greatlearning

“Life Insurance Sales”

Submitted in Partial Fulfillment of requirements for the Award of certificate of Post Graduate Program in Business Analytics and Business Intelligence

Capstone Project Report

Submitted to

GREAT LAKES

INSTITUTE OF MANAGEMENT

Global Mindset - Indian Roots

Submitted by

Nishitha Ramesh

Under the guidance of

Mr. Jay Narayan Das

Batch – (PGPDSBA.B.Oct’19)

Year of Completion (March’ 2022)

CERTIFICATE

This is to certify that the participants Nishitha Ramesh who is a student of Great Lakes Institute of Management, has successfully completed her project on “Life Insurance Sales”

This project is the record of authentic work carried out by them during the academic year 2021-2022.

Mentor’s Name and Sign

Program Director

Name :

Date :

Place: Bangalore

Table of Contents

1. Introduction

a. 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.

b. Need of the study/project

Insurance sector is highly data-driven industry. Every day a new company is formed and thus the competition is increasing exponentially. In order to stay ahead of the curve, around 86% of the companies are investing in insurance data analytics to optimize their mechanisms. It can be observed that the probability of the insurance companies achieving their long-term goals increases significantly by unleashing the power of the data that is collected over the years.

c. Business/social opportunity

Different types of analytics can be done based on the requirement of each company. Following are the business opportunities that can be obtained by the data analysis in the insurance sectors.

i. Improving Employee Performance and Satisfaction: By analyzing the data about the employee performance, they can be rewarded with bonuses which will increase the employee satisfaction. The companies can also provide trainings for performance improvements for employee who have comparatively low performance. By doing this the company can benefit from the high productivity of its employees.

ii. Improving Customer Satisfaction: By analyzing the perspective customer data, the companies can predict the needs of the customers and thus increase the potential to make a sale when compared to a company following the conventional methods of selling. The existing customer data can be used to find the insights and thus improve customer satisfaction.

iii. Lead Generation: By analyzing the data on the internet, the companies can deep dive into the customer behavior and up-sell or cross-sell opportunities in the market.

iv. Risk Analysis and Fraud Detection: By storing the previous fraudulent customers data and doing a predictive analysis on the new claim to calculate the risk percentage, frauds can be prevented. This data can also be used to recognize if any patterns or trends exists when a new insurance claim is made thus avoid risks and loses.

2. Exploratory data analysis

a. Understanding how data was collected

The Sales data is the Life Insurance data of a leading insurance company. This data has information about the customers and the performance of agents.

b. Visual inspection of data

i. Descriptive details of variables:

Sl No.	Variable	Description
1.	CustID	Unique customer ID
2.	AgentBonus	Bonus amount given to each agent in last month
3.	Age	Age of customer
4.	CustTenure	Tenure of customer in organization
5.	Channel	Channel through which acquisition of customer is done
6.	Occupation	Occupation of customer
7.	EducationField	Field of education of customer
8.	Gender	Gender of customer
9.	ExistingProdType	Existing product type of customer
10.	Designation	Designation of customer in their organization
11.	NumberOfPolicy	Total number of existing policies of a customer
12.	MaritalStatus	Marital status of customer
13.	MonthlyIncome	Gross monthly income of customer
14.	Complaint	Indicator of complaint registered in last one month by customer
15.	ExistingPolicyTenure	Max tenure in all existing policies of customer
16.	SumAssured	Max of sum assured in all existing policies of customer
17.	Zone	Customer belongs to which zone in India. Like East, West, North and South
18.	PaymentMethod	Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly
19.	LastMonthCalls	Total calls attempted by company to a customer for cross sell
20.	CustCareScore	Customer satisfaction score given by customer in previous service call

ii. Dimension of the data: The Sales data set contains 20 variables with 4520 record entries.

```
In [31]: #Dimension of data set
         sales.shape

Out[31]: (4520, 19)
```

Figure 1: Dimension of the Sales data set

c. Understanding of attributes

a. Occupation is a categorical variables with values Small business, Large business, Free Lancer and Salaried. There are eight object fields which indicate that there are 8 categorical variables.
b. The variable Customer ID (CustID) is a continuous integer variable. Similar to this there are a total of five integer variables.
c. The variable Monthly Income (MonthlyIncome) indicates the discrete values which has the values of each customer. Similar to this there a total of seven float variables.

```
In [4]: #Information about the variables
        sales.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CustID                4520 non-null   int64
 1   AgentBonus            4520 non-null   int64
 2   Age                   4251 non-null   float64
 3   CustTenure            4294 non-null   float64
 4   Channel               4520 non-null   object
 5   Occupation            4520 non-null   object
 6   EducationField        4520 non-null   object
 7   Gender                4520 non-null   object
 8   ExistingProdType      4520 non-null   int64
 9   Designation           4520 non-null   object
 10  NumberOfPolicy        4475 non-null   float64
 11  MaritalStatus         4520 non-null   object
 12  MonthlyIncome         4284 non-null   float64
 13  Complaint             4520 non-null   int64
 14  ExistingPolicyTenure  4336 non-null   float64
 15  SumAssured            4366 non-null   float64
 16  Zone                  4520 non-null   object
 17  PaymentMethod         4520 non-null   object
 18  LastMonthCalls        4520 non-null   int64
 19  CustCareScore         4468 non-null   float64
dtypes: float64(7), int64(5), object(8)
memory usage: 706.4+ KB
```

