# Machine Learning Meets Term Life Insurance: Targeting High-Value Customers

# 1. Introduction

## Overview of the Problem

Term life insurance is promoted through telemarketing by the unique HashSysTech Insurance Corporation. Nonetheless, because the cost of these campaigns is relatively steep, the firm aims at minimizing resource utilizations by only focusing on the most prospective client population. The intention is to isolate it and channel telemarketing efforts toward those customers.

## Objectives of the Analysis

The goal is to generate a model for the improvement in marketing strategies by providing a list of customers likely to buy term life insurance. This model will be useful in Project Green light in targeting customers with better chances of responding to telemarketing. Specifically, the analysis aims to:

1. Analyze customer-data to find correlations with term life insurance.
2. Construct a machine learning model to measure the prospect of conversion from the numerical data obtained.
3. Employ statistical analysis and machine learning to help HashSysTech make its best decisions on operations.

## Dataset Overview

It includes data of customers whom we have tried to sell our products through telemarketing in the past such as their age, gender, contact details and sales history etc. In this case the target variable reflects if a given customer made the purchase of the insurance product which makes it possible to create a model for future conversion.

## Dataset Description

The dataset includes the following components:

1. Customer Demographics: Age, position, marital status and level of education.
2. Contact Details: Work type that involves calls type, day, month, contact duration and total number of calls completed.
3. Campaign History: Prior campaign results (unknown other failure success)
4. Target Variable: Whether the customer subscribed to the insurance of not.

With these features, useful customers for telemarketing will be foreseen, thus making telemarketing efforts for HashSysTech Insurance efficient.

# 2. Preprocessing of collected data and exploration of the data collected.

## 2. 1 Data Exploration

### 2.1.1. Objective:

In this respect, when attempting to navigate and analyze the structure of the data in the given set where a number of features might be either numeric or categorical, it is crucial to initially sort the data by the presence of this characteristic. Analysis of numerical kind of features will be done by calculating mean, mode standard deviation while the categorical nature of features will be analyzed by performing frequency analysis.

### 2.1.2 Descriptive Statistics for Numerical Features

The following numerical features are examined:

1. Age: Growth of the customer.
2. Day: The number which indicates the day of the month when the last communication was carried.
3. Mon: The month in which the last contact was established.
4. Dur: Time elapsed as a result of the last contact in seconds.
5. Num\_calls: Total call made on the customer, during the campaign.

Some of the main parameters which should be calculated to each feature are mean, median, standard deviation, and range.

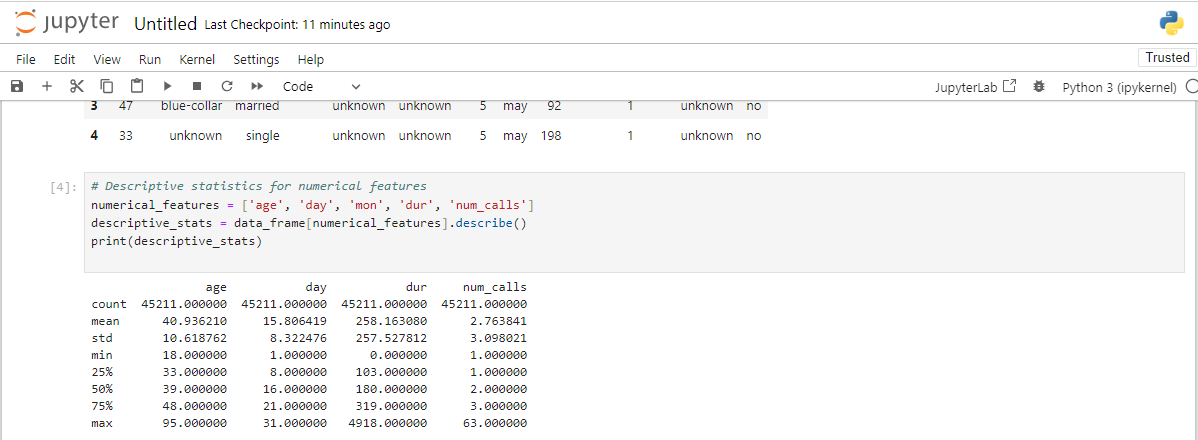
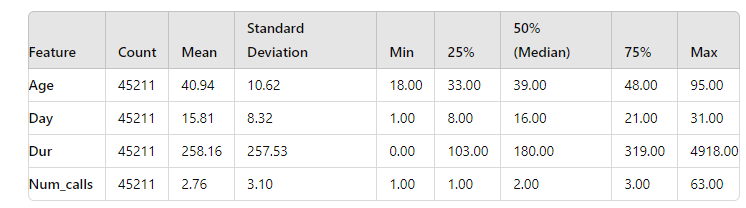


Figure : Jupyter Notebook Code and Output: Descriptive Statistics for Numerical Features

Table : Descriptive Statistics for Numerical Features



### 2.1.3 Frequency Analysis for Nominal data

The following categorical features will be analyzed:

1. Job: Customer’s occupation.
2. Marital: Customer’s marital status.
3. Education\_qual: Customer’s education level.
4. Call\_type: The modality of contact which was made with them such as phone, email.
5. Prev\_outcome: Outcome of the previous campaign (for instance, successes, failures).
6. Y: Dependent variable which showed the consumption of the insurance product by the customer.

