# Capstone Presentation

**RAMESH PANDEY**
**PGP-DSBA Online**
**Date: 28/07/2023**

# Business Problem Understanding

The life insurance sector in India is a significant part of the country's financial industry and plays a crucial role in the overall economy.
The number of service providers are approx 24 and the Premium collected in FY20 is approx 7.3 trillion rupees.

India's life insurance market is one of the largest in the world in terms of the number of policies sold. So it’s imperative to focus on this industry and make it grow. To make it grow we need more agents. And for more agents performing well 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.
The bonus for agents is a crucial factor in motivating and incentivizing them to perform better, which directly impacts the company's overall productivity and revenue.

# Business Problem Understanding

Constraints:

Data Availability: The success of the prediction relies on the availability and quality of data related to agent performance, customer interactions, demographics, and other relevant attributes. So, the data could be more in numbers and could be more diverse and balanced.

Model Interpretability: Understanding the factors influencing the predictions can aid in justifying bonus distributions. So, it was little difficult to judge which are important factors.

# Business Problem Understanding

Scope:

Agent Bonus Prediction: The primary focus of the project is to build a predictive model that accurately estimates the bonus amount for each agent based on historical data. The model's predictions will help the company make informed decisions on bonus allocation.

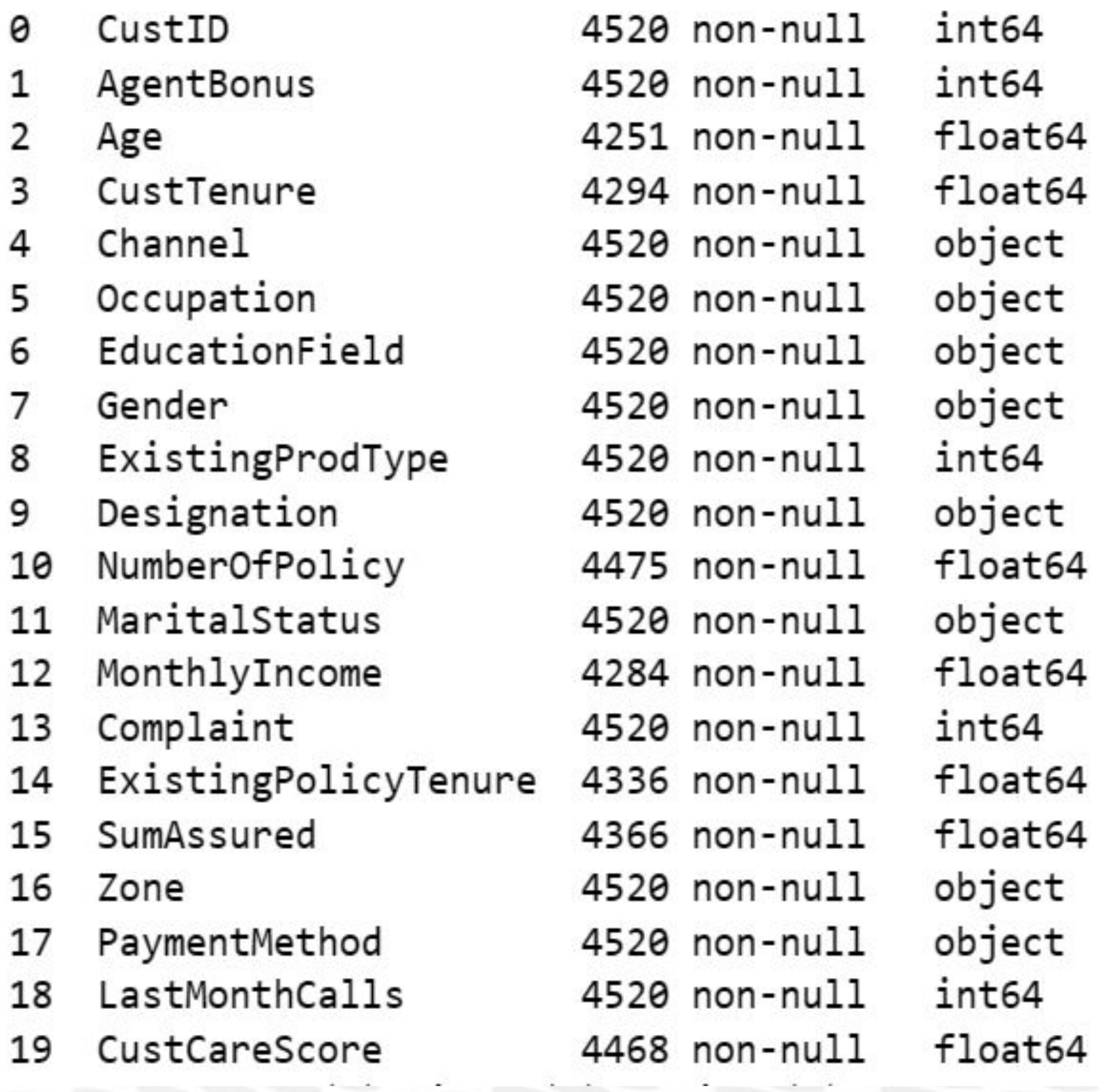
Engagement and Upskill Programs: The predicted bonus amounts will be used to tailor engagement activities and upskill programs for agents. High-performing agents may be offered additional incentives or recognition, while low-performing agents may receive training and support to improve their performance.

