**1**

**Life Insurance Sales – Capstone**

**Project Notes – 2**

**SARATH KUMAR V**

**2**

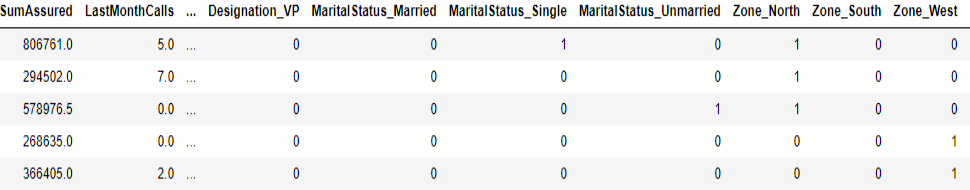
*Table of Contents*

**1. Model Building and Interpretation ................................................................................................. 3 2. Split X and y into training and test set in 75:25 ratio……………………………………………………3 3. Building a Multiple Linear Regression Model…….………………....................................................5 4. VIF Calculation.........................................................................................................................................7 5. Stats-model Implementation……………..……..………………………………………………………………..9**

**6. Model Tuning…………………………………….………….……………………………………………………………13 7. Feature Importance……………………………………………………………………………………………………14 8. Recommendations………………………………….……………...…………………………………………………..15**

*Table of Figures*

**Figure 1 – Linear Regression Scatterplot..................................................................................... 8 Figure 2 - Principal Components vs Variance Ratio.................................................................13 Figure 3 - PCA Heatmap ……………………………………………………………………………………………15**

**3** **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.

· 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

The coefficient for CustTenure is

The coefficient for ExistingProdType is The coefficient for NumberOfPolicy is The coefficient for MonthlyIncome is The coefficient for Complaint is

The coefficient for ExistingPolicyTenure is The coefficient for SumAssured is

The coefficient for LastMonthCalls is The coefficient for CustCareScore is The coefficient for Channel\_Online is

The coefficient for Channel\_Third Party Partner is The coefficient for Occupation\_Large Business is The coefficient for Occupation\_Salaried is

The coefficient for Occupation\_Small Business is The coefficient for EducationField\_Engineer is The coefficient for EducationField\_MBA is

The coefficient for EducationField\_Post Graduate is The coefficient for EducationField\_Under Graduate is The coefficient for Gender\_Male is

The coefficient for Designation\_Executive is The coefficient for Designation\_Manager is

The coefficient for Designation\_Senior Manager is The coefficient for Designation\_VP is

The coefficient for MaritalStatus\_Married is The coefficient for MaritalStatus\_Single is The coefficient for MaritalStatus\_Unmarried is The coefficient for Zone\_North is

The coefficient for Zone\_South is The coefficient for Zone\_West is

The coefficient for PaymentMethod\_Monthly is The coefficient for PaymentMethod\_Quarterly is The coefficient for PaymentMethod\_Yearly is

The intercept for our model is 1092.3485100144962

21.64543636236496 22.620905021409023 46.508784274329514 6.254332127798309 0.03188513622751349 33.0503807570841 40.22901549596465 0.003548018281339438 -2.308709717687992 7.559056565466554 22.691900907509453 3.4952779925482345 -616.8600099371561 -474.9729637586688 -581.6372411869505 26.675848148157876 -177.27368717977166 -92.6094978672669 2.331225272073949 25.187256483000322 -493.36122500604984 -481.4192660702273 -277.42121914512296 -2.956791388368395 -48.20378324641499 29.658243912402032 -188.87907531620797 62.35415312785426 193.51057687776427 49.998087081147155 141.95193527244763 112.02879394979776 -79.92080455281895



**Training Testing**

**R-Squared** 0.8068152802160813 0.7825646087670782

**4**

**RMSE** 600.5900784990952 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')

**5**

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

Intercept 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 dtype: float64

1092.348510 21.645436 22.620905 46.508784 6.254332 0.031885 33.050381 40.229015 0.003548

-2.308710 7.559057 22.691901 3.495278

-616.860010 -474.972964 -581.637241 26.675848 -177.273687 -92.609498 2.331225 25.187256 -493.361225 -481.419266 -277.421219 -2.956791 -48.203783 29.658244 -188.879075 62.354153 193.510577 49.998087 141.951935 112.028794 -79.920805

**6**



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

**7**



**VIF Value**

Age VIF CustTenure VIF

ExistingProdType VIF NumberOfPolicy VIF MonthlyIncome VIF Complaint VIF ExistingPolicyTenure VIF SumAssured VIF LastMonthCalls VIF CustCareScore VIF Channel\_Online VIF