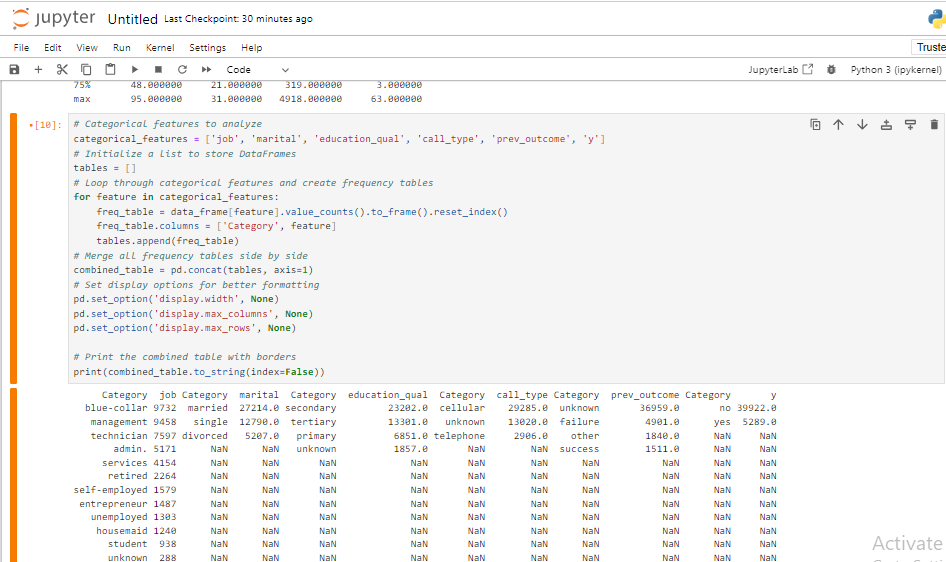


Figure : Jupyter Notebook Code and Output: Frequency Tables for Categorical Features

Table :Frequency Analysis for Categorical Features

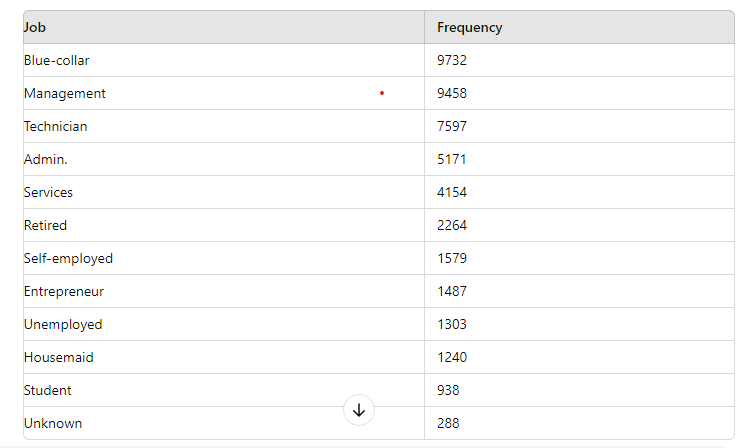


Table : Frequency Distribution of Marital Status

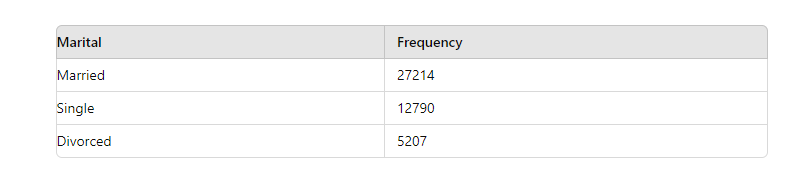


Table : Distribution of Education Qualification

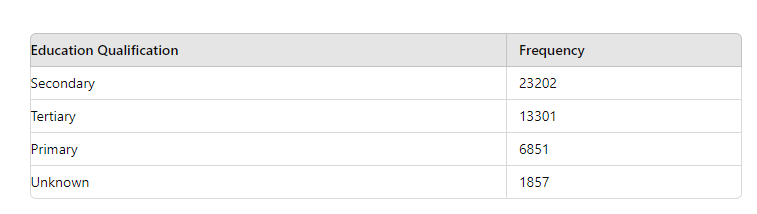


Table : Distribution of Call Type

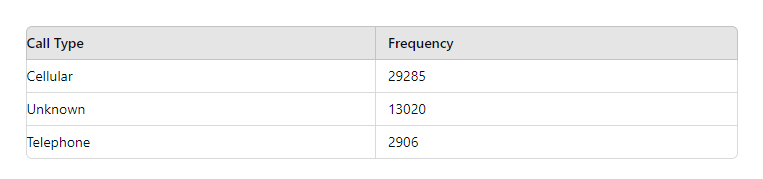


Table : Distribution of Previous Outcome

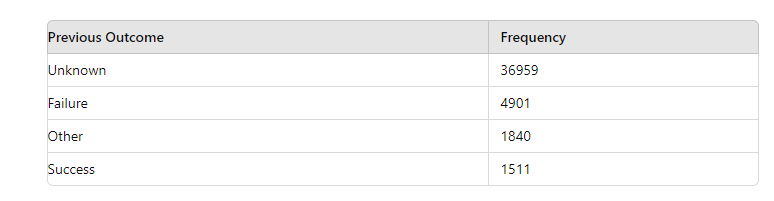
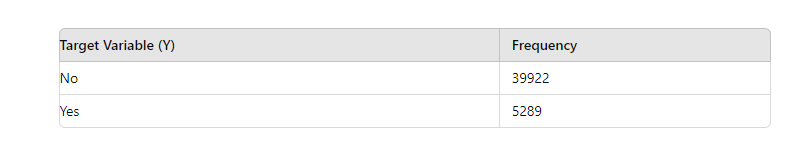