# Business Problem Understanding

Objectives:

1. Enhance Agent Performance
2. Improve Business Efficiency
3. Customer Satisfaction
4. Competitive Advantage

# Modelling Approach Used & Why



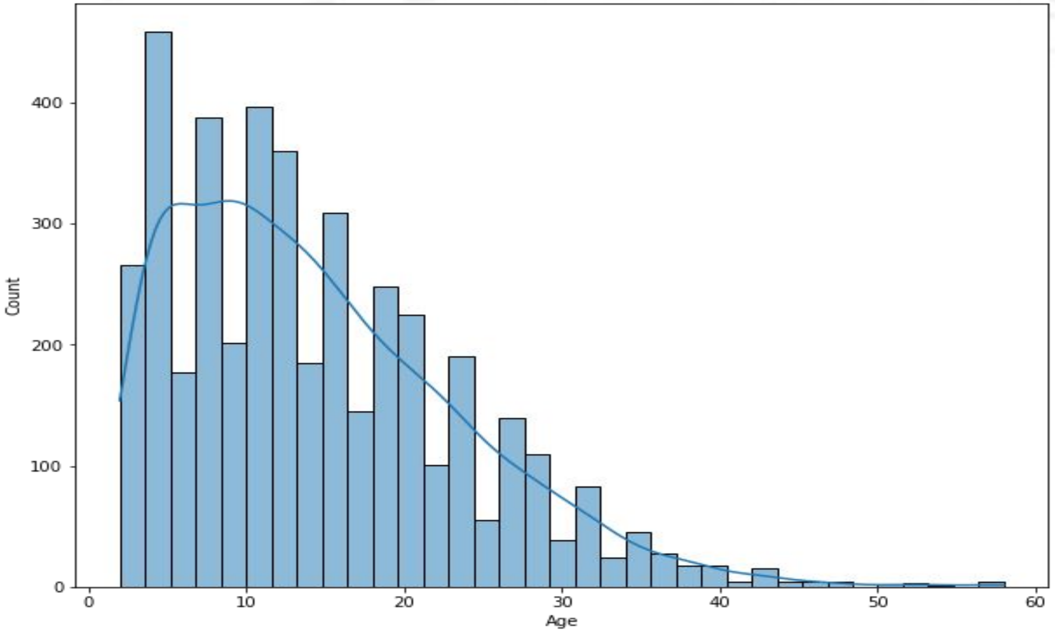
```
0   CustID               4520 non-null   int64
1   AgentBonus           4520 non-null   int64
2   Age                  4251 non-null   float64
3   CustTenure           4294 non-null   float64
4   Channel              4520 non-null   object
5   Occupation           4520 non-null   object
6   EducationField       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
```

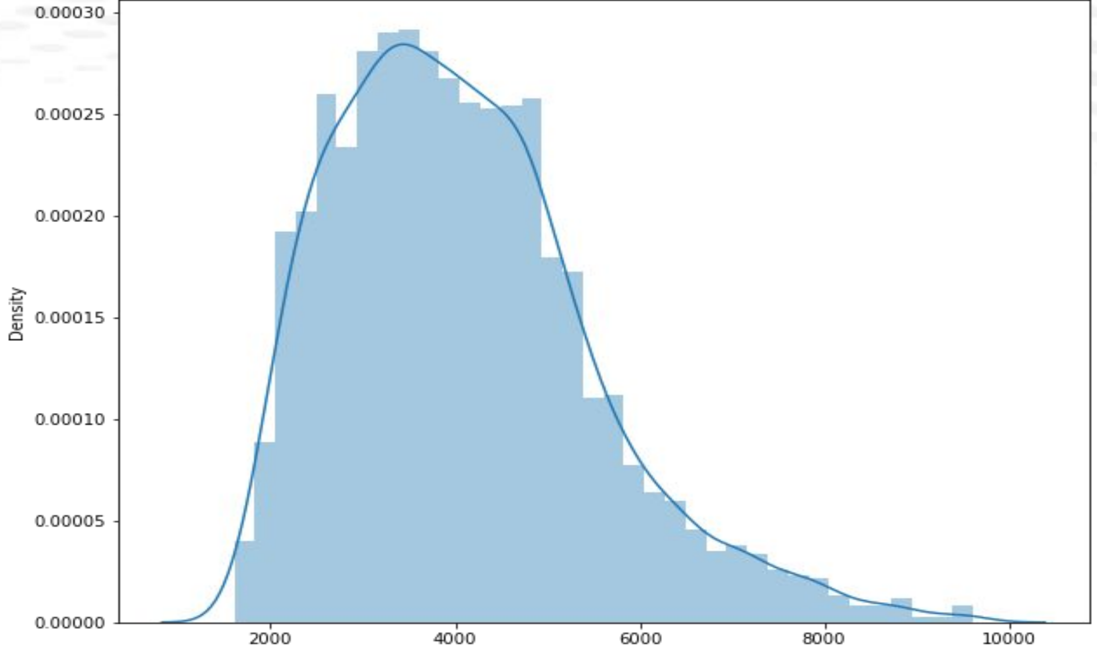
From the data perspective, I can see that there are many missing data and some data don’t have correct data type like CustCareScore. So, these needs to be rectified.

