x

|  |
| --- |
| **Life Insurance Sale** |
| Capstone Project  By Oindrila Mandal |
| This document presents the insights and analysis of data sets of a leading life insurance company which helps them in recommending appropriate rewards and training program for their agents. |
|  |

[**Introduction** 2](#_Toc84747937)

[Problem Statement 2](#_Toc84747938)

[Constraint 2](#_Toc84747939)

[Scope 2](#_Toc84747940)

[Need of this project 2](#_Toc84747941)

[**Exploratory Data Analysis** 3](#_Toc84747942)

[Important findings from EDA 3](#_Toc84747943)

[Univariate analysis of some important continuous variable 3](#_Toc84747944)

[Univariate analysis of some important categorical variable 5](#_Toc84747945)

[Bivariate and Multivariate analysis 8](#_Toc84747946)

[Correlation Matrix 13](#_Toc84747947)

[**Data Preprocessing and Data Cleaning** 14](#_Toc84747948)

[Missing value and Outlier treatment 14](#_Toc84747949)

[Variable Transformation 15](#_Toc84747950)

[Variable Removed 16](#_Toc84747951)

[**Model Building** 16](#_Toc84747952)

[Data Splitting 16](#_Toc84747953)

[Model Selection 17](#_Toc84747954)

[Linear Regression 18](#_Toc84747955)

[Ridge Regression 19](#_Toc84747956)

[LASSO Regression 20](#_Toc84747957)

[Random Forest Regressor 20](#_Toc84747958)

[ANN Regressor 21](#_Toc84747959)

[Effort to improve model performance 21](#_Toc84747960)

[Random Forest Grid Search 22](#_Toc84747961)

[ANN Grid Search 22](#_Toc84747962)

[**Model validation** 23](#_Toc84747963)

[**Interpretation / Recommendation** 25](#_Toc84747964)

[Implication on the Business 25](#_Toc84747965)

[Recommendation 25](#_Toc84747966)

# **Introduction**

While we are blessed with the gift of speed in this fast-paced life, our health is lagging far behind in the race. In our modern lifestyle Life insurance policies become a part of our life. Insurance plans are beneficial to anyone looking to protect their family, assets/property and themselves from financial risk/losses.

Life insurance agents represents the company (policy provider) to the customer (policy holder). Their duties include attracting new clients and answering their questions, helping clients choose the best policy and making sure all legal requirements are met and work with clients and beneficiaries to process insurance claims promptly.

## Problem Statement

A leading life insurance company wants to predict appropriate Agent bonus for their Agent based on their 5420 customer records and design engagement programs targeted at its high performing Agents and Upskill program for its low performing agents.

The data set comprising of customer details, policy details and customer satisfaction score for their agents. Target variable Agent Bonus need to be predicted based on 18 predictors. Agent Bonus has high spread with minimum value 1605 and maximum value 9608.

## Constraint

The available data are customer centric information whereas agent related information is missing for prediction of Agent bonus. High dispersion of target column.

## Scope

Analyze 18 different factors influencing agent’s performance from the historical customer data, Model building, predict Agent Bonus and derived insights.

## Need of this project

Identify high performing agent and low performing agent and help the company to design proper activity program and award for high performing agent and proper training program for their low performing agent instead of letting them go.

This will help agent to be motivated and get confidence which will translate in same way when they will execute and delivery different programs as well as to their customer while they will try to sale different policies. This can increase sales of policies.

They can refer other people to join the company as agent which will help to create job opportunity for many people. All these will benefit company revenue, marketing and advertisement of the company, customer satisfaction and agent delight.

# **Exploratory Data Analysis**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

## Important findings from EDA

* Spotted Null variables
* Identify variables having outliers.
* Large dispersion in Target variable
* Data set comprises of continuous, categorical and nominal variable.
* Spotted messy columns name
* Identify unwanted column.