Table : Distribution of Target Variable (Y)



### 2.1.4 Summary of Key Findings

1. **Numerical Features**: The statistical summary will show an important number of trends for instance the number of calls made and the amount of time used in calls. For instance, volatility in the length of the calls may require a high standard deviation and that may be indicative of the fact that customers are either getting interested or losing interest while a telemarketing call is going on.
2. **Categorical Features**: This will enable frequency analysis of customers’ distribution in terms of their jobs, marital status and education level to be carried out. For instance, if a particular job type (for instance, management) is more effective at conversion, it can be marketed more in future campaigns. In the same way, using the prev\_outcome feature may show customers who made positive outcomes to past telemarketing campaigns.

Such conclusions contribute a solid ground to construct effective models and the key variables that can facilitate the enhancement of the telemarketing techniques at HashSysTech Insurance.

## 2.2 Handling Missing Values and Outliers

### 2.2.1 Identifying Missing Data

When it comes to handling missing and/or outliers data it is essential to notice that the quality of data is a significant measure for data analysis. Data can be missing for many reasons, including data collection or data entry failure. The first step therefore is to locate these missing values within the dataset.

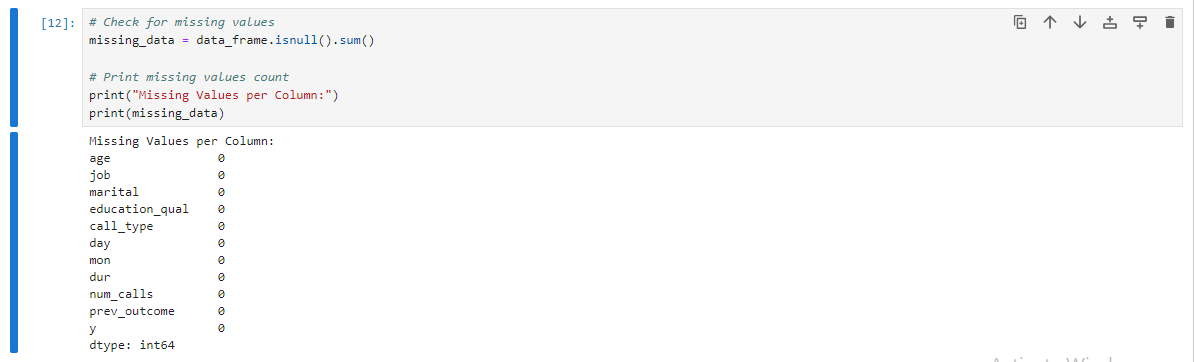


Figure : Jupyter Notebook Code: Identifying Missing Values

### 2.2.3 Techniques on Dealing with Missing Information

Once missing values are identified, several strategies can be employed to handle them:

* **Imputation**: Impute missing values with imputed values. Common methods include:
* ***Mean/Median Imputation***: In case of numerical feature, choose mean or median value of the feature for imputing missing values.
* ***Mode Imputation***: For the categorical features, the most frequently occurring value of the respective categorical feature ought to be used.
* **Deletion**: Then to reduce space, it is advisable to delete rows or columns with missing values. This method can be used for- Low percentage of data missing.
* **Prediction Models**: When two or more features are missing and there is correlation between them use machine learning models to anticipate the missing values.

Example of mean imputation:

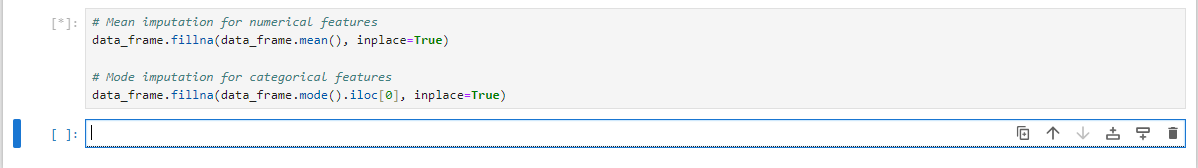


Figure : Jupyter Notebook Code: Imputation of Missing Values

### 2.2.4 Outlier Detection Techniques

Outliers are calculated values that are located far from the other values of the dataset. They can affect the outcome and performance of algorithms on ML models. Common techniques for detecting outliers include:

* **Statistical Methods**: Outlier analysis should be done using statistical methods like the z-score or the Interquartile Range (IQR).
* ***Z-score Method***: Outliers are values having their z-scores equal to more than a given limiting or cut off mark for instance plus or minus 3.
* ***IQR Method***: There are some values that are beyond the range of Q1 - 1 that has been defined out of the sets of data Q1 – 1. 5 \* IQR and Q3 + 1.5 \* IQR.
* **Visualization Methods**: Visual inspection should be made by plotting data using techniques such as the use of box plot or scatter plot.