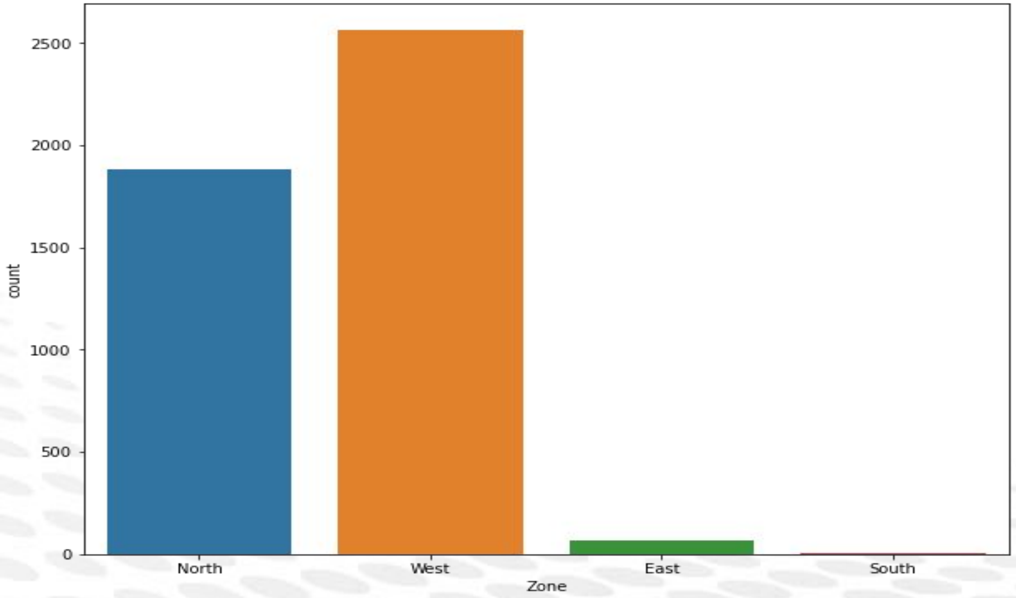
Also using boxplots I can see that there are outliers in most of the features.

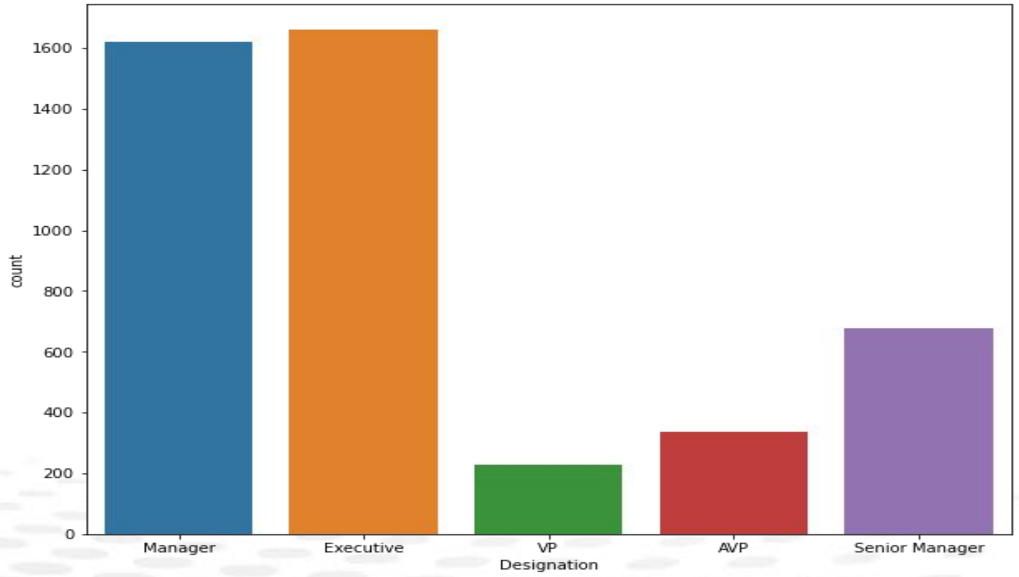
I have treated outliers by replacing lower side outliers to minimum value (which are mostly not present in the data) and upper side outliers to 1.5 times IQR which is equal to 75 percentile minus 25 percentile.

