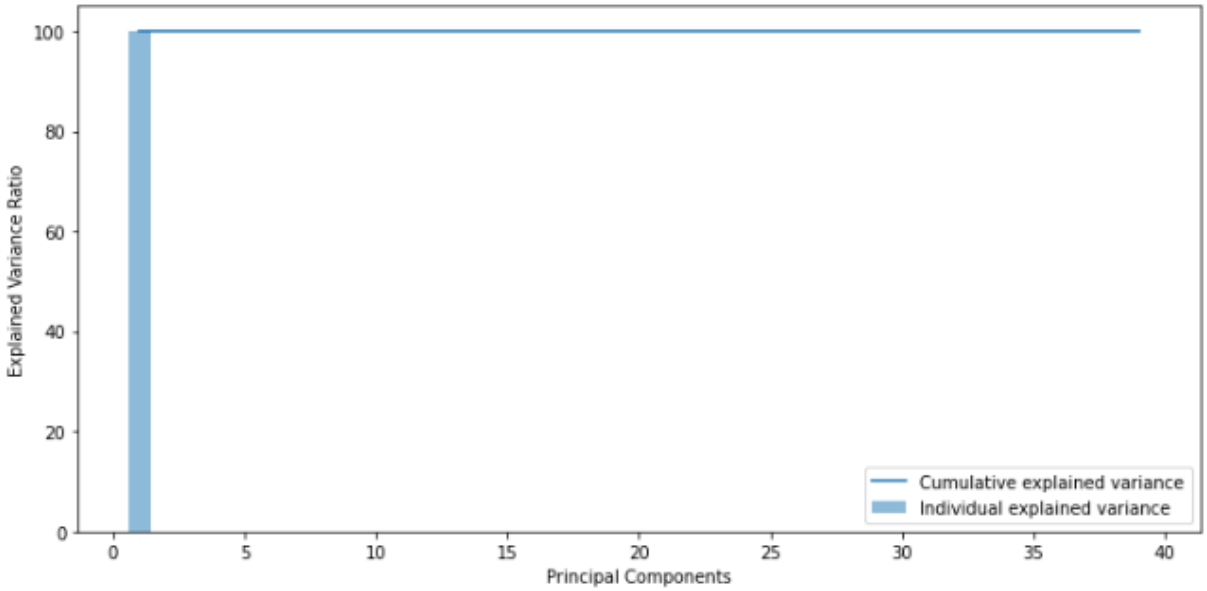
	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	608.208934	590.392992	0.803620	0.798161
Decision Tree Regressor	0.000000	760.245991	1.000000	0.665318
Random Forest Regressor	190.483906	519.251378	0.980738	0.843873
ANN Regressor	497.488121	606.842724	0.868612	0.786756

Below are the observations and the analysis of Scores we have from different models on Train and Test data

- So far we have observed that our Linear Model is performing better when compared to other models as the variation between the Train and Test data is very minimum.
- If we compare the models, we find that most of the other models are in overfitting zone with respect to Linear Regression model.
- We have observed that we are dealing with overfitting problem, and for that we can use hyperparameter tuning like

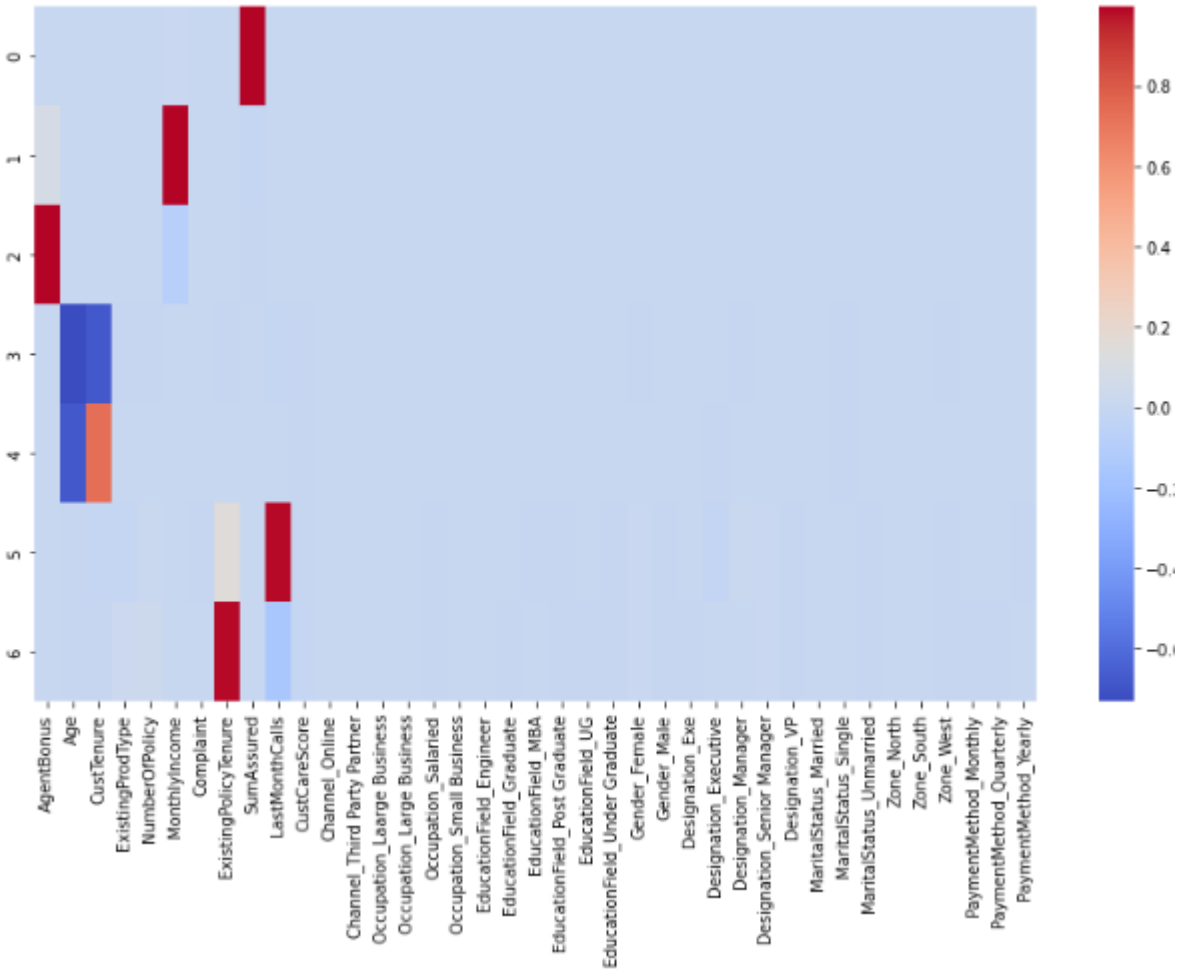
Checking if PCA can be applied

```
Cumulative Variance Explained [ 99.97526512  99.99912394  99.99999975  99.99999985  99.99999995
  99.99999997  99.99999998  99.99999999  99.99999999  99.99999999
  99.99999999 100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.        ]
```