Figure 2: Information about variables

d. Univariate analysis

i. Agent Bonus:

a) The value for Agent Bonus (AgentBonus) variable ranges from 1605.00 to 9608.00 with an average value of 4077.84.

```
In [7]: sales.describe()["AgentBonus"]

Out[7]: count    4520.000000
        mean     4077.838274
        std      1403.321711
        min      1605.000000
        25%      3027.750000
        50%      3911.500000
        75%      4867.250000
        max      9608.000000
        Name: AgentBonus, dtype: float64
```

Figure 3: Agent Bonus Description

b) From the below figure we can see that the Agent bonus is almost equally distributed among all the percentiles i.e. there are no sudden jumps in the bonus values.

```
0.5% agents recived bonus lower than    1755.19
  1% agents recived bonus lower than    1876.38
  5% agents recived bonus lower than    2158.0
 10% agents recived bonus lower than    2418.0
 90% agents recived bonus lower than    5917.1
 95% agents recived bonus lower than    6755.500000000002
 99% agents recived bonus lower than    8234.44000000001
99.5% agents recived bonus lower than   8757.215
```

Figure 4: Percentile Distribution of Agent Bonus Range

c) From the below bar plot of the Agent Bonus, we can say that maximum number of agents have received the bonus amount between 3000 – 5000.

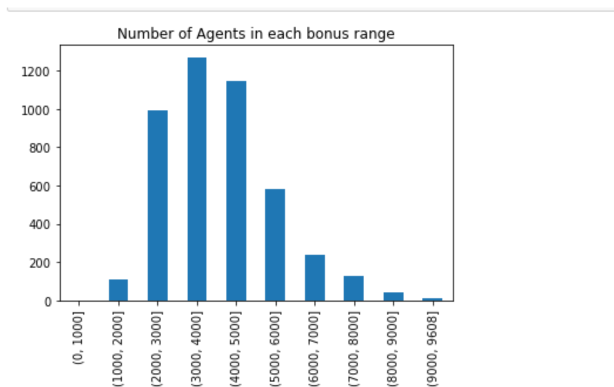

Figure 5: Plot of Number of Agents in the Agent Bonus Range

ii. Age:

a) The value of Age varies from 2 to 58 with an average of 14.49.

```
In [10]: sales.describe()["Age"]

Out[10]: count    4251.000000
         mean       14.494707
         std         9.037629
         min         2.000000
         25%         7.000000
         50%        13.000000
         75%        20.000000
         max        58.000000
         Name: Age, dtype: float64
```

Figure 6: Description of Age variable

b) From the below bar plot of the Age we can say that maximum customers belong to the age group 5 to 15.

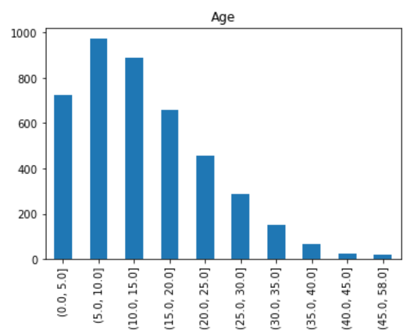

Figure 7: Plot to show the Age distribution

iii. Occupation:

a) After cleaning the Occupation of the customer column has four unique categories namely,
   i. Large Business
   ii. Small Business
   iii. Salaried
   iv. Free Lancer

b) The plot of this variable shows that most of the customers are 'Salaried'.

```
Before ['Salaried' 'Free Lancer' 'Small Business' 'Laarge Business'
 'Large Business']
After ['Salaried' 'Free Lancer' 'Small Business' 'Large Business']
```

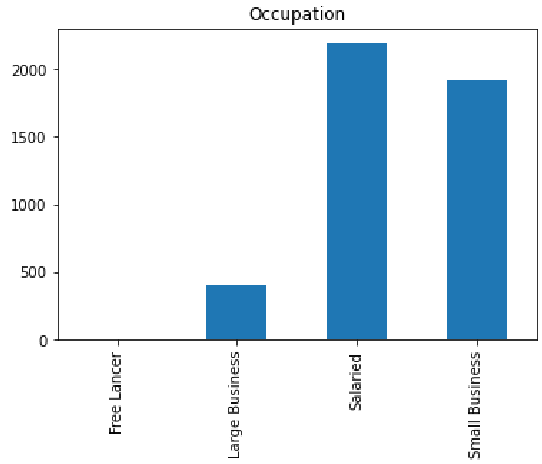

Figure 8: Plot of Occupation

iv. **Education Field:**

a) After cleaning the Education Field of the customer column has four unique categories namely,

   i. Graduate
   ii. Post Graduate
   iii. Under Graduate
   iv. Diploma

b) The plot of this variable shows that most of the customers are 'Graduates'.

```
Before ['Graduate' 'Post Graduate' 'UG' 'Under Graduate' 'Engineer' 'Diploma'
 'MBA']
After ['Graduate' 'Post Graduate' 'Under Graduate' 'Diploma']
```

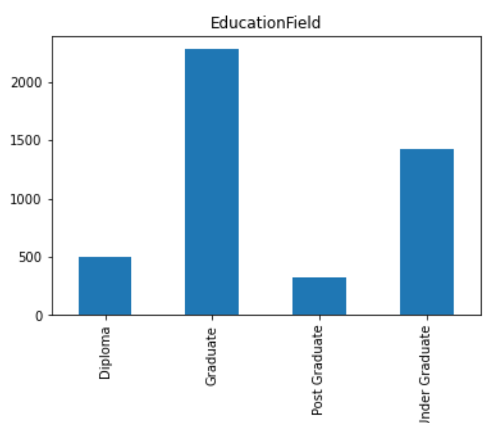

Figure 9: Plot of Education Field

v. Gender:

a) After cleaning the Gender of the customer column has two unique categories namely,

   i. Male
   ii. Female

b) The plot of this variable shows that most of the customers are ‘Male’.