# Modelling Approach Used & Why









# Modelling Approach Used & Why

I have used median imputation for numerical column as they had outliers in them and mode imputation for categorical columns.

The mean age of customer is around 15 years which means that the insurance is mainly taken by parents for their kids.
The median is 13 which is slightly deviated from the mean.
customer tenure is approx 14.5 years that means customers tends to stay for a longer period with the company.
But there are customers who have been associated for just 2 years and maximum is 57 years.

There is skewness in the data and it can affect the accuracy and interpretability of the model and can lead to biased results.

# Modelling Approach Used & Why

Yes the data seem to be imbalance.

If we see the column of Zone, the data consists most of the customers and agents from West and North and very less from South and East, so it will be difficult to predict about these zones.

Similarly data is imbalanced as per the gender as more data is about male and less about female.

Similarly for Occupation, we don’t have much data about freelancers, so we can’t comment whether freelancers don’t purchase the insurance or we are not targeting them.

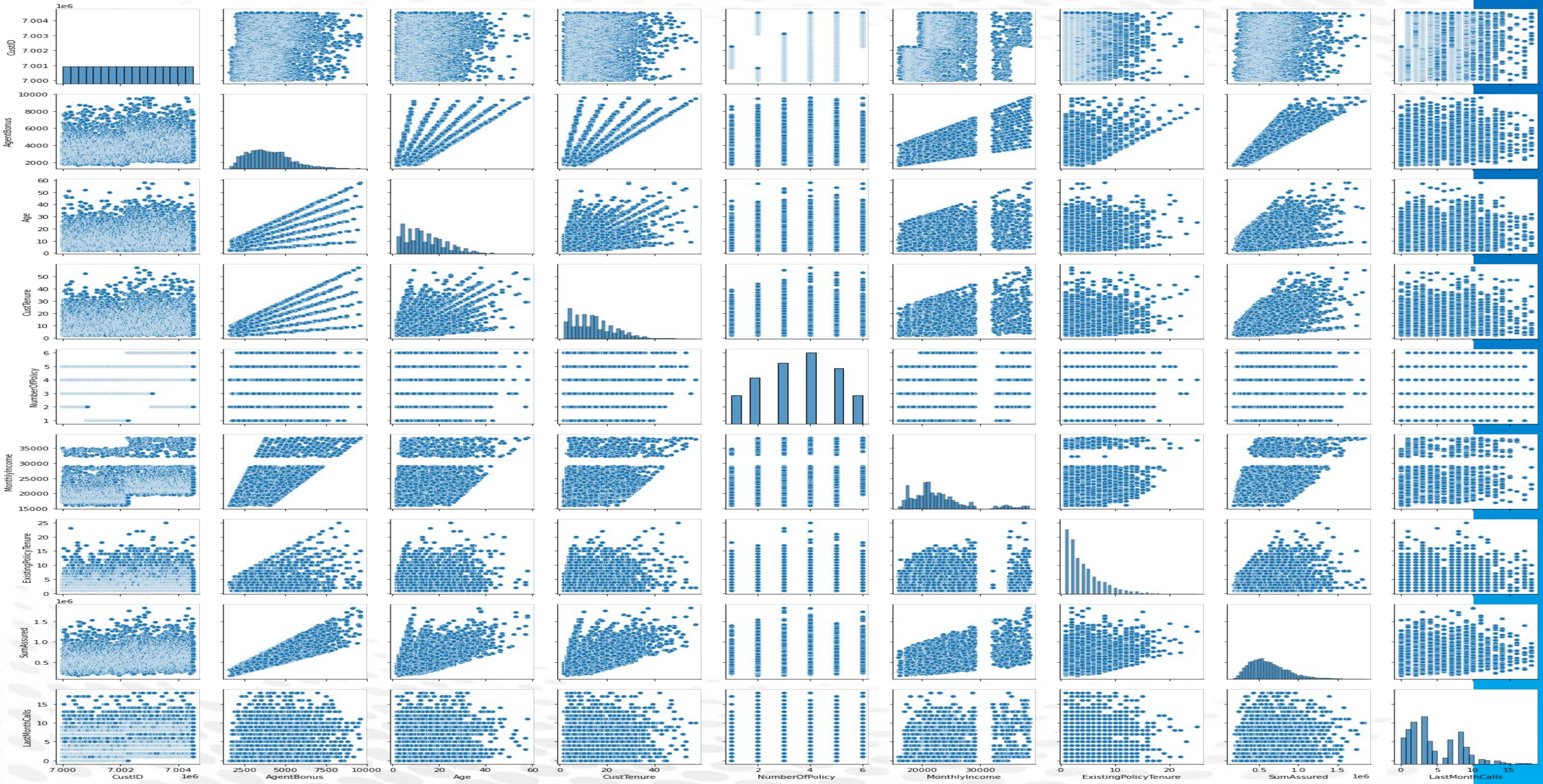
# Modelling Approach Used & Why

Hence, its difficult to predict the bonuses on these features whether we should give more bonus to North and West zone agents or to agents having male customer or female customer.