## Univariate analysis of some important continuous variable

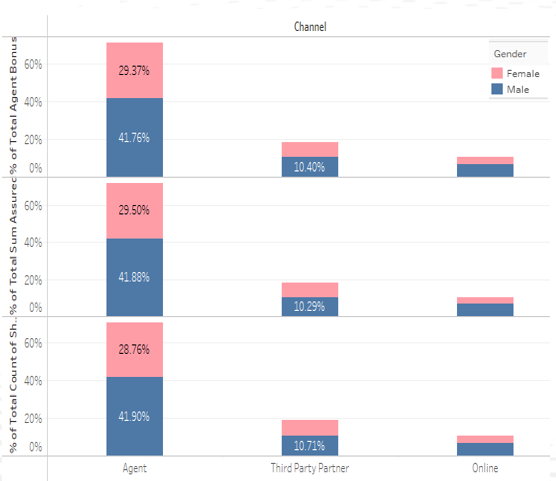
|  |  |
| --- | --- |
|  | AgentBonus is continuous variable. It is right skewed i.e. the mean is larger than median value.  Outliers are there for Agent Bonus. |
|  | Monthly Income is continuous variable. It is right skewed i.e. the mean is larger than median value.  Outliers are there for Monthly Income |
|  | Sum assured is continuous variable. It is right skewed i.e. the mean is larger than median value.  Outliers are there for Sum Assured. |
|  |  |
|  | Customer age is continuous variable. It is right skewed i.e. the mean is larger than median value.  Outliers are there for Customer Age. |

## Univariate analysis of some important categorical variable

|  |  |
| --- | --- |
|  | There are three channels: Agent, Third Party Partner and online  Agent Channel has highest count  Online channel has lowest count |
|  | There are five type of designation holder customers:  Executive has the highest count.  Manager has the second highest count.  VP has the lowest count |
|  | There are four types of payment method.  Half yearly payment method has the highest count  Quarterly has the lowest count. |
|  | There are four types of occupation.  Salaried has the highest count.  Small Business has the second highest count.  Free lancer has the negligible count. |
|  | There are four types of marital status customer.  Married customer has the highest count.  Unmarried has the lowest count. |
|  | There are four type of education field.  Graduate customer has the highest count.  Under graduate has the second highest count.  MBA has the lowest count. |

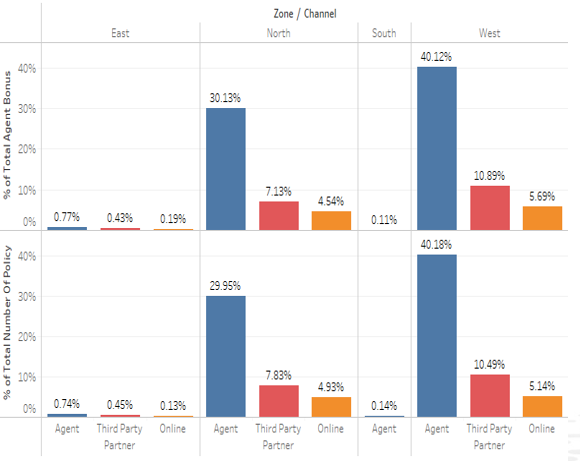
## Bivariate and Multivariate analysis

### Relation among Total agent Bonus, Total Sum assured, Total customer, Channel and Gender:



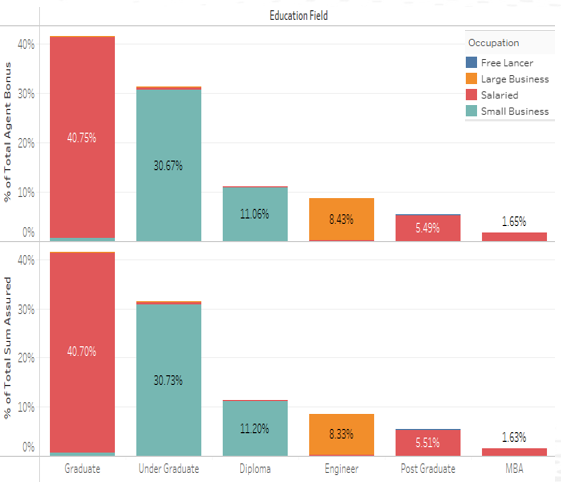
* **70%** of policies are sold through Agent
* **60%** of customers are Male, **40%** customers are Female
* Total amount of Agent Bonus is highest for Agent.
* Maximum total Sum Assured policies sold through Agent Channel
* A smaller number of policies sold through Online Channel and Third-party partner