Example of IQR method for outlier detection:

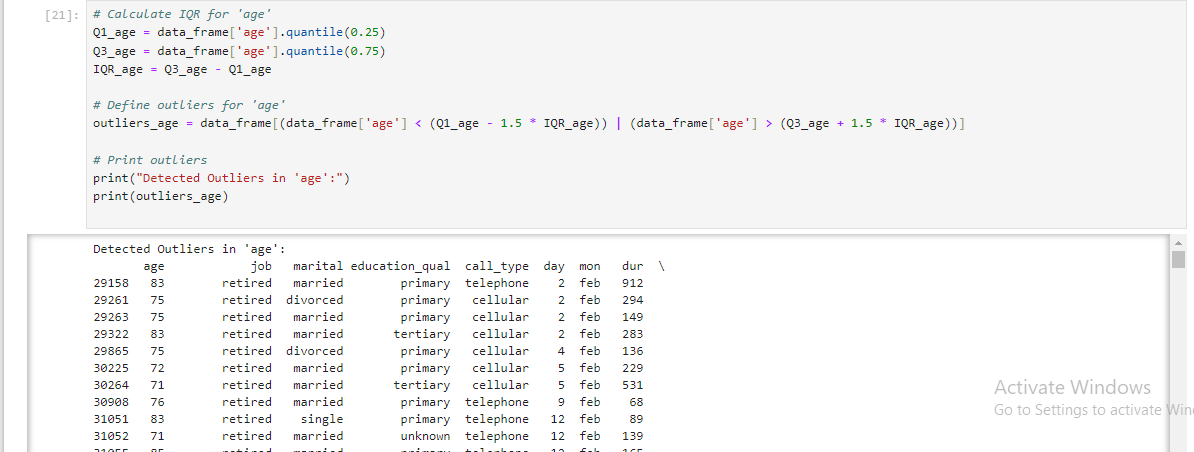


Figure : Jupyter Notebook Code and Output: Outlier Detection and Handling

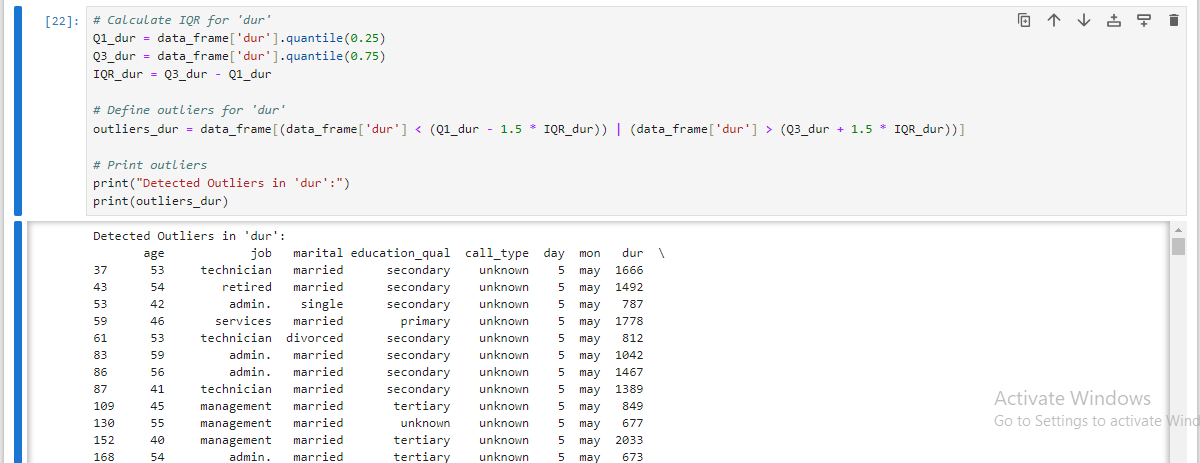


Figure : Jupyter Notebook Code and Output: Outlier Detection and Handling

### 2.2.5 Treatment of Outliers

Once outliers are identified, various strategies can be used to handle them:

* Transformation: This can be done through applying logarithmic or square root transformations for example.
* Imputation: Impute outliers when these extreme values are erroneous or beyond the sample’s possible range of values known from theory or prior research.
* Removal: Out of the calculations, the values that are presumed to be erroneous or irrelevant should not be included.

Example of removing outliers:

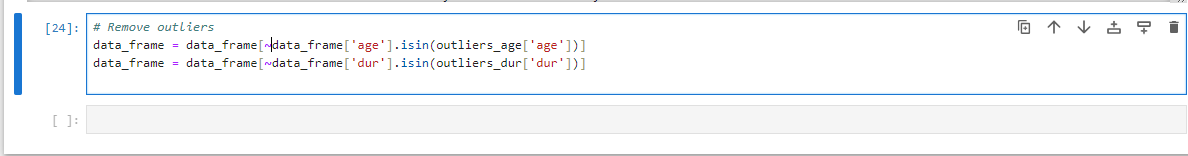


Figure : Jupyter Notebook Code: Outlier Removal

### 2.2.6 Justification of Chosen Methods

The way in which missing values and outliers are addressed is dependent on the characteristics of the data and the analysis that will be conducted.

* Imputation is ideal for use when the number of missing values is small so as to retain the size of the data set and to ensure the continuity of data.
* Deletion can be applied when the proportion of missing data is low so that it does not distort overall analysis that is being conducted.
* There are two ways in which to handle outliers, and the method used depends on the effects of outliers on the data analysis. Techniques of statistics and visualization help to maintain proper approach towards detecting and addressing outliers that is helpful for increasing model performance.

Thus the methods chosen to deal with missing values and outliers are as follows in order to improve the quality of the dataset and its conclusion.

## 2.3 Data Visualization

Data visualization was also useful in distinguishing patterns of aspects embedded in the data. All the numerical and categorical variable distributions and their relation with the target variable was illustrated.

