***Life-Insurance Sale Capstone***

***Final Report***

**Tushar Babbar**

**(PGP DSBA Dec20)**

*Table of Contents*

1. **Problem Statement: Life Insurance Data ......................**...................................................................................... **3**
2. **Need for this Study/Project**.................................................................................................................................................3
3. **Why is this (agent bonus) important for the business/company?** ....................................................3
4. **Data Report/Dictionary**..........................................................................................................................................................4
5. **Performing Exploratory Data Analysis (EDA)**………………………………………………………………..4
6. **Checking for Unique Categorical Values**………………………………………………………………7
7. **Univariate/Bivariate Analysis**………………………………………………………………………………8
8. **Categorical Variable’s Univariate Analysis**…………………………………………………………13
9. **Bivariate Analysis ( Pairplot/Heatmap)**……………………………………………………………..18

**10. Model Building and Interpretation ............................................................................................... 20**

**11. Split X and y into training and test set in 75:25 ratio…………………………………………………20**

**12. Building a Multiple Linear Regression Model…….……………….................................................22**

**13. VIF Calculation.......................................................................................................................................24**

**14. Stats-model Implementation……………..……..…………………………………………………………….…25**

**15. Model Tuning………………………………….………….…………………………………………………………..…30**

1. **Feature Importance…………………………………………………………………………………………...……31**
2. **Recommendations………………………….……………...……………………………………………………....34**

*Table of Figures*

**Figure 1 – Univariate Analysis – Numerical Columns** ....................................................................... 8

**Figure 2 - Univariate Analysis – Categorical Columns** .....................................................................13

**Figure 3 - Categorical variable w.r.t Target Variable (Bivariate)**………………………………..15

**Figure 4 - Bivariate Analysis- Numerical Data Pairplot**...................................................................18

**Figure 5 - Bivariate Analysis- Numerical Data Correlation Heatmap**......................................19

**Figure 6 – Linear Regression Scatterplot..................................................................................... 28**

**Figure 7 - Principal Components vs Variance Ratio.................................................................29**

**Figure 8 - PCA Heatmap ……………………………………………………………………………………………30**

**Problem Statement: Life Insurance Data**

* The dataset belongs to a leading life insurance company.
* The company wants to predict the bonus for its agents so that it may design appropriate engagement activity for their high performing agents and upskill programs for low performing agents.

**Need for this Study/Project**

* With this problem we want to better understand how the insurance company agents are performing, it’s not to underpay or overpay, as the payment is regulated by IRDA.
* With the predictions it’s better for the company to understand where they need to focus more as for agents selling less policies the company needs some booster training performs. As the policies are as good as the agents portray it to be to the potential customer.
* While the agents performing good i.e. selling more policies there needs to be a way to reward them, to make their contribution known so that they perform the same and even better in future.

**Why is this (agent bonus) important for the business/company?**

* A company is as good as their employers.
* For a Life Insurance Company, their agents are the best way to make the companies policies, aims, and perks known to the customer. Once the customer is intrigued by the policy delivery by the agent, it is easier to convince the customer hence improving the sales and thereby motivating the agent as well.
* With this, the market share of the company will gain more ground dominating the potential opponents.
* Moreover, the agents can be classified into categories giving the company better insight where the need to put more effort.
* The customer feedback can help the company develop improved and updated policies/products. Meeting customer needs.
* Hereby, the easiest way to retain their agents.
* Overall, multiplying and adding to company’s profit.

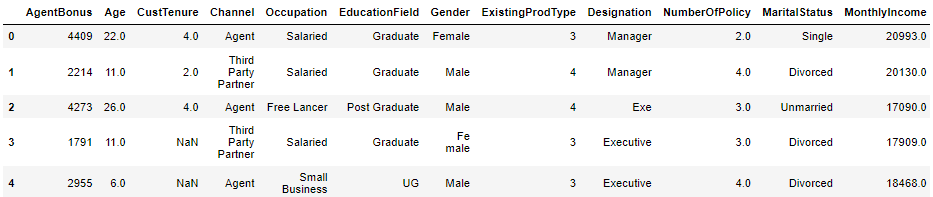
**Data Report/Dictionary**

The following data is provided by Great Learning cover the Life Insurance Sales made by the company, the data dictionary consists of:

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

**Performing Exploratory Data Analysis (EDA).**

**Head of the Data**



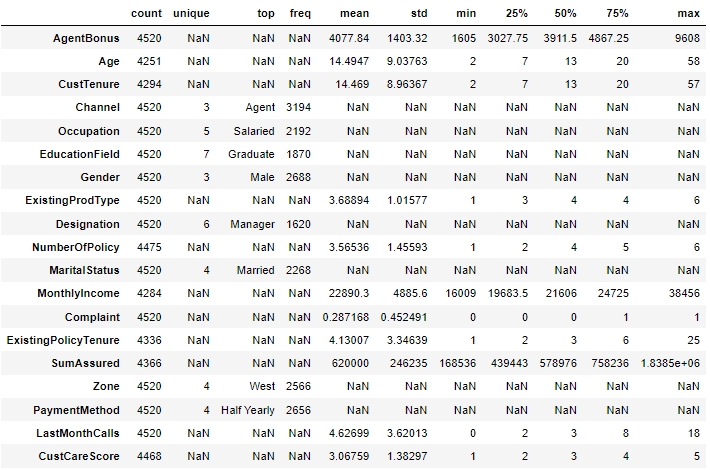
* I’ve removed CustID as it is irrelevant to agent bonus.
* Head gives us the idea of what the basic dataset looks like.
* Complete list of all variables is not presented.