### Relation among Total agent Bonus, Total Number of policies, Zone, and Channel:



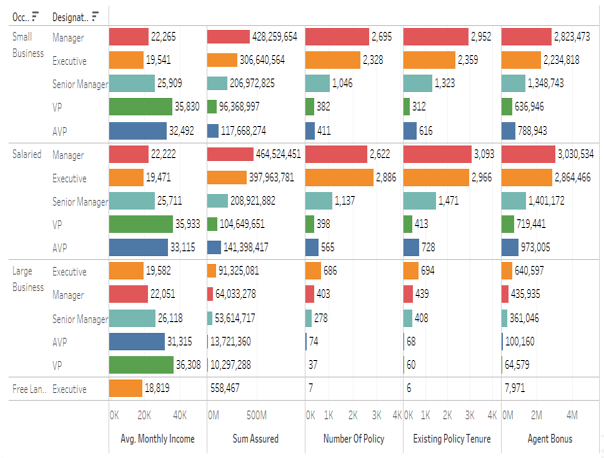
* **56%** policies sold in West zone
* **43%** policies sold in North zone
* In west zone total Agent bonus is highest
* Only **0.14%** policies sold in South zone.
* Policies are sold in South zone only through Agents
* Number of policies and the total Agent bonus are negligible in South and East Zone.

### Relation among Total agent Bonus, Total Sum assured, Education Field, Occupation:



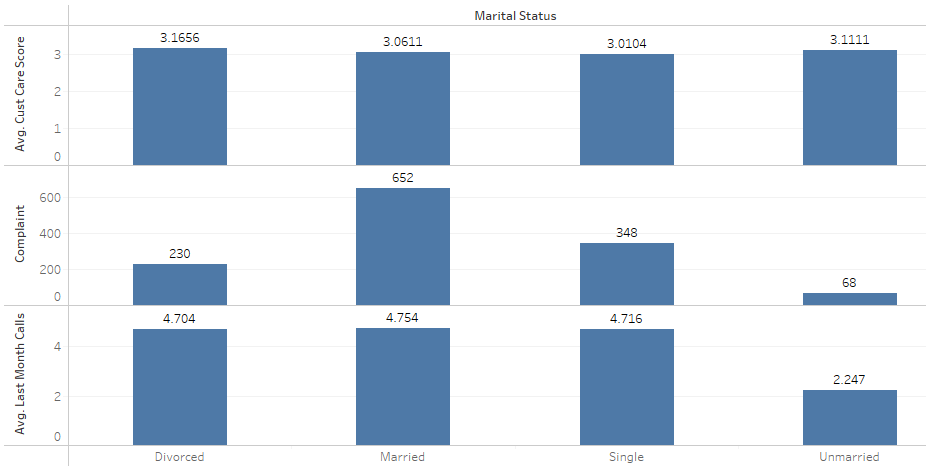
* **40%** of Agent Bonus is for Graduate Customer.
* **30%** of Agent Bonus is for Under Graduate customer
* Total amount of sum assured value is highest for Graduate customer.
* Agent having Salaried people as their customer getting highest bonus.
* Engineer, Post Graduate and MBA Customers have less amount of sum assured policies.
* Agent Bonus is less for the customer who have done PG degree.