Channel\_Third\_Party\_Partner VIF Occupation\_Large\_Business VIF Occupation\_Salaried VIF Occupation\_Small\_Business VIF EducationField\_Engineer VIF EducationField\_MBA VIF EducationField\_Post\_Graduate VIF EducationField\_Under\_Graduate VIF Gender\_Male VIF Designation\_Executive VIF Designation\_Manager VIF Designation\_Senior\_Manager VIF Designation\_VP VIF MaritalStatus\_Married VIF MaritalStatus\_Single VIF MaritalStatus\_Unmarried VIF Zone\_North VIF

Zone\_South VIF Zone\_West VIF

PaymentMethod\_Monthly VIF PaymentMethod\_Quarterly VIF PaymentMethod\_Yearly VIF

= 1.33 = 1.32 = 4.36 = 1.12 = 4.17 = 1.01 = 1.11 = 1.73 = 1.2 = 1.03 = 1.05 = 1.04

= 153.84 = 427.21 = 434.53 = 18.0 = 2.0

= 17.68 = 2.73 = 1.03 = 7.73 = 5.43 = 2.73 = 1.84 = 1.92 = 1.88 = 1.34 = 19.18 = 1.12 = 19.15 = 2.13 = 1.11 = 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%.***

**8**



**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 Age CustTenure

ExistingProdType NumberOfPolicy MonthlyIncome Complaint ExistingPolicyTenure SumAssured LastMonthCalls CustCareScore Channel\_Online EducationField\_Engineer EducationField\_MBA

EducationField\_Post\_Graduate Gender\_Male Designation\_Manager Designation\_Senior\_Manager MaritalStatus\_Married MaritalStatus\_Single MaritalStatus\_Unmarried Zone\_South

Zone\_West PaymentMethod\_Monthly PaymentMethod\_Quarterly PaymentMethod\_Yearly dtype: float64

-235.677149 22.256764 23.459540 -32.270239 3.179880 0.062588 32.347109 40.038106 0.003593 1.657254 9.045225 29.871935 -20.287296 -97.213875 10.231469 15.950300 -124.840296 -24.565951 -54.039328 16.120937 -205.556385 144.726473 -5.727819 13.015562 34.504220 4.557490

**9** 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 stats model again.**

· **To ideally bring down the values to lower levels we can drop one of the variables that is highly correlated.**

· ***Dropping variables would bring down the multi collinearity level down***

**Training**

**Testing**

**RMSE (LM2)** 607.0547411435514

629.0548786960638

**RMSE (LM1)** 600.5900784990952

621.5274260080358

**10**

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**

**Training**

**Testing**

**RMSE (LM2)** 602.6246250878111

620.4861930401804

**RMSE (LM1)** 600.5900784990952

621.5274260080358

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

**11**

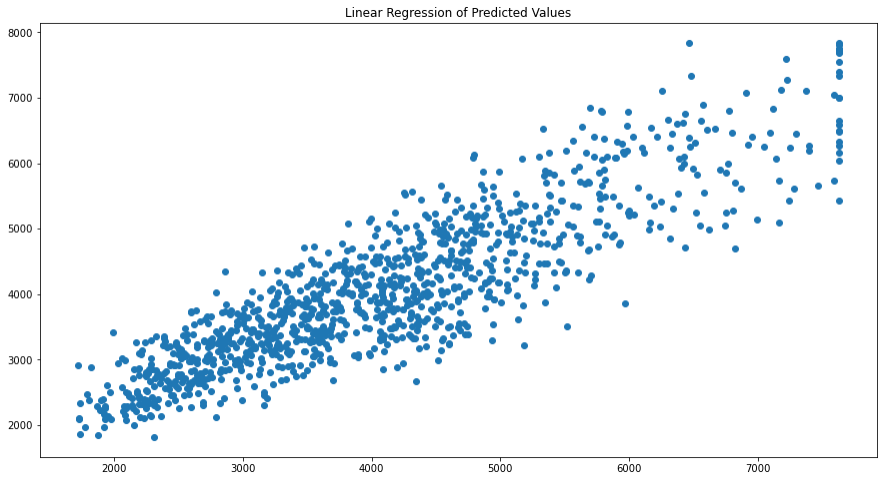


Figure (1) - Linear Regression Scatterplot

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.***

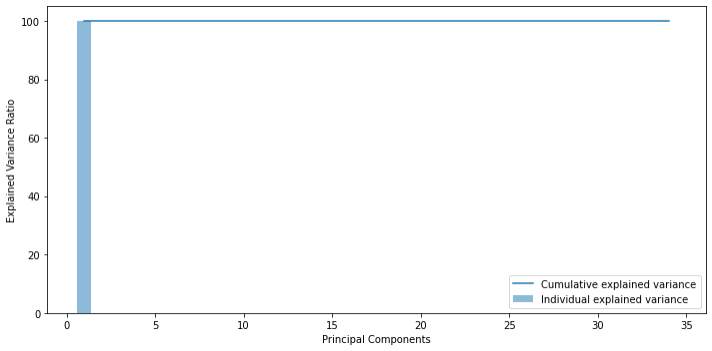
**Linear Regression Decision Tree Regressor Random Forest Regressor ANN Regressor**

