February 2022

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

Life Insurance Sales - Capstone

Final Business Report

Thakur Arun Singh

Contents

[Predicting Life Insurance Company Bonus for Its Agents 1](#_Toc95080450)

[Problem Statement 2](#_Toc95080451)

[Problem 1 4](#_Toc95080452)

[Problem 2 5](#_Toc95080453)

[Problem 3 8](#_Toc95080454)

[Problem 4 11](#_Toc95080455)

[Problem 5 15](#_Toc95080456)

[Problem 6 25](#_Toc95080457)

[Appendix – Graphical Representations 27](#_Toc95080458)

# Predicting Life Insurance Company Bonus for Its Agents

Capstone Project Report Submitted by

Thakur Arun Singh

## Problem Statement

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 up skill programs for low performing agents

**List of Tables and Figures**

The dataset contains the following information about 4520 policy holders.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.No** | **Columns** | **Description** | **Data Type** |
| 1 | CustID | Unique customer ID | int64 |
| 2 | AgentBonus | Bonus amount given to each agents in last month | int64 |
| 3 | Age | Age of customer | float64 |
| 4 | CustTenure | Tenure of customer in organization | float64 |
| 5 | Channel | Channel through which acquisition of customer is done | object |
| 6 | Occupation | Occupation of customer | object |
| 7 | EducationField | Field of education of customer | object |
| 8 | Gender | Gender of customer | object |
| 9 | ExistingProdType | Existing product type of customer | int64 |
| 10 | Designation | Designation of customer in their organization | object |
| 11 | NumberOfPolicy | Total number of existing policy of a customer | float64 |
| 12 | MaritalStatus | Marital status of customer | object |
| 13 | MonthlyIncome | Gross monthly income of customer | float64 |
| 14 | Complaint | Indicator of complaint registered in last one month by customer | int64 |
| 15 | ExistingPolicy Tenure | Max tenure in all existing policies of customer | float64 |
| 16 | SumAssured | Max of sum assured in all existing policies of customer | float64 |
| 17 | Zone | Customer belongs to which zone in India. Like East, West, North and South | object |
| 18 | PaymentMethod | Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly | object |
| 19 | LastMonthCalls | Total calls attempted by company to a customer for cross sell | int64 |
| 20 | CustCareScore | Customer satisfaction score given by customer in previous service call | float64 |

**Summary Statistics**

At a glance we have 7 floating point features, 5 integer variables, and 8 categorical text features. For the numeric variables, we have the following statistic summary:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **unique** | **top** | **freq** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **AgentBonus** | 4520 | NaN | NaN | NaN | 4077.84 | 1403.32 | 1605 | 3027.75 | 3911.5 | 4867.25 | 9608 |
| **Age** | 4251 | NaN | NaN | NaN | 14.4947 | 9.03763 | 2 | 7 | 13 | 20 | 58 |
| **CustTenure** | 4294 | NaN | NaN | NaN | 14.469 | 8.96367 | 2 | 7 | 13 | 20 | 57 |
| **Channel** | 4520 | 3 | Agent | 3194 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Occupation** | 4520 | 5 | Salaried | 2192 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **EducationField** | 4520 | 7 | Graduate | 1870 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Gender** | 4520 | 3 | Male | 2688 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **ExistingProdType** | 4520 | NaN | NaN | NaN | 3.68894 | 1.01577 | 1 | 3 | 4 | 4 | 6 |
| **Designation** | 4520 | 6 | Manager | 1620 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **NumberOfPolicy** | 4475 | NaN | NaN | NaN | 3.56536 | 1.45593 | 1 | 2 | 4 | 5 | 6 |
| **MaritalStatus** | 4520 | 4 | Married | 2268 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **MonthlyIncome** | 4284 | NaN | NaN | NaN | 22890.3 | 4885.6 | 16009 | 19683.5 | 21606 | 24725 | 38456 |
| **Complaint** | 4520 | NaN | NaN | NaN | 0.287168 | 0.452491 | 0 | 0 | 0 | 1 | 1 |
| **ExistingPolicyTenure** | 4336 | NaN | NaN | NaN | 4.13007 | 3.34639 | 1 | 2 | 3 | 6 | 25 |
| **SumAssured** | 4366 | NaN | NaN | NaN | 620000 | 246235 | 168536 | 439443 | 578976 | 758236 | 1.84E+06 |
| **Zone** | 4520 | 4 | West | 2566 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **PaymentMethod** | 4520 | 4 | Half Yearly | 2656 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **LastMonthCalls** | 4520 | NaN | NaN | NaN | 4.62699 | 3.62013 | 0 | 2 | 3 | 8 | 18 |
| **CustCareScore** | 4468 | NaN | NaN | NaN | 3.06759 | 1.38297 | 1 | 2 | 3 | 4 | 5 |

**Abbreviations**

No abbreviations were used in the entire report.

**Executive Summary**

Academics and practitioners have studied over the years models for predicting firm’s various aspects, using statistical and machine-learning approaches. We are going to discuss one of the various aspects. An earlier sign that company employees are dissatisfied is the firm policies, work load, pay scale and bonuses. We are going to dive deeper and predict the bonus for its employees, so that we may designed an appropriate engagement activity for their high performing employees and up skill programs for low performing employees.

### Problem 1

Introduction - What did you wish to achieve while doing the project?

**Resolution**

**Project Introduction**

Insurance is an instrument available to individuals and organizations to reduce the exposure of financial risk. It is a contractual obligation between two parties, wherein one party (the insurer) agrees to pay another party (the insured) for the agreed financial amount subject to happening of an agreed event. For this, the insured pays amount, known as premium to the insurer in exchange of the protection to the financial amount as agreed upon. The contract of insurance is based on 7 key principles –

* Utmost Good Faith
* Insurable Interest
* Proximate Cause
* Indemnity
* Subrogation
* Contribution
* Loss Minimization

Premium paid by customer is a major source of revenue for the insurance companies. Default in premium payments results in significant revenue losses and hence insurance companies put their efforts to minimize the leakage in revenue. Life Insurance company spend heavy amounts in establishing marketing set ups and pay hefty first year commissions. That increases the cost of acquisition of a new customer. Studies have shown that $1 paid towards customer retention increases profits by more than $5 spent on new customer acquisition.

All Life Insurance companies offer various incentive plans for their employees, to boost their sales and to have higher balance of insurance plans, **“The higher the Insurance amount the higher the bonus pay out”**. However, not all companies get successful in the above Mantra. Some companies fail to have a better plan for bonus payouts.

**Project Objective**

An Insurance company is interested to predict the probability of a customer premium. This will help in strategizing the agent force to reach out to policy holders in advance to follow up for payment of premium and it will also help in predicting the agent bonus design appropriate engagement activity for their high performing employees and also have training programs for non/low performing employees

This is achieved by identifying the patterns of the default from the historical data & predict the default in premium payment by employing appropriate model/s, from the armory of machine learning and predictive analytics.

**Data Source**

The dataset contains the following information about 4520 policy holders.

**Project Approach**

The work that we have completed:

* Merged data from other sources like, demographic information, account details etc. for further deep analysis
* Data Quality and preparation activities were performed like missing value treatment, imputation, data type conversions for homogeneity in the data set
* Performed EDA on the data to understand the data and to determine any outlier and treatment for the same
* Also used ANNOVA to understand if the model performance can be improved

We have created multiple models and applied them on different sets of data as required. All the different models which were created were then evaluated using the AUC / F1 score at the end for the testing data set. Based on this an optimal model was chosen. Eventually we also found the feature importance for the most optimal model.

Various tree based as well as distance based models were built as part of this exercise using the different data sets as elaborated earlier. These models were built using sklearn and statsmodel libraries. There were various constraints, biggest one being the Type 2 error, which we had to minimize, as bonus variable was the main objective of this exercise. This will discuss in detail later in the report

At the end of the project, we also provided business recommendations.

### Problem 2

EDA - Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. - Both visual and non-visual understanding of the data.

**Resolution:**

* First we import all the necessary libraries in Python, and then import the data file which is ‘LifeInsuranceSales’. Once we import the file we confirm whether the data has been uploaded correctly or not using ‘head’ function. Using this function we can view the data and all the columns and headers whether they are aligning correctly or not.
* Then using the ‘shape’ function we can understand how many row and columns are there in our data set.
* To check the data type of all the columns and also to check the null values, ‘info’ function. Has been used.
* To see the detail description of the data such as, Count, Mean, Median, Min, Max, Standard Deviations etc,
* Using the ‘isnull’ function, one can understand if there are any null values in the data set. And we do not have any null values in the existing data set.
* Using the ‘dups’ function we check for the duplicates and there were no duplicate values.
* We also identified the unique values in categorical data.

**Variable Rationalization**

In order to study the data better, I performed a preliminary variable reduction in the beginning itself. At this stage, we reduced the variable on the following criteria:

* Redundant Variables
* Business relevance
* Correlated Variables
* Clubbed Variables

**Important Variables Identification**

For determining the important variables for predicting accurate agent bonus, subject matter expertise is the method followed. Those variables are:

• Percentage Premium paid by cash or credit

• Age of the policy holder

• Monthly income of the policy holder

• Sum assured, in Insurance industry terminology policy amount

• Existing policy tenure till the date of collection of data

**Note: (Please refer to the graphical representation at the end of the document)**

**Unimportant Variables Identification**

On checking the relationship between monthly income and categorical variables (below variables) in EDA, intuitively it appears to be unimportant variables, however significance can be interpreted post model validation, therefore retained for the purpose of running models:

* Marital Status
* Occupation
* Designation
* Education Field
* Channel

**Note: (Please refer to the graphical representation at the end of the document)**

**Correlation:**

Top 5 strong correlations:

* Sum Assured & Agen tBonus
* Monthly Income & Agent Bonus
* Customer Tenure & Agent Bonus
* Age & Agent Bonus
* Monthly Income & Sum Assured

**Note: (Please refer to the graphical representation at the end of the document)**

**Insights:**

* Agent bonus, majority of the bonus falls between 3,000 to 4,000.
* Looking at the data we can see that the data collected with a wide verity of age range from 18 years to 58 years.
* There is good mix of gender where we have 40% Female 60% and 60% Male
* 50% of the data consist of married people
* From the entire data we have about 35% of the people who are at Manager Level.
* About 49% of the people who took Life Insurance policy are salaried employees.
* 42% of the people have 4 existing policies
* 97% of the business comes from North and West Zone
* Only 11% of the premium is paid between Monthly and Quarterly

**The key observations (summary) based on exploratory analysis are as follows:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.No** | **Columns** | **Description** | **Data Type** |
| 1 | CustID | Unique customer ID | int64 |
| 2 | AgentBonus | Bonus amount given to each agents in last month | int64 |
| 3 | Age | Age of customer | float64 |
| 4 | CustTenure | Tenure of customer in organization | float64 |
| 5 | Channel | Channel through which acquisition of customer is done | object |
| 6 | Occupation | Occupation of customer | object |
| 7 | EducationField | Field of education of customer | object |
| 8 | Gender | Gender of customer | object |
| 9 | ExistingProdType | Existing product type of customer | int64 |
| 10 | Designation | Designation of customer in their organization | object |
| 11 | NumberOfPolicy | Total number of existing policy of a customer | float64 |
| 12 | MaritalStatus | Marital status of customer | object |
| 13 | MonthlyIncome | Gross monthly income of customer | float64 |
| 14 | Complaint | Indicator of complaint registered in last one month by customer | int64 |
| 15 | ExistingPolicy Tenure | Max tenure in all existing policies of customer | float64 |
| 16 | SumAssured | Max of sum assured in all existing policies of customer | float64 |
| 17 | Zone | Customer belongs to which zone in India. Like East, West, North and South | object |
| 18 | PaymentMethod | Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly | object |
| 19 | LastMonthCalls | Total calls attempted by company to a customer for cross sell | int64 |
| 20 | CustCareScore | Customer satisfaction score given by customer in previous service call | float64 |

### Problem 3

Data Cleaning and Pre-processing - Approach used for identifying and treating missing values and outlier treatment (and why) - Need for variable transformation (if any) - Variables removed or added and why (if any)

**Resolution:**

**Research Approach**

In the subsequent sections, we will create a predictive model based on logistic regression and other machine learning models to understand and predict the agent bonus

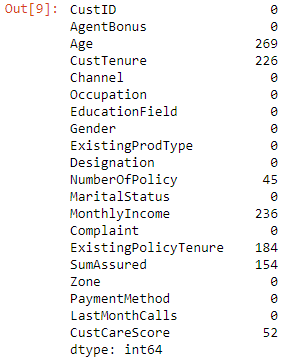
**Data preparation**

**Data Cleaning and Pre-processing:**

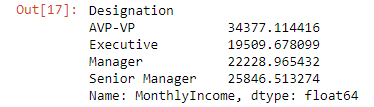
First we need to convert categorical string variables into number. Secondly, missing values needed to be treaded. We are imputing missing values with K-Nearest Neighbors.