### Relation among Total agent Bonus, Total Number of policies, Zone, and Channel:



* Manager and Executive of Salaried /Small Business occupation are Prime Customer
* Maximum number of policies are sold to Salaried/Small Business Manager and Executive customer
* Agent Bonus and Total Sum assured amount are highest for them.
* A smaller number of policies are sold to AVP and VP customer.
* Agent Bonus is lowest for AVP and VP customer.

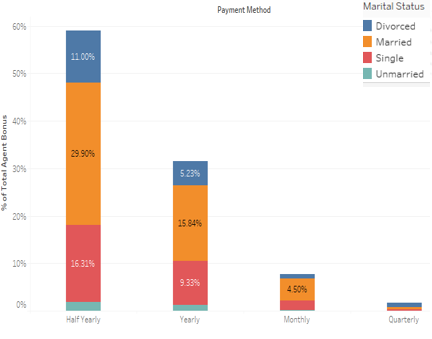
### Relation among Customer care score, Complaint:



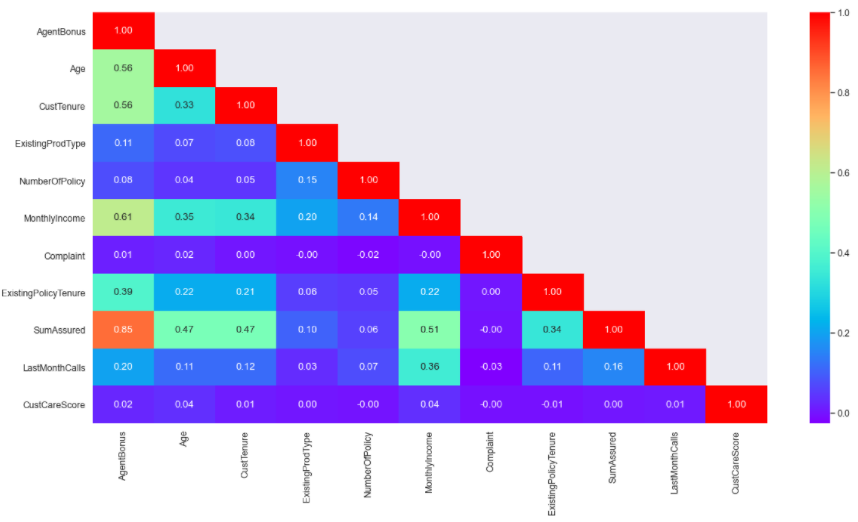
* Average Customer satisfaction score is almost same for all type of customers.
* Married customer raised highest number of complaints.
* Average number of last month’s calls for cross sell is same for Divorced, Married and Single customer.

### Relation among Agent Bonus, Payment Method and Marital Status:

* 51% of Agent Bonus is for Married Customer
* Maximum customer opts for Half yearly payment



## Correlation Matrix



* Agent Bonus is highly correlated with Sum Assured.
* Agent Bonus has moderate correlation with Monthly Income, Customer Tenure and Customer Age. That means there is some relationship between them but there are some other variables affecting them. So, the direct relationship is not strong but it is certainly noticeable.
* Agent Bonus has very weak correlation with Customer satisfaction score and the number of Complaints.
* Sum assured amount has moderate correlation with Customer Age and Customer tenure.
* We cannot see any multicollinearity.

# **Data Preprocessing and Data Cleaning**

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Data cleaning is one of the steps of Data Preprocessing. The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

## Missing value and Outlier treatment

There are 1166 entries having null value in this data sets. The presence of missing value in data set will drastically impact the quality of machine learning model while we will train the model with the dataset having missing values. Many machine learning algorithms do not support missing value. So, the handling of missing data is very important during the preprocessing of the dataset.

To handle missing value, we will fill the data with the most probable value.

From Boxplot we can see, there are huge number of outliers present in data set.  Outliers increase the variability in your data, which decreases statistical power. Many [parametric statistics](https://en.wikipedia.org/wiki/Parametric_statistics), like mean, correlations, and every statistic based on these is sensitive to outliers.