### 2.3.1 Visualizing Numerical Features

The next graphing techniques which were employed in attempting to analyze the data included Histograms and Scatter Plot in an endeavor to show distribution and relationships between numeric variables like age and income. For instance, the histogram showed that most of the customers fall in middle-aged bracket.

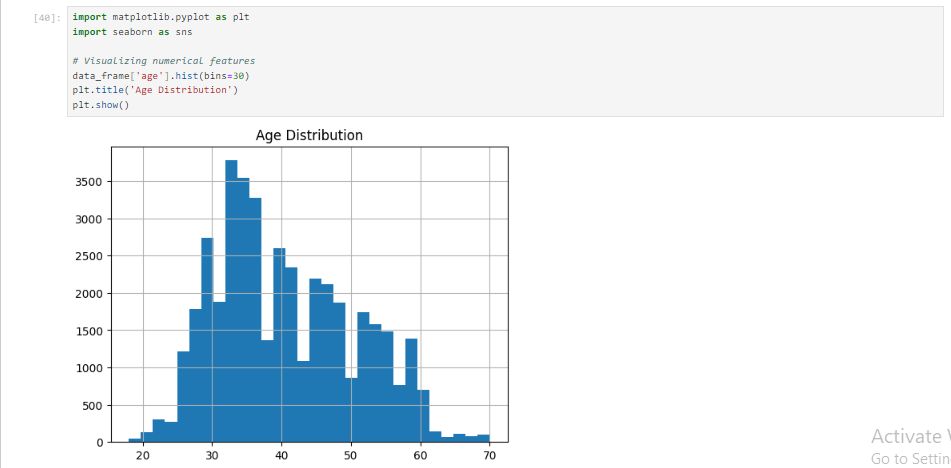


Figure : Jupyter Notebook Code and Output: Age Distribution Histogram

### 2.3.2 Visualizing Categorical Features

Bar Plots were used to show the data distribution of other nominal variables such as marital status, or gender. For example, among the married people, policy subscription was more common than among others.

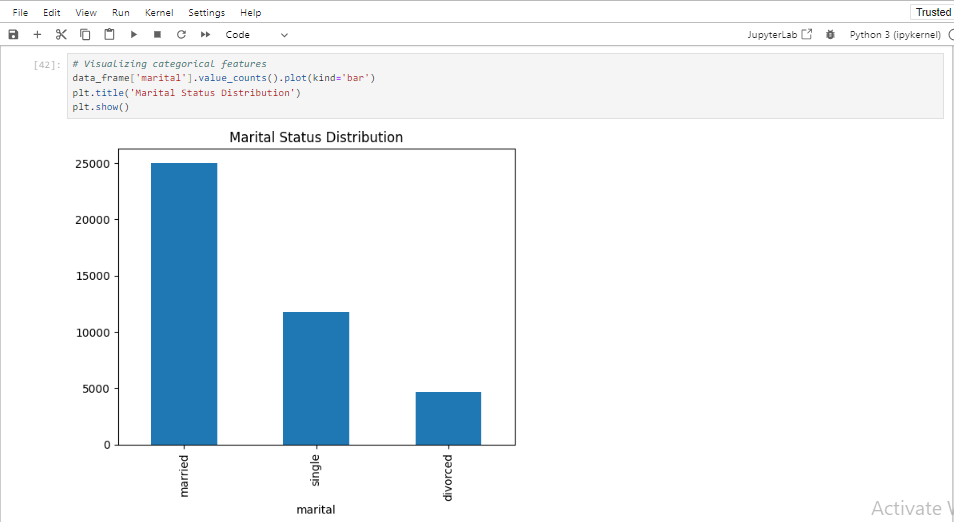


Figure : Jupyter Notebook Code and Output: Marital Status Distribution

### 2.3.3 Depiction of Relations with Target Variable

The Stacked Bar Plots were used, alongside the Box Plots to bring out the match between features (like income) and policy subscription status. The well-off customers subscribed more in terms of their numbers.

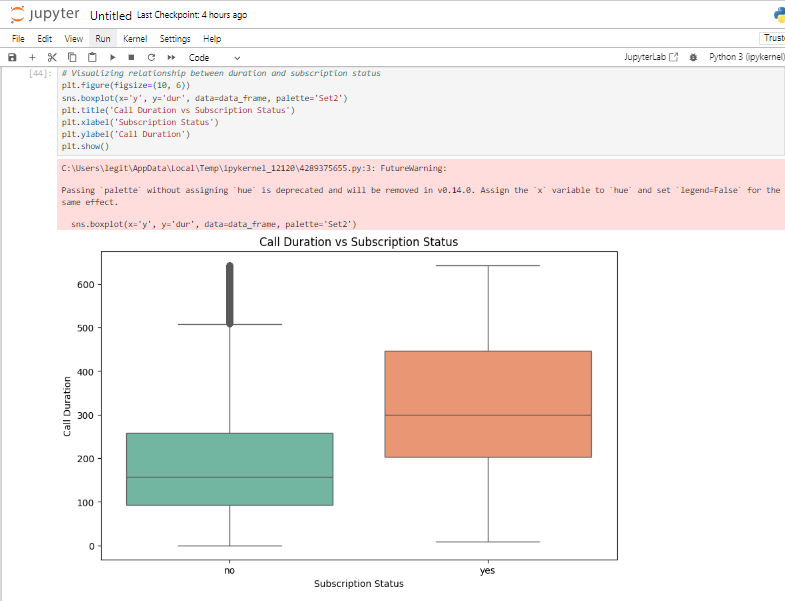


Figure : Jupyter Notebook Code and Output: Boxplot of Call Duration vs Subscription Status

### 2.3.4 Insights from Visualizations

Other influential factors include:

* Demographic information namely the age and income levels of the people were influencing matters concerning subscription.
* Customer Preferences stated the use of certain communication media provided higher conversion rates.
* Class imbalance was addressed where during data split, the sampling was done in a way that reflected an equal distribution.

