Life-Insurance Sale Capstone

Project Notes – 2

Tushar Babbar

(PGP DSBA Dec20)

Table of Contents

Table of Figures

Model Building and Interpretation

- Regression uses numerical variables,
- But we have a lot of categorical variables we wish to use in our models further,
- And since most of the categorical variables have categories more than 2, therefore applying one-hot encoding.
- One-Hot encoding takes every level of the category and turns it into a variable with two level (yes/no).

The data looks like this after one-hot encoding.

SumAssured	LastMonthCalls	...	Designation_VP	MaritalStatus_Married	MaritalStatus_Single	MaritalStatus_Unmarried	Zone_North	Zone_South	Zone_West
806761.0	5.0	...	0	0	1	0	1	0	0
294502.0	7.0	...	0	0	0	0	1	0	0
578976.5	0.0	...	0	0	0	1	1	0	0
268635.0	0.0	...	0	0	0	0	0	0	1
366405.0	2.0	...	0	0	0	0	0	0	1

- Building our Linear Regression Model with the unprocessed data above.
- Keep in mind, this data holds no outliers as they were removed in EDA – PN1

Split X and y into training and test set in 75:25 ratio

```
The coefficient for Age is                                  21.64543636236496
The coefficient for CustTenure is                           22.620905021409023
The coefficient for ExistingProdType is                     46.508784274329514
The coefficient for NumberOfPolicy is                       6.254332127798309
The coefficient for MonthlyIncome is                        0.03188513622751349
The coefficient for Complaint is                            33.0503807570841
The coefficient for ExistingPolicyTenure is                 40.22901549596465
The coefficient for SumAssured is                           0.003548018281339438
The coefficient for LastMonthCalls is                       -2.308709717687992
The coefficient for CustCareScore is                        7.559056565466554
The coefficient for Channel_Online is                       22.691900907509453
The coefficient for Channel_Third Party Partner is          3.4952779925482345
The coefficient for Occupation_Large Business is            -616.8600099371561
The coefficient for Occupation_Salaried is                  -474.9729637586688
The coefficient for Occupation_Small Business is            -581.6372411869505
The coefficient for EducationField_Engineer is              26.675848148157876
The coefficient for EducationField_MBA is                   -177.27368717977166
The coefficient for EducationField_Post Graduate is         -92.6094978672669
The coefficient for EducationField_Under Graduate is        2.331225272073949
The coefficient for Gender_Male is                          25.187256483000322
The coefficient for Designation_Executive is                -493.36122500604984
The coefficient for Designation_Manager is                  -481.4192660702273
The coefficient for Designation_Senior Manager is           -277.42121914512296
The coefficient for Designation_VP is                       -2.956791388368395
The coefficient for MaritalStatus_Married is                -48.20378324641499
The coefficient for MaritalStatus_Single is                 29.658243912402032
The coefficient for MaritalStatus_Unmarried is              -188.87907531620797
The coefficient for Zone_North is                           62.35415312785426
The coefficient for Zone_South is                           193.51057687776427
The coefficient for Zone_West is                            49.998087081147155
The coefficient for PaymentMethod_Monthly is                141.95193527244763
The coefficient for PaymentMethod_Quarterly is              112.02879394979776
The coefficient for PaymentMethod_Yearly is                 -79.92080455281895
The intercept for our model is 1092.3485100144962
```

	R-Squared	RMSE
Training	0.8068152802160813	600.5900784990952
Testing	0.7825646087670782	621.5274260080358

Checking the same using statsmodel, to get more insights on p-value, r-squared and adjusted r-squared value.

Before we move to statsmodel,

- We need to rename some columns created after encoding as they have some spaces which will not be accepted my statsmodel.

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')
```

Building a Multiple Linear Regression Model, with 'AgentBonus' as the independent variable and all other variables as dependent variables - LINEAR MODEL 1 (LM1)