Instead of deleting all rows having outliers we will replace outliers with the nearest good data.

### Approach of missing value and outlier treatment

There is no null value in Target column. But Outlier present in Target column. We will not treat outlier value for target column as this will lead to data loss. We will treat Null value and outlier for other columns using KNN Imputer.

Why KNN Imputer?

KNN is an algorithm that is useful for matching a point with its closest k neighbors in a multi-dimensional space. It can be used for data that are continuous, discrete, ordinal and categorical which makes it particularly useful for dealing with all kind of missing data.

Steps for imputation:

* Separate the target column and predictors from data set
* Convert all outlier value into null value
* Scaled all the continuous variable with standard scaler. Scaling is required as KNN is a distance-based algorithm.
* Concatenate scaled predictors and target column
* Impute all the null values with KNN imputer.

## Variable Transformation

There are a number of categorical predictors in our dataset.

Machine learning algorithms and deep learning neural networks require that input and output variables are numbers. This means that categorical data must be encoded to numbers before we can use it to fit and evaluate a model.

* We have encoded all the categorical value with category code of the categorical.
* Codes are an array of integers which are the positions of the actual values in the categories array.

Rename variable:

There are some messy values for some categorical variable ('Laarge Business' in Occupation, 'UG' in EducationField, 'Fe male' in Gender, 'Exe' in Designation), we have replaced these values with correct value.

## Variable Removed

In this analysis ‘CustId’ is not an important variable. This is just a unique Id. We have removed this column from our data set.

# **Model Building**

A machine learning model is built by learning and generalizing from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfil its purpose.

## Data Splitting

### Independent variable- target Variable

The target variable of a dataset is the [feature](https://www.datarobot.com/wiki/feature/) of a dataset about which we want a deeper understanding.

To build a mathematical relationship between Input and Output variable, we separate our independent variable/input variable and Target variable. ‘X’ denotes input feature set where ‘y’ denotes target column.

### Train and Test Split

In the development of machine learning models, it is desirable that the trained model perform well on new, unseen data. In order to replicate the new, unseen data, the available data is subjected to data splitting. The dataset is split to 2 portions (train-test split). Particularly, the first portion is the larger data subset that is used as the training set (here it is 70% of the original data) and the second is normally a smaller subset and used as the testing set (the remaining 30% of the data).

## Model Selection

In our data set Input X, and output y are labelled data. Our Target variable y is continuous variable.

That means our target variable is parameterized by [mean](https://en.wikipedia.org/wiki/Mean) and [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation). We need to choose such models which are based upon parametric statistics.

Hence, we are going to build Supervised Regression Model.

* Linear Regression
* Ridge Regression
* Lasso Regression
* Random Forest Regressor
* Artificial Neural Network Regressor

### **Model Building Steps**

* Separate our independent variable/input variable and Target variable
* Split the data into Train and Test data set in 70:30 ratio
* Training set is used to build model
* Such trained model is applied on the testing set to make predictions.
* Select best model on the basis of the model’s performance on the testing set.
* Hyperparameter optimization also performed.
* Calculate R2, Adjusted R2, RMSE, MAPE for Train and Test data to check model performance.

## Linear Regression

Linear Regression gives us an equation between independent variables and target variable [AgentBonus].