Principal Components Vs Explained Variance Ratio

We see that cumulative variance is 99% which is almost about 100% so we need not perform PCA here.



PCA Heatmap

We see that not much can be observed about the components from the heatmap, Hence dropping the need to perform PCA as all the variables holds almost the good deal of importance in the predictions.

Now we are going for Model tuning but before that we have to understand what Model Tuning is and why is it used.

Model Tuning is basically the experimental process of finding the optimal values of hyper parameters to maximize the model performance. The purpose of model tuning a model is just to ensure that it performs at its best. This process involves adjusting various elements of the model to achieve optimal results. By fine tuning the model, we can actually maximise its performance without overfitting and reduce the variance error in our model.

Basically there are different types of Model tuning but here we are going to consider 'Grid Search' technique.

Grid Search on Decision Tree

```
{'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 50}
```

Grid Search on Random Forest

```
{'max_depth': 10, 'max_features': 6, 'min_samples_leaf': 3, 'min_samples_split': 30, 'n_estimators': 300}
```

Grid Search on ANN

```
{'activation': 'relu', 'hidden_layer_sizes': 500, 'solver': 'adam'}
```

Different Models with different Scores (After Hyper parameter Tuning)

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	608.208934	590.392992	0.803620	0.798161
Decision Tree Regressor	495.236822	573.495484	0.869798	0.809549
Random Forest Regressor	540.850048	583.163906	0.844709	0.803073
ANN Regressor	497.488121	606.842724	0.868612	0.786756

Observations drawn after hyper parameter tuning are as follows:

- We have observed that most of the variables in this data set has now moved out of overfitting zone.
- We have also observed that our model which is Linear Regression is still stable and the difference between Train and Test data is minimal.

So based on the observations we can say that Linear Regression is the most suitable and stable model throughout.

But if model accuracies is something to be watched for then we can clearly say that 'Random Forest' model is by far better than any other model as the difference between the Train and Test data is less than 5%.

Feature Importance:

We see the 'Sum_assured' as our most important feature and on the second hand Zone_South being the least.

Interpretations

- The Company basically wants to predict the ideal bonus for agents and the level of engagement of high and low performing agents respectively.
- The Customers who has high designations has a higher chances to buy more policies like if the Designation is Vice President of any company then that person is going to buy more policies.
- Hence, for high performing agents we can create a healthy contest and for low performing agents what better can be done is, we can train them, or suggest them to purchase or get policies with high sum assured as it is very significant to our model.
- There is one more important feature which is Customer Tenure where the agents need to focus on the customers policy tenure which is ranging between 8 – 20 and this is where the majority of customer falls.
- Now keeping an eye on those customers who has higher monthly incomes because the greater monthly incomes the Customers has, the greater possibilities of Customers to buy higher valued policies.
- And through the models, the agents who are higher performing, for them we have a few variables which is quite significant for them. For eg, 'Sum_Assured'.

Recommendations

- Firstly, we have to see for high performing agents, for them we can create a healthy contest with a threshold, like if they achieve the target like the desired sum_assured then they are eligible for certain perks and incentives like exotic family vacation packages, latest gadgets and many more.
- Secondly, we have to look for low performing agents, for them we can upgrade or introduce some feedbacks or up skill certain programs to train them generating higher sum assured policies, reaching out to certain people to ultimately becoming the higher or top performers.
- Apart from all these, we need some more data or predictors like Premium Amount, this will help us to solve the business problem even better as well as we will be having more variables to test upon, thereby we are going to have more accurate results in real time problems or data like this.
- We can also add another predictor as our Customers geographical location or Regions not just the zones but also the people who usually lives in rural areas or remote areas are less likely to buy a policy whereas those living in highly developed location are likely to be belonging to the upper class should be targeted.
- Similarly, there can be another predictor as well like 'Agent ID' which can be introduced that will helps us to make it easier to observe the high and low performing agents and their trends.
- The amount or the policy premium collected from the customers acts as a very good predictor in terms of analysing the agent bonus, which gives us the real insights towards the monetary business agent who is doing on regular basis.