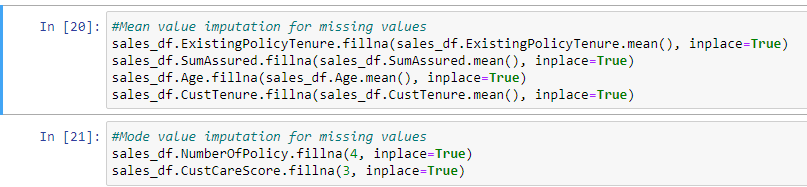
**Missing Value Treatment**

The data has lot of missing data. Therefore, we need to treat the missing values and it is required. First we check for missing values : We can see that there are lot of missing values in multiple columns

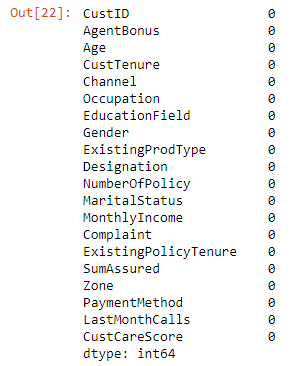
****

Then we are substituting Missing values for MonthlyIncome





Once the imputation is done we can check for data and there are no missing values



**Outlier treatment**

The data contains outliers in variables

* Monthly Income
* Customer Tenure
* Existing Policy Tenure
* Sum Assured
* Number of policy
* Customer Score

However, with a view that outliers exist in real time data and imputation or capping or removal results in in data loss – Yet, some outliers were treated and removed from the data set.

**Note: (Please refer to the graphical representation at the end of the document)**

**Normality Test of the Continues Variables**

To check whether the data is unbalanced, we checked using Shapiro-wilk to test the normality of the continues variables.

The Shapiro-Wilk test for normality is available when using the Distribution platform to examine a continuous variable. The null hypothesis for this test is that the data are normally distributed. If the p-value is greater than 0.05, then the null hypothesis is not rejected.

H0- Data is normal H1 - Data is not normal

**Note: (Please refer to the graphical representation at the end of the document)**

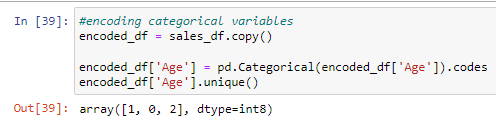
**Variable Transformation**

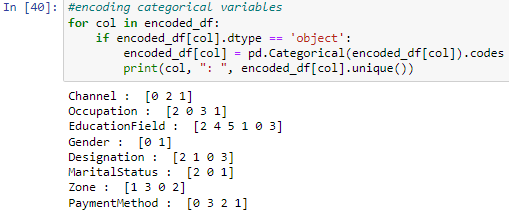
For the purpose of model building and from thenceforth, there were few variables which were transformation below are the transformed variables.

Categories created for age:



Encoding categorical variables





**Addition of new variables**

There were 2 new variables added in the data set. We will consider 3 ultimate clusters as that is giving us very fewer negative silhouette widths than 4 clusters

Note: Positive silhouette width suggests that the observation belong to the correct cluster, negative would be opposite.

### Problem 4

Model building - Clear on why was a particular model(s) chosen. - Effort to improve model performance.

**Resolution:**

**Tools and Techniques used**

* Python for Data preparation, as data set is very large (4520 records),
* Training & Testing data split - 70: 30

**Binning**

Binning the outliers is the method used to classify data into categories to smoothen the presence of outliers. These bins would be useful in providing insights of the category or categories where customers might default.

**One hot encoding**

The features are encoded using a one-hot (aka ‘one-of-K’ or ‘dummy’) encoding scheme. This creates a binary column for each category and returns a sparse matrix or dense array (depending on the sparse parameter)

**Data split into test and train**

The data containing the target variable (Agent Bonus) is being split into Train data and Test data. The purpose of creating subset of training and test data set is to create a model based on train data and validate the built model using the test data set.

The industry best practice states

* The training and testing data sets are too be different to avoid overfitting.
* 70% into train data and 30% test data is usually an acceptable division.

KFold Cross Validation has been adopted to bring in the benefits of multiple random splits of dataset into training and testing. This is a powerful tool to prevent overfitting of the model. This is used to determine the optimal parameters of the model.

**Models Used**

As the objective of this project is to predict the agent bonus it is a classification problem.

The dataset contains target variable “Agent Bonus”, wherein “0” represents that the agent will not be renewed and “1” that agent will be rewarded. Therefore, supervised learning algorithm needs to be applied for this prediction.

Different supervised learning algorithms in classification problems that are applied are:

1. Logistic Regression:

Logistic Regression is a classification algorithm that estimates discrete values like yes/no, true/false, 0 or 1 etc. This model is most useful for understanding the influence of several independent variables on a single outcome variable. It works very well on linearly separable classes, making use of odds ratio and sigmoid function.

Interpretation

By using Logistic Regression, we have predicted the agent bonus who will be getting higher bonus (on train data) who with an accuracy of over 80.81% which seems good.

Also we have predicted the agent bonus will be getting higher bonus (on test data) with an accuracy of over 81.28% which seems good.

1. K-Nearest Neighbours (KNN):

KNN is a simple supervised learning algorithm that is used for solving both regression and classification problems. This is called a lazy learner because it computes the maximum points of K nearest neighbours for a given new data point.

1. Ensemble methods:

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. Bagging, Boosting and Stacking are different ensemble methods that are used.

**Models Built**

Various tree based as well as distance based models were built as part of this exercise using the different data sets as elaborated earlier. These models were built using sklearn and statsmodel libraries. There were various constraints, biggest one being the Type 2 error, which we had to minimize, as bonus variable was the main objective of this exercise. This will discuss in detail later in the report.

**Model Tuning method**

Various model tuning approaches were followed. Primarily we made use of GridSearchCV function with cv = 3 for model hyper parameter tuning. Also we had to tweak the threshold values to maximize the recall values. Threshold tweaking was required as we had a typical problem of recall precision trade off.

Various different approaches were followed to create multiple models. As mentioned earlier we had created multiple data sets like tree, tree\_scaled, tree\_smote, tree\_smote\_scaled, linear, linear\_scaled,linear\_smote, linear\_smote\_scaled etc.

We have also creating two generic functions which will be used to evaluate various models and also to tweak their threshold to maximize the recall.

Usage 1 – APPLY\_EVAL

This is used to train the model, apply the model on test set and then output all the performance metrics like confusion matrix, Classification report, AUC curve etc.

Usage 1 – TWEAK\_THRESHOLD

This is used to tweak the threshold, once the best model has been selected after hyper parameter tuning. Threshold is tweaked to maximize the recall.

Logic 2 – APPLY\_EVAL

X\_train, X\_test,y\_train & y\_test are input to the function along with the model and param grid for GCV. Model is trained, tuned then validated against the test set and performance metrics are generated.

Logic 2 – TWEAK\_THRESHOLD

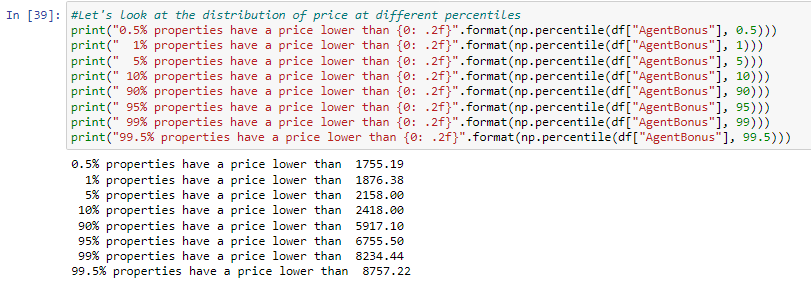
Threshold tweaking is done by calculating performance metrics like recall for all the values of probabilities between 0 and 1, and a step size of 0.1. Threshold with best AUC score is selected.

|  |  |  |
| --- | --- | --- |
| **TREE / LINEAR** | **ENSEMBLE MODELLING** | **SCALED / UNSCALED** |
| 1. We used two data sets.   * Tree * Linear   2.Tree - For Tree based models like CART, Random Forest etc.  3. Linear - For distance based models like Kmeans, LDA etc. | 1. Various ensemble models were also used apart from regular models.  2. Both Bagging and Boosting approaches were tried, evaluated and compared to determine the best model for our purpose. | 1. Some of the models were sensitive to scaling e.g. SVM, KMeans etc.  2. On the other hand we had models like Logit and other tree based models which are scaling agnostic, we used unscaled data set there. |

We have created multiple models as part of the Agent bonus prediction. The models include descriptive models like KMeans where we try to segment the gain insights and also predictive classification models like Random Forest, Gradient Boosting model, Logistic regression in order to predict bonus. Combined they can provide prescriptive analysis to the life insurance company and help them with the strategies.

Various permutation and combinations were tried for various models.

* We have included the distribution of price at different percentiles



* Let's create a range variable to understand how many records we have in different slabs

**Note: (Please refer to the graphical representation at the end of the document)**

Log transformation of the AgentBonus variable looks to be slightly more symmetrically distributed. We can use a log of the AgentBonus variable as our target variable in the regression model, to check if performance is better than the AgentBonus feature used without any transformation.

Sum assured is highly correlated to Agent Bonus - we can see it in the below table.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AgentBonus** | **Age** | **CustTenure** | **ExistingProdType** | **NumberOfPolicy** | **MonthlyIncome** | **Complaint** | **ExistingPolicyTenure** | **SumAssured** | **LastMonthCalls** | **CustCareScore** |
| **AgentBonus** | 1 | 0.5523 | 0.5558 | 0.113 | 0.0793 | 0.5667 | 0.0143 | 0.3491 | 0.8449 | 0.1997 | 0.0232 |
| **Age** | 0.5523 | 1 | 0.3235 | 0.0735 | 0.0468 | 0.328 | 0.0203 | 0.1915 | 0.4662 | 0.1169 | 0.0343 |
| **CustTenure** | 0.5558 | 0.3235 | 1 | 0.0828 | 0.0487 | 0.3184 | 0.0043 | 0.1928 | 0.4682 | 0.1177 | 0.0115 |
| **ExistingProdType** | 0.113 | 0.0735 | 0.0828 | 1 | 0.1499 | 0.1906 | -0.003 | 0.0593 | 0.1037 | 0.0332 | 0.0041 |
| **NumberOfPolicy** | 0.0793 | 0.0468 | 0.0487 | 0.1499 | 1 | 0.1335 | -0.016 | 0.0505 | 0.0638 | 0.0751 | -0.001 |
| **MonthlyIncome** | 0.5667 | 0.328 | 0.3184 | 0.1906 | 0.1335 | 1 | -0.005 | 0.1425 | 0.4607 | 0.3374 | 0.0356 |
| **Complaint** | 0.0143 | 0.0203 | 0.0043 | -0.003 | -0.016 | -0.005 | 1 | 0.0027 | -2E-04 | -0.026 | -0.004 |
| **ExistingPolicyTenure** | 0.3491 | 0.1915 | 0.1928 | 0.0593 | 0.0505 | 0.1425 | 0.0027 | 1 | 0.3018 | 0.0965 | -0.007 |
| **SumAssured** | 0.8449 | 0.4662 | 0.4682 | 0.1037 | 0.0638 | 0.4607 | -2E-04 | 0.3018 | 1 | 0.158 | 0.0033 |
| **LastMonthCalls** | 0.1997 | 0.1169 | 0.1177 | 0.0332 | 0.0751 | 0.3374 | -0.026 | 0.0965 | 0.158 | 1 | 0.0064 |
| **CustCareScore** | 0.0232 | 0.0343 | 0.0115 | 0.0041 | -0.001 | 0.0356 | -0.004 | -0.007 | 0.0033 | 0.0064 | 1 |