### **Model Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Columns | Coefficient | std err | t | P>|t| |
| Intercept | 4000.7948 | 65.706 | 60.889 | 0 |
| SumAssured | 796.9786 | 14.877 | 53.571 | 0 |
| MonthlyIncome | 326.9745 | 13.719 | 23.833 | 0 |
| PaymentMethod | -10.6123 | 8.339 | -1.273 | 0.203 |
| Age | 175.081 | 13.204 | 13.259 | 0 |
| CustTenure | 182.26 | 13.114 | 13.898 | 0 |
| Channel | -3.5723 | 14.347 | -0.249 | 0.803 |
| Gender | 39.6666 | 23.077 | 1.719 | 0.086 |
| Zone | 6.0594 | 11.208 | 0.541 | 0.589 |
| ExistingPolicyTenure | 106.3021 | 12.18 | 8.728 | 0 |
| MaritalStatus | 1.7845 | 14.876 | 0.12 | 0.905 |
| Occupation | -12.4728 | 20.469 | -0.609 | 0.542 |
| Designation | -37.5265 | 12.883 | -2.913 | 0.004 |
| Complaint | 15.7792 | 25.211 | 0.626 | 0.531 |
| EducationField | 11.6807 | 7.436 | 1.571 | 0.116 |
| LastMonthCalls | -7.2581 | 12.078 | -0.601 | 0.548 |
| CustCareScore | 15.4328 | 8.277 | 1.865 | 0.062 |
| ExistingProdType | -69.9682 | 17.809 | -3.929 | 0 |

From coefficient we get,

When Sum Assured increases by 1 unit, Agent Bonus increases by 796.98 units, keeping all other predictors constant. similarly, when Monthly Income of Customer increases by 1 unit, Agent Bonus increases by 327 units, keeping all other predictors constant.

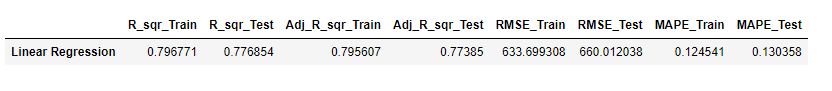
There are also some negative co-efficient values, for instance, Existing Prod type has its corresponding co-efficient as -69.96. This implies, when the customer continues with same policy for long term, the Agent Bonus decreases by 69.96 units, keeping all other predictors constant.

From above p-value and coefficient we can conclude that there are some variables having significant p-value i.e. p-value<0.05 and higher coefficient value. These variables have significant influence on the prediction of Agent bonus.

From Variation Inflation factor, p-value, coefficient value we get our important variable for this analysis. These variables are:

SumAssured, MonthlyIncome, Age, CustTenure

### **Model Evaluation:**



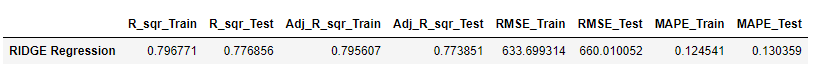
From above result we can conclude that,

* Our Linear Regression Model is able to explain only 79% variability in the response variable around its mean for Train data and can explain only 77% variability for Test data.
* Our model score is not satisfactory.
* Adjusted R square is similar with R squared
* There is not so much difference in RMSE train and test value. RMSE has higher value.
* There is no such difference in MAPE train and Test value.

## Ridge Regression

Here we able to reduce the magnitude of the coefficient of predictors. Some values reach to zero although they are not absolute zero.

### **Model Evaluation:**

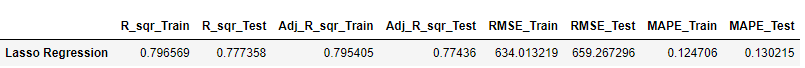


* From above result, we can see there is no such improvement in model score in according to our Linear Regression Model.
* All the evaluation score is quite similar with our previous model, i.e. Linear Regression Model.
* Our Ridge Regression Model is able to explain only 79% variability in the response variable around its mean for Train data and can explain only 77% variability for Test data.
* So, we can say that our model score is not satisfactory.

## LASSO Regression

Here we able to reduce the magnitude of the coefficient of predictors. Some values reduced to absolute zero.

### **Model Evaluation:**

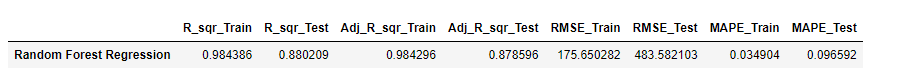


* From above result, we can see there is no such improvement in model score in according to our Linear Regression Model.
* All the evaluation score is quite similar with our previous models, i.e. Linear Regression Model and Ridge Regression Model.
* Our LASSO Regression Model is able to explain only 79% variability in the response variable around its mean for Train data and can explain only 77% variability for Test data.
* So, we can say that our model score is not satisfactory.