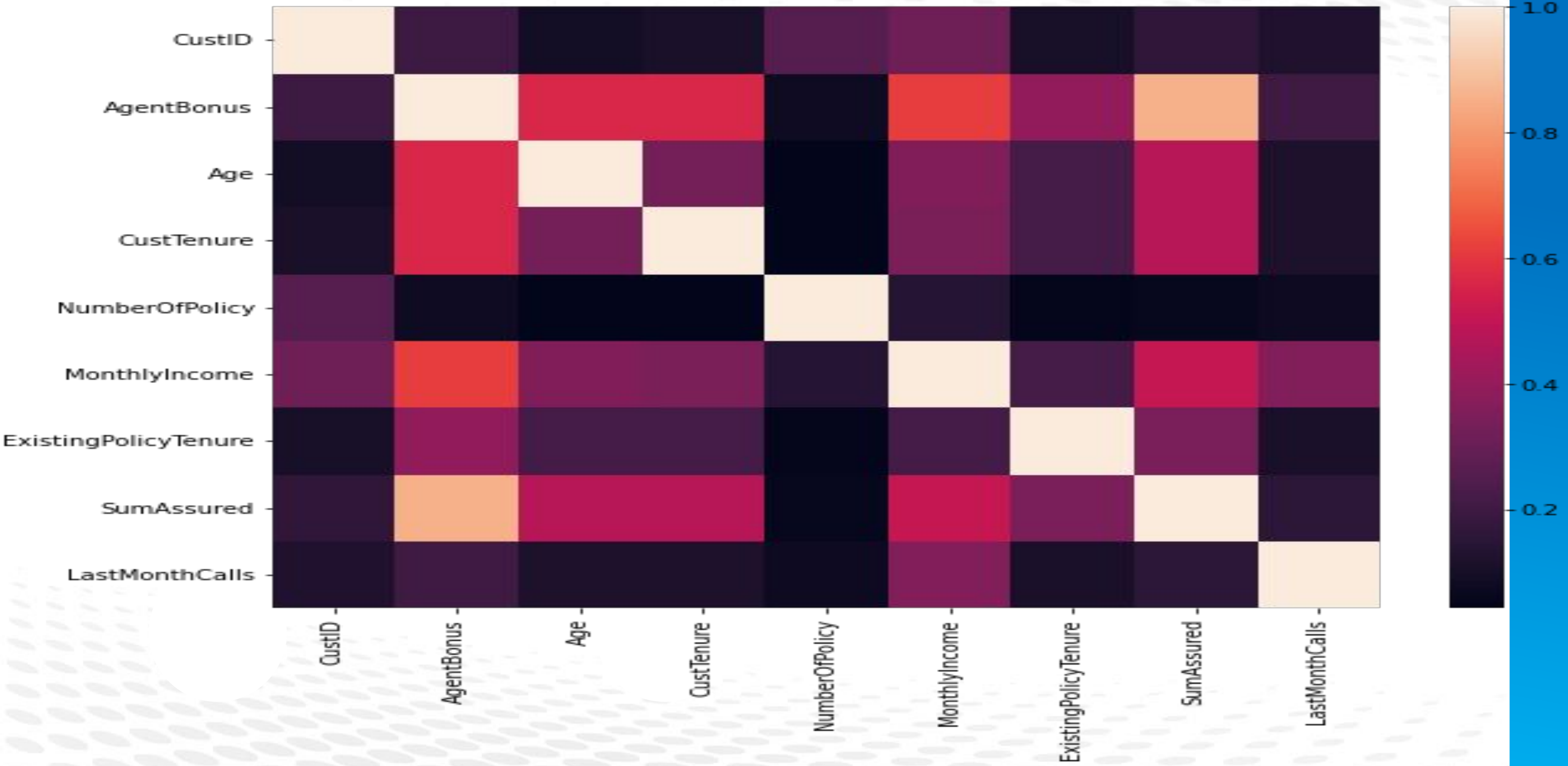
The company can take steps to ensure fairness and motivation among agents. This could include adjusting the bonus allocation methodology, setting clear performance metrics, or implementing reward systems that consider individual agent performance relative to their peers.

We can create artificial data using SMOTE so as to create balanced data, but creating data for all columns may lead to bias in the data. Hence, we need to proceed with given data to do the analysis.