Let's check if being a SumAssured has any bonus impact –

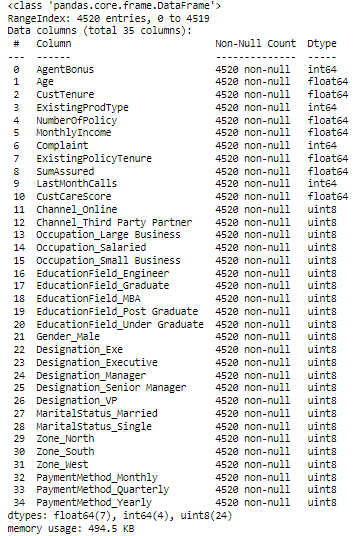
**Note: (Please refer to the graphical representation at the end of the document)**

Above box plots gives us the zone wise distribution

Below table shows StandardScaler

Then we apply Zscore



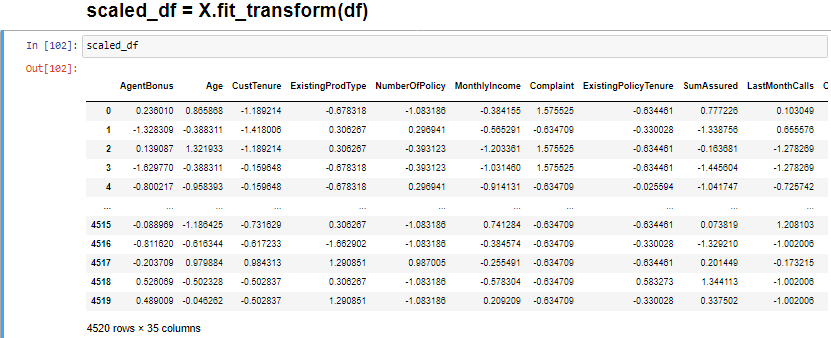


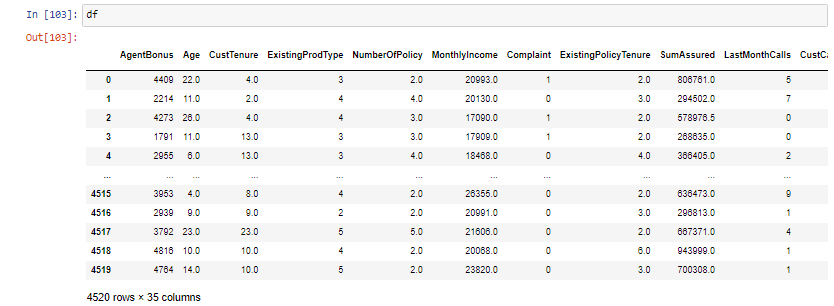
### Problem 5

Model validation - How was the model validated ? Just accuracy, or anything else too?

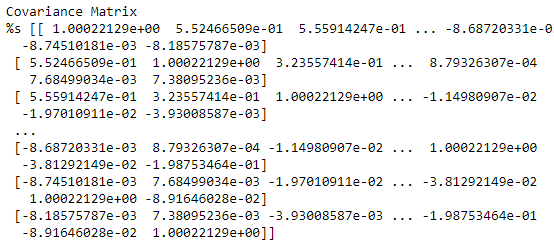
**Resolution:**

New data set - scaled\_df = X.fit\_transform(df)

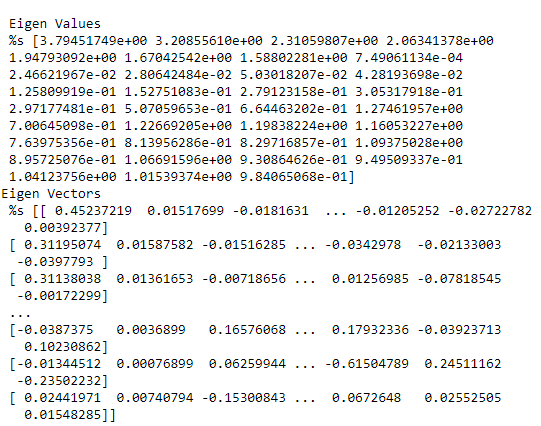
****

****

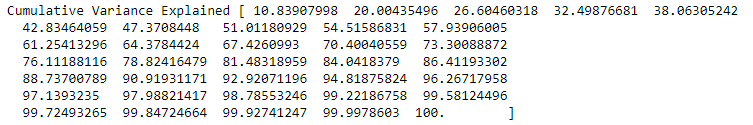
**We create the Covariance Matrix**



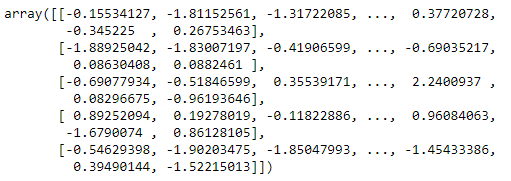
**Step 2- Get eigen values and eigen vector**



We also performed Cumulative Variance



**Note: (Please refer to the graphical representation at the end of the document)**





Visually we can observe that there is steep drop in variance explained with increase in number of PC's.

We will proceed with 5 components here. But depending on requirement 90% variation or 5 components will also do well.

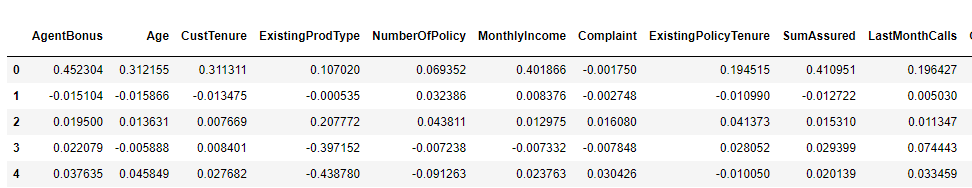
Cumulative sum of variance explained with [n] features



Below graph shows the PCA Analysis.

**Note: (Please refer to the graphical representation at the end of the document)**

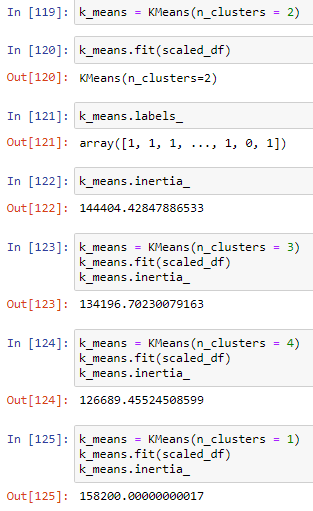
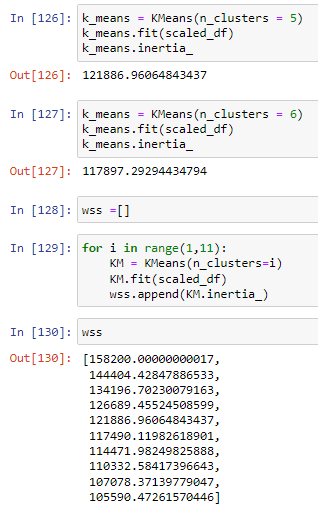
Below table gives the snapshot of scaled data frame



**Note: (Please refer to the graphical representation at the end of the document)**

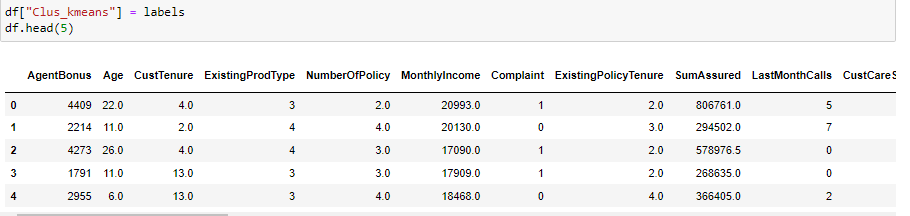
The above heat map and the color bar basically represent the correlation between the various features and the principal component itself. Component 2 looks more related to aspect - We can label it as aspect property. Depending on relations ship, we could go ahead and label relationship with features.

Here we start with KMeans clustering

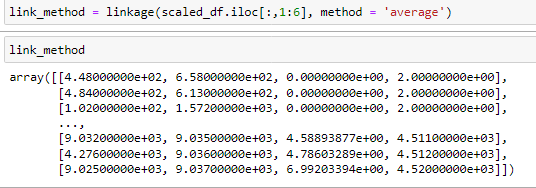
**Note: (Please refer to the graphical representation at the end of the document)**

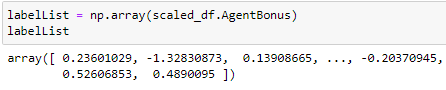
The above graph shows the WSS



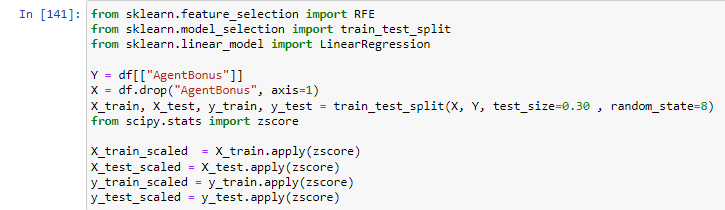
After the clustering we prepare

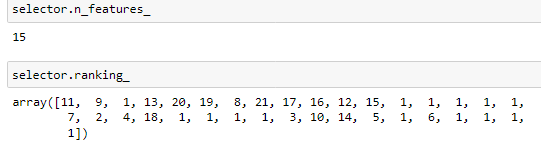






Now we create Regression Model





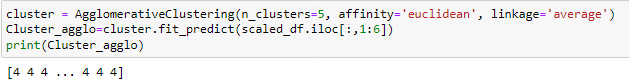
Below table gives us the feature and the rank



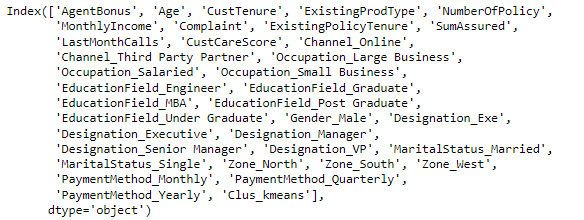
Most important features are Designation, Occupation, South zone, Payment Method, Existing Prod Type, Education Field,

Zone west and North, Existing Policy Tenure, Cust Tenure, Age

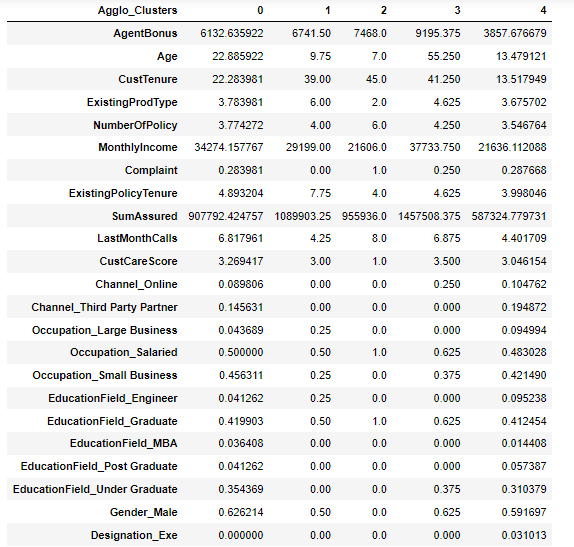
We have also created Agglomerative Clustering

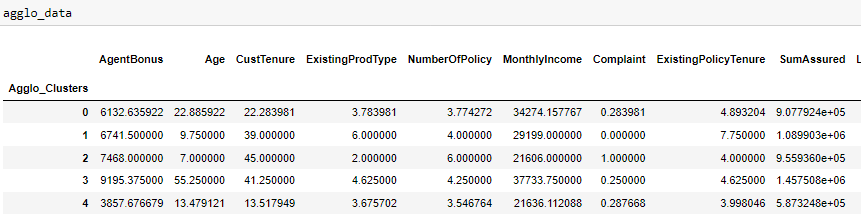


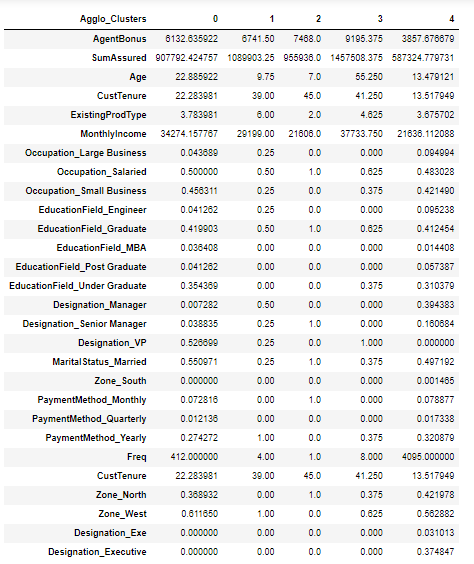
We then drop df.drop(columns=['Agglo\_CLusters'],inplace=True)