## Random Forest Regressor

Random forest is a type of supervised learning algorithm that uses ensemble methods (bagging) to solve both regression and classification problems. This Algorithm Consists of a large number of individual decision trees at training time and outputting the mean/mode of prediction of the individual trees.

### **Model Evaluation:**

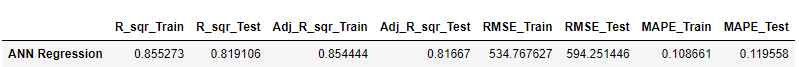
****

* From above result we can conclude that,
* Our Random Forest Regression Model is able to explain 98.4% variability in the response variable around its mean for Train data and can explain 88% variability for Test data.
* Our model is performing well on the training data but is not performing well on the Test data. This model is Overfit.
* Adjusted R square is similar with R squared
* There is significance difference in RMSE train and test value, which is not a good fit.
* There is significant difference in MAPE train and Test value which is also not a good fit.

## ANN Regressor

Regression is method dealing with linear dependencies, neural networks can deal with nonlinearities. So, if data has some nonlinear dependencies, neural networks should perform better than regression.

### **Model Evaluation:**



* From above result we can conclude that,
* Our ANN Regression Model is able to explain 85.5% variability in the response variable around its mean for Train data and can explain 82% variability for Test data.
* Our model score is satisfactory although it has higher RMSE value for Train and Test data.
* Adjusted R square is similar with R squared
* There is slight difference in RMSE train and test value. RMSE has higher value.
* There is no such difference in MAPE train and Test value.

## Effort to improve model performance

After building all model, we can say Random Forest and ANN model may improve their performance if we tune the model.

We have evaluated each model only on the training set which lead to [overfitting](https://elitedatascience.com/overfitting-in-machine-learning) for random Forest Regressor.

Hyperparameter tuning is the best method to determine optimal model as it try many different combinations to evaluate performance of the model.

An overfit model may look impressive on the training set, but will be useless in a real application. Therefore, the standard procedure for hyperparameter optimization accounts for overfitting through [cross validation](http://scikit-learn.org/stable/modules/cross_validation.html).

## Random Forest Grid Search

A model has to match the business objectives hence various permutation and combination has been carried on to refine the model.

Best parameter for this model is:

max\_depth=12, max\_features=9, min\_samples\_leaf=25,

min\_samples\_split=75, n\_estimators=500, random\_state=123

### **Model Evaluation:**



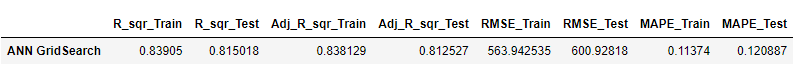
* From above result we can conclude that,
* After hyper tuning using Grid Search Random Forest model is able to explain 88.1% variability in the response variable around its mean for Train data and can explain 85% variability for Test data.
* Our model is performing well on the training data as well as on the Test data. This model is good fit.
* Adjusted R square is similar with R squared
* There is less difference in RMSE train and test value and it has lower value that means predicted value is closer to actual fit which is good fit.
* There is minimum difference in MAPE train and Test value which is a good fit.

## ANN Grid Search

Best parameter for this model is:

'activation': 'relu', 'hidden\_layer\_sizes': 100, 'solver': 'adam'

### **Model Evaluation:**



* From above result we can conclude that,
* After hyper tuning using Grid search ANN model is able to explain 84% variability in the response variable around its mean for Train data and can explain 82% variability for Test data.
* Our model is performing well on the training data as well as on the Test data. This model is also a good fit
* Adjusted R square is similar with R squared
* There is less difference in RMSE train and test value but it has higher value which is not a good fit.
* There is minimum difference in MAPE train and Test value which is a good fit.

