LIFE INSURANCE CASE STUDY

CAPSTONE Project Notes -2

AUGUST 14, 2022

ALIND KHANNA
GREAT LAKES

Table of Contents

DATA DICTIONARY

Data	Variable	Discerption
Sales	CustID	Unique customer ID
Sales	AgentBonus	Bonus amount given to each agents in last month
Sales	Age	Age of customer
Sales	CustTenure	Tenure of customer in organization
Sales	Channel	Channel through which acquisition of customer is done
Sales	Occupation	Occupation of customer
Sales	EducationField	Field of education of customer
Sales	Gender	Gender of customer
Sales	ExistingProdType	Existing product type of customer
Sales	Designation	Designation of customer in their organization
Sales	NumberOfPolicy	Total number of existing policy of a customer
Sales	MaritalStatus	Marital status of customer
Sales	MonthlyIncome	Gross monthly income of customer
Sales	Complaint	Indicator of complaint registered in last one month by customer
Sales	ExistingPolicyTenure	Max tenure in all existing policies of customer
Sales	SumAssured	Max of sum assured in all existing policies of customer
Sales	Zone	Customer belongs to which zone in India. Like East, West, North and South
Sales	PaymentMethod	Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly
Sales	LastMonthCalls	Total calls attempted by company to a customer for cross sell
Sales	CustCareScore	Customer satisfaction score given by customer in previous service call

LIFE INSURANCE CASE STUDY

Introduction

Insurance is one of the vital products for both business and human life. It provides necessary financial support in case of uncertainties also it 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.

Model Building and Interpretation

- Given problem seems more of a regression problem, since there are continuous variables involved.
- We see there are some categorical variables in given data set since they have to be converted to numerical, as regression uses only numerical variables.

Since most of the cat variable have more than 2 categories, we apply One-Hot Encoding

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	...

5 rows × 39 columns

...	Designation_VP	MaritalStatus_Married	MaritalStatus_Single	MaritalStatus_Unmarried	Zone_North	Zone_South	Zone_West	PaymentMethod_Monthly	Pa
...	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	

Before proceeding towards model building, we are renaming variables.

COLUMN NAMES

```
Index(['Age', 'CustTenure', 'ExistingProdType', 'NumberOfPolicy',
       'MonthlyIncome', 'Complaint', 'ExistingPolicyTenure', 'SumAssured',
       'LastMonthCalls', 'CustCareScore', 'Channel_Online',
       'Channel_Third Party Partner', 'Occupation_Large Business',
       'Occupation_Salaried', 'Occupation_Small Business',
       'EducationField_Engineer', 'EducationField_MBA',
       'EducationField_Post Graduate', 'EducationField_Under Graduate',
       'Gender_Male', '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', 'AgentBonus'],
      dtype='object')
```

RENAMED COLUMNS (SPACES REMOVED)

```
Index(['Age', 'CustTenure', 'ExistingProdType', 'NumberOfPolicy',
       'MonthlyIncome', 'Complaint', 'ExistingPolicyTenure', 'SumAssured',
       'LastMonthCalls', 'CustCareScore', 'Channel_Online',
       'Channel_Third_Party_Partner', 'Occupation_Large_Business',
       'Occupation_Salaried', 'Occupation_Small_Business',
       'EducationField_Engineer', 'EducationField_MBA',
       'EducationField_Post_Graduate', 'EducationField_Under_Graduate',
       'Gender_Male', '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', 'AgentBonus'],
      dtype='object')
```

- As we are using the same data, we used in PN1 and in EDA we have already treated null values, outliers in data hence we could continue with Model Building exercise.

MODEL BUILDING:

- First step to start model building is divide data into Train and Test data.