**Shape of the dataset**

Total rows in the dataset: 4520

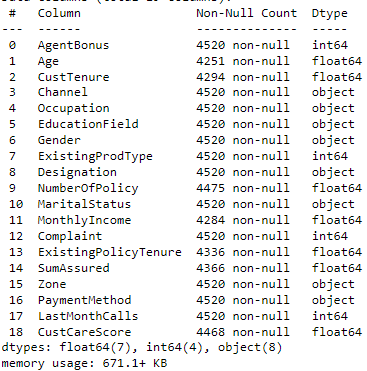
Total columns in the dataset: 19

**Descriptive Statistics of the Columns**



* The table includes the complete description for all variable with categorical variables included.
* The description includes, variable count, unique values, top frequently occurring categories like Agent-3194, mean, standard deviation, minimum, 25%, 50%(median), 75%, and maximum values present in the respective variables.
* Hence the ‘NaN’ here is observed for Categorical Variables as a string object cannot have numeric values.
* This we will change by encoding the data in future if needed.
* We can also observe the missing values as the count is not constant for all the variables.
* The unique is only present for categorical variables which hold a specific category
* Example: Gender has male and female hence it should hold unique value of 2 but later we observed some subcategories needs to be renamed.

**Info of the parameters**



* We have 7 parameters having ‘float’ data type.
* We have 4 parameters having ‘integer’ data type.
* We have 8 parameters having ‘object’ data type.
* Age is shown as float, however we will later observe is its needed to change it to int or not, it won’t make any difference in our observations.
* We can clearly observe some missing values.
* Further count of missing values is provided below.
* CustID 0
* AgentBonus 0
* Age 269
* CustTenure 226
* Channel 0
* Occupation 0
* EducationField 0
* Gender 0
* ExistingProdType 0
* Designation 0
* NumberOfPolicy 45
* MaritalStatus 0
* MonthlyIncome 236
* Complaint 0
* ExistingPolicyTenure 184
* SumAssured 154
* Zone 0
* PaymentMethod 0
* LastMonthCalls 0
* CustCareScore 52
* **Number of duplicate rows = 0**
* The **Missing values** can affect the prediction’s hence need to be treated, hence the missing values are

imputed with the **median values** in the respective column.

**Checking for Unique Categorical Values.**

CHANNEL has 3 Unique Values.

Online 468

Third Party Partner 858

Agent 3194

Name: Channel, dtype: int64

OCCUPATION has 5 Unique Values.

Free Lancer 2

Laarge Business 153

Large Business 255

Small Business 1918

Salaried 2192

Name: Occupation, dtype: int64

EDUCATIONFIELD has 7 Unique Values.

MBA 74

UG 230

Post Graduate 252

Engineer 408

Diploma 496

Under Graduate 1190

Graduate 1870

Name: EducationField, dtype: int64

GENDER has 3 Unique Values.

Fe male 325

Female 1507

Male 2688

Name: Gender, dtype: int64

DESIGNATION has 6 Unique Values.

Exe 127

VP 226

AVP 336

Senior Manager 676

Executive 1535

Manager 1620

Name: Designation, dtype: int64

MARITALSTATUS has 4 Unique Values.

Unmarried 194

Divorced 804

Single 1254

Married 2268

Name: MaritalStatus, dtype: int64

ZONE has 4 Unique Values.

South 6

East 64

North 1884

West 2566

Name: Zone, dtype: int64

PAYMENTMETHOD has 4 Unique Values.

Quarterly 76

Monthly 354

Yearly 1434

Half Yearly 2656

* Here it can be observed that subcategories highlighted with a different colour shows

an error in naming convention hence have to be renamed.

* Example: ‘Laarge’ and ‘Large’ Business can be put in the same category, the same for

‘UG’ and ‘Under Graduate’, ‘Graduate’ and ‘Post Graduate’, ‘Fe male’ and ‘Female’, and

‘Exe’ and ‘Executive’**.**

**Univariate/Bivariate Analysis**

**AgentBonus**

|  |  |
| --- | --- |
|  |  |

Figure 1(a) Distplot/Histplot - AgentBonus

* **The distribution of "AgentBonus" seems to be positively/right skewed.**
* **The data ranges from 1605 to 9600.**
* **The box plot holds many outliers.**

**Age:**

|  |  |
| --- | --- |
|  |  |

Figure 1(b) Distplot/Histplot - Age

* **The distribution of "Age" seems to be positively/right skewed.**
* **The data ranges from 2 to 58.**
* **The box plot holds many outliers.**

**CustTenure:**

|  |  |
| --- | --- |
|  |  |

Figure 1(c) Distplot/Histplot - CustTenure

* **The distribution of "CustTenure" seems to be positively/right skewed.**
* **The data ranges from 2 to 57.**
* **The box plot holds many outliers.**

**ExistingProdType:**

|  |  |
| --- | --- |
|  |  |

Figure 1(d) Distplot/Histplot - ExistingProdType

* **The distribution of "ExistingProdType" seems to be slightly left skewed.**
* **The data ranges from 1 to 6.**
* **The box plot holds outliers.**

**NumberOfPolicy**

|  |  |
| --- | --- |
|  |  |

Figure 1(e) Distplot/Histplot - NumberofPolicy

* **The distribution of "NumberOfPolicy" seems to be slightly left skewed.**
* **The data ranges from 1 to 6.**
* **The box plot has no outliers.**

**MonthlyIncome**

|  |  |
| --- | --- |
|  |  |

Figure 1(f) Distplot/Histplot - MonthlyIncome

* **The distribution of "MonthlyIncome" seems to be positively/right skewed.**
* **The data ranges from 16000 to 38500.**
* **The box plot holds many outliers**.