# Modelling Approach Used & Why

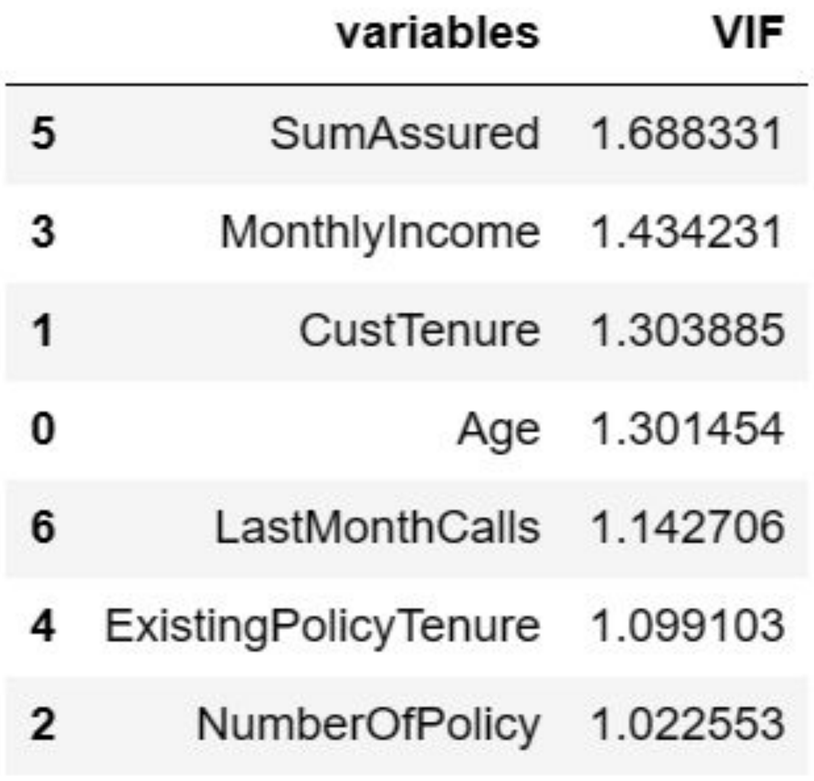


# Modelling Approach Used & Why

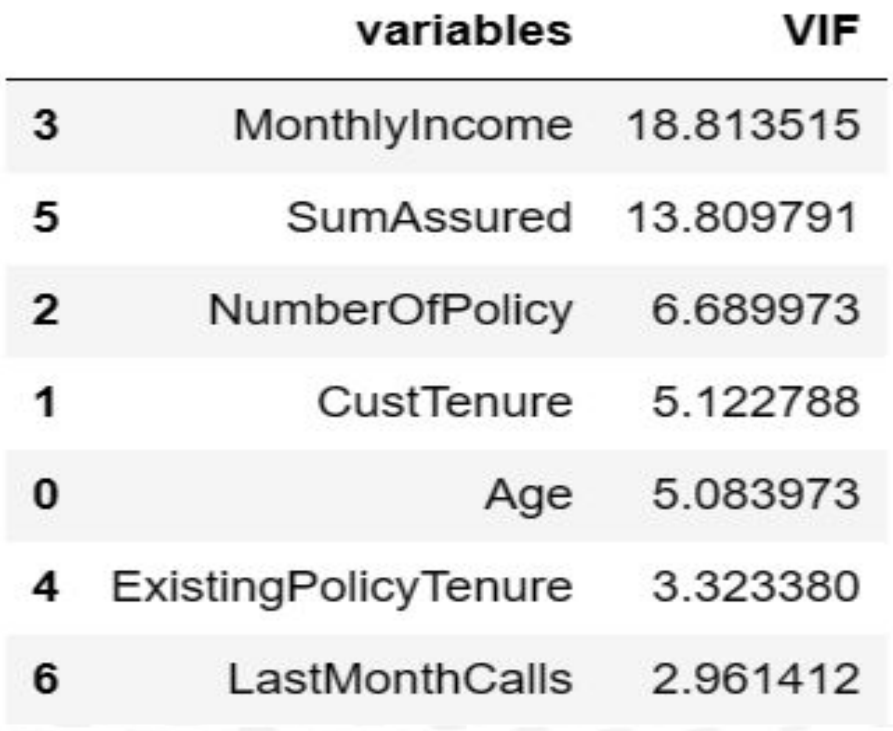


# Modelling Approach Used & Why

We can’t remove any feature as without scaling I am getting VIF score for features greater than 5 which are correlated with our target variable.



|  | variables | VIF |
|---|---|---|
| 5 | SumAssured | 1.688331 |
| 3 | MonthlyIncome | 1.434231 |
| 1 | CustTenure | 1.303885 |
| 0 | Age | 1.301454 |
| 6 | LastMonthCalls | 1.142706 |
| 4 | ExistingPolicyTenure | 1.099103 |
| 2 | NumberOfPolicy | 1.022553 |



|  | variables | VIF |
|---|---|---|
| 3 | MonthlyIncome | 18.813515 |
| 5 | SumAssured | 13.809791 |
| 2 | NumberOfPolicy | 6.689973 |
| 1 | CustTenure | 5.122788 |
| 0 | Age | 5.083973 |
| 4 | ExistingPolicyTenure | 3.323380 |
| 6 | LastMonthCalls | 2.961412 |

# Modelling Approach Used & Why

I have built various models to predict the outcomes. I have used regressor model as the output is continuous in nature.