```
Intercept                        1092.348510
Age                                21.645436
CustTenure                         22.620905
ExistingProdType                   46.508784
NumberOfPolicy                      6.254332
MonthlyIncome                       0.031885
Complaint                          33.050381
ExistingPolicyTenure               40.229015
SumAssured                          0.003548
LastMonthCalls                     -2.308710
CustCareScore                       7.559057
Channel_Online                     22.691901
Channel_Third_Party_Partner         3.495278
Occupation_Large_Business        -616.860010
Occupation_Salaried              -474.972964
Occupation_Small_Business        -581.637241
EducationField_Engineer            26.675848
EducationField_MBA               -177.273687
EducationField_Post_Graduate      -92.609498
EducationField_Under_Graduate       2.331225
Gender_Male                        25.187256
Designation_Executive            -493.361225
Designation_Manager              -481.419266
Designation_Senior_Manager       -277.421219
Designation_VP                     -2.956791
MaritalStatus_Married             -48.203783
MaritalStatus_Single               29.658244
MaritalStatus_Unmarried          -188.879075
Zone_North                         62.354153
Zone_South                        193.510577
Zone_West                          49.998087
PaymentMethod_Monthly             141.951935
PaymentMethod_Quarterly           112.028794
PaymentMethod_Yearly              -79.920805
dtype: float64
```

OLS Regression Results

Dep. Variable:	AgentBonus	R-squared:	0.807
Model:	OLS	Adj. R-squared:	0.805
Method:	Least Squares	F-statistic:	424.7
Date:	Sun, 05 Dec 2021	Prob (F-statistic):	0.00
Time:	23:49:42	Log-Likelihood:	-26499.
No. Observations:	3390	AIC:	5.307e+04
Df Residuals:	3356	BIC:	5.327e+04
Df Model:	33		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1092.3485	467.264	2.338	0.019	176.198	2008.499
Age	21.6454	1.420	15.245	0.000	18.862	24.429
CustTenure	22.6209	1.428	15.840	0.000	19.821	25.421
ExistingProdType	46.5088	23.229	2.002	0.045	0.964	92.054
NumberOfPolicy	6.2543	7.560	0.827	0.408	-8.569	21.078
MonthlyIncome	0.0319	0.005	5.954	0.000	0.021	0.042
Complaint	33.0504	23.172	1.426	0.154	-12.381	78.482
ExistingPolicyTenure	40.2290	4.066	9.894	0.000	32.257	48.201
SumAssured	0.0035	5.88e-05	60.294	0.000	0.003	0.004
LastMonthCalls	-2.3087	3.109	-0.743	0.458	-8.405	3.787
CustCareScore	7.5591	7.644	0.989	0.323	-7.429	22.547
Channel_Online	22.6919	34.552	0.657	0.511	-45.054	90.438
Channel_Third_Party_Partner	3.4953	26.973	0.130	0.897	-49.389	56.380
Occupation_Large_Business	-616.8600	453.438	-1.360	0.174	-1505.902	272.182
Occupation_Salaried	-474.9730	428.923	-1.107	0.268	-1315.949	366.003
Occupation_Small_Business	-581.6372	436.329	-1.333	0.183	-1437.134	273.860
EducationField_Engineer	26.6758	155.095	0.172	0.863	-277.414	330.766
EducationField_MBA	-177.2737	123.966	-1.430	0.153	-420.330	65.783
EducationField_Post_Graduate	-92.6095	87.381	-1.060	0.289	-263.934	78.715
EducationField_Under_Graduate	2.3312	36.703	0.064	0.949	-69.631	74.293
Gender_Male	25.1873	21.339	1.180	0.238	-16.652	67.027
Designation_Executive	-493.3612	59.744	-8.258	0.000	-610.500	-376.222
Designation_Manager	-481.4193	50.448	-9.543	0.000	-580.330	-382.508
Designation_Senior_Manager	-277.4212	48.283	-5.746	0.000	-372.088	-182.755
Designation_VP	-2.9568	63.911	-0.046	0.963	-128.266	122.352
MaritalStatus_Married	-48.2038	28.749	-1.677	0.094	-104.572	8.164
MaritalStatus_Single	29.6582	31.785	0.933	0.351	-32.662	91.978
MaritalStatus_Unmarried	-188.8791	59.461	-3.177	0.002	-305.462	-72.296
Zone_North	62.3542	91.992	0.678	0.498	-118.011	242.720
Zone_South	193.5106	285.551	0.678	0.498	-366.362	753.383
Zone_West	49.9981	91.518	0.546	0.585	-129.439	229.435
PaymentMethod_Monthly	141.9519	56.403	2.517	0.012	31.363	252.541
PaymentMethod_Quarterly	112.0288	85.052	1.317	0.188	-54.730	278.787
PaymentMethod_Yearly	-79.9208	33.879	-2.359	0.018	-146.346	-13.496

Omnibus:	126.575	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	141.177
Skew:	0.474	Prob(JB):	2.21e-31
Kurtosis:	3.315	Cond. No.	5.53e+07

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.53e+07. This might indicate that there are
strong multicollinearity or other numerical problems.

RMSE – value - 600.5900784990948

The variation in R-squared and Adjusted R-squared is not too significant

VIF Value