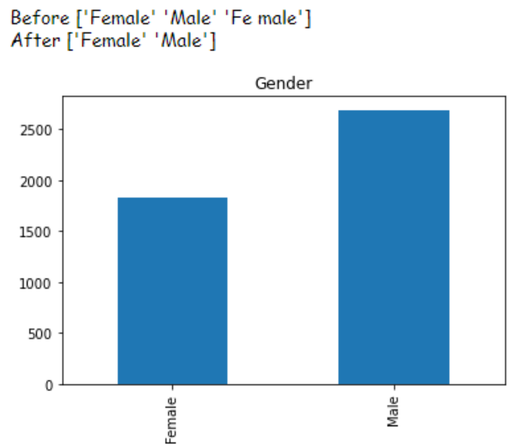

Figure 10: Plot of Gender distribution

vi. Designation:

a) After cleaning the Designation of the customer column has five unique categories namely,

   i. Manager
   ii. Exe
   iii. VP
   iv. AVP
   v. Senior Manager

b) The plot of this variable shows that most of the customers are ‘Executives’ and ‘Managers’.

Before ['Manager' 'Exe' 'Executive' 'VP' 'AVP' 'Senior Manager']
After ['Manager' 'Exe' 'VP' 'AVP' 'Senior Manager']

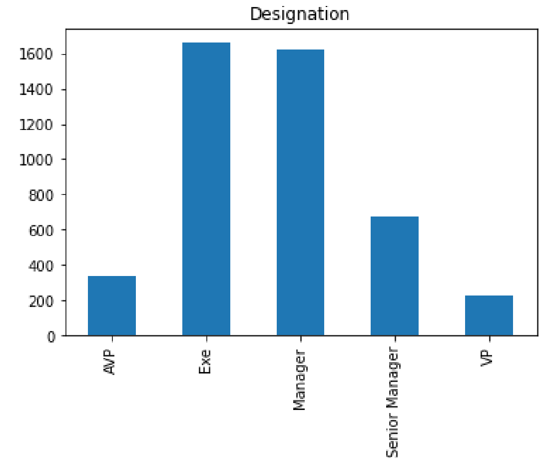

Figure 11: Plot of Designation distribution

vii. **Payment Method**: The plot of this variable shows that most of the customers are following the 'Half Yearly' and 'Yearly' payment methods.

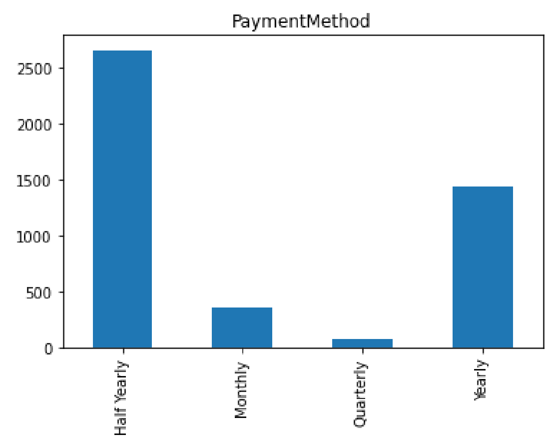

Figure 12: Plot of Payment Method Distribution

viii. **Zone**: The plot of this variable shows that most of the customers are from 'West' and 'North' zones.

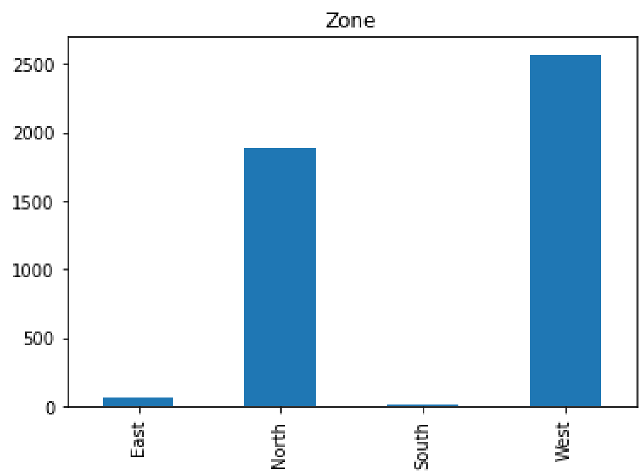

Figure 13: Plot of Zone distribution

ix. Marital Status: The plot of this variable shows that most of the customers are of ‘Married’ Martial status.