To build the models, first I have split the data into Train and Test set with the train size as 0.7 of the total data and Test data as 0.3 of total data.

Shape of the data are as:

Input Train - (3164, 51)

Input Test - (1356, 51)

Output Train - (3164,)

Output Test - (1356,)

# Modelling Approach Used & Why

The first model used is linear regression model as it is simple and widely used for numerical target. Since, I didn’t get a better accuracy, so I tried other models.

The next model is  Decision Tree model as it can capture complex interactions between features, which might not be captured well by linear regression. The accuracy of the Decision tree model is not that much, hence I tried other model for better accuracy, like Ensemble models.

I used ensemble learning method as it combines multiple decision trees to improve accuracy and reduce overfitting.

The next model built is Random Forest model. The model has given better accuracy than Decision Tree.

# Modelling Approach Used & Why

To further increase the accuracy I tried Gradient Boost as the model.
This is the best model as it has given the highest accuracy among the non tuned models.
I have tried to tune all the models.
Linear model was tuned using L1 and L2 regularizations, i.e., Lasso and Ridge regularisation models but the accuracy was approx same as that of Linear model.
The best tuned Decision tree model is the model_dt_6 model with max_depth = 7.
The best tuned Decision tree model is the model_rf5 model with n_estimators=100, max_depth = 6 and min_sample_split =5.
Till now best tuned Gradient Boost model is the best_model1 model with ‘learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 100.

# Insights from Analysis

## Before Tuning:

| | Linear | DT | RF | GB |
|---|---|---|---|---|
| Train | 0.8021 | 1.0 | 0.9788 | 0.8727 |
| Test | 0.7986 | 0.7175 | 0.8563 | 0.8491 |
| R2 | 0.7986 | 0.7175 | 0.8563 | 0.8491 |
| MSE | 407518.4029 | 571543..9624 | 290651.9734 | 305311.5037 |

## After Tuning:

| | Ridge | Lasso | DT | RF | GB |
|---|---|---|---|---|---|
| Train | 0.8021 | 0.8021 | 0.8568 | 0.8588 | 0.9288 |
| Test | 0.7986 | 0.7987 | 0.8161 | 0.8350 | 0.8603 |
| R2 | 0.7986 | 0.7987 | 0.8161 | 0.8350 | 0.8603 |
| MSE | 407507.8720 | 407256.6183 | 371980.1317 | 333686.0254 | 282701.9276 |