```
Age  VIF                                  =  1.33
CustTenure  VIF                           =  1.32
ExistingProdType  VIF                     =  4.36
NumberOfPolicy  VIF                       =  1.12
MonthlyIncome  VIF                        =  4.17
Complaint  VIF                            =  1.01
ExistingPolicyTenure  VIF                 =  1.11
SumAssured  VIF                           =  1.73
LastMonthCalls  VIF                       =  1.2
CustCareScore  VIF                        =  1.03
Channel_Online  VIF                       =  1.05
Channel_Third_Party_Partner  VIF          =  1.04
Occupation_Large_Business  VIF            =  153.84
Occupation_Salaried  VIF                  =  427.21
Occupation_Small_Business  VIF            =  434.53
EducationField_Engineer  VIF              =  18.0
EducationField_MBA  VIF                   =  2.0
EducationField_Post_Graduate  VIF         =  17.68
EducationField_Under_Graduate  VIF        =  2.73
Gender_Male  VIF                          =  1.03
Designation_Executive  VIF                =  7.73
Designation_Manager  VIF                  =  5.43
Designation_Senior_Manager  VIF           =  2.73
Designation_VP  VIF                       =  1.84
MaritalStatus_Married  VIF                =  1.92
MaritalStatus_Single  VIF                 =  1.88
MaritalStatus_Unmarried  VIF              =  1.34
Zone_North  VIF                           =  19.18
Zone_South  VIF                           =  1.12
Zone_West  VIF                            =  19.15
PaymentMethod_Monthly  VIF                =  2.13
PaymentMethod_Quarterly  VIF              =  1.11
PaymentMethod_Yearly  VIF                 =  2.31
```

- Wherever VIF score > 5, multicollinearity is present
- Multicollinearity is detected for `Occupation_Large_Business, Occupation_Salaried, Occupation_Small_Business, EducationField_Engineer, EducationField_Post_Graduate, Designation_Executive, Designation_Manager(can be omitted), Zone_North, Zone_West.`

We still find we have multi collinearity in the dataset, to drop these values to a further lower level we can drop columns after performing stats model.

- **From stats model we can understand the features that do not contribute to the Model**
- ***We can remove those features after that the Vif Values will be reduced. Ideal value of VIF is less than 5%.***

Calculating VIF again after dropping variables having vif>5

```
Age  VIF                                =  1.32
CustTenure  VIF                         =  1.31
ExistingProdType  VIF                   =  3.53
NumberOfPolicy  VIF                     =  1.11
MonthlyIncome  VIF                      =  1.7
Complaint  VIF                          =  1.01
ExistingPolicyTenure  VIF               =  1.11
SumAssured  VIF                         =  1.71
LastMonthCalls  VIF                     =  1.17
CustCareScore  VIF                      =  1.02
Channel_Online  VIF                     =  1.02
EducationField_Engineer  VIF            =  1.11
EducationField_MBA  VIF                 =  1.03
EducationField_Post_Graduate  VIF       =  1.13
Gender_Male  VIF                        =  1.02
Designation_Manager  VIF                =  1.18
Designation_Senior_Manager  VIF         =  1.25
MaritalStatus_Married  VIF              =  1.92
MaritalStatus_Single  VIF               =  1.87
MaritalStatus_Unmarried  VIF            =  1.33
Zone_South  VIF                         =  1.01
Zone_West  VIF                          =  1.02
PaymentMethod_Monthly  VIF              =  1.92
PaymentMethod_Quarterly  VIF            =  1.09
PaymentMethod_Yearly  VIF               =  2.06
```

Running statsmodel again after dropping the necessary variables above - LINEAR MODEL 2 (LM2)