We have split data into 75:25 ratio

Data shape After Test/Train Split

```
Train data (3390, 18)
Test Data (1130, 18)
```

Insights of R Square & RMSE

	R square	RMSE
Train Data	0.809	596
Test Data	0.781	623.1

LINEAR REGRESSION MODEL

- In the first iteration towards building linear regression model, we used all of the independent variables.

```
AgentBonus
Age
CustTenure
ExistingProdType
NumberOfPolicy
MonthlyIncome
Complaint
ExistingPolicyTenure
SumAssured
LastMonthCalls
CustCareScore
Channel_Online
Channel_Third Party Partner
Occupation_Large Business
Occupation_Salaried
Occupation_Small Business
EducationField_Engineer
EducationField_MBA
EducationField_Post Graduate
EducationField_Under Graduate
Gender_Male
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
```

LM1 summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             AgentBonus   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.799
Method:                 Least Squares   F-statistic:                     410.4
Date:                Sun, 14 Aug 2022   Prob (F-statistic):               0.00
Time:                        07:01:24   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
Kurtosis:                       3.430   Cond. No.                     1.94e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.94e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

RMSE value – 608.92 The variation in R-squared and Adjusted R-squared is not too significant

ITERATION 2: In iteration 2 we consider only those independent variables for which P value is less than 0.05, therefore we drop all redundant variables to bring down multicollinearity levels.

LM2 results summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             AgentBonus   R-squared:                       0.806
Model:                            OLS   Adj. R-squared:                  0.805
Method:                 Least Squares   F-statistic:                     1399.
Date:                Sat, 11 Dec 2021   Prob (F-statistic):               0.00
Time:                        13:22:00   Log-Likelihood:                -26511.
No. Observations:                3390   AIC:                         5.304e+04
Df Residuals:                    3379   BIC:                         5.311e+04
Df Model:                          10
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                    643.6161    129.776      4.959      0.000     389.168     898.064
Age                           21.8786      1.416     15.451      0.000      19.102      24.655
CustTenure                    22.7193      1.424     15.955      0.000      19.927      25.511
MonthlyIncome                  0.0372      0.004      8.473      0.000       0.029       0.046
ExistingPolicyTenure          40.1752      4.037      9.951      0.000      32.259      48.091
SumAssured                     0.0036   5.85e-05     60.654      0.000       0.003       0.004
Designation_Executive       -427.4484     52.722     -8.108      0.000    -530.818    -324.079
Designation_Manager         -436.7599     45.193     -9.664      0.000    -525.367    -348.152
Designation_Senior_Manager  -258.6449     43.277     -5.977      0.000    -343.496    -173.794
MaritalStatus_Married        -67.6078     21.235     -3.184      0.001    -109.243     -25.973
MaritalStatus_Unmarried     -226.2434     55.495     -4.077      0.000    -335.050    -117.437
==============================================================================
Omnibus:                      128.393   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              143.854
Skew:                           0.475   Prob(JB):                     5.79e-32
Kurtosis:                       3.341   Cond. No.                     9.23e+06
==============================================================================
```

VIF values

```
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
```

We see in above case, there are many variables which exhibit multicollinearity, having VIF values more than 5, so we drop those variables.

VIF values (After dropping variables)