# **Model validation**

To evaluate the model, we have calculated

**R Square:** It determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable). Higher value R Squared is better fit.

**Adjusted R square:** The adjusted R-squared is a modified version of R-squared that adjusts for predictors that are not significant in a regression model. Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.

**RMSE:** It is defined as the square root of the average squared error. RMSE is an absolute measure of fit. RMSE can be interpreted as the standard deviation of the unexplained variance. Lower values of RMSE indicate better fit.

**MAPE:** MAPE is the sum of the individual absolute errors divided by the demand (each period separately). It is the average of the percentage errors which expresses accuracy as a percentage of the error. Lower values of MAPE indicate better fit.

### **Interpretation of the most optimum model**

Final comparison of all evaluation score for all model:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **R\_sqr\_Train** | **R\_sqr\_Test** | **Adj\_R\_sqr\_Train** | **Adj\_R\_sqr\_Test** | **RMSE\_Train** | **RMSE\_Test** | **MAPE\_Train** | **MAPE\_Test** |
| Linear Regression | 0.79677 | 0.7769 | 0.795607 | 0.77385 | 633.699 | 660.01 | 0.1245 | 0.13036 |
| Random Forest Regression | 0.98439 | 0.8802 | 0.984296 | 0.878596 | 175.65 | 483.58 | 0.0349 | 0.09659 |
| ANN Regression | 0.85527 | 0.8191 | 0.854444 | 0.81667 | 534.768 | 594.25 | 0.1087 | 0.11956 |
| RIDGE Regression | 0.79677 | 0.7769 | 0.795607 | 0.773851 | 633.699 | 660.01 | 0.1245 | 0.13036 |
| Lasso Regression | 0.79657 | 0.7774 | 0.795405 | 0.77436 | 634.013 | 659.27 | 0.1247 | 0.13022 |
| RandomForest GridSearch | 0.88111 | 0.8546 | 0.880425 | 0.852679 | 484.698 | 532.7 | 0.0992 | 0.10883 |
| ANN GridSearch | 0.83905 | 0.815 | 0.838129 | 0.812527 | 563.943 | 600.93 | 0.1137 | 0.12089 |

From above result we can say,

* Linear Regression, Lasso Regression, Ridge Regression are poor performed model
* Random Forest without tuning and XG Boost without tuning are overfit models.
* ANN without Tuning, Random Forest with tuning and ANN with tuning models are good performing model.
* But among these three model Random Forest with Grid search has the highest R squared or model score value, whereas lowest RMSE value than all other model. That means, predicted value is closer to actual value. Random forest with tuning performing well for both training and test data.
* So, I will say, Random Forest with hyper parameter tuning using Grid search is the optimum model.

# **Interpretation / Recommendation**

## Implication on the Business

* As Random Forest with hyper parameter tuning using Grid search is performing well on train data as well as test data and it has a good accuracy score, the Insurance company can use this model for prediction of Agent Bonus.
* This model can predict agent bonus more accurately which will help company to design appropriate engagement activity for their high performing agents and upskill programs for low performing agents.
* This model will help company to identify high performing agent and low performing agent.

## Recommendation

* Agent is the primary channel, company should take all necessary steps to retain high performing Agent, as they are the key factor for the company Growth.
* To increase the number of Female customers they can introduce special scheme to attract them.
* To attract younger customer, they can introduce long term policy having higher sum assured value with lower premium.
* To increase sale in South and East zone, they should find out the pain area and can also design Upskill program for the agent of these zone to enhance their skill.
* The Company should ensure the satisfaction of the customer, holding Managerial and Executive position in Small Business or in Salaried Occupation.
* They can introduce higher sum assured value with lower premium policy to them as they have average monthly income.
* VP and AVP customers have higher monthly income. They have potential to buy good policy. To increase the sale among them Agent should understand their needs and requirement and suggest scheme accordingly.