```
Intercept                      -235.677149
Age                              22.256764
CustTenure                       23.459540
ExistingProdType                -32.270239
NumberOfPolicy                    3.179880
MonthlyIncome                     0.062588
Complaint                        32.347109
ExistingPolicyTenure             40.038106
SumAssured                        0.003593
LastMonthCalls                    1.657254
CustCareScore                     9.045225
Channel_Online                   29.871935
EducationField_Engineer         -20.287296
EducationField_MBA              -97.213875
EducationField_Post_Graduate     10.231469
Gender_Male                      15.950300
Designation_Manager            -124.840296
Designation_Senior_Manager      -24.565951
MaritalStatus_Married           -54.039328
MaritalStatus_Single             16.120937
MaritalStatus_Unmarried        -205.556385
Zone_South                      144.726473
Zone_West                        -5.727819
PaymentMethod_Monthly            13.015562
PaymentMethod_Quarterly          34.504220
PaymentMethod_Yearly              4.557490
dtype: float64
```

This time we are getting a negative intercept

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              AgentBonus   R-squared:                       0.803
Model:                             OLS   Adj. R-squared:                  0.801
Method:                  Least Squares   F-statistic:                     547.2
Date:                 Sat, 11 Dec 2021   Prob (F-statistic):               0.00
Time:                         00:31:07   Log-Likelihood:                -26535.
No. Observations:                3390   AIC:                         5.312e+04
Df Residuals:                    3364   BIC:                         5.328e+04
Df Model:                          25
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                     -235.6771     93.849     -2.511      0.012    -419.684     -51.670
Age                             22.2568      1.431     15.552      0.000      19.451      25.063
CustTenure                      23.4595      1.437     16.323      0.000      20.642      26.277
ExistingProdType               -32.2702     21.099     -1.529      0.126     -73.638       9.097
NumberOfPolicy                   3.1799      7.601      0.418      0.676     -11.723      18.083
MonthlyIncome                    0.0626      0.003     18.138      0.000       0.056       0.069
Complaint                       32.3471     23.352      1.385      0.166     -13.438      78.132
ExistingPolicyTenure            40.0381      4.095      9.777      0.000      32.009      48.067
SumAssured                       0.0036    5.9e-05     60.886      0.000       0.003       0.004
LastMonthCalls                   1.6573      3.097      0.535      0.593      -4.414       7.729
CustCareScore                    9.0452      7.700      1.175      0.240      -6.051      24.142
Channel_Online                  29.8719     34.341      0.870      0.384     -37.460      97.204
EducationField_Engineer        -20.2873     38.882     -0.522      0.602     -96.521      55.947
EducationField_MBA             -97.2139     90.008     -1.080      0.280    -273.689      79.262
EducationField_Post_Graduate    10.2315     22.269      0.459      0.646     -33.430      53.893
Gender_Male                     15.9503     21.457      0.743      0.457     -26.119      58.020
Designation_Manager           -124.8403     23.744     -5.258      0.000    -171.395     -78.286
Designation_Senior_Manager     -24.5660     32.955     -0.745      0.456     -89.180      40.048
MaritalStatus_Married          -54.0393     28.999     -1.864      0.062    -110.896       2.818
MaritalStatus_Single            16.1209     32.012      0.504      0.615     -46.645      78.887
MaritalStatus_Unmarried       -205.5564     59.836     -3.435      0.001    -322.876     -88.237
Zone_South                     144.7265    273.767      0.529      0.597    -392.041     681.493
Zone_West                       -5.7278     21.280     -0.269      0.788     -47.451      35.996
PaymentMethod_Monthly           13.0156     54.141      0.240      0.810     -93.137     119.168
PaymentMethod_Quarterly         34.5042     85.134      0.405      0.685    -132.416     201.425
PaymentMethod_Yearly             4.5575     32.348      0.141      0.888     -58.866      67.981
==============================================================================
Omnibus:                      160.583   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              188.423
Skew:                           0.522   Prob(JB):                     1.21e-41
Kurtosis:                       3.494   Cond. No.                     1.72e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

As it can be observed above the P-value for multiple variables are greater than our alpha i.e 0.05, depicting multicollinearity present therefore we will drop the variables and perform the statsmodel again.

- **To ideally bring down the values to lower levels we can drop one of the variable that is highly correlated.**
- ***Dropping variables would bring down the multi collinearity level down***

	RMSE (LM2)	**RMSE (LM1)**
Training	607.0547411435514	600.5900784990952
Testing	629.0548786960638	621.5274260080358

Since for model 2 our RMSE value has increased, it is not an optimal way to choose the new model. Not a significant change in R-squared either.
Removing variables until all the insignificant variables are removed.

```
                            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:                        00:44:36   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
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.23e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The overall P value is less than alpha, so rejecting H0 and accepting Ha that atleast 1 regression co-efficient is not 0. Here all regression co-efficients are not 0

We can see all variables are having p-value < 0.05 and the r-squared value hasn't changes much either

	RMSE (LM2)	RMSE (LM1)
Training	602.6246250878111	600.5900784990952
Testing	620.4861930401804	621.5274260080358

Since for model 2 our RMSE value has increased, it is not an optimal way to choose the new model.