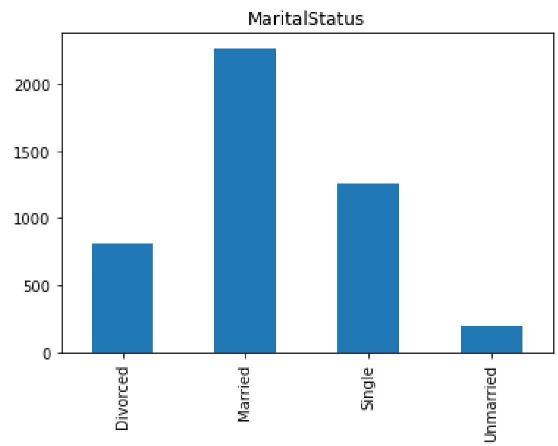

Figure 14: Plot of Martial Status distribution

e. Bivariate analysis

i. Agent Bonus Vs Age:
From the below graph we can see that the Agent Bonus and Age are directly proportional i.e. as the Age increases the Agent Bonus value increases.

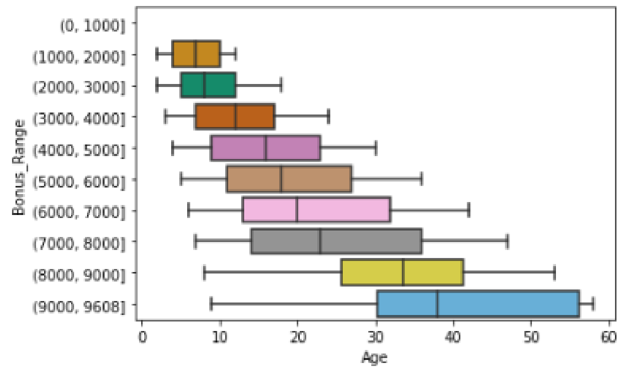

Figure 15: Plot of Agent Bonus Vs Age

ii. Agent Bonus Vs Customer Tenure:

From the below graph we can see that the Customer Tenure and Agent Bonus are directly proportional i.e. as the Agent Bonus increases the Customer Tenure value increases.

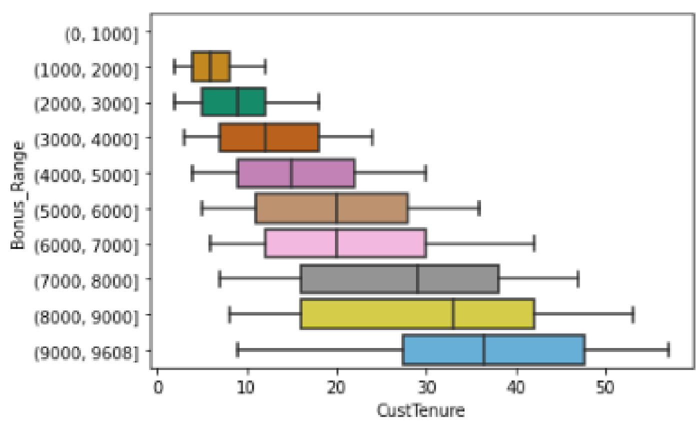

Figure 16: Plot of Agent Bonus Vs Customer Tenure

iii. Agent Bonus Vs Existing Product Type:

From the below graph we can see that the Existing Product Type is 3 or 4 for most of the Agent Bonus.

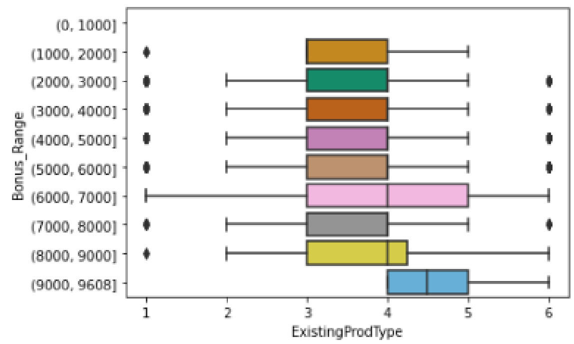

Figure 17: Plot of Agent Bonus Vs Existing Product Type

iv. Agent Bonus Vs Number of Policy:

From the below graph we can see that the Number of Policies is between 2 and 5 for most of the Agent Bonus.

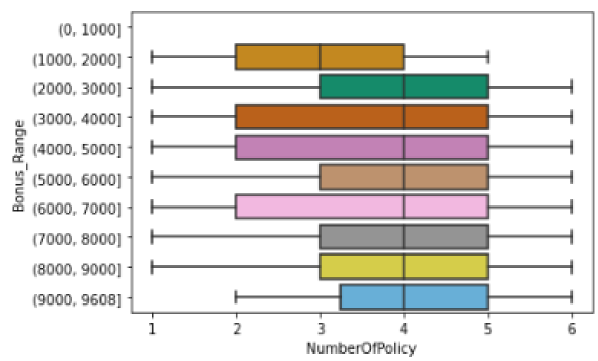

Figure 18: Plot of Agent Bonus Vs Number of Policy

v. Agent Bonus Vs Monthly Income:

For the Agent Bonus from 3000 – 6000 the Monthly income is not proportionality distributed. This is indicated by the high number of outliers.