**Complaint**

|  |  |
| --- | --- |
|  |  |

Figure 1(g) Distplot/Histplot - Complaint

* **The distribution of "Complaint" seems to be positively/right skewed.**
* **The data ranges from 0 to 1.**
* **The box plot holds no outliers.**

**ExistingPolicyTenure**

|  |  |
| --- | --- |
|  |  |

Figure 1(h) Distplot/Histplot - ExistingPolicyTenure

* **The distribution of "ExistingPolicyTenure" seems to be positively/right skewed.**
* **The data ranges from 1 to 25.**
* **The box plot holds many outliers.**

**SumAssured**

|  |  |
| --- | --- |
|  |  |

Figure 1(i) Distplot/Histplot - SumAssured

* **The distribution of "SumAssured" seems to be positively/right skewed.**
* **The data ranges from 1.68 \* 105 to 1.83 \* 105.**
* **The box plot holds many outliers.**

**LastMonthCalls**

|  |  |
| --- | --- |
|  |  |

Figure 1(j) Distplot/Histplot - LastMonthCalls

* **The distribution of "LastMonthCalls" seems to be positively/right skewed.**
* **The data ranges from 0 to 18.**
* **The box plot holds outliers.**

**CustCareScore**

|  |  |
| --- | --- |
|  |  |

Figure 1(k) Distplot/Histplot – CustCareScore

* **The distribution of "CustCareScore" seems to be slightly left skewed.**
* **The data ranges from 1 to 5.**
* **The box plot holds no outliers**

**Skewness**

AgentBonus 0.822348

Age 0.998425

CustTenure 0.981002

ExistingProdType -0.401100

NumberOfPolicy -0.108161

MonthlyIncome 1.434315

Complaint 0.941129

ExistingPolicyTenure 1.601730

SumAssured 1.002018

LastMonthCalls 0.810417

CustCareScore -0.138120

* We can observe skewness in the data with ExistingProdType, NumberofPoilicy and

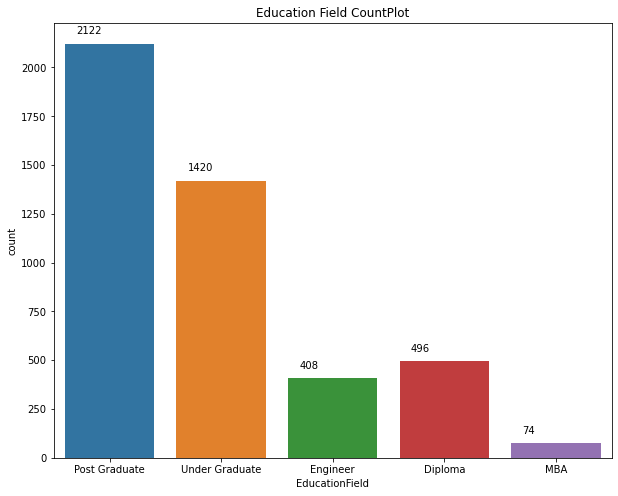
CustCareScore being negatively skewed.

* Rest all other parameters holds positive skewness the max being for

ExistingPolicyTenure.

**Categorical Variable’s Univariate Analysis**

**Education Field**

****

Post Graduate 0.47

Under Graduate 0.31

Diploma 0.11

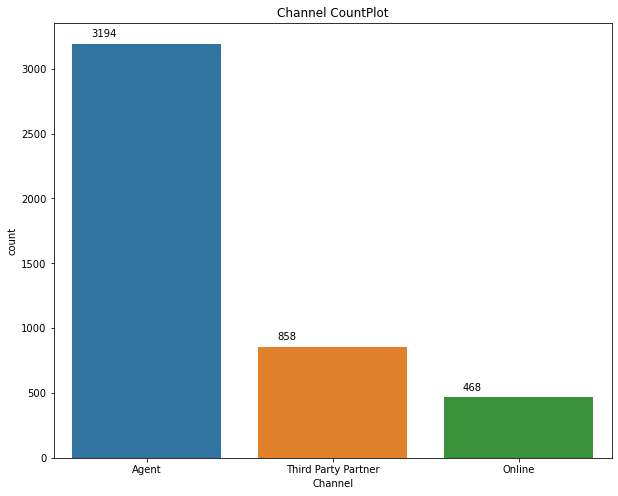
Engineer 0.09

MBA 0.02

Most Customers approached are Post

Graduates having 47% weightage.

Figure 2(a) Count Plot - EducationField

****

**Channel**

Agent 0.71

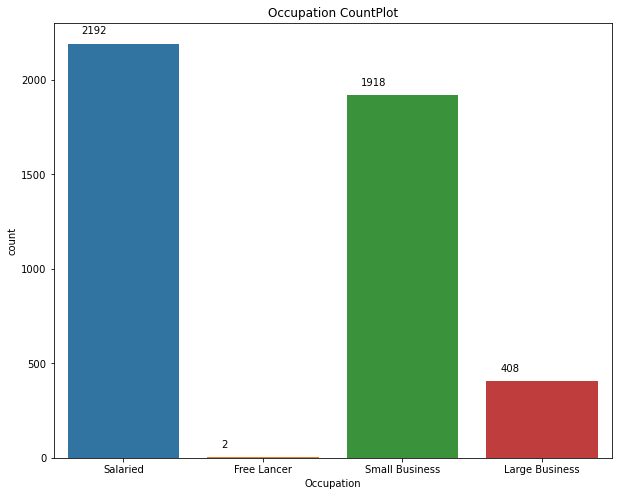
Third Party Partner 0.19

Online 0.10

Acquisition of a customer is mostly done

Via an Agent having 71% weightage.