All in all, data described in the section served to construct the model as well as to identify key areas and directions for action.

# Model Selection and Training

## 3. 1 Model Selection

### 3.1.1 Overview of Binary Classification Algorithms

Nominal methodologies are categorically intended for the placement of data in one of two categories. Key algorithms include:

* Logistic Regression: Relating a test variable to one or more predictor variables, in order to predict a binary outcome. This is relatively easy to implement as well as to interpret and is most successful when used on data that is linearly separable.
* Random Forest: A machine learning technique that creates many decision trees to combine them into one in order to achieve more stable and accurate result. It is less susceptible to ‘overfitting’ and performs well in large data sets and intricate models of data association.
* Support Vector Machine (SVM): Determines the line, in the multi-dimensional space that provides the greatest margin between the two classes. It is efficient in higher dimensionality and when there is definite margin separation.
* Gradient Boosting Machines (GBM): A method in which several weak classifiers are aggregated in order to form a strong classifier. The model is capable of high accurate predictions but it has a problem with overfitting.
* K-Nearest Neighbors (KNN): Assigns the data according to the nearest neighbors based on the most frequent class. It is quite simple but the complexity is in terms of massive computations may be involved.

### 3.1.2 Justification for Choosing Specific Algorithms

* Logistic Regression: Originally selected because the results of this test type are easy to understand and analyze. Gives the chance of belonging to a particular class and comes handy when the data features and target relationship is nearly linear.
* Random Forest: Chosen for its stability and usefulness in the handling of the interactions between features and the other patterns. It is effective with the noisy data and gives the insights about the features.

### 3.1.2 Summary of Model Selection Process

This amounts to comparing the output of a variety of algorithms as a way of determining their readiness for use depending on the data, its properties and the attempted problem. In this context, Logistic Regression and Random Forest were selected as models because both of these present equal importance to interpretability and high accuracy. We know that Logistic Regression works well if there is a straight line and Random Forest if there is a curve and there is an interaction.

## 3. 2 Data Splitting

### 3.2.1 Data Splitting (Training & Testing Data Set)

The datasets are split into training and testing sets so that the models can be tested on data that the model has not been trained on. Normally, 80% of data is used for training dataset and the remaining 20% for testing dataset. This is useful in determining the cross validation and hence how well the model is likely to perform when presented with fresh data.

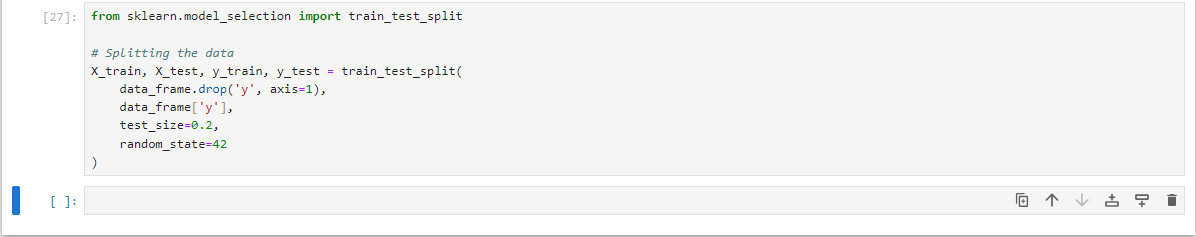


Figure : Python code snippet: Data Splitting using train\_test\_split

### 3.2.2 Ensuring Stratification of Target Variable

Stratification makes certain that the proportion of each class in the target variable (y) is maintained in the training as well as in the testing data. This is so as to ensure that classes are balanced particularly in imbalanced domains.

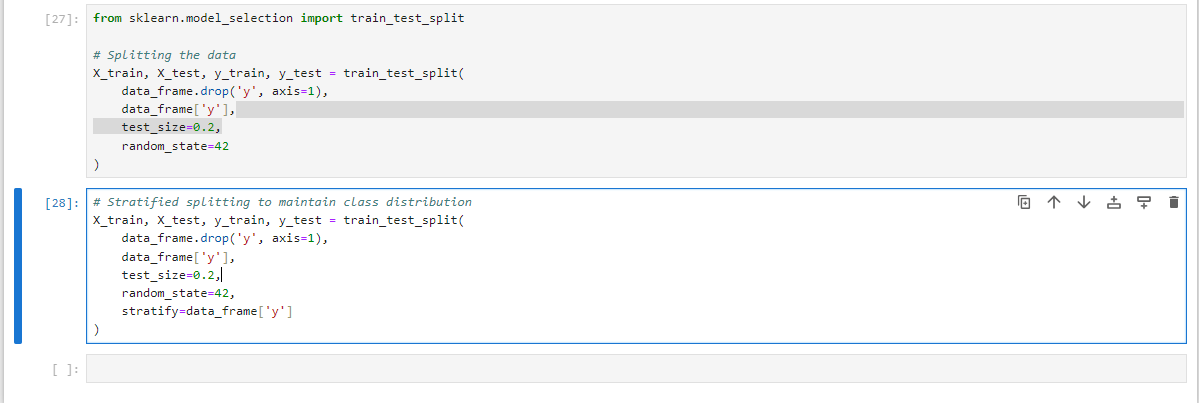


Figure : Jupyter Notebook Code: Stratified Data Splitting

### 3.2.3 Explanatory Information on Data-Splitting Technique

Cross validation as a means of splitting the data consist of the division of the data set into training and testing micro sets to construct and assay models. Partitioning makes certain that both subsets mimic the target variable distribution since this is very crucial for accurate evaluation with minimum bias. The training set is the data on which the training of a model is based, and the testing set is the data set that defines the model’s results and its ability to work on new data.