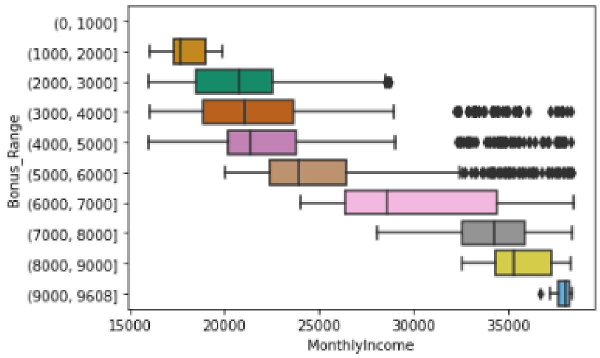

Figure 19: Plot of Agent Bonus Vs Monthly Income

vi. Agent Bonus Vs Existing Policy Tenure:

From the below graph we can see that the value of the Existing Policy Tenure lies between the range of around 3 to 13.

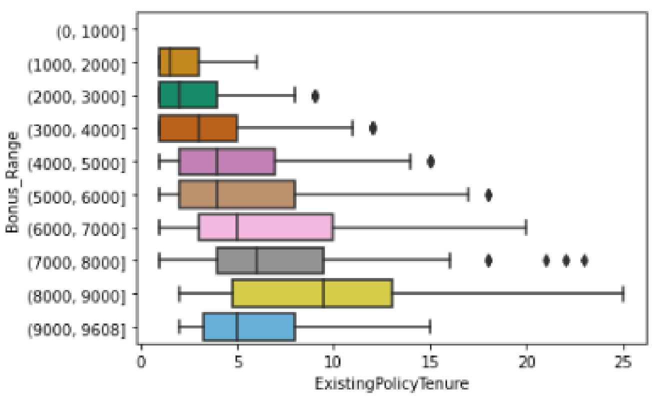

Figure 20: Plot of Agent Bonus Vs Existing Policy Tenure

vii. Agent Bonus Vs Sum Assured:

From the below graph we can see that the Sum Assured is directly proportional to the Agent Bonus i.e. the Agent Bonus increases as the Sum Assured to the customer increases.

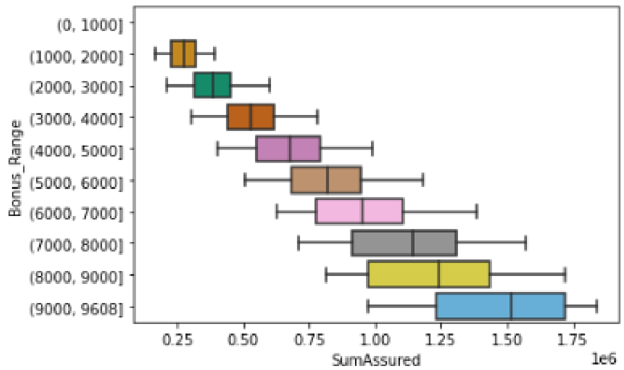

Figure 21: Plot of Agent Bonus Vs Sum Assured

viii. Agent Bonus Vs Last Month Calls:

From the below graph we can say that number of calls made lies between 0 – 10 for the Agent Bonus Ranges.

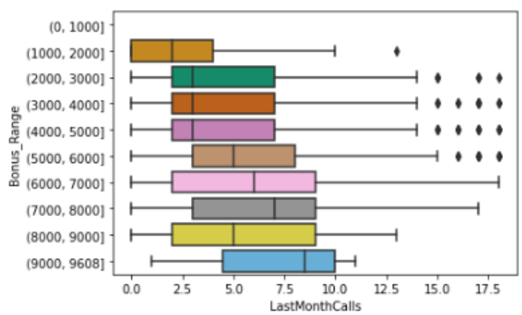

Figure 22: Plot of Agent Bonus Vs Last Month Calls

ix. Correlation Values between Variables.

From the below Correlation values we can see that only few variables are highly correlated with each other.

Following is the set of most correlated variables.

a) Agent Bonus and Sum Assured – 0.85
b) Agent Bonus and Monthly Income – 0.61
c) Agent Bonus and Age – 0.56
d) Agent Bonus and Customer Tenure – 0.56
e) Age and Sum Assured – 0.47
f) Customer Tenure and Sum assured – 0.47
g) Monthly income and Sum assured – 0.51

	AgentBonus	Age	CustTenure	ExistingProdType
AgentBonus	1.000000	0.559481	0.561344	0.113023
Age	0.559481	1.000000	0.328627	0.070555
CustTenure	0.561344	0.328627	1.000000	0.079891
ExistingProdType	0.113023	0.070555	0.079891	1.000000
NumberOfPolicy	0.076448	0.042143	0.045021	0.150923
MonthlyIncome	0.612196	0.354162	0.344911	0.198468
Complaint	0.014281	0.021888	0.003807	-0.003486
ExistingPolicyTenure	0.392415	0.216259	0.214984	0.057066
SumAssured	0.854257	0.474434	0.474610	0.102597
LastMonthCalls	0.199708	0.114670	0.115993	0.033191
CustCareScore	0.022860	0.035694	0.011145	0.003813

	NumberOfPolicy	MonthlyIncome	Complaint
AgentBonus	0.076448	0.612196	0.014281
Age	0.042143	0.354162	0.021888
CustTenure	0.045021	0.344911	0.003807
ExistingProdType	0.150923	0.198468	-0.003486
NumberOfPolicy	1.000000	0.136518	-0.016416
MonthlyIncome	0.136518	1.000000	-0.004510
Complaint	-0.016416	-0.004510	1.000000
ExistingPolicyTenure	0.049673	0.215130	0.002061
SumAssured	0.060359	0.506208	-0.000256
LastMonthCalls	0.074069	0.359623	-0.026320
CustCareScore	-0.002265	0.036553	-0.003835