Below table shows the grouping by Agglo\_Clusters



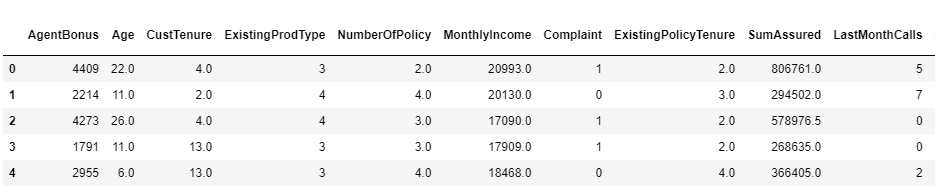




Now we get the silhouette score

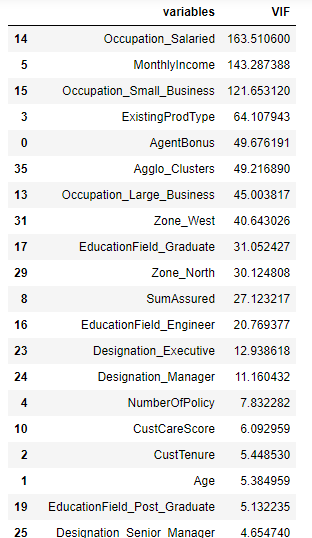


Below table shows the data with Sil\_width





Next we calculate variance inflation factor

Predict mileage (mpg) for a set of attributes not in the training or test set

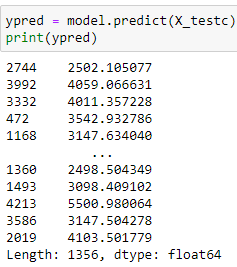
Since this is regression, plot the predicted y value vs actual y values for the test data

A good model's prediction will be close to actual leading to high R

Now we get the value of coefficient of determination

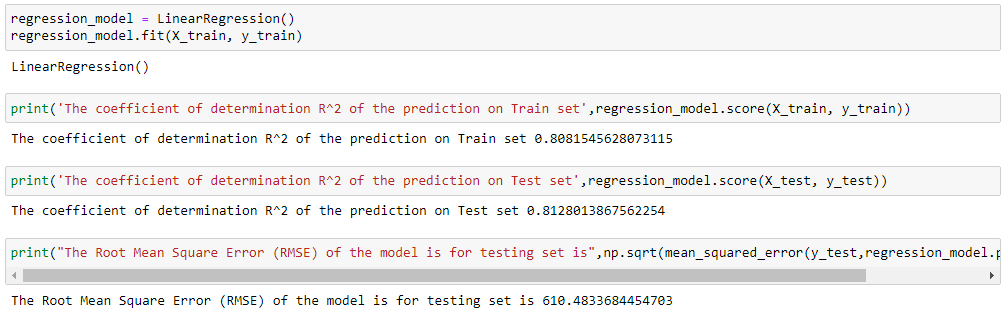


Then we get the Predictions on test set



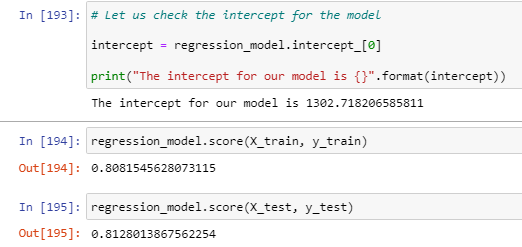
The Root Mean Square Error

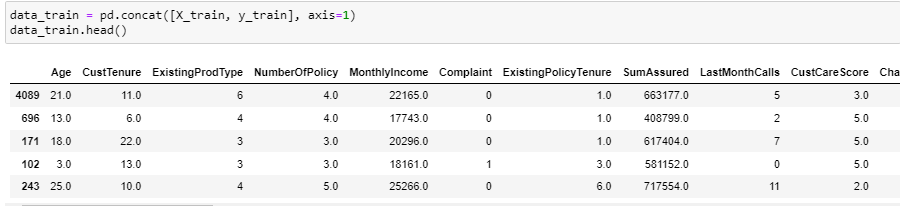


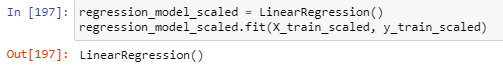


Let us explore the coefficients for each of the independent attributes

Now let us check the intercept for the model



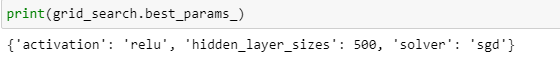




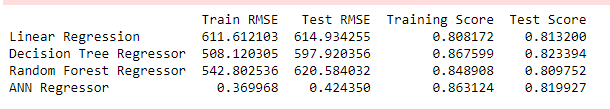
Let us explore the coefficients for each of the independent attributes

Predict mileage (mpg) for a set of attributes not in the training or test set

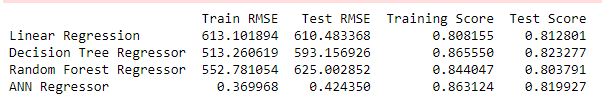
Since this is regression, plot the predicted y value vs actual y values for the test data. A good model's prediction will be close to actual leading to high R and R2 values



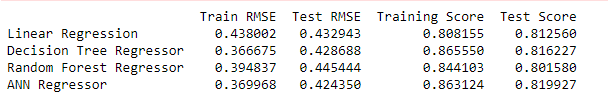
**best\_params\_annr={'activation': 'relu', 'hidden\_layer\_sizes': 500, 'solver': 'sgd'}**



**Without tuning**



**Final Output**



Model score - R2 or coeff of determinant

R^2=1–RSS / TSS

0.8131119381670872

**We can see that the scaled output has a better score.**

Now we see stats model formula

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 1.95e+07. 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 at least 1 regression co-efficient is not 0. Here all regression co-efficients are not 0

### Problem 6

Final interpretation / recommendation - Very clear and crisp on what recommendations do you want to give to the management / client.

**Resolution:**

**Actionable Insights and Recommendations**

The final interpretation is as follows:

The objective we have here is that an Insurance company is looking for a practicable model to predict the agent bonus for the high performing agents. This will help in strategizing the agent force to reach out to policy holders in advance to follow up for payment of premium

1. Based on the Variable importance of the Logistic Regression model, the insurance company is suggested to orient its agents force to contact policy holders for Renewal premium Monthly and Quarterly.
2. Age between 30-40 years are higher. As we saw in the EDA, the mean age of a policyholder is around 35 years and agent can target these age group.
3. Zones: Business need to focus on expanding their reach in East Zone and South Zone
4. Agents should focus more on AVPs as their number is really high all the metrics.
5. When Age increases by 1 unit, AgentBonus increases by 23.04 units, keeping all other predictors constant. similarly, when MonthlyIncome increases by 1 unit, AgentBonus increases by 0.03 units, keeping all other predictors constant.
6. There are also some negative co-efficient values. Occupation\_Large\_Business has its corresponding co-efficient as -39.31. This implies, when the Occupation is large business, the AgentBonus decreases by 39.31 units, keeping all other predictors constant.

The End

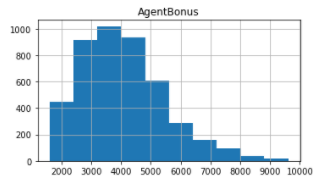
Thakur Arun Singh

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

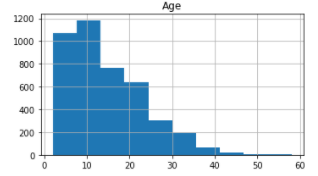
### Appendix – Graphical Representations

|  |  |  |
| --- | --- | --- |
| Title | Artifact/Location | Remarks |
| Source of Data |  | Data File for Insurance Premium |
| Data Dictionary |  | Insurance Premium Renewals (provided with dataset) |
| Modified Data Source(for Model Validation) |  | Variable description and rationale behind selection |
| Project Notes 1 |  | As per Capstone Project guidelines/instructions |
| Project Notes 2 |  | As per Capstone Project guidelines/instructions |
| Project Presentation |  | As per Capstone Project  guidelines/instructions |
| Python Code for Reference |  | Complete Python-code used in EDA, Model Building, etc. |

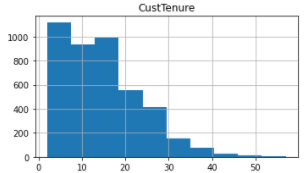
**Agent Bonus:** From the below graph we can see that the majority of the bonus falls between 3,000 to 4,000.



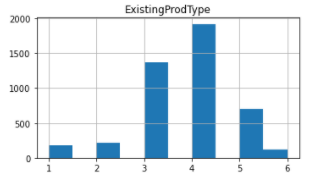
**Age:** Below the graph shows that the average age of the customer is about 30-35 years.

****

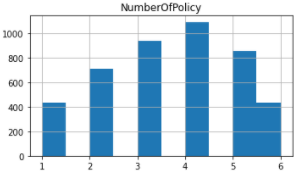
**Customer Tenure:** Below graph shows that there are over 1000 customer who are loyal customers for at least 10 years.



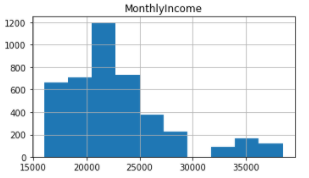
**Existing Product Type:** Below graph shows that there at about 2000 people who hold at least 4 products



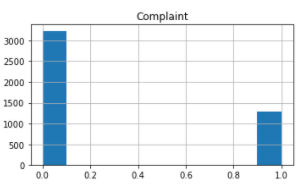
**Number of Policies:** Below graph shows that at least 1000 people hold 4 policies each.

****

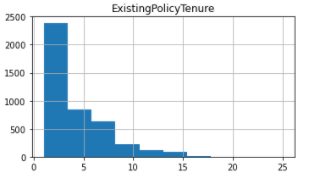
**Monthly Income:** About 1200 people have a monthly income between 20,000 – 25,000.



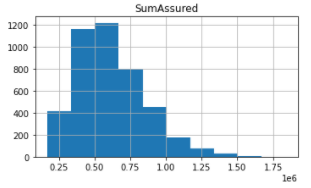
**Complaints:** There were about 3000+ people who had complained between 0-1 times.

****

**Existing Policy Tenure:** There are about 2500 customer who took policies at least a year ago.



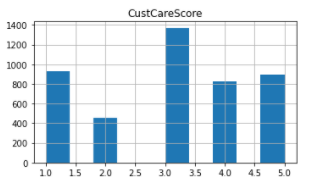
**Sum Assured:** The major count of the sum assured is between 400K and 600K



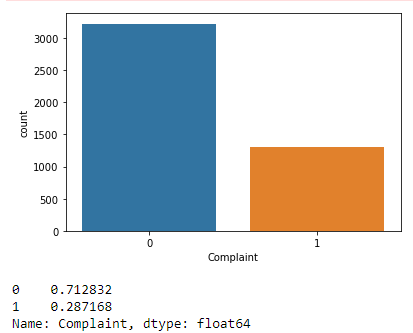
**Last Month Calls:** We have received at least 2 calls from nearly 1400 customers

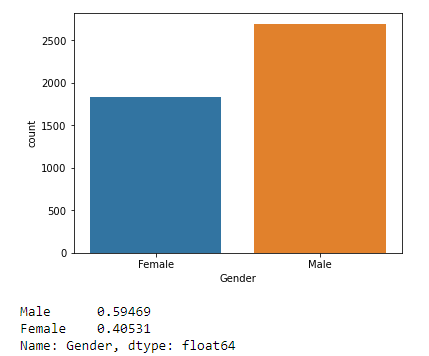
****

**Customer Care Score:** Marjory of the customer lies between the Customer care score of 3 – 3.5.

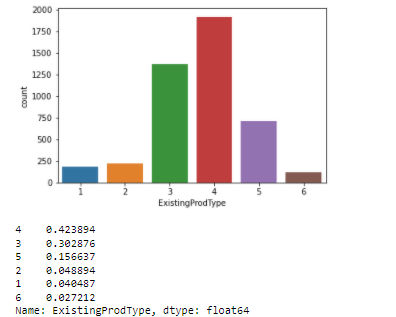


**Complaints:** Below graph shows the count of the customers who complained and not complained.

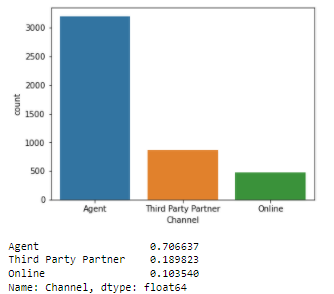


**Gender:** Below graph gives me the count of Gender

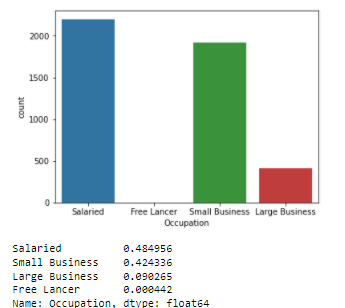
**Existing Policy Type:** Below Count Plot shows the existing policy types