Figure 2(b) Count Plot - Channel

****

**Occupation**

Salaried 0.48

Small Business 0.42

Large Business 0.09

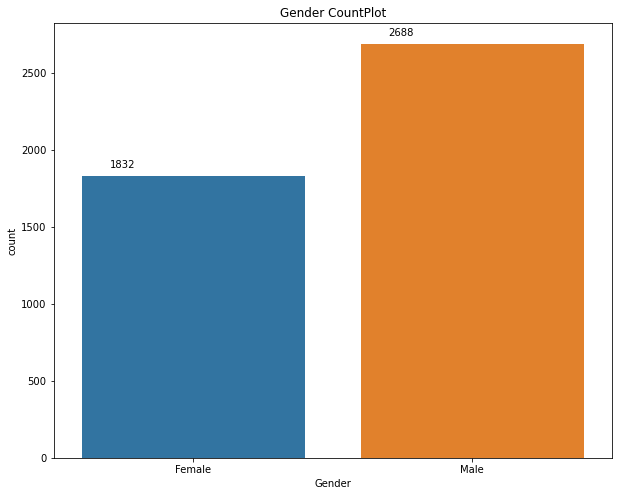
Free Lancer 0.00

Most customers have Salaried Occupations

Around 48%.

Here freelancers have a minute

weightage**.** Figure 2(c) Count Plot - Occupation

****

**Gender**

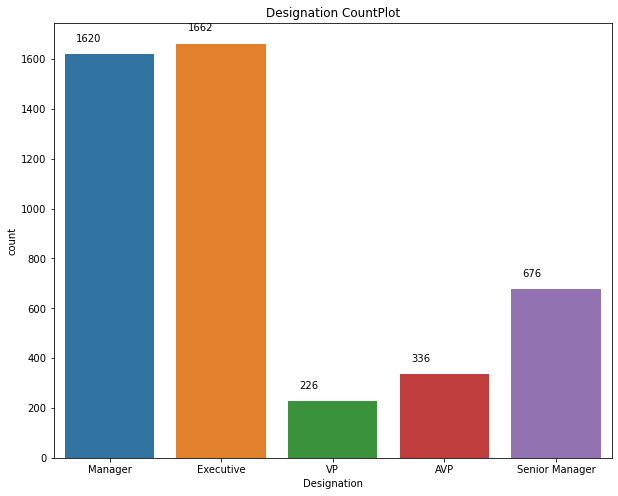
Male 0.59

Female 0.41

Approximately 59% of customers

Are males.

Figure 2(d) Count Plot - Gender

****

**Designation**

Executive 0.37

Manager 0.36

Senior Manager 0.15

AVP 0.07

VP 0.05

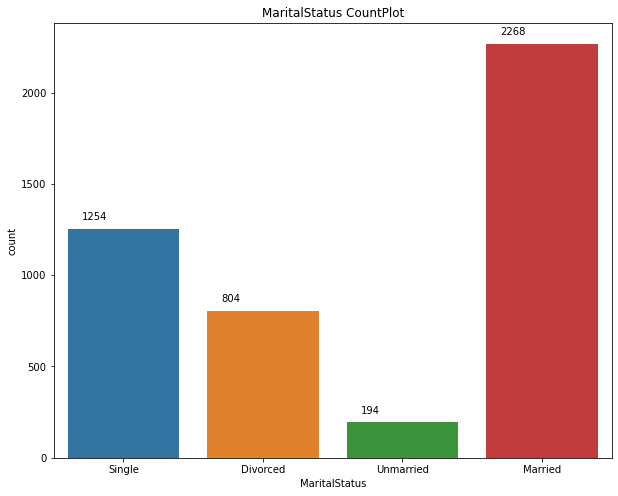
Most customers are either a

Executive or Managers having

Weightage of 37% and 36%

Respectively.

Figure 2(e) Count Plot - Designation

**Marital Status**

Married 0.50

Single 0.28

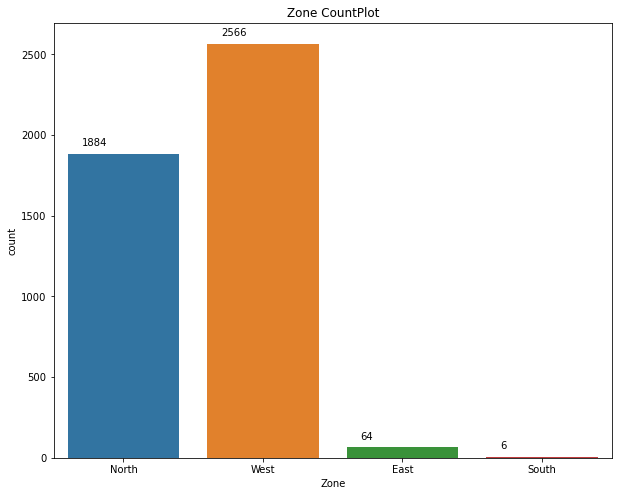
Divorced 0.18

Unmarried 0.04

**Around 50% of the customers**

**Are married.**

Figure 2(f) Count Plot -Marital Status

****

**Zone**

West 0.57

North 0.42

East 0.01

South 0.00

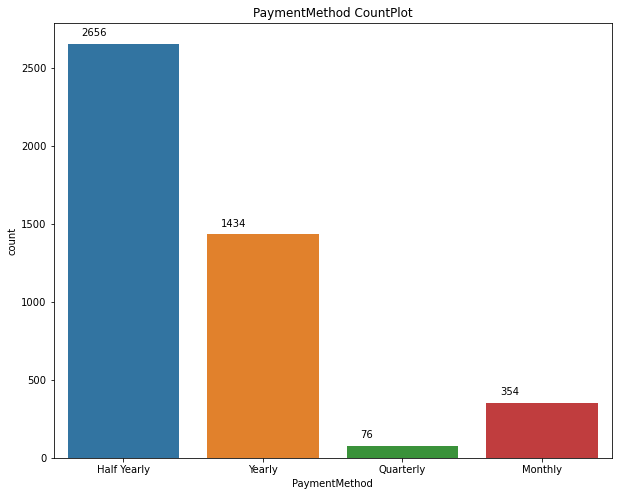
West Zone brings the most

Customers with 57% weightage.