	ExistingPolicyTenure	SumAssured	LastMonthCalls
AgentBonus	0.392415	0.854257	0.199708
Age	0.216259	0.474434	0.114670
CustTenure	0.214984	0.474610	0.115993
ExistingProdType	0.057066	0.102597	0.033191
NumberOfPolicy	0.049673	0.060359	0.074069
MonthlyIncome	0.215130	0.506208	0.359623
Complaint	0.002061	-0.000256	-0.026320
ExistingPolicyTenure	1.000000	0.339366	0.107888
SumAssured	0.339366	1.000000	0.156674
LastMonthCalls	0.107888	0.156674	1.000000
CustCareScore	-0.005679	0.002911	0.005934

	CustCareScore
AgentBonus	0.022860
Age	0.035694
CustTenure	0.011145
ExistingProdType	0.003813
NumberOfPolicy	-0.002265
MonthlyIncome	0.036553
Complaint	-0.003835
ExistingPolicyTenure	-0.005679
SumAssured	0.002911
LastMonthCalls	0.005934
CustCareScore	1.000000

Figure 23: Correlation Matrix Value

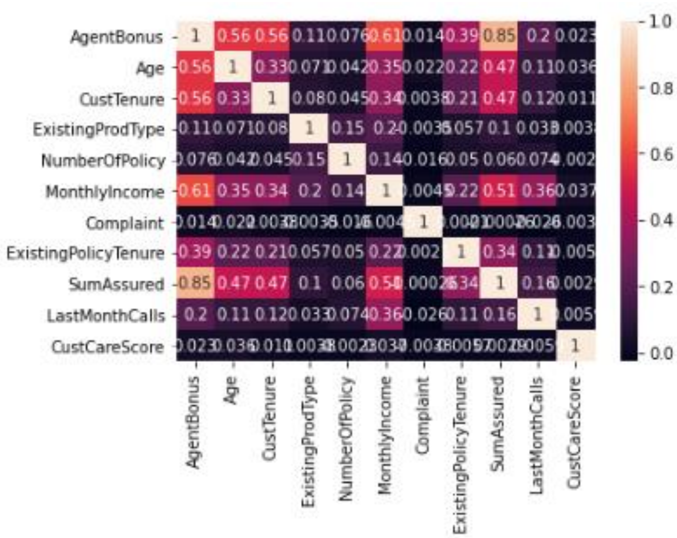

Figure 24: Heat Map of Correlation Matrix Values.

3. Data Cleaning and Pre-processing

a. Missing Value: Not all variables have values for all the records. Some of the values are missing for some variables. The list of these variables and the missing value count is as shown below. In order to remove replace these values the KNN Imputation method is used. Using the common values from the nearest neighbors, the value of the missing datapoint is added.

```
In [5]: sales_na = sales.isna().sum()
        sales_na[sales_na.values > 0].sort_values(ascending = False)

Out[5]: Age                   269
        MonthlyIncome         236
        CustTenure            226
        ExistingPolicyTenure  184
        SumAssured            154
        CustCareScore          52
        NumberOfPolicy         45
        dtype: int64
```

Figure 25: Missing Value Count

b. Removal of unwanted variables

i. The variable Customer ID (CustID) can be dropped as per the initial Exploratory Data Analysis (EDA) as this variable is a continuous variable which doesn't provide much information towards our data modeling.

ii. The variable Age can be dropped as the variable values are not matching the level of Education the customer has.

4. Model building and interpretation

A. Various models

The data set is divided into Train set and Test set in the ratio 67:33.

```
In [40]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=4)

In [41]: print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)

         (3028, 17) (1492, 17) (3028,) (1492,)
```

Figure 26: Train_Test Split

I. Multiple Linear Regression:

a. Model 1: Full Model where all the Independent variables are used for building the model. The predicted value for Agent Bonus will be a combination of all the following independent Variables.