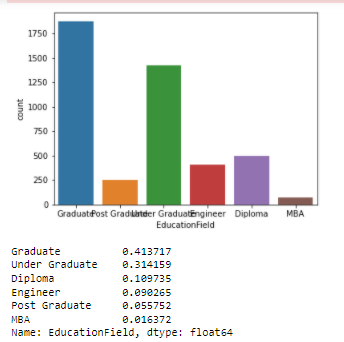
Blow graph shows the channel data



Below graph shows the occupation



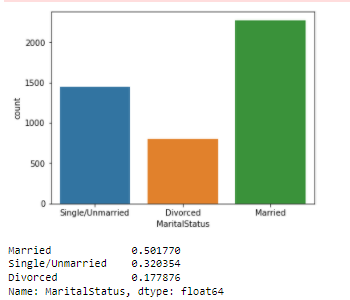
Below graph shows the education field



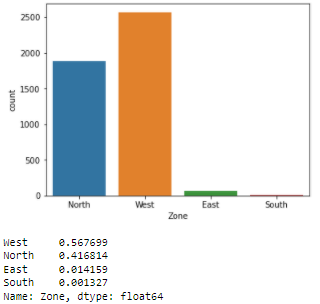
Below graph shows the designation



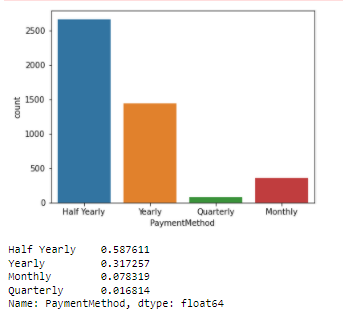
Below graph shows the marital status



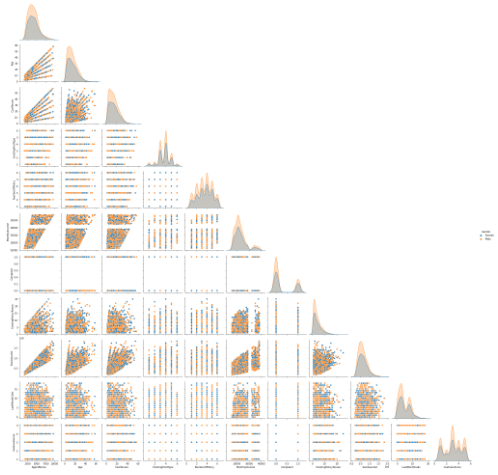
Below graph shows the various Zones



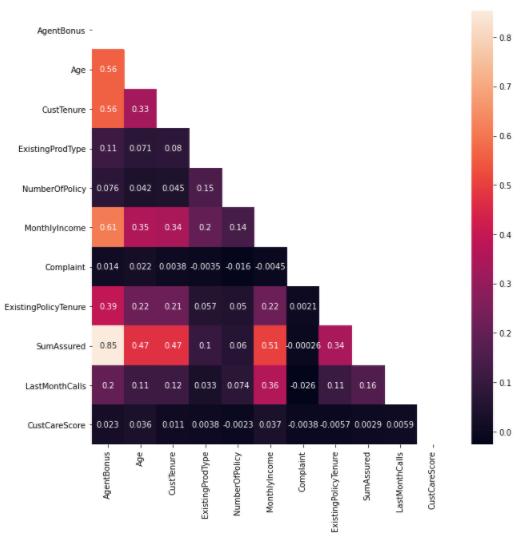
Below graph shows various payment methods



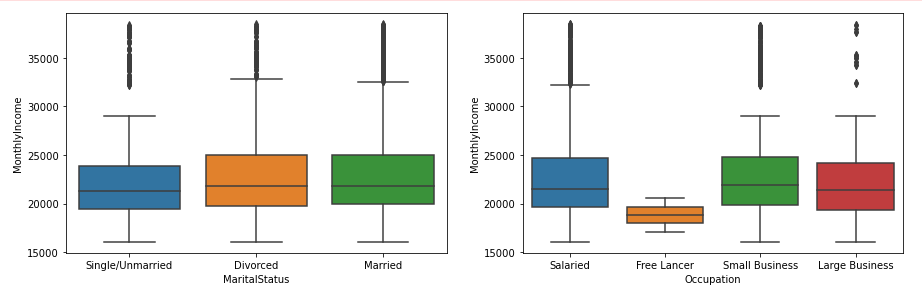
Below Pair plot shows the pairwise relationships in a dataset

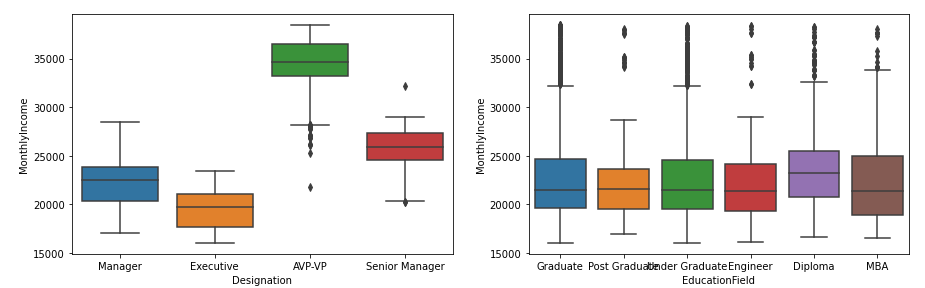


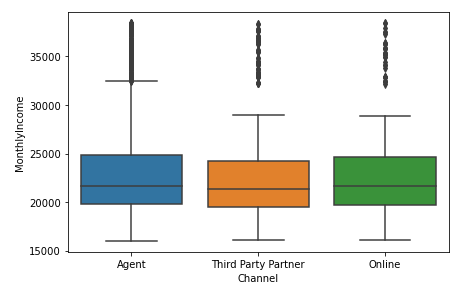
Below Correlation / Heat Map shows Strong Correlation



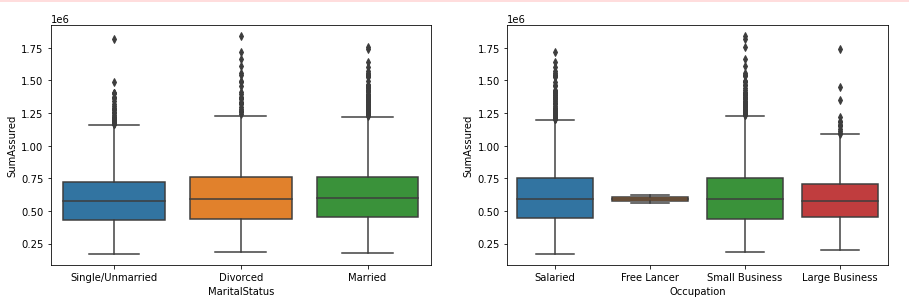
Below box plots shows relationship between MonthlyIncome & categorical variables

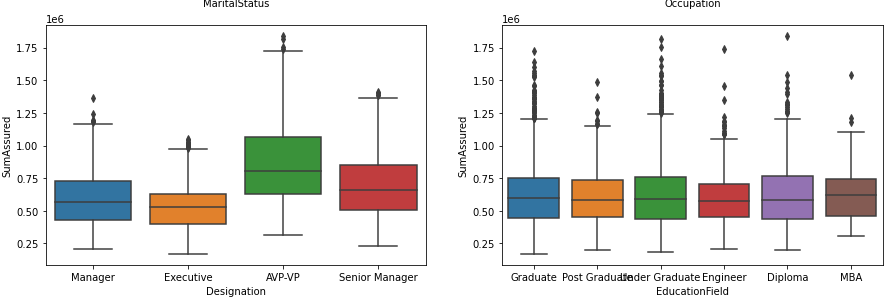


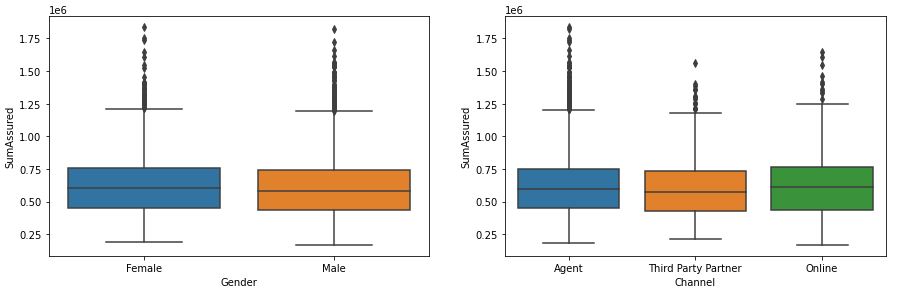




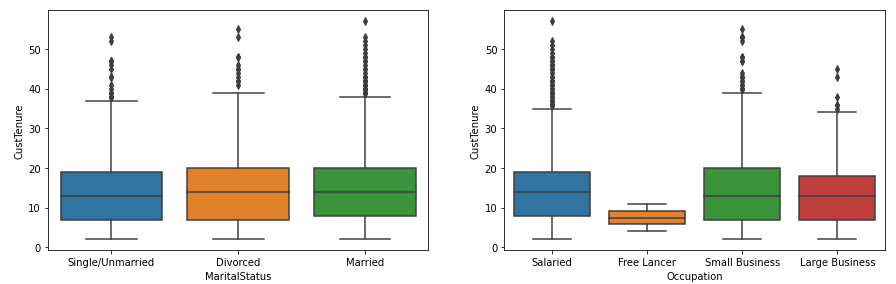
Customer Designation creates clear groups for MonthlyIncome of the customer so Missing Values in MonthlyIncome will be filled considering means of every group

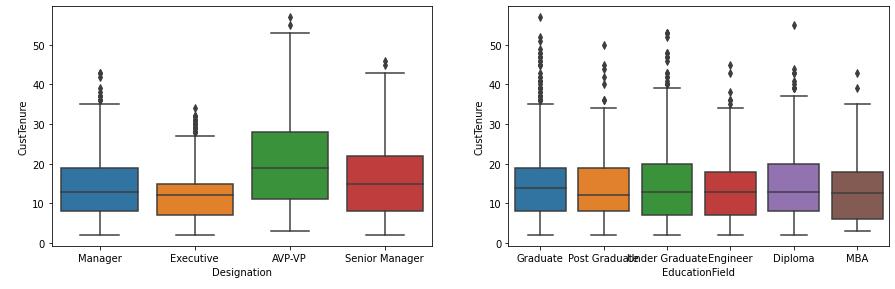


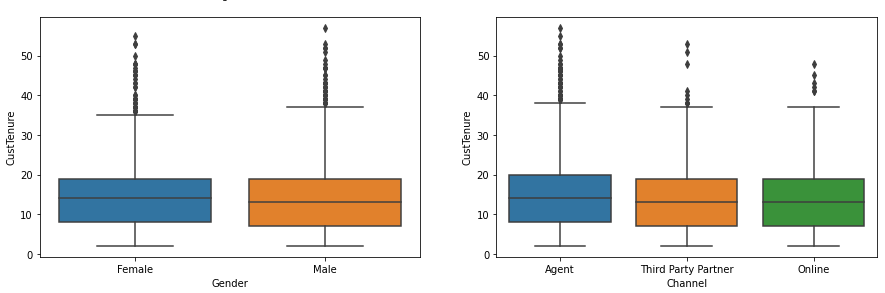




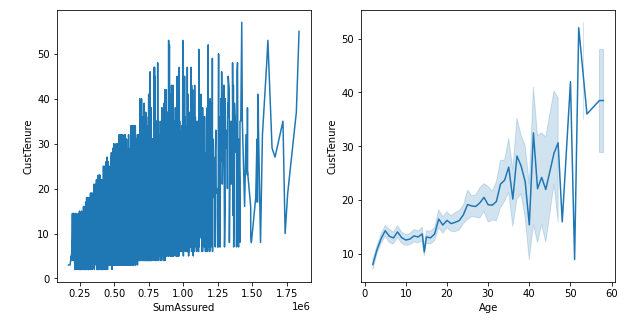
Below box plots shows the relationship between CustTenure & categorical variables