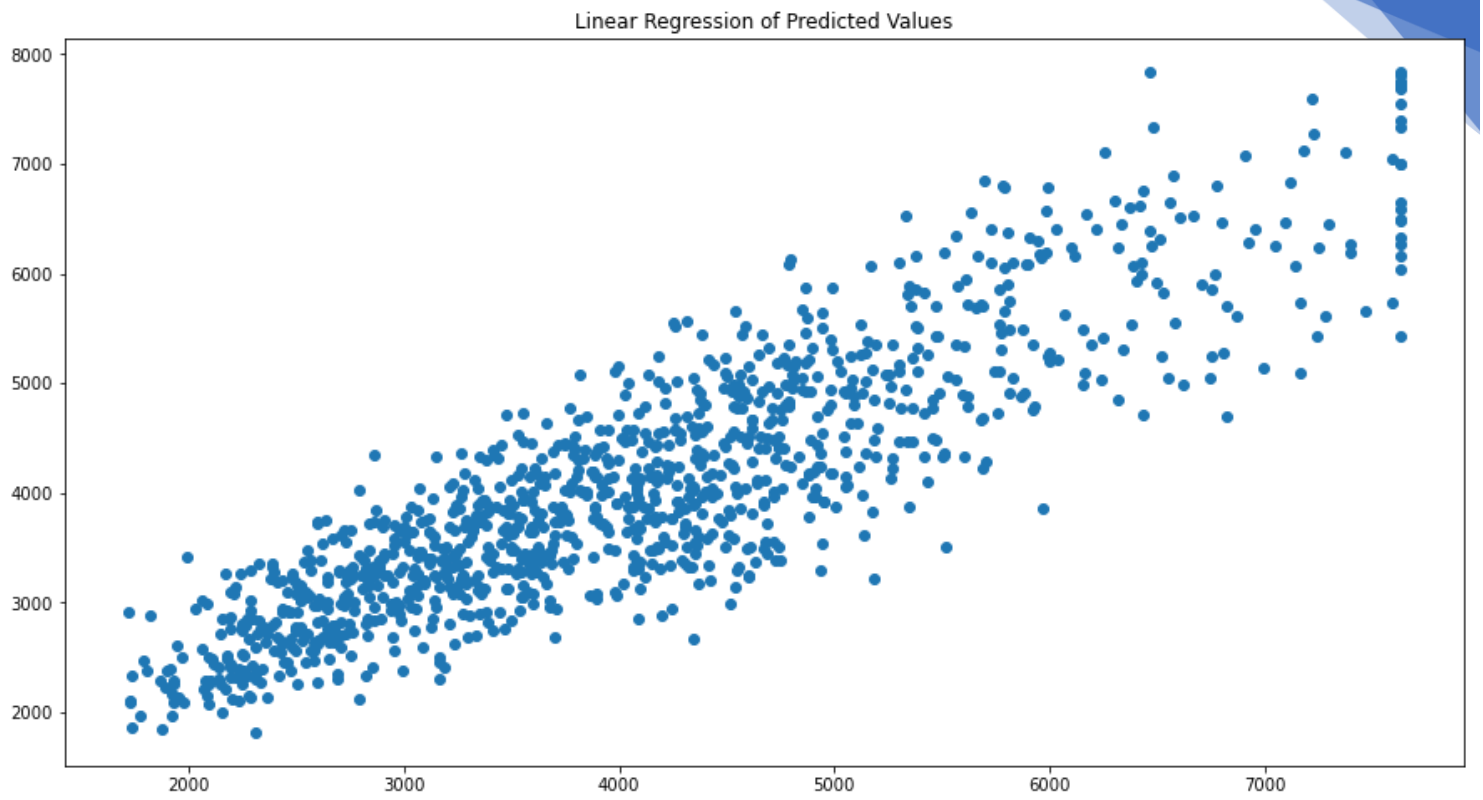

Figure (1) - Linear Regression Scatterplot

The variables are following a linear trend with a little homoscedasticity.

Comparing Linear Regression Model with Other models like Random Forest, Artificial Neural Network and Decision Trees – With base parameter values are no hyperparameter tuning the parameters.

We are scaling the data for ANN. Without scaling it will give very poor results. Computations becomes easier

Scaling is done as some variables with greater weight will affect the predictions more, hence scaling is done to bring all variables in a common range e.g., 0 to 1. Due to which the predictions can be unbiased and not biased to one specific variable with higher weights. For e.g., age and sum assured.