- CustTenure
- Channel
- Occupation
- EducationField
- Gender
- ExistingProdType
- Designation
- NumberOfPolicy
- MaritalStatus
- MonthlyIncome
- Complaint
- ExistingPolicyTenure
- SumAssured
- Zone
- PaymentMethod
- LastMonthCalls
- CustCareScore

```
                    OLS Regression Results
==============================================================================
Dep. Variable:          AgentBonus   R-squared:                       0.801
Model:                         OLS   Adj. R-squared:                  0.800
Method:              Least Squares   F-statistic:                     713.2
Date:             Sun, 13 Feb 2022   Prob (F-statistic):               0.00
Time:                     17:58:57   Log-Likelihood:                -23782.
No. Observations:             3028   AIC:                         4.760e+04
Df Residuals:                 3010   BIC:                         4.771e+04
Df Model:                       17
Covariance Type:         nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept              114.8151     88.911      1.291      0.197     -59.517     289.148
CustTenure              27.5504      1.500     18.371      0.000      24.610      30.491
Channel                  8.4652     14.388      0.588      0.556     -19.746      36.676
Occupation             -17.6190     20.261     -0.870      0.385     -57.347      22.109
EducationField          -2.3760     12.230     -0.194      0.846     -26.355      21.603
Gender                  -9.5290     23.276     -0.409      0.682     -55.167      36.109
ExistingProdType       -33.6060     13.455     -2.498      0.013     -59.988      -7.224
Designation            -47.0478     13.238     -3.554      0.000     -73.003     -21.092
NumberOfPolicy          -3.1377      8.041     -0.390      0.696     -18.904      12.629
MaritalStatus            0.5097     14.906      0.034      0.973     -28.717      29.736
MonthlyIncome            0.0548      0.003     17.824      0.000       0.049       0.061
Complaint               78.8959     25.146      3.138      0.002      29.591     128.201
ExistingPolicyTenure    38.6118      3.737     10.331      0.000      31.284      45.940
SumAssured               0.0038   5.96e-05     63.552      0.000       0.004       0.004
Zone                    12.2704     11.313      1.085      0.278      -9.912      34.453
PaymentMethod           14.7780      9.637      1.533      0.125      -4.118      33.674
LastMonthCalls           1.8616      3.355      0.555      0.579      -4.717       8.440
CustCareScore            8.4188      8.349      1.008      0.313      -7.952      24.790
==============================================================================
Omnibus:                      179.232   Durbin-Watson:                   2.017
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              211.664
Skew:                           0.618   Prob(JB):                     1.09e-46
Kurtosis:                       3.386   Cond. No.                     5.22e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.22e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 27: Multiple Linear Regression Model 1

b. Model 2: Based on the value of P (P>0.56), the following variables are selected as they are significant.
   - CustTenure
   - Channel
   - Occupation

- ExistingProdType
- Designation
- MonthlyIncome
- Complaint
- ExistingPolicyTenure
- SumAssured
- Zone
- PaymentMethod
- LastMonthCalls
- CustCareScore

```
                    OLS Regression Results
==============================================================================
Dep. Variable:          AgentBonus   R-squared:                       0.801
Model:                         OLS   Adj. R-squared:                  0.800
Method:              Least Squares   F-statistic:                     933.8
Date:             Sun, 13 Feb 2022   Prob (F-statistic):               0.00
Time:                     17:58:58   Log-Likelihood:                -23782.
No. Observations:             3028   AIC:                         4.759e+04
Df Residuals:                 3014   BIC:                         4.768e+04
Df Model:                       13
Covariance Type:         nonrobust
========================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept            102.5648     82.519      1.243      0.214     -59.234     264.364
CustTenure            27.5579      1.498     18.392      0.000      24.620      30.496
Channel                8.6225     14.373      0.600      0.549     -19.560      36.805
Occupation           -19.4089     18.015     -1.077      0.281     -54.732      15.914
ExistingProdType     -34.1885     13.260     -2.578      0.010     -60.189      -8.188
Designation          -46.7476     13.190     -3.544      0.000     -72.610     -20.885
MonthlyIncome          0.0547      0.003     17.871      0.000       0.049       0.061
Complaint             79.4859     25.099      3.167      0.002      30.273     128.699
ExistingPolicyTenure  38.6737      3.725     10.383      0.000      31.370      45.977
SumAssured             0.0038   5.96e-05     63.620      0.000       0.004       0.004
Zone                  12.2184     11.275      1.084      0.279      -9.889      34.326
PaymentMethod         14.9948      9.605      1.561      0.119      -3.838      33.827
LastMonthCalls         1.8605      3.345      0.556      0.578      -4.699       8.420
CustCareScore          8.3875      8.338      1.006      0.315      -7.962      24.737
==============================================================================
Omnibus:                      178.903   Durbin-Watson:                   2.018
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              211.172
Skew:                           0.618   Prob(JB):                     1.39e-46
Kurtosis:                       3.383   Cond. No.                     4.84e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.84e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 28: Multiple Linear Regression Model 2

c. Model 3: Further to improve the model performance, the following variables are selected based on P value(P>0.4).
    - CustTenure
    - Occupation
    - ExistingProdType
    - Designation
    - MonthlyIncome
    - Complaint
    - ExistingPolicyTenure
    - SumAssured
    - Zone
    - PaymentMethod
    - LastMonthCalls

- CustCareScore

```
                         OLS Regression Results
==============================================================================
Dep. Variable:           AgentBonus   R-squared:                       0.801
Model:                          OLS   Adj. R-squared:                  0.800
Method:               Least Squares   F-statistic:                      1104.
Date:              Sun, 13 Feb 2022   Prob (F-statistic):               0.00
Time:                      17:58:58   Log-Likelihood:                -23783.
No. Observations:              3028   AIC:                         4.759e+04
Df Residuals:                  3016   BIC:                         4.766e+04
Df Model:                        11
Covariance Type:          nonrobust
======================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept           106.8576     82.145      1.301      0.193     -54.209     267.924
CustTenure           27.5391      1.498     18.387      0.000      24.602      30.476
Occupation          -19.5966     18.008     -1.088      0.277     -54.907      15.713
ExistingProdType    -34.0218     13.252     -2.567      0.010     -60.005      -8.039
Designation         -45.6730     13.083     -3.491      0.000     -71.325     -20.021
MonthlyIncome         0.0549      0.003     18.336      0.000       0.049       0.061
Complaint            78.9995     25.076      3.150      0.002      29.832     128.167
ExistingPolicyTenure 38.7252      3.723     10.403      0.000      31.426      46.024
SumAssured            0.0038   5.95e-05     63.656      0.000       0.004       0.004
Zone                 12.3500     11.271      1.096      0.273      -9.750      34.450
PaymentMethod        14.7512      9.598      1.537      0.124      -4.068      33.571
CustCareScore         8.5382      8.328      1.025      0.305      -7.790      24.867
==============================================================================
Omnibus:                      179.146   Durbin-Watson:                   2.017
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              211.496
Skew:                           0.618   Prob(JB):                     1.19e-46
Kurtosis:                       3.383   Cond. No.                     4.82e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.82e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 29:Multiple Linear Regression Model 3

II. k Nearest Neighbor:

a. Model 1: The first model is with 10 neighbors. The value of agent bonus is selected based on the values of 10 nearest neighbors.