Here freelancers have a minute

weightage.

Figure 2(g) Count Plot - Zone

****

**PaymentMethod**

Half Yearly 0.59

Yearly 0.32

Monthly 0.08

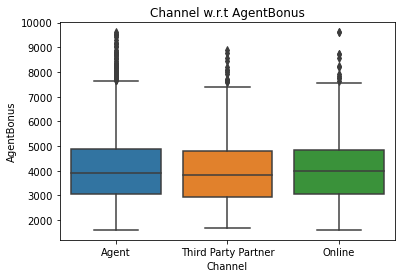
Quarterly 0.02

Around 59% of Customers went

For half-yearly payment plan

Figure 2(h) Count Plot - PaymentMethod

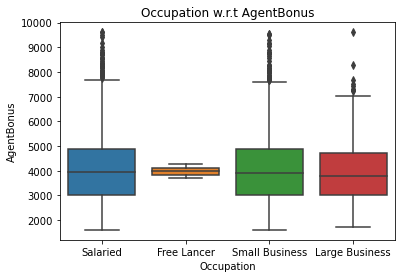
**Categorical Variables Bivariate Analysis w.r.t Agent Bonus**

* Agent Bonus has a lot of outlier values for every

channel with almost similar mean values for all 3

channels.

Figure 3(a) Boxplot – Channel w.r.t AgentBonus

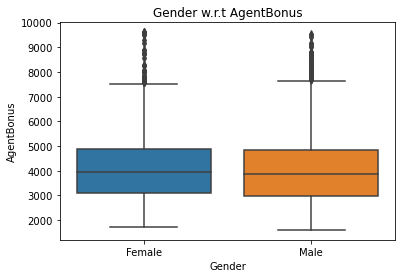
****

Almost similar mean value for all Occupations.

NO outliers present for Free Lancer

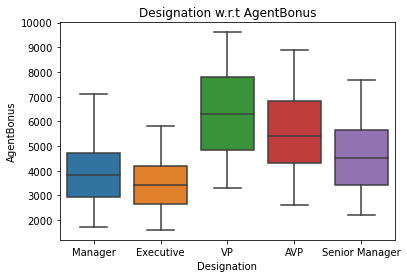
Could be because we have only 2 data points for Free Lancer.

Figure 3(b) Boxplot – Occuapation w.r.t AgentBonus

* ****Agent Bonus has a lot of outlier values

for both Genders with almost similar mean values for both Male and Female.

Figure 3(c) Boxplot – Genderl w.r.t AgentBonus

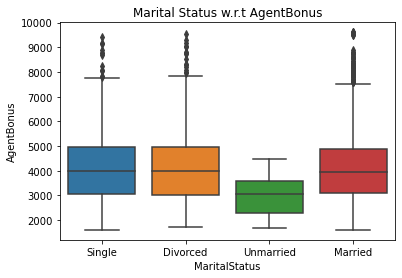
****

No outliers present.

VP Designation has the highest mean

As compared to other Designations.

Figure 3(d) Boxplot – Desgnationl w.r.t AgentBonus

****

* Agent Bonus has a lot of outlier

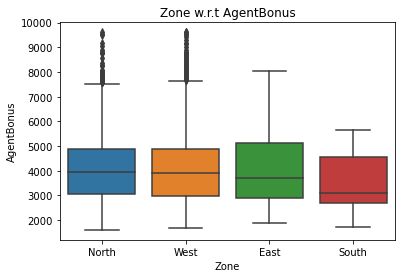
values for all MaritalStatus except

Unmarried customers.

* With almost similar mean values for all 3

customers except unmarried.

Figure 3(e) Boxplot – MaritalStatus w.r.t AgentBonus

****

Outliers present only for North and West Zones.

Both having almost

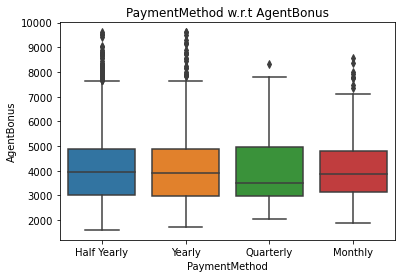
Similar means.

No outliers present in East and

South Zones possibly due to less

Customer traffic from those Zones.

Figure 3(f) Boxplot – Zone w.r.t AgentBonus

****

* Outliers present for all

Payment methods chosen by the customer.

* Quarterly paying customers having the lowest mean.

Figure 3(g) Boxplot – Channel w.r.t AgentBonus

**Pairplot**

*A pair plot plots the relationships between all numeric variables in a dataset. The diagonal below is the histogram for each variable and shows the distribution. From the below plot, we can observe if there are relationships between every two pair of variables.*