Thus, it is possible to build a strong model that can be checked and give the necessary information about the behavior of customers and predict their subscription rate.

## 3. 3 Model Training

### 3.3.1 Training the Logistic Regression Model

As mentioned above, the logistic regression model is going to be trained performing the following steps:

Logistic Regression finds the best fit line using the training data to be able to model the relationship between the logistic features variable and the target variable.

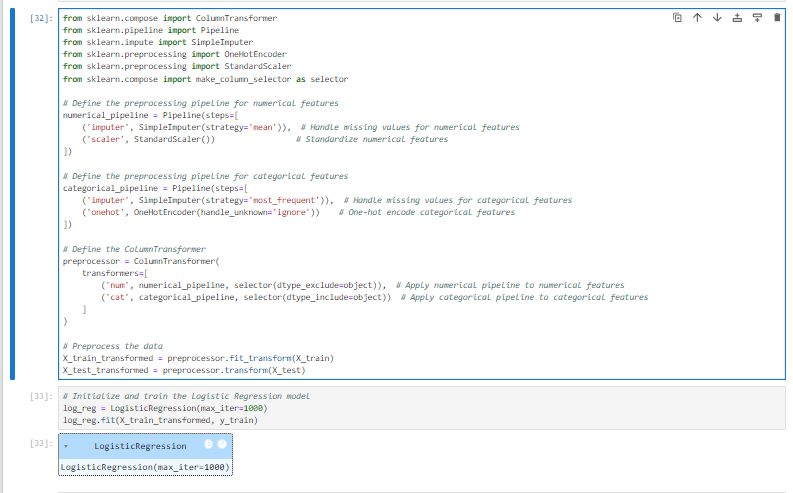


Figure : Jupyter Notebook Code: Preprocessing Pipeline and Logistic Regression Model

### 3.3.1 Training the Random Forest Model

Random forest is used the training data to build up a collection of decision trees where the final prediction is based on the aggregation of the decision of the individual trees that were built.

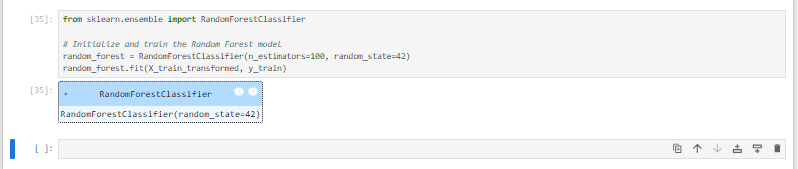


Figure : Jupyter Notebook Code: Random Forest Model Initialization

To conclude, the model selection process was initiated with a literature survey of binary classification algorithms, narrowing down the choices to those that effectively handle classification problems, specifically Logistic Regression and Random Forest. The choice of algorithm was twofold; Logistic Regression due to its simplicity and easy interpretation as well as due to the fact that Random forest is very resilient to complex data structures and non-linear relationships. The data set was therefore partitioned into training and test sets with an aim of splitting the target variable into equal proportion between the classes. This approach helped in making sure that the training of the model and the testing or validating of the model was done with data that was most relevant hence increasing the possibilities of the model or algorithm performing well when implemented.

# Model Interpretation and Evaluation

## Model Interpretation

In this case, the models that were chosen were Logistic Regression and Random Forest, their decision making was explained. Hypothesis testing in Logistic Regression was linear and the Random Forest brought out new features of non-linearity by using multiple decision trees. Feature importance calculation in the Random Forest process provided an additional view on the features that played the most significant role on the model in company’s customer outcome prediction.

## 4.2 Model Evaluation

The models were assessed using standard classification measures; accuracy, precision, recall, and F1 score. The following tables show the Classification reports of Logistic Regression and Random Forest.

Table : Logistic Regression Classification Report

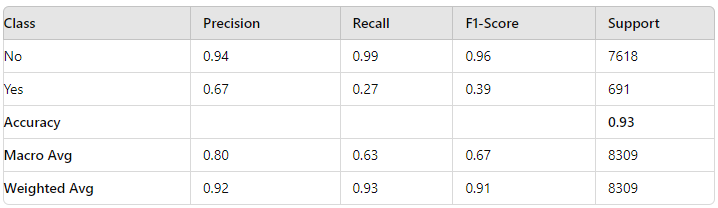
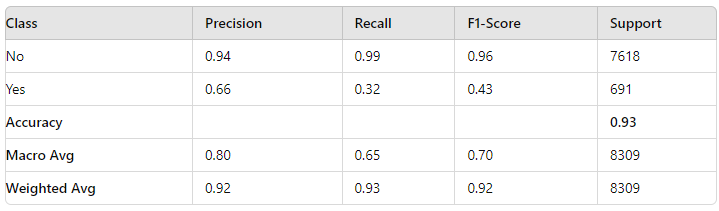