```
In [45]: knn_model = KNeighborsRegressor(n_neighbors=10)
         knn_model.fit(X_train, y_train)
         y_pred_KNN1 = knn_model.predict(X_test)
```

Figure 30: KNN with k = 10

b. Model 2: The first model is with 5 neighbors. The value of agent bonus is selected based on the values of 5 nearest neighbors.

```
In [46]: knn_model = KNeighborsRegressor(n_neighbors=5)
         knn_model.fit(X_train, y_train)
         y_pred_KNN2 = knn_model.predict(X_test)
```

Figure 31:KNN with k = 5

c. Model 3: The first model is with 3 neighbors. The value of agent bonus is selected based on the values of 3 nearest neighbors.

```
In [47]: knn_model = KNeighborsRegressor(n_neighbors=3)
         knn_model.fit(X_train, y_train)
         y_pred_KNN3 = knn_model.predict(X_test)
```

Figure 32:KNN with k = 3

B. Model Tuning

1) Grid Search:

Based on the grid search for KNN model, the best value for k is predicted as 3.

```
In [48]: parameters = {"n_neighbors": range(1, 50)}
         gridsearch = GridSearchCV(KNeighborsRegressor(), parameters)
         gridsearch.fit(X_train, y_train)

Out[48]: GridSearchCV(estimator=KNeighborsRegressor(),
                      param_grid={'n_neighbors': range(1, 50)})

In [49]: gridsearch.best_params_

Out[49]: {'n_neighbors': 3}

In [50]: test_preds_grid = gridsearch.predict(X_test)
         test_mse = mean_squared_error(y_test, test_preds_grid)
         test_rmse = sqrt(test_mse)
         test_rmse

Out[50]: 406.14398107406316
```

Figure 33: Grid Search on kNN

5. Testing predictive model against the test set using various appropriate performance metrics

I. R^2 Value

For all the three MLR models, the R^2 value didn't change significantly. The value remained at 0.80.

II. Mean Squared Error (MSE)

The following are the values for the MSE for all the models.

- There is no significant improvement in the MSE for MLR
- The value of MSE for KNN changes significantly.

```
--------------------------------------------------------------------------------
Mean Squared error
--------------------------------------------------------------------------------
Multiple Linear Regression 1 :  401396.40331891744
Multiple Linear Regression 2 :  401236.35000007466
Multiple Linear Regression 3 :  401154.3793931682
--------------------------------------------------------------------------------
k Nearest Neighbours (KNN with k = 10):  315280.0762021136
k Nearest Neighbours (KNN with k = 5):  228250.66163804493
k Nearest Neighbours (KNN with k = 3):  164952.933362689
--------------------------------------------------------------------------------
```

Figure 34: Mean Square Error (MSE)

III. Root Mean Square Error (RMSE)

The following are the values for the MSE for all the models.

- There is no significant improvement in the RMSE for MLR
- The value of RMSE for KNN changes significantly.

```
--------------------------------------------------------------------------------
Root Mean Squared error
--------------------------------------------------------------------------------
Multiple Linear Regression 1 :  633.55852398884
Multiple Linear Regression 2 :  633.4321984238524
Multiple Linear Regression 3 :  633.3674915822315
--------------------------------------------------------------------------------
k Nearest Neighbours (KNN with k = 10):  561.4980642906203
k Nearest Neighbours (KNN with k = 5):  477.7558598678251
k Nearest Neighbours (KNN with k = 3):  406.14398107406316
--------------------------------------------------------------------------------
```

Figure 35: Root Mean Square Error (RMSE)

6. Final interpretation / recommendation

The following are the customer profile recommendations for agents to increase their bonus,

1) Customers with customer tenure more than 15 years.
2) Customers with policies of product type 3,4 or 5 (Mainly 4).
3) Customers with 4 or 5 policies.
4) Customers with monthly income between 23,000 – 32,000
5) Customers with existing policies with policy tenure of 4 – 10 months.
6) Sum assured to the customer should be between 75,000 – 1,50,000
7) Agent should make at least 6 - 7 calls per month.
8) Customers with the following characteristics are more in number. Hence, the probability of selling the policy is higher to a customer with a similar characteristic.
    a) Monthly income type – Salaried
    b) Education level – Graduates
    c) Gender – Male
    d) Designation – Executive / Manager
    e) Mode of payment – Half yearly / Yearly
    f) Zone – North / West
    g) Marital Status – Married.