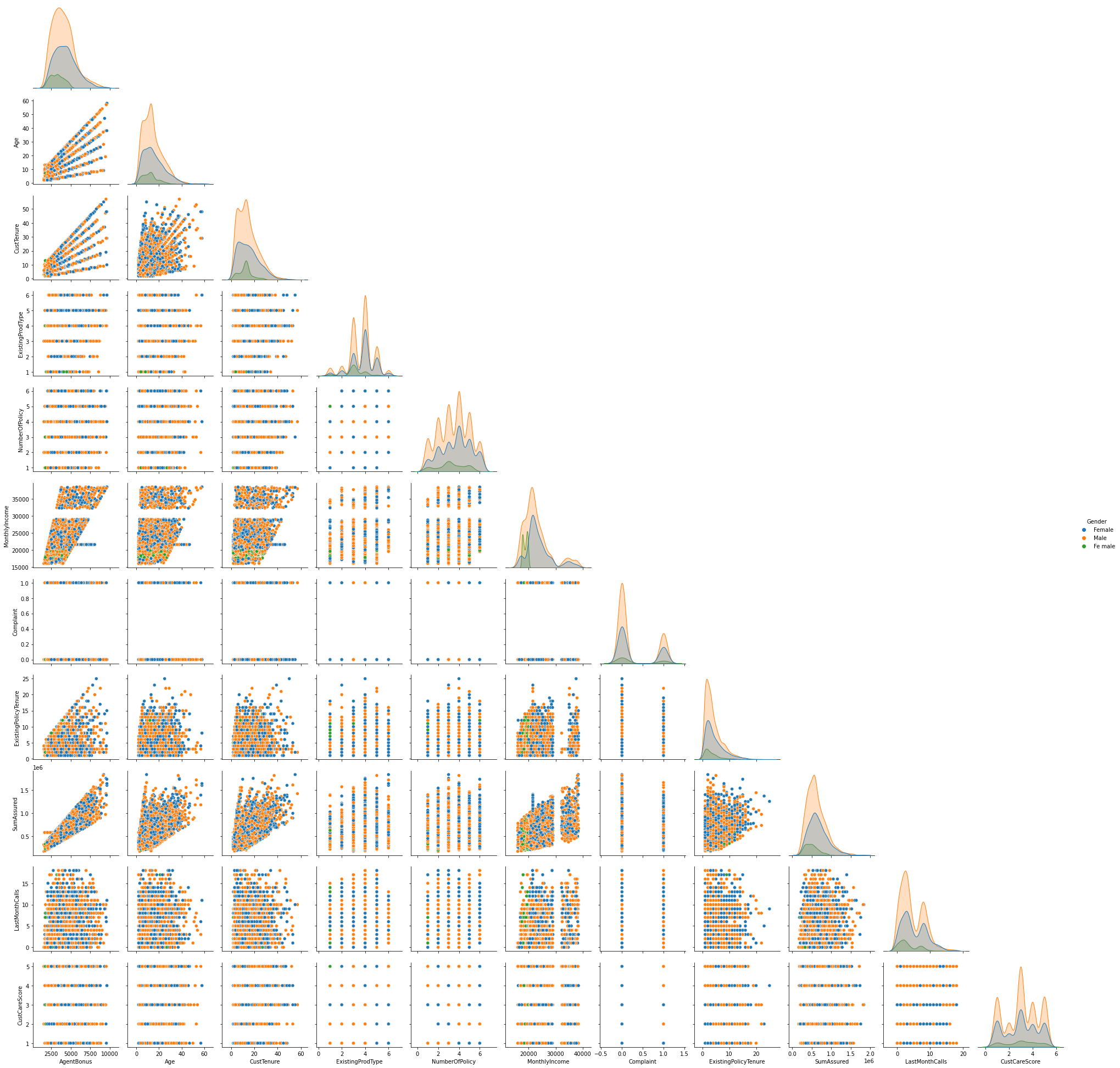
****

Figure 4 – Pairwise Distribution Plot

**Correlation Heatmap.**

1. *The correlation coefficient shown in the table below shows the degree of correlation between the two variables represented in X axis and Y axis. It varies between -1 (maximum negative correlation) to +1 (maximum positive correlation).*

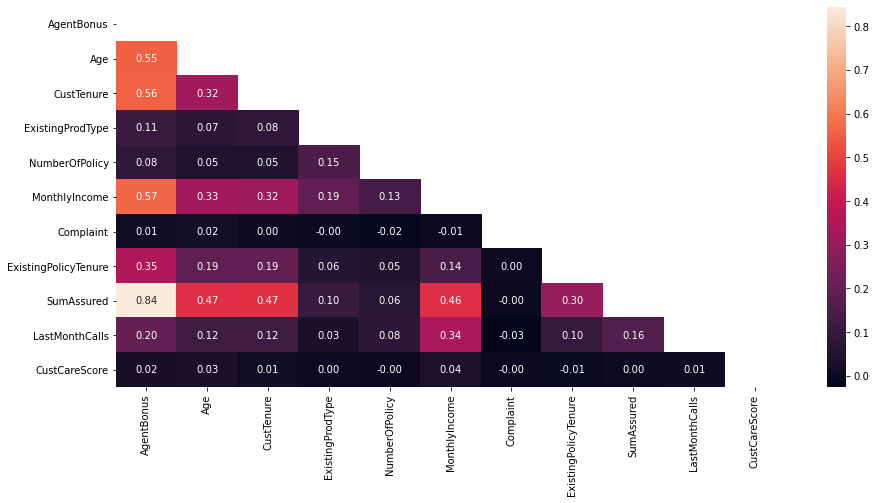
****

Figure 5 - Correlation Heatmap

* Here the lighter colors depict high correlation and darker colors depict low

correlation.

* We can observe that there is almost no multicollinearity in the data.
* Multicollinearity refers to more variables affect our dependant variables, here from the graph above

only SumAssured makes the cut as a variable affecting the AgentBonus

* Complaint and CustCareScore have almost no correlation with any other parameter,

hence dropping these columns will not make a difference as they hold no weightage in predictions for our dependant variable, i.e AgentBonus where these columns ultimately are ignored in the prediction, hence are removed..

* AgentBonus and SumAssured have high correlation with each other of 0.84.

**Business insights from EDA**

1. Outlier Removal is performed but it does not seem as the correct approach as some variables like

SumAssured are allowed to have some outliers however our model will be affected if outliers are not removed as we will use Linear Regression for our optimal model, where outliers will produce a biased result with Linear Regression and to prevent that from happening we’ll go with the outliers removed .

1. We can add new variables like Premium which will become another variable having direct correlation with AgentBonus and will make it easier to observe the high performing and the low performing

agents as the ones who bring in more premium and good for the firm and performing well and those

incurring low premium needs to be focused more on.