```
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
```

Comparing Linear Model Iterations results:

	RMSE(LM1)	**RMSE(LM2)**
Train Data	608.92	609.65
Test Data	633.95	632.83

We see there is no significant change in R square or RMSE vales in both iterations, this could not be optimal way of choosing best model.

We need to check for different models Random Forest, Artificial Neural Network and Decision Tree with base parameters and compare results for choosing optimum model.

Data Scaling

- We scale the data in order to bring the values in the common range, which helps us making our decisions unbiased.
- We observe that age, sum assured are carrying higher weights, so in order to make our decision based on them we have rationalize the data and brought to common scale using data scaling
- Scaling doesn't impact coefficient of attributes or its intercept values.
- Data Scaling, also helps us reducing multicollinearity

Comparing Score Running Different Models (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

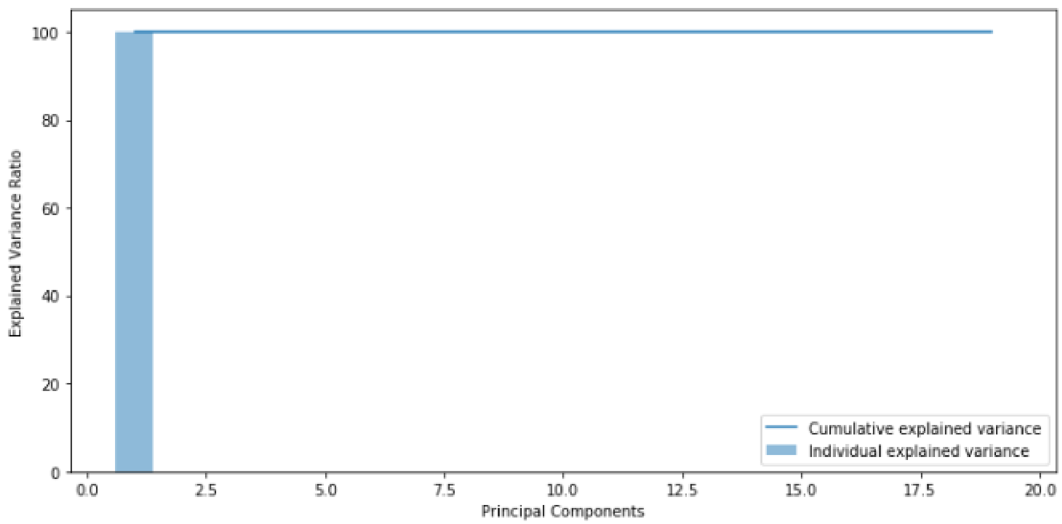
Observations and Score Analysis

- Based on model comparison score we find most other models are in overfitting zone.
- Linear Regression is performing better when compared to others as variation between test and train data is minimal.
- In order to solve the overfitting problem, we have used the hyperparameter tuning based on grid search.

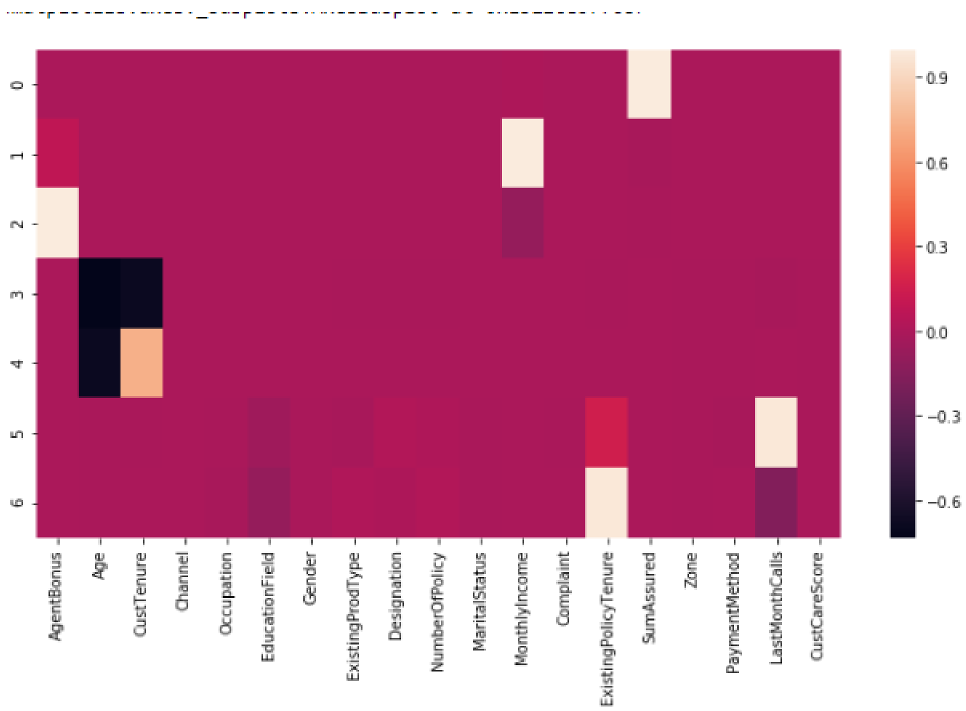
Checking if PCA can be applied here