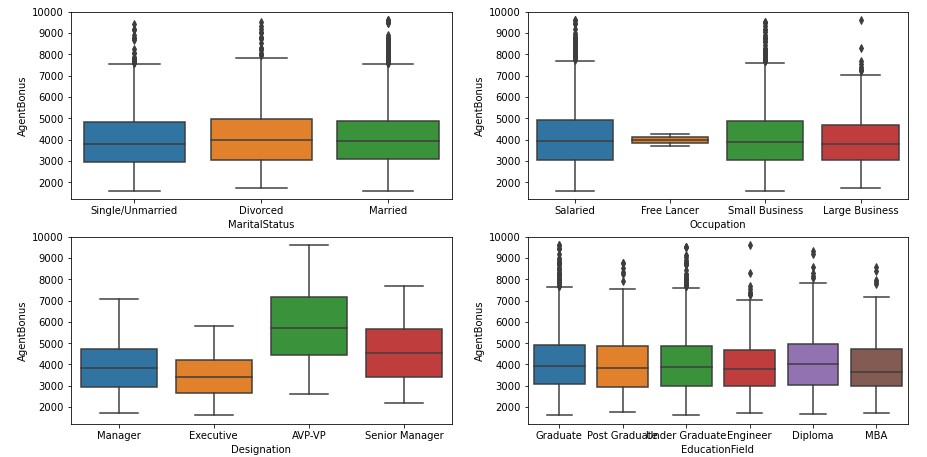


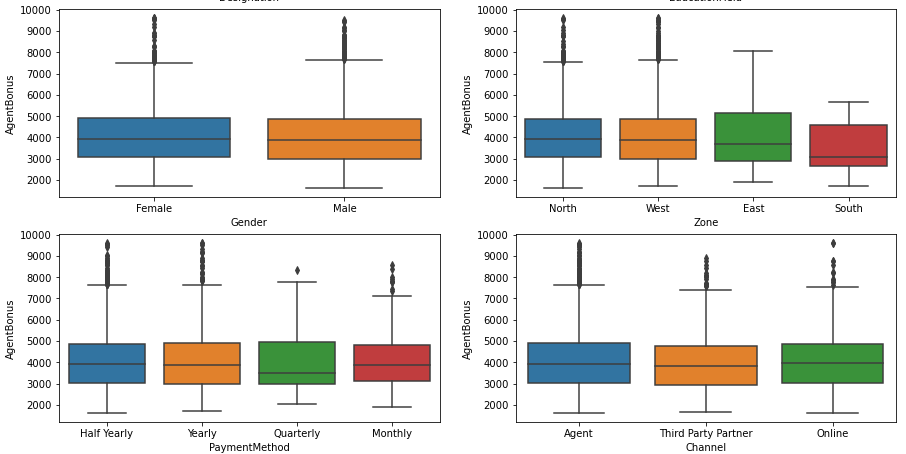


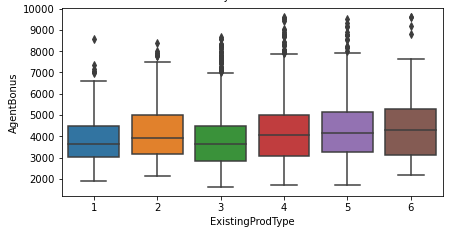
Below graphs shows the relationship between Customer tenure VS Sum Assured and age



Below Box plots shows relationship between AgentBonus & categorical variables

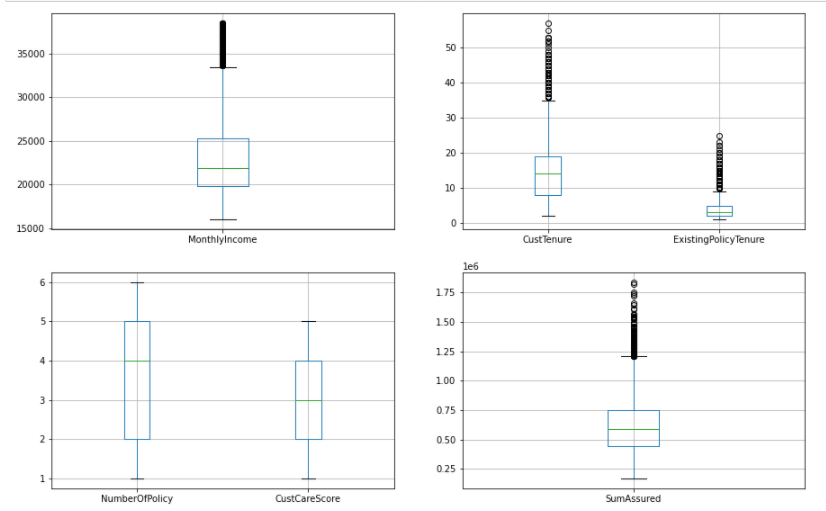






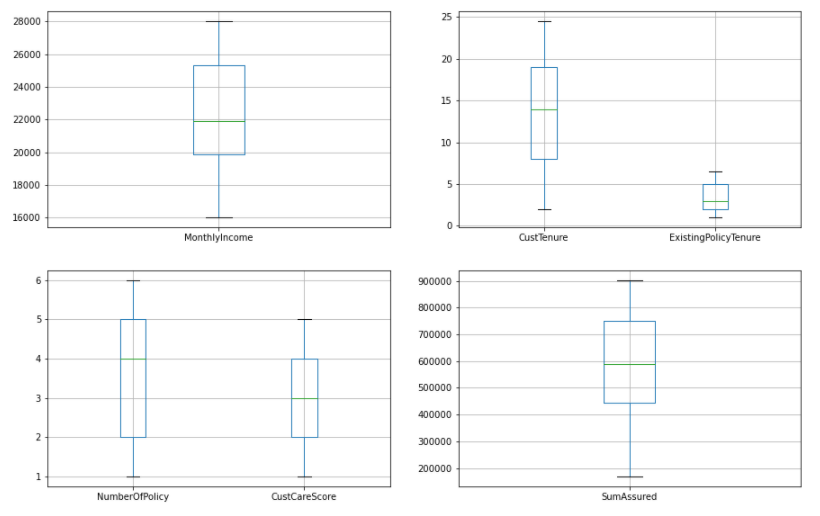
**Outlier treatment**

First we check for the outliers and box plots shows the outliers

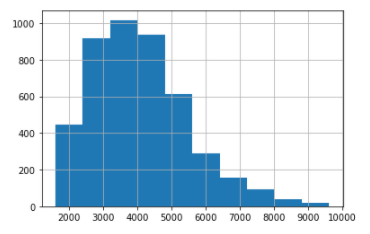


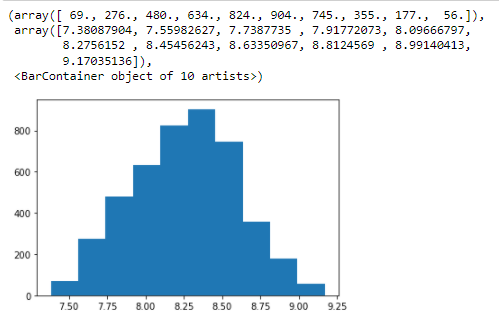
Based on the above graph we remove the below outliers

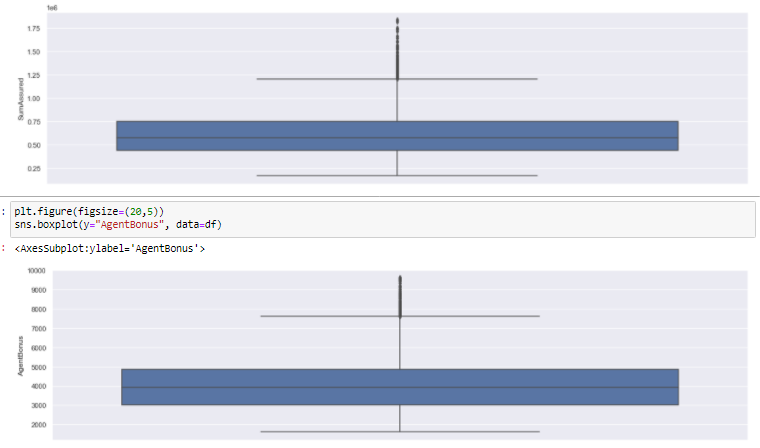
After removing the outliers, we can see from the below box plots that there are no outliers

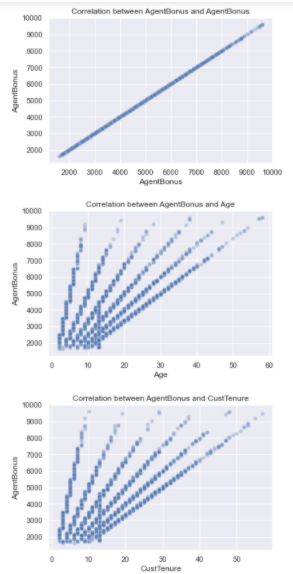


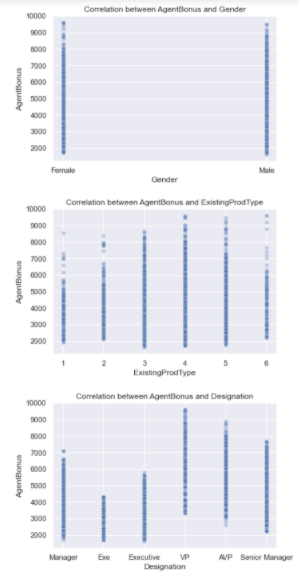
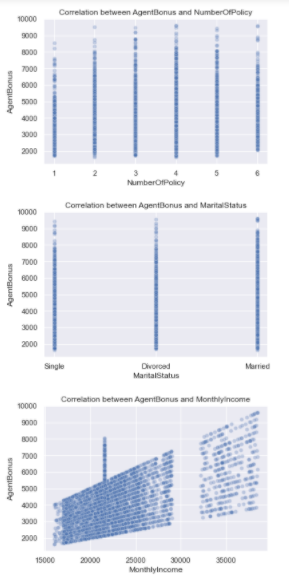
**Normality Test of the Continues Variables**

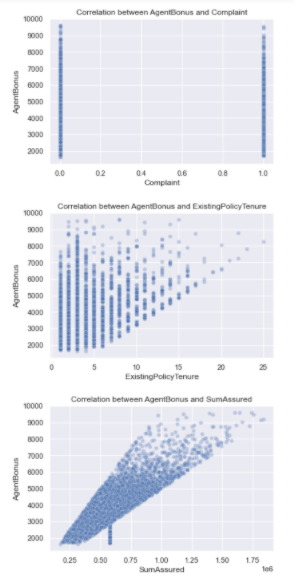


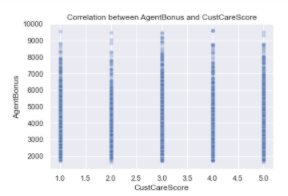
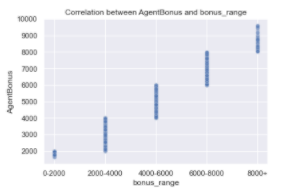