# Insights from Analysis

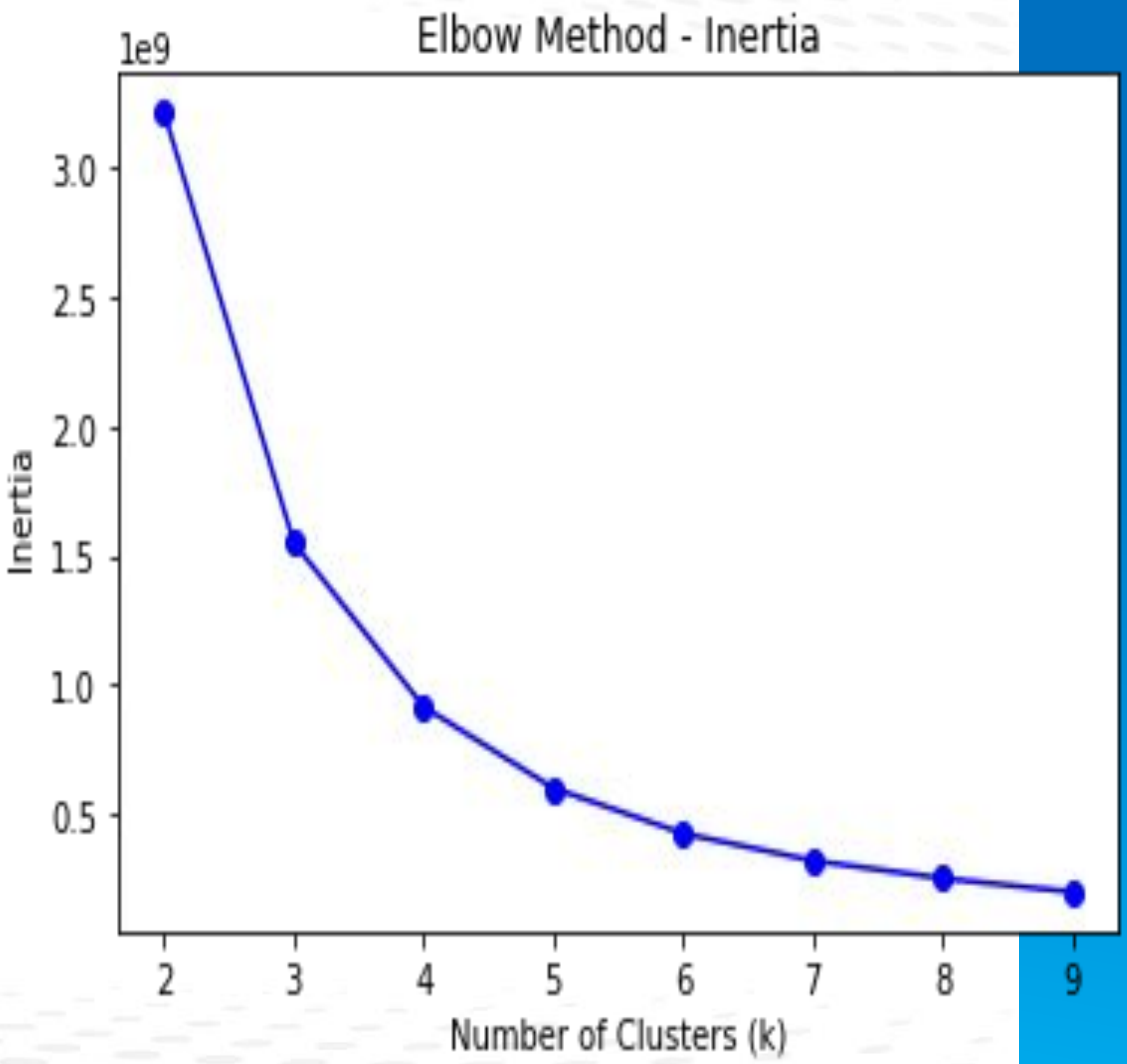
The best model is the Grid Search CV with Gradient Boost model.
The best parameters for the model are:
Till now best tuned Gradient Boost model is the best_model1 model with ‘learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 100.
The error value is 282701.9276 and the train and test accuracy are 0.9228 and  0.8603 respectively.
The top 5 most important feature for this model is as follows:
SumAssured with score of 0.7639430591559299
MonthlyIncome with score of 0.07108005105367851
Age with score of 0.06849981811994652
CustTenure with score of 0.05981710542023578
ExistingPolicyTenure with score of 0.008937615539527446.

# Insights from Analysis

The clustering has been done using k means and to find the number of clusters to be formed, I have used Elbow method and divided data into 4 parts as

| | |
|---|---|
| 0 | 1547 |
| 2 | 1378 |
| 3 | 1211 |
| 1 | 384 |

Cluster 1 has the highest average age which is approx 23 years with the highest average agent bonus which is 7147 and the highest average customer tenure which is approx 23 years.



# Insights from Analysis

Cluster 2 has the least average age which is approx 9 years with the least average agent bonus which is 2603 and the least average customer tenure which is approx 9 years.

Cluster 3 has the average age which is approx 18 years with the average agent bonus which is 5103 and the average customer tenure which is approx 18 years.

Cluster 0 has the average age which is approx 13 years with the average agent bonus which is 3826 and the average customer tenure which is approx 13 years.

So the cluster have been divided based on age where 2 clusters 3 and 1 have similar age of customer and its tenure but average agent bonus have a difference of approx 400. This might be due to the Designation of customers. Cluster 3 has more VP and AVP and hence higher bonus for agents.

# Recommendations

Some key insights and corresponding recommendations are as follows:

For high performing agents we can give high bonuses as reward so that they can continue to perform well.

For low performing agents we can keep some upskill programs where we can give ideas and relevant knowledge and target people so that they can also perform well and earn good bonuses.

Agents with high Sum Assured customer gets higher bonuses so the agents should target or convince people to go for high sum assured plans.

# Recommendations

Longer tenure and experience have a positive correlation with higher bonuses. So, the agents who are not performing well can be asked to target customers who purchases plans for long duration.

Existing policy tenure, Monthly income and Age, significantly influence bonuses. So, agents should be trained to target such customers.

Identifying top-performing agents and providing them with additional incentives can lead to improved overall performance.

Offering financial planning and skill development programs to agents with lower incomes and education can help them improve their performance.

# Recommendations

Since the mean age of customers is 15 years, so we can target married people with children to opt for company’s insurance.

We can increases the benefits for customers who remain for longer period of time with the company.

Since, product type 4 is more popular, we can try to sell it to different customers specially the married customers.

We can focus on finding the issue for low customer score and depending on that conduct training for the agents.

# Recommendations

We can target more salaried and small business owners to take up insurance and also focus on why freelancers are not taking insurance or whether we are not targeting them and depending on that focus on onboarding freelancers and large business owner.

We can focus more on VP and AVP by giving some personalized insurance policy, as they might bring more revenue as they might be opting for higher premium.

Focus more on South and East zone and find why there are less customers from here. Get more agents there and get more customers.