1. However, adding new variables are not as simple as it sounds as here we have 4520 rows that needs

to have a value which will add to the predictions and if we are not careful enough, the new variable introduced will add more variance to our predictions and can be biased too, which ultimately can affect the model, hence it is not recommended unless you have extreme and thorough domain knowledge.

1. With this we’ve completed the EDA in the coming exercises we’ll build the model as this is a Classification problem, Regression Techniques for model building will be our go-to approach.
2. The data from the EDA can be said to be highly unbalanced eg: Zone, South has less weightage similar for Occupation- Freelancer, more data is needed or upscale the data, similar can be the case with

EducationField\_MBA where we need to have enough data to not make bias decisions which can be

done by upscaling the data which will add another problem where the data would be repeatable and not accurate enough to give accurate predictions.

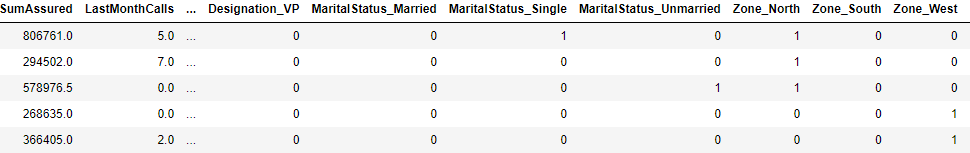
**We might have to convert some categorical variables by encoding them into numeric values for out**

**model Building. Stay Tuned to find more.**

**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 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'],

**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

* **Here the variables with a high value are less significant and do not affect or add to the predictions of dependant variable here AgentBonus.**
* **The variables with low value mean they are highly significant to the predictions hence don’t**

**require a high value to balance the weightage it adds to the dependant variable.**

* **And as the value becomes closer to zero the more significant the variable becomes like here**

**SumAssured which we know for a fact is highly significant and is also proved by our EDA**.

**More information about variable significance will be provided in the end with the final equation.**

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

==============================================================================

Here, R-squared (R2) is a statistical measure that **represents the proportion of the variance for a dependent variable** that's explained by an independent variable or variables in a regression model. Hence a higher R-squared value means the data is capturing maximum variance hence the higher the value, the better the results.

**RMSE – value - 600.5900784990948**

**R squared – value - 0.807**

**Adjusted R squared – value - 0.805**

The variation in R-squared and Adjusted R-squared is not too significant and we have a high value for both, hence a good model.

**Variance Inflation Factor(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.

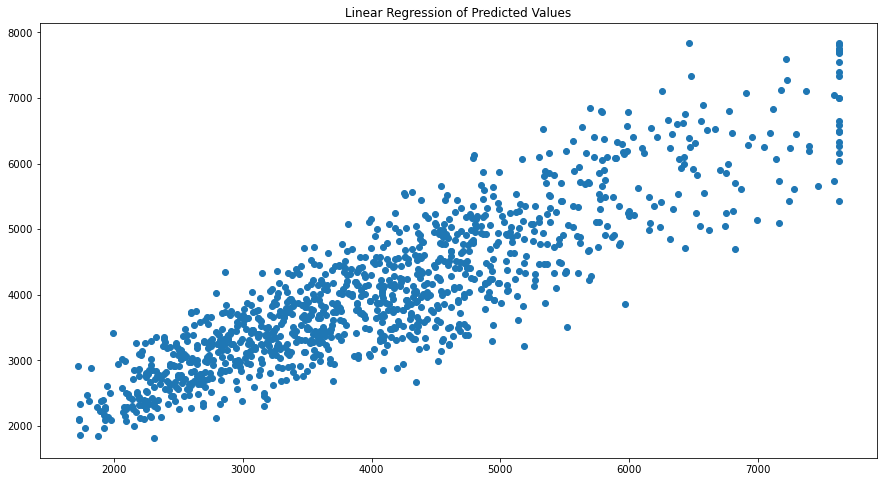
### The overall P value is less than alpha, so rejecting H0 and accepting Ha that atleast 1 regression co-efficient is not 0. Here all regression co-efficients are not 0

**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.**

* **Modelling approach used here is Linear Regression, which is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.**



**Figure 6 - Linear Regression Scatterplot**

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

**Model Outputs (Without Model Tuning):**

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

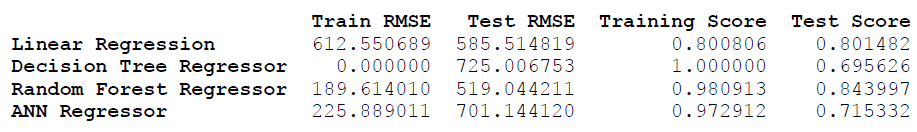
Scaling is done as some variables with greater weight will affect the predictions more, hence scaling is done to bring all variables in a common range e.g., 0 to 1. Due to which the predictions can be unbiased and not biased to one specific variable with higher weights. For e.g., age and sum assured.

## 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.***



**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.

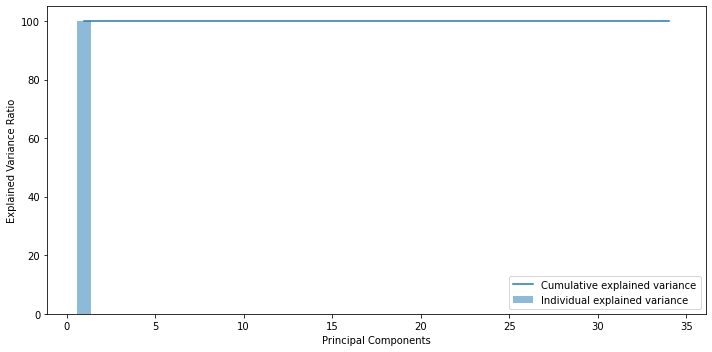
100. 100. 100. 100. 100.

100. 100. 100. 100. 100.