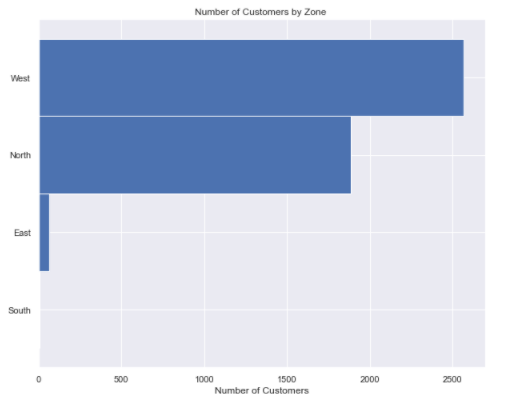


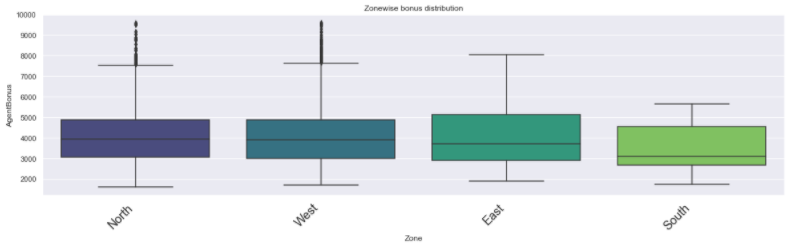
 

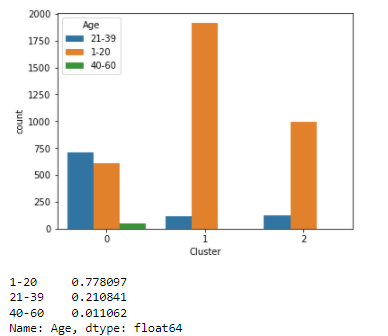
 

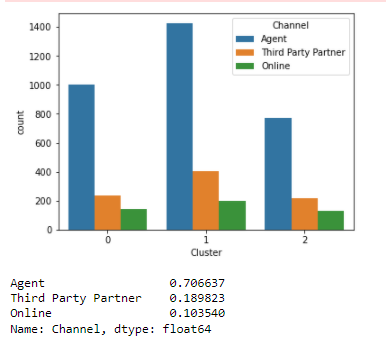
 

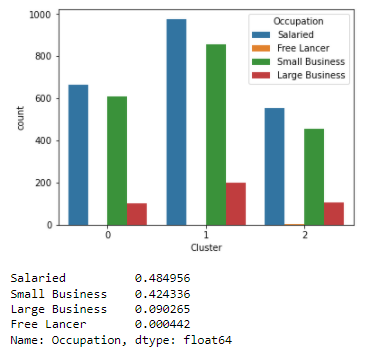


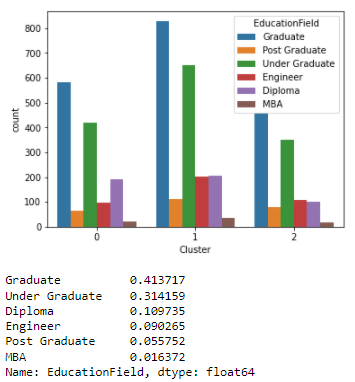


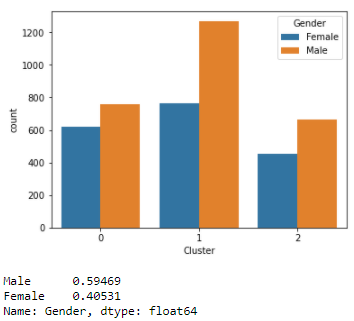
Below graphs shows the findings with variables VS Clusters

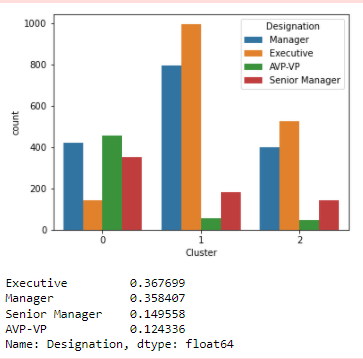


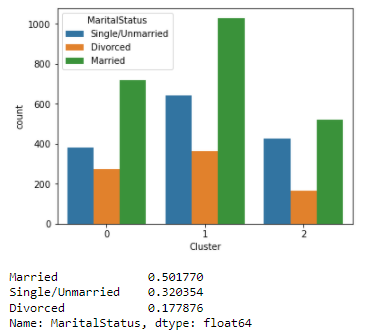


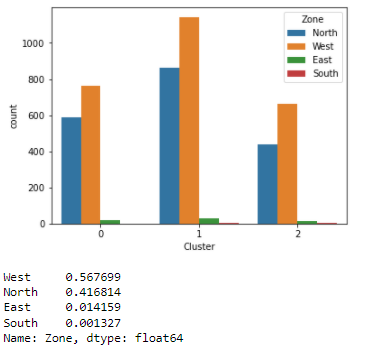


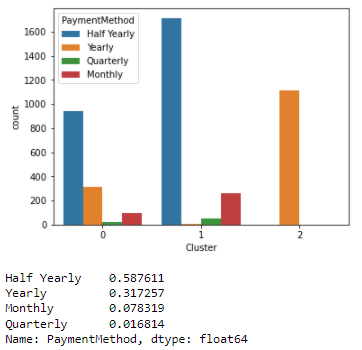




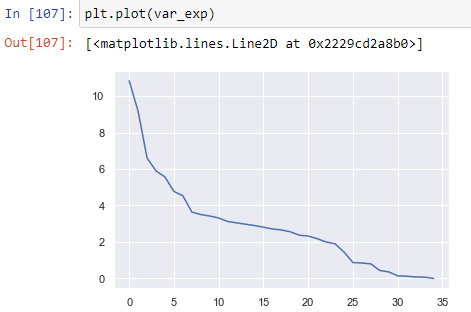




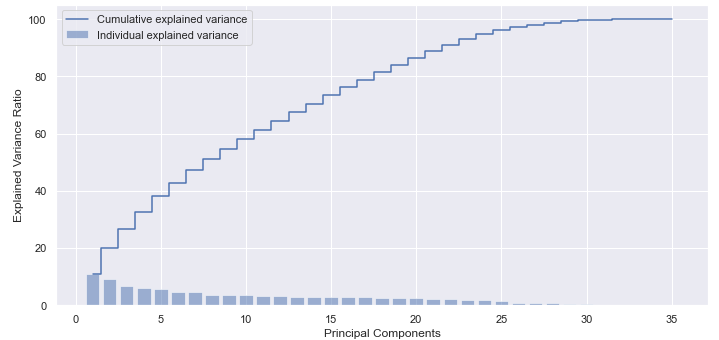




Cumulative variance



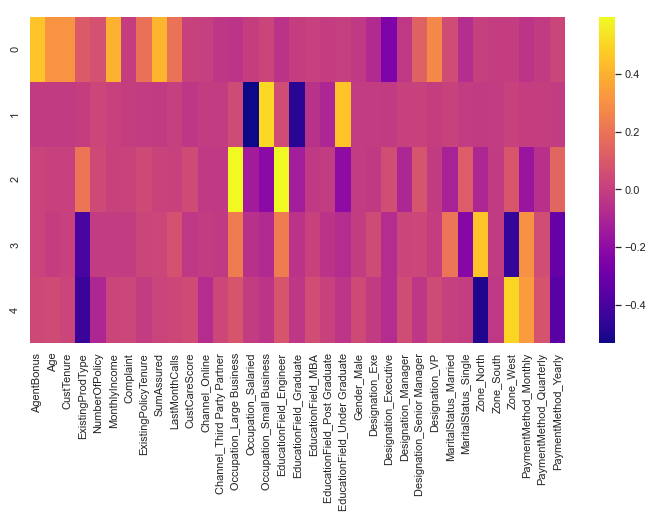
The below figure shows Individual explained variance and Cumulative explained variance plotted against Explained Variance Ratio Principal Components

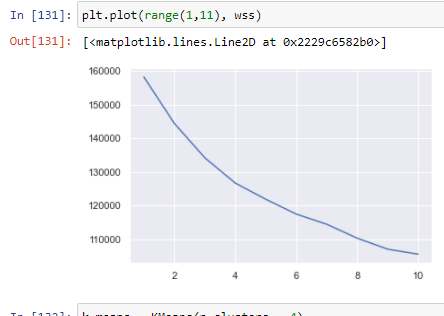


Then using scikit learn PCA. It does all the above steps and maps data to PCA dimensions in one shot

**NOTE** - we are generating only 4 PCA dimensions (dimensionality reduction from 18 to 4)

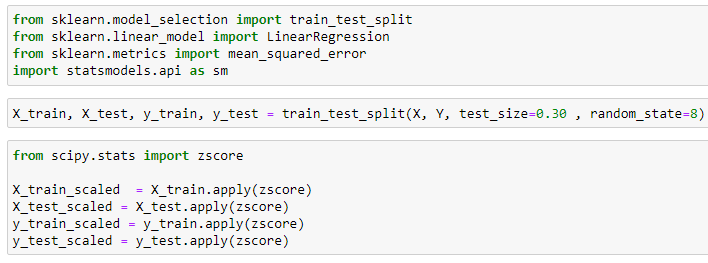






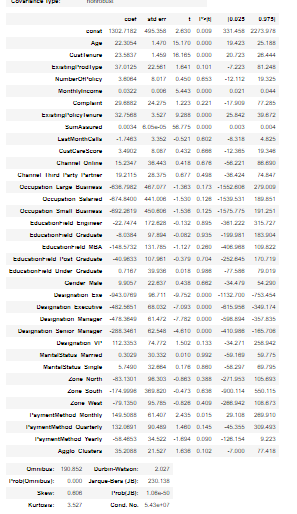
Here, we see that the value of VIF is high for many variables. Here, we may drop variables with VIF more than 5 (very high correlation) & build our model

Now we split the data



**Using Statsmodels OLS**





Notes:

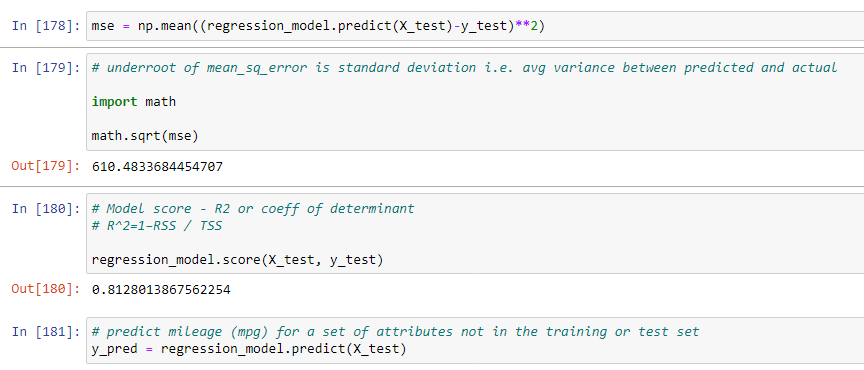
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

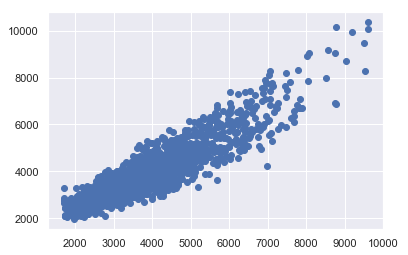
[2] The condition number is large, 5.43e+07. This might indicate that there are

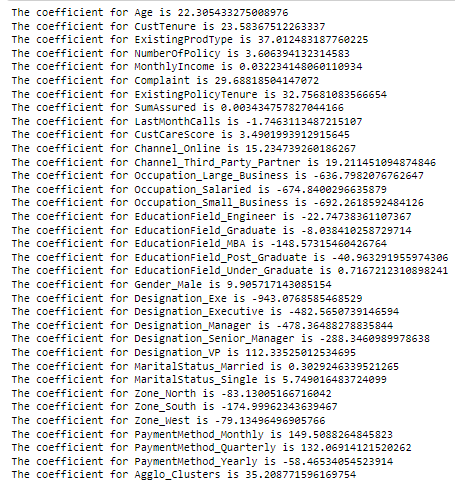
strong multicollinearity or other numerical problems.