**Train RMSE** 612.550689 0.000000 189.614010 225.889011

**Test RMSE** 585.514819 725.006753 519.044211 701.144120

**Training Score** 0.800806 1.000000 0.980913 0.972912

**Test Score** 0.801482 0.695626 0.843997 0.715332

**12**

**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

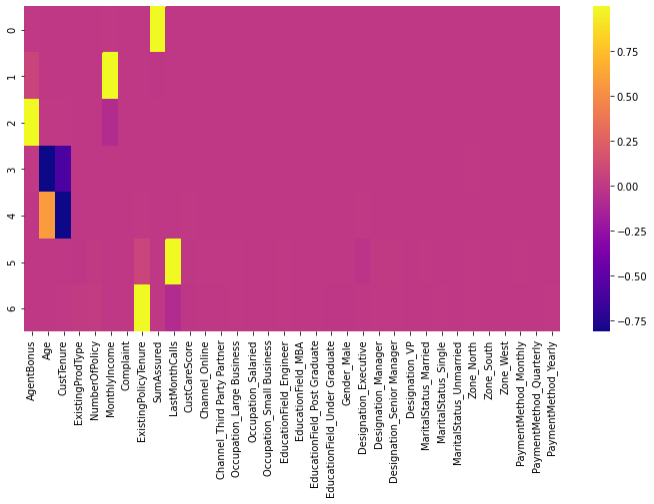
99.99999997 99.99999998 99.99999999 99.99999999 99.99999999 99.99999999 100. 100. 100. 100.

100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 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**

**13**



**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}

**14** **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'}

**Linear Regression Decision Tree Regressor Random Forest Regressor ANN Regressor**

**Train RMSE** 612.550689 495.463438 527.410585 28.117642

**Test RMSE** 585.514819 569.694730 572.885614 670.444991

**Training Score** 0.800806 0.869679 0.852331 0.999580

**Test Score** 0.801482 0.812065 0.809954 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, 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:**

SumAssured CustTenure Age MonthlyIncome

ExistingPolicyTenure Designation\_Executive Designation\_VP LastMonthCalls Designation\_Manager Designation\_Senior Manager ExistingProdType NumberOfPolicy MaritalStatus\_Unmarried CustCareScore

Zone\_North MaritalStatus\_Single MaritalStatus\_Married Gender\_Male

Channel\_Third Party Partner Complaint

Zone\_West EducationField\_Post Graduate Occupation\_Salaried EducationField\_Under Graduate

Imp 0.428155 0.155577 0.144097 0.113766 0.038903 0.032743 0.027304 0.010814 0.010730 0.007526 0.004708 0.004006 0.003666 0.002908 0.001236 0.001231 0.001103 0.001099 0.001056 0.001049 0.001029 0.000941 0.000940 0.000844



PaymentMethod\_Yearly Occupation\_Small Business Channel\_Online PaymentMethod\_Monthly EducationField\_Engineer Occupation\_Large Business PaymentMethod\_Quarterly EducationField\_MBA Zone\_South

**15**

0.000832 0.000793 0.000773 0.000698 0.000623 0.000546 0.000171 0.000131 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) \* ExistingProdTy pe + (6.25) \* NumberOfPolicy + (0.03) \* MonthlyIncome + (33.05) \* Complaint + (40.23) \* ExistingPolicyTenure + (0.0) \* SumAssured + (-2.31) \* LastMonthCalls + (7.56) \* CustCar eScore + (22.69) \* Channel\_Online + (3.5) \* Channel\_Third\_Party\_Partner + (-616.86) \* O ccupation\_Large\_Business + (-474.97) \* Occupation\_Salaried + (-581.64) \* Occupation\_Sma ll\_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) \* MaritalSta tus\_Married + (29.66) \* MaritalStatus\_Single + (-188.88) \* MaritalStatus\_Unmarried + (6 2.35) \* Zone\_North + (193.51) \* Zone\_South + (50.0) \* Zone\_West + (141.95) \* PaymentMet hod\_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.**