```
Cumulative Variance Explained [ 99.97526511  99.99912392  99.99999973  99.99999984  99.99999993
  99.99999996  99.99999997  99.99999998  99.99999998  99.99999999
  99.99999999  99.99999999 100.         100.         100.
 100.         100.         100.         100.        ]
```

- Since cumulative variance is almost 99%, hence there is no need to perform PCA



Principal Components vs variance ratio



PCA heatmap

- Not much can be observed about the components from the heatmap, therefore dropping the need to perform PCA as almost all these variables hold a good deal of significance in the predictions

MODEL TUNNING

Next step, we go for grid search for hyper parameter tuning to observe is there any significant difference in observed results.

Grid Search on Decision Tree

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

Grid Search on Random Forest Tree

```
{'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'}
```

Comparing Score Running Different Models (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 after Hyper Parameter Tuning

- We observe most of the variables have moved out of overfitting zone.
- Also, we observe that Linear Regression model is still stable with train-test difference on minimal.

Based, on the observations we can say Linear Regression is most stable model throughout,

If model accuracies are some want to be watched for then we can say Random Forest model does better job and as there is less than 5 percent difference between test-train data.

Feature Importance:

We see sum assured as most important feature with Zone_South being the least.

	Imp
SumAssured	0.430643
CustTenure	0.147939
Age	0.135314
MonthlyIncome	0.122517
ExistingPolicyTenure	0.035902
Designation_VP	0.031614
Designation_Executive	0.023948
Designation_Manager	0.013429
LastMonthCalls	0.012173
Designation_Senior Manager	0.007222
Designation_Exe	0.004794
ExistingProdType	0.004702
MaritalStatus_Unmarried	0.003864
NumberOfPolicy	0.003687
Gender_Female	0.002636
CustCareScore	0.002612
EducationField_UG	0.001531
EducationField_Under Graduate	0.001157
MaritalStatus_Married	0.001078
Zone_North	0.001078
Zone_West	0.001074
MaritalStatus_Single	0.001051
Gender_Male	0.001030
Complaint	0.001022
Channel_Third Party Partner	0.000927
PaymentMethod_Yearly	0.000889
Occupation_Salaried	0.000817
Occupation_Small Business	0.000794
EducationField_Graduate	0.000775
EducationField_Engineer	0.000725
Channel_Online	0.000642
PaymentMethod_Monthly	0.000556
Occupation_Large Business	0.000525
EducationField_Post Graduate	0.000510
Occupation_Laarge Business	0.000323
EducationField_MBA	0.000290
PaymentMethod_Quarterly	0.000204
Zone_South	0.000005

Interpretation and Business Recommendations.

• Company wants to predict the ideal bonus for agent and level of engagement of high and low performing agents respectively.

• Through the model, for high performing agent we will find variable significance, for eg, Sum Assured is highly significant here.

• If the Designation is VP the person buys more policy or high value policies.

• Therefore, for high and low performing agents, we will train them, suggesting them to purchase or get policies with high sum assured as it is very significant to our model.

• Another important feature is Customer tenure where the agents need to focus on the customers who've a tenure ranging between 8-20 this where the majority of the customer are.

• Focusing on customers with greater monthly incomes as greater the monthly income, greater is the possibility of the customer buying a higher valued policy.

Recommendations.

• For High Performing Agents we can create a healthy contest with a threshold.

• Where, if they achieve the desired sum assured, they are eligible for certain incentives like latest gadgets, exotic family vacation packages and some extra perks as well.

• For low performing agents, we can introduce certain feedback upskill programs to train them into closing higher sum assured policies, reaching certain people to ultimately becoming top/high performers.

• Apart from this, we need more data/predictors like Premium Amount, this will help us to solve the business problem even better as well have more variables to test upon thereby having more accurate results in real time problems like this.

• I also feel another predictor can be added as customers geographical location or Region and not just the zones as people living in rural areas are less likely to buy a policy whereas those living in a highly developed location are likely to be belonging to the upper class and should be targeted.

• Similarly, another predictor can be AgentID can be introduced which will make it easier to observe the high and low performing agent trend

• Premium collected from customer is, also a very good predictor in terms of analysing agent bonus, this gives real insight towards the monetary business agent is doing on regular basis.