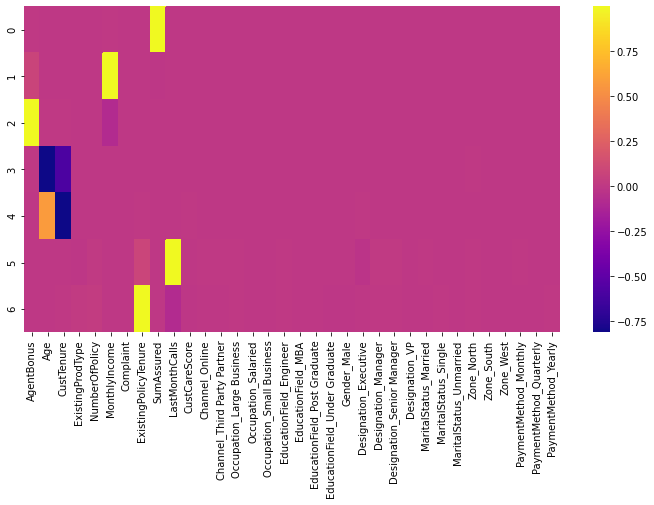
100. 100. 100. 100. 100.

100. 100. 100. 100. ]

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

****

**Figure 7 – Principal Components vs Variance Ratio**

****

**Figure 8 – 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),

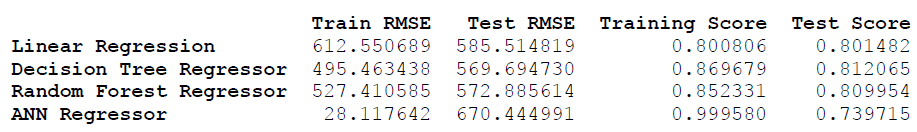
param\_grid={'activation': ['tanh', 'relu'],

'hidden\_layer\_sizes': [500, (100, 100)],

'solver': ['sgd', 'adam']})

Best parameters - {'activation': 'tanh', 'hidden\_layer\_sizes': 500, 'solver': 'adam'}

**Model Outputs (With Model Tuning):**



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.**

**MODEL SELECTION**

* **From the previous results, it is evident that Linear Regression is a better model.**
* **Why Linear Regression?**
  + **Post removal of variables causing multicollinearity, Linear Regression provided a good R-squared value and similarly a high adjusted R squared value. Hence a good percentage of variance can be successfully explained by our model.**
  + **A very important factor being the train and test set accuracy scores are ~80% and consistent.**
  + **Unlike other models where overfitting and inconsistency in the performance metrics can be observed. Linear Regression model does not show these inconsistencies in the observation.**

**(Here by overfitting we mean, the model is performing very good for training set and giving poor results for the testing set)**

* + **The LR model makes it easier to understand the model, multicollinearity in the data. Also, unlike other model its computational time is quick therefore we can run it multiple times whereas ANN and Random Forests needs capable machines as they are very time consuming models. Might have to wait for hours and in our case they still don’t perform better than LR.**

**Note: 100 % accuracy cannot be achieved in real life data as there is always some unexplainable factors and noise that’s always present in our data.**

**MODEL EVALUATION**

**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

* **From the equation the variables with a low or no coefficient value depicts that the variable is very important to the independent variable’s prediction. As the coefficients value increase it shows the variable has become comparatively less significant.**

**The variable significance can be explained using the \* method, where \* depicts highly significant, \*\* less significant, and \*\*\* and \*\*\*\* least significant.**

|  |  |
| --- | --- |
| **Variables** | **Significance** |
| **SumAssured, MonthlyIncome** | **\*** |
| **LastMonthCalls, CustCareScore, Channel\_Third\_Party\_Partner, EducationField\_Under\_Graduate, Designation\_VP,**  **NumberOfPolicy** | **\*\*** |
| **Age, CustTenure, Channel\_Online, EducationField\_Engineer, Gender\_Male, MaritalStatus\_Single, Complaint, ExistingPolicyTenure, MaritalStatus\_Married, Zone\_West, Zone\_North, PaymentMethod\_Yearly, EducationField\_Post\_Graduate** | **\*\*\*** |
| **Occupation\_Large\_Business, Occupation\_Salaried, Occupation\_Small\_Business, EducationField\_MBA, Designation\_Executive, Designation\_Manager, Designation\_Senior\_Manager, MaritalStatus\_Unmarried, Zone\_South, Paymentmethod\_Monthly, PaymentMethod\_Quaterly** | **\*\*\*\*** |

* **R-Squared Obtained from final Linear Regression Model: 0.806**
* **Adjusted R-Squared Obtained from final Linear Regression Model: 0.805**
* **Decision Trees, Random Forest, and ANN (Before Hyperparameter Tuning) :**
  + **It can be observed that all the 3 models have overfitting problems where we have ideal accuracies of ~100% for our training set. However the models are performing poorly on our testing set having accuracies ~70% – 84%. There is a major accuracy difference between the training and testing set which is not acceptable for predictions.**
  + **If the accuracy difference is greater than 6-10% it is advised to not accept the model as the predictions can be unreliable.**
* **Decision Trees, Random Forest, and ANN (After Hyperparameter Tuning) :**
  + **After Hyperparameter Tuning Decision Trees and Random Forest models showed no overfitting errors.**
  + **The training accuracies were ~85% and testing accuracies were ~80%.**
  + **ANN still showed no improvement in results and was still overfitting.**
* **Although the Decision Trees and Random Forest were performing good, I went with Linear Regression as it gave more stable results and Variable importance could be calculated more easily from the Linear Regression Equation and stats-model performed to predict the results.**

**Insights from Analysis.**

* **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 and highly correlated to our target variable.**
* **SumAssured is highly significant as the agent performing good is the one which is getting more profit for the company selling more or high value policies.**
* **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.**
* **From the Linear Regression Equation we can find insights and remove all the least significant variables**

**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.**