SCALING

- **Scaling can be useful to reduce or check the multi collinearity in the data, so if scaling is not applied, I find the VIF – variance inflation factor values very high. Which indicates presence of multi collinearity**
- ***These values are calculated after building the model of linear regression. To understand the multi collinearity in the model***
- ***The scaling had no impact in model score or coefficients of attributes nor the intercept.***

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	612.550689	585.514819	0.800806	0.801482
Decision Tree Regressor	0.000000	725.006753	1.000000	0.695626
Random Forest Regressor	189.614010	519.044211	0.980913	0.843997
ANN Regressor	225.889011	701.144120	0.972912	0.715332

Here Linear Regression is the best performing model with almost same Training and Testing Accuracies.

On the other hand, we can observe that the other three models namely, Decision Tree, Random Forest, and ANN are Overfitting the model, i.e. the model is performing better while training but poorly while testing.

To fix this we will use Hyperparameter Tuning, this will be done by performing grid search .

Checking if PCA can be applied here.

```
Cumulative Variance Explained [ 99.97511098  99.99912638  99.99999976  99.99999986  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.        ]
```

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

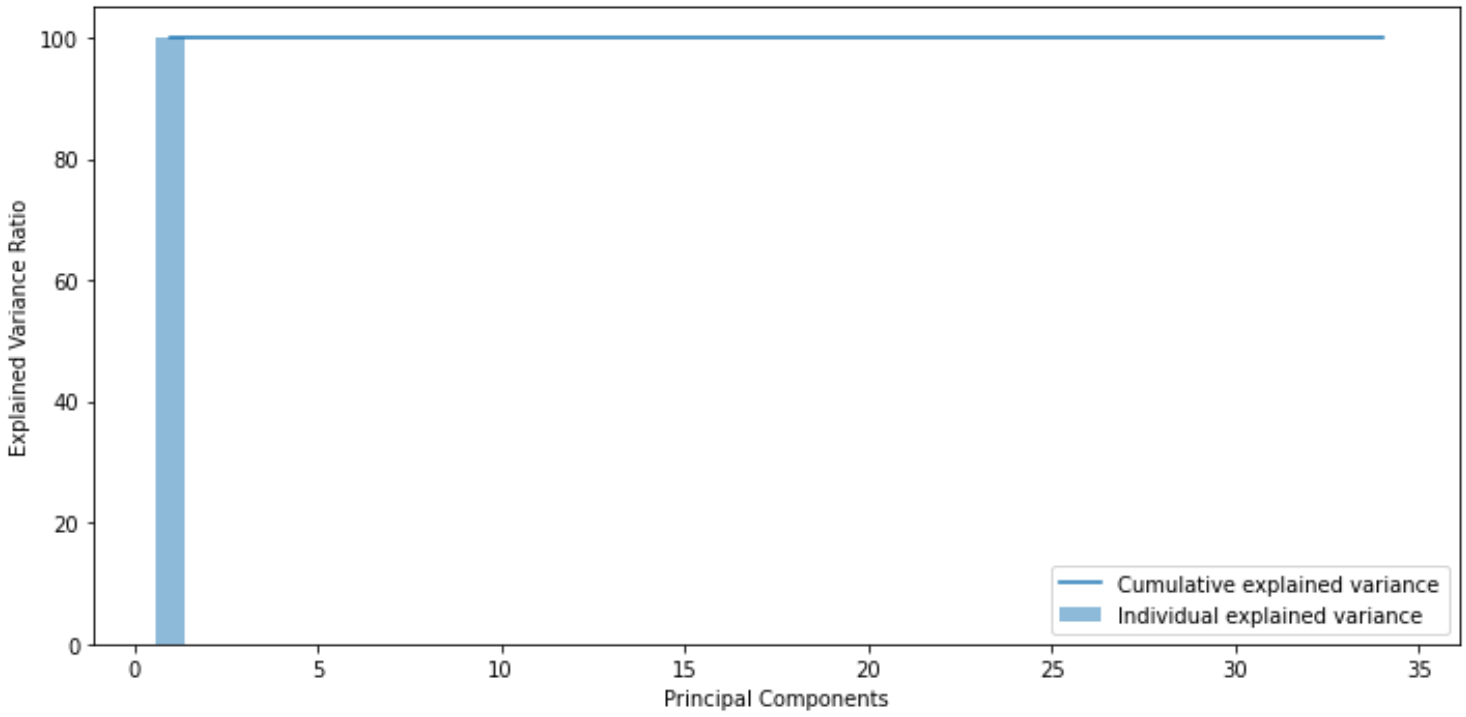

Figure 2 – Principal Components vs Variance Ratio

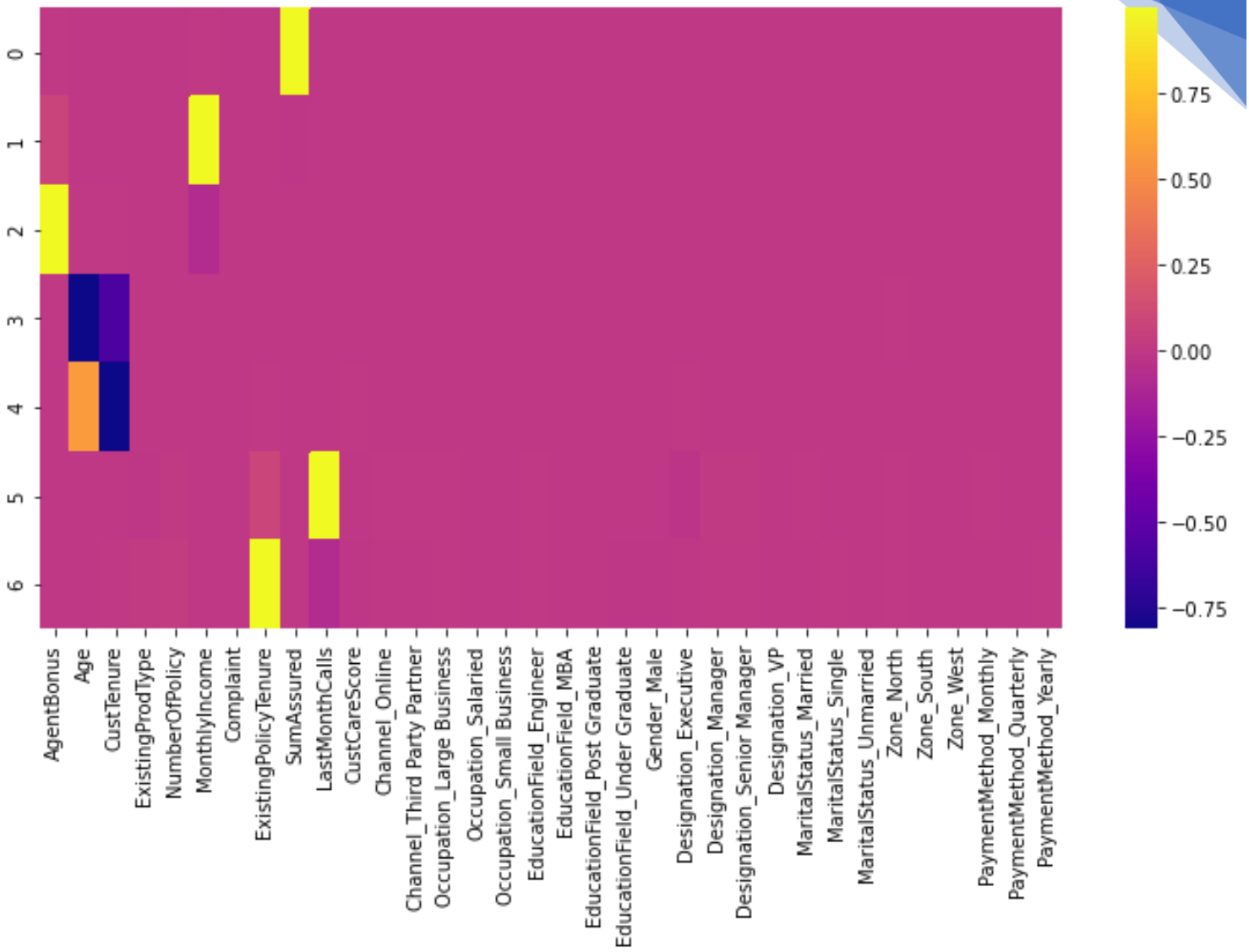

Figure 3 – PCA Heatmap

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 TUNING

We will perform grid search for hyperparameter tuning and check if that makes a difference in our accuracies.