Then we create another Regression Model

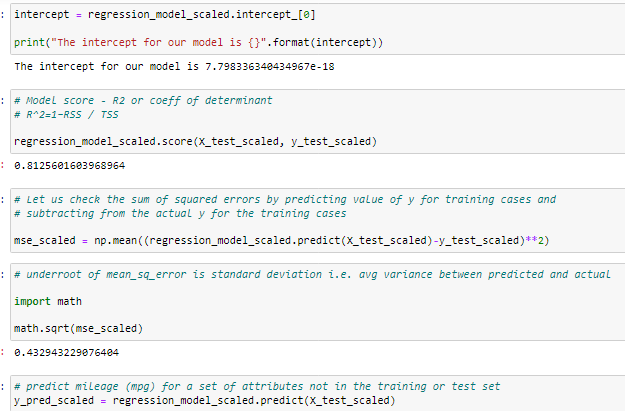


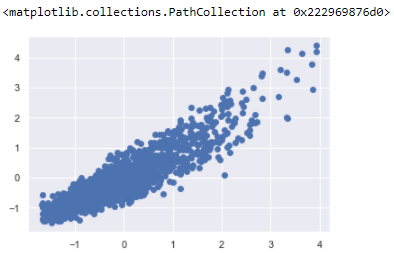








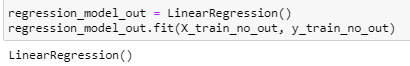




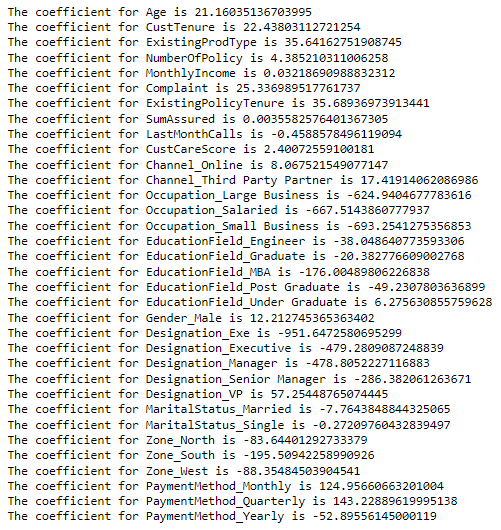
We again check the variance inflation factor

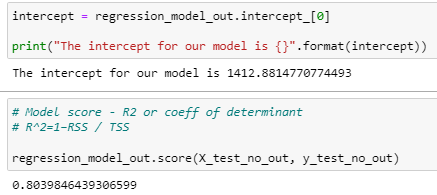


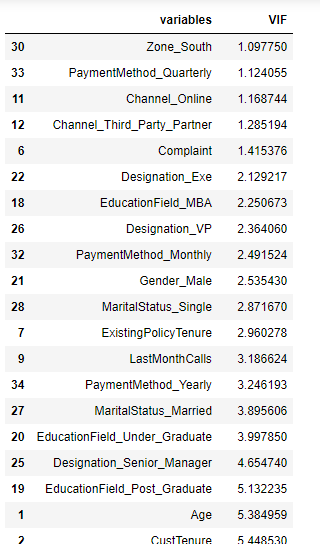
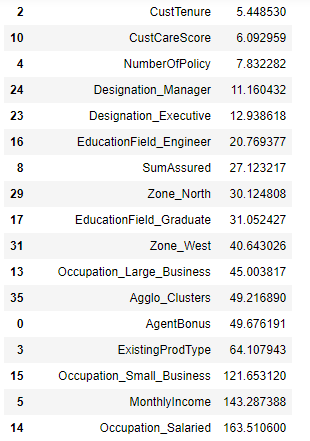
Let’s invoke the LinearRegression function and find the best fit model on training data.



Now let us explore the coefficients for each of the independent attributes





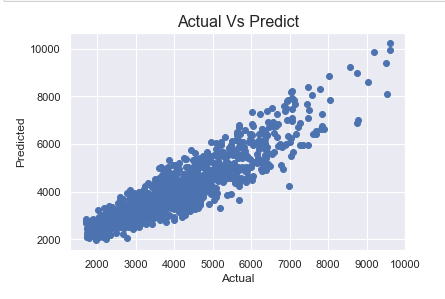
The final Linear Regression equation is

AgentBonus = b0 \* Intercept + b1 \* Zone\_South + b2 \* PaymentMethod\_Quarterly + b3 \* Channel\_Online + b4 \* Channel\_Third\_Party\_Partner + b5 \* Complaint + b6 \* Designation\_Exe + b7 \* Designation\_VP + b8 \* EducationField\_MBA + b9 \* PaymentMethod\_Monthly + b10 \* Gender\_Male + b11 \* MaritalStatus\_Single + b12 \* ExistingPolicyTenure + b13 \* LastMonthCalls + b14 \* PaymentMethod\_Yearly + b15 \* Designation\_Senior\_Manager + b16 \* MaritalStatus\_Married + b17 \* EducationField\_Under\_Graduate + b18 \* EducationField\_Post\_Graduate + b19 \* Age + b20 \* CustTenure + b21 \* CustCareScore + b22 \* NumberOfPolicy + b23 \* Designation\_Manager + b24 \* Designation\_Executive + b25 \* EducationField\_Engineer + b26 \* SumAssured + b27 \* Zone\_North + b28 \* EducationField\_Graduate + b29 \* Occupation\_Large\_Business + b30 \* Zone\_West + b31 \* ExistingProdType + b32 \* MonthlyIncome + b33 \* Occupation\_Small\_Business

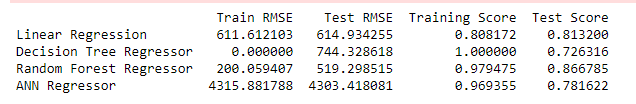
AgentBonus = (843.69) \* Intercept + (-0.74) \* Zone\_South + (121.99) \* PaymentMethod\_Quarterly + (16.45) \* Channel\_Online + (8.64) \* Channel\_Third\_Party\_Partner + (31.43) \* Complaint + (-897.0) \* Designation\_Exe + (19.69) \* Designation\_VP + (-134.73) \* EducationField\_MBA + (179.77) \* PaymentMethod\_Monthly + (3.74) \* Gender\_Male + (13.12) \* MaritalStatus\_Single + (35.55) \* ExistingPolicyTenure + (-2.62) \* LastMonthCalls + (-71.13) \* PaymentMethod\_Yearly + (-261.65) \* Designation\_Senior\_Manager + (-16.47) \* MaritalStatus\_Married + (15.54) \* EducationField\_Under\_Graduate + (-76.12) \* EducationField\_Post\_Graduate + (23.04) \* Age + (24.0) \* CustTenure + (6.01) \* CustCareScore + (1.52) \* NumberOfPolicy + (-447.69) \* Designation\_Manager + (-469.53) \* Designation\_Executive + (-39.86) \* EducationField\_Engineer + (0.0) \* SumAssured + (-44.05) \* Zone\_North + (-63.73) \* EducationField\_Graduate + (-39.31) \* Occupation\_Large\_Business + (-44.97) \* Zone\_West + (54.11) \* ExistingProdType + (0.03) \* MonthlyIncome + (-92.01) \* Occupation\_Small\_Business

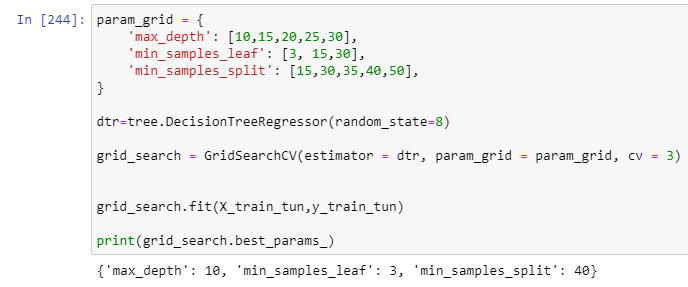
When Age increases by 1 unit, AgentBonus increases by 23.04 units, keeping all other predictors constant. similarly, when MonthlyIncome increases by 1 unit, AgentBonus increases by 0.03 units, keeping all other predictors constant.

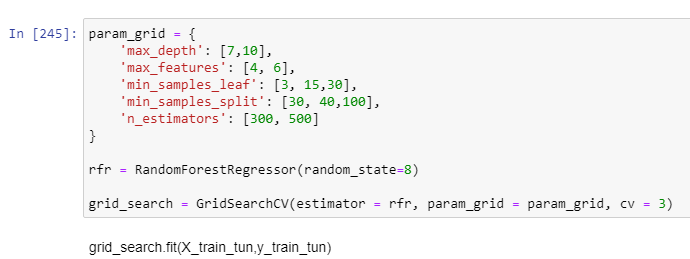
There are also some negative co-efficient values. Occupation\_Large\_Business has its corresponding co-efficient as -39.31. This implies, when the Occupation is Large business, the AgentBonus decreases by 39.31 units, keeping all other predictors constant.



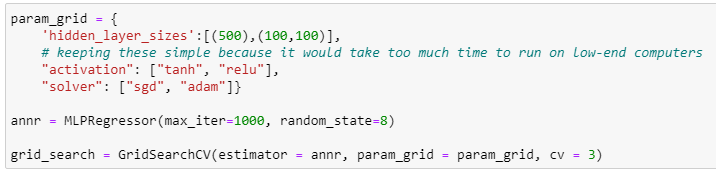
We are scaling the data for ANN. Without scaling it will give very poor results. Computations becomes easier.

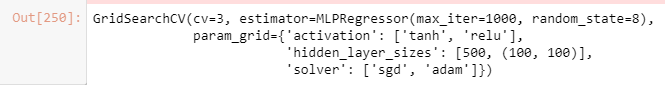


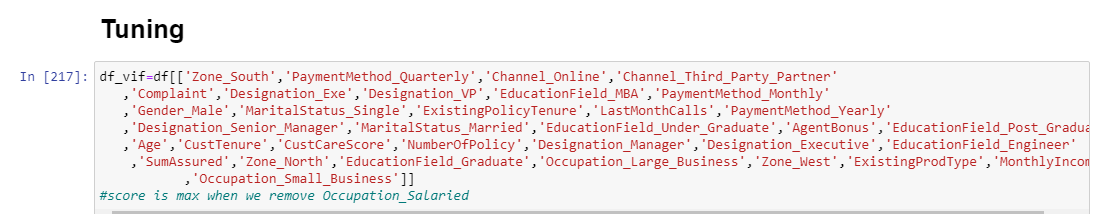


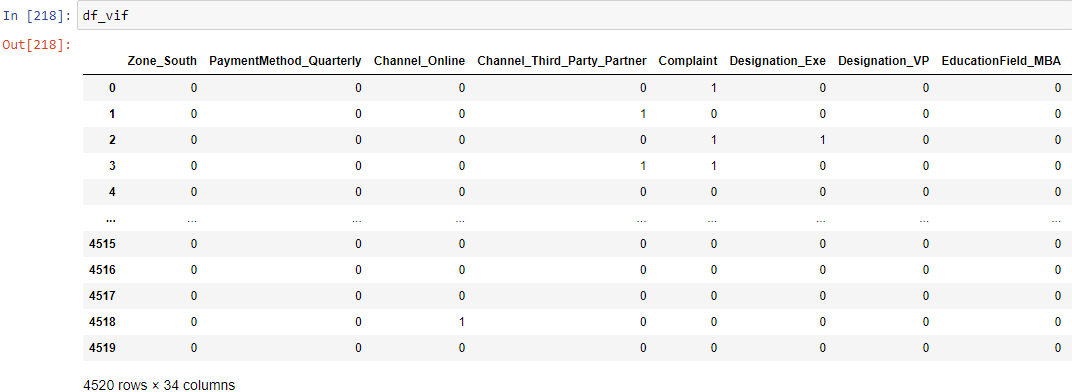




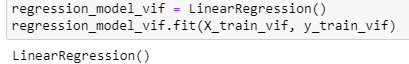




****

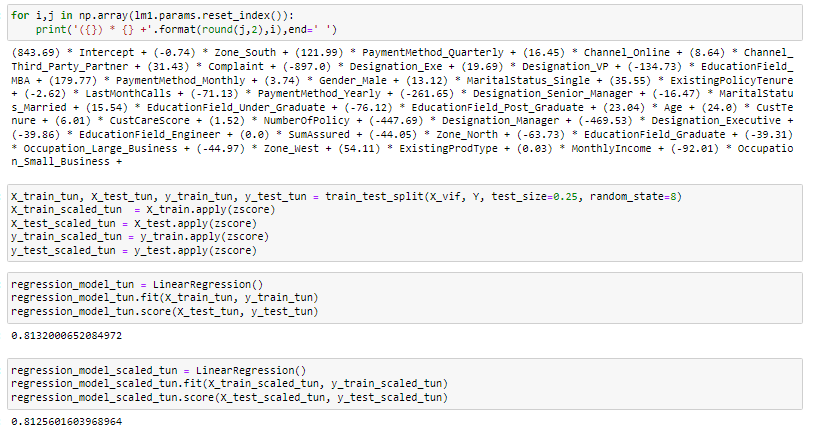
****

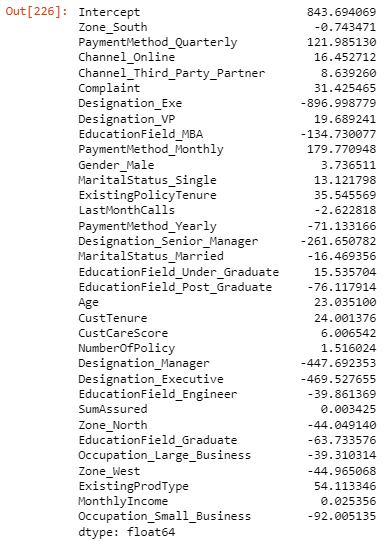
We now invoke the LinearRegression function and find the best fit model on training data



Let us explore the coefficients for each of the independent attributes







lm1.summary

