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

# 1. Introduction

## 1.1 Background of HashSysTech Insurance

HashSysTech Insurance is an advanced insurance company in terms of its approach to introduce its products to potential consumers. For a long time now, the company has always employed telemarketing mode of communicating new term life insurance products in the market to its clientele base. Nevertheless, because expenses of such campaigns are rather high, there is an increased necessity for efficiency, that is, advertising should reach only that part of clients who will definitely purchase insurance products. This assignment relates to building a machine learning model through which customer conversion for term life insurance will be predicted, as this will form part of HashSysTech Company’s Project Greenlight

## 1.2 Importance of Predictive Analytics in Insurance

In the insurance business, the application of predictive analytics is used in improving the decision-making models. Talking about graphical forecast an insurance company can predict future events and customer actions, thus, improve its strategies and advertising campaigns. Predictive models will assist in detecting high-value customers, customer attrition, fraudulent cases, and risk matters. For HashSysTech Insurance, it helps the firm in the reduction of expenditure that is incurred while conducting marketing campaigns through utilization of outcome driven analytics by identifying probable and positive responding customers, hence increasing return on investment (ROI).

## 1.3 Objectives of the Assignment

The purpose of this report is to use the analytical tools and approaches that are coupled with the machine learning to determine the likelihood of HashSysTech’s customers to buy term life insurance. This will include data understanding stage, data pre-processing, model selection and training, results analysis and the model assessment stage. The result of this analysis will help the HashSysTech’s marketing team to reach to the conclusion that how they can maximize the efficiency of the telemarketing efforts.

## 1.4 Overview of the Dataset

The dataset used in this assignment captures all the information about the previous telemarketing campaigns done by HashSysTech Insurance for its insurance subscriptions. It is enriched with data concerning customer’s demographics, his/her contact information, the