Grid Search on Decision Tree

Best parameters - `{'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 40}`

Grid Search on Random Forest

```
GridSearchCV(cv=3, estimator=RandomForestRegressor(random_state=123),
             param_grid={'max_depth': [7, 10], 'max_features': [4, 6],
                         'min_samples_leaf': [3, 15, 30],
                         'min_samples_split': [30, 50, 100],
                         'n_estimators': [300, 500]})

Best Parameters - {'max_depth': 10, 'max_features': 6, 'min_samples_leaf': 3,
                   'min_samples_split': 30, 'n_estimators': 500}
```

Using Grid Search for ANN

```
GridSearchCV(cv=3, estimator=MLPRegressor(max_iter=10000, random_state=123),
             param_grid={'activation': ['tanh', 'relu'],
                         'hidden_layer_sizes': [500, (100, 100)],
                         'solver': ['sgd', 'adam']})

Best parameters - {'activation': 'tanh', 'hidden_layer_sizes': 500, 'solver': 'adam'}
```

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	612.550689	585.514819	0.800806	0.801482
Decision Tree Regressor	495.463438	569.694730	0.869679	0.812065
Random Forest Regressor	527.410585	572.885614	0.852331	0.809954
ANN Regressor	28.117642	670.444991	0.999580	0.739715

After Hyperparameter tuning it can be observed the problem of overfitting is removed for most of the models however some overfitting can be observed in ANN.

Apart from this, we can observe Linear Regression is still the most stable having not much variation between training and testing sets.

If you're looking for more stable Model, definitely go for Linear Regression model, else Decision Tree and Random Forest can be chosen for higher accuracy and are good models as the there's only 5% fluctuations between training and testing model. Random forest is the better choice between the Regressors as random forest is the more advanced version of decision trees where we can further tweak the parameters according to the needs.

Feature Importance from the model can be observed here:

	Imp
SumAssured	0.428155
CustTenure	0.155577
Age	0.144097
MonthlyIncome	0.113766
ExistingPolicyTenure	0.038903
Designation_Executive	0.032743
Designation_VP	0.027304
LastMonthCalls	0.010814
Designation_Manager	0.010730
Designation_Senior Manager	0.007526
ExistingProdType	0.004708
NumberOfPolicy	0.004006
MaritalStatus_Unmarried	0.003666
CustCareScore	0.002908
Zone_North	0.001236
MaritalStatus_Single	0.001231
MaritalStatus_Married	0.001103
Gender_Male	0.001099
Channel_Third Party Partner	0.001056
Complaint	0.001049
Zone_West	0.001029
EducationField_Post Graduate	0.000941
Occupation_Salaried	0.000940
EducationField_Under Graduate	0.000844

```
PaymentMethod_Yearly            0.000832
Occupation_Small Business       0.000793
Channel_Online                  0.000773
PaymentMethod_Monthly           0.000698
EducationField_Engineer         0.000623
Occupation_Large Business       0.000546
PaymentMethod_Quarterly         0.000171
EducationField_MBA              0.000131
Zone_South                      0.000003
```

Sum Assured is the most important feature here, Zone_South being the least important.

The Equation

(1092.35) * Intercept + (21.65) * Age + (22.62) * CustTenure + (46.51) * ExistingProdType + (6.25) * NumberOfPolicy + (0.03) * MonthlyIncome + (33.05) * Complaint + (40.23) * ExistingPolicyTenure + (0.0) * SumAssured + (-2.31) * LastMonthCalls + (7.56) * CustCareScore + (22.69) * Channel_Online + (3.5) * Channel_Third_Party_Partner + (-616.86) * Occupation_Large_Business + (-474.97) * Occupation_Salaried + (-581.64) * Occupation_Small_Business + (26.68) * EducationField_Engineer + (-177.27) * EducationField_MBA + (-92.61) * EducationField_Post_Graduate + (2.33) * EducationField_Under_Graduate + (25.19) * Gender_Male + (-493.36) * Designation_Executive + (-481.42) * Designation_Manager + (-277.42) * Designation_Senior_Manager + (-2.96) * Designation_VP + (-48.2) * MaritalStatus_Married + (29.66) * MaritalStatus_Single + (-188.88) * MaritalStatus_Unmarried + (62.35) * Zone_North + (193.51) * Zone_South + (50.0) * Zone_West + (141.95) * PaymentMethod_Monthly + (112.03) * PaymentMethod_Quarterly + (-79.92) * PaymentMethod_Yearly

Interpretation and Business Recommendations.

- **Company wants to predict the ideal bonus and what is the engagement for high and low performing agents respectively.**
- **From the model, the 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.**