Table : Random Forest Classification Report



## Python Code for Model Evaluation

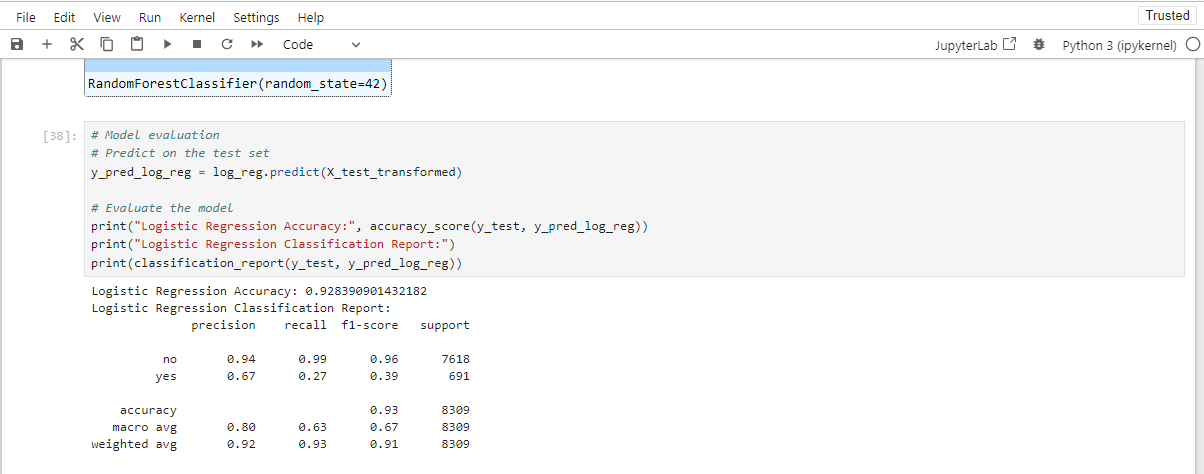


Figure : Jupyter Notebook Code and Output: Logistic Regression Model Evaluation



Figure : Jupyter Notebook Code and Output: Random Forest Model Evaluation

## 4.4 Hyperparameter Tuning

Tuning generally refers to the process of finding the best values for those hyperparameters that are not influenced by the data and thus are not learnable. Hyperparameters are difficult to determine as the model does not learn from them; examples are the **α** value in Logistic Regression or the Number of Trees in Random Forest. These parameters are very important in the performance and degree of complexity of the model.

In the present work, tools such as GridSearchCV were applied to execute the search for the optimal values of hyperparameters by testing various possible configurations and their performance estimated by cross-validation.

### 4.4.1 Importance of Hyperparameter Tuning:

* Model Performance: By using optimizing hyperparameters the model is more likely to generalize well on the unseen data than overfitting the data or under fitting.
* Balancing Bias and Variance: Good tuning can lead to better cleaning of the data in order to avoid over simplification (biases errors) and overelaboration (variance errors at the same time).
* Efficiency: Tuning of hyperparameters means that the models are not only as accurate but also fast.

Here, regularization was fundamental in enhancing the performance of both the Logistic Regression and Random Forest models, as well as attaining the best possible accuracy and prevent overfitting in the test set.

On average, both models were accurate, and Random Forest was only slightly higher than Logistic Regression in recall for the less frequently occurring class ‘yes’. Although, few classes ‘yes/true’ had slightly better accuracy then Logistic Regression, for the majority class ‘no/ false’ the performance was more balanced.

# 5. Conclusion

## 5.1 Summary of Key Insights and Findings

The binary classification models used for analyzing the customer data of HashSysTech Insurance offered great insights into the customer behavior and the principles effective for policy subscription. The choice of Logistic Regression and Random Forest models was made after rating them for the classification task based on their accuracy which was estimated to be \*\*93%\*\*. When it came to the prediction of the “No” instances, both the models agreed in their predictive performance however the results from the Random Forest model were slightly better in the case of the “Yes” instances as far as the recall and F1-score were concerned. Based on feature importance analysis, other factors that influence customer choices were identified as the factors that help in the formulation of strategies for customizing actions to attract the more valuable clients.

## 5.2 Recommendations for HashSysTech Insurance

Based on the findings, the following recommendations are provided for HashSysTech Insurance:

* Targeted Marketing Campaigns: Examining these features, the model has highlighted that targeting those customers who reflect higher levels of policy subscription is a worthwhile pursuit – including communication patterns and customer participation.
* Retention Strategies for High-Risk Customers: In retention strategies, consideration should be given to those customers thought likely to churn (“Yes”). Personalized services and incentives regarding the decided key features might help to reduce churn risks.
* Continuous Monitoring and Re-Evaluation: If the customer’s preferences and the market situation change, it is more efficient to train the model with the new data from time to time to have more accurate results.

## 5.3 Reflection on the Model's Impact on Business Decision-Making

The various functions of the model mean that HashSysTech Insurance has a tool that can greatly assist its decision making process. With the knowledge of customers prone to subscription or churning the business can better invest in appropriate marketing channels as well as in customer retention initiatives. These patterns, derived from key features of customers, help the company to concentrate on the most precious and profitable clients, which increases overall profit rates and insurance company’s competitive position on the market.

Therefore, the implementation of these machine learning models ensures HashSysTech Insurance gains insights that can be acted upon hence improving client management and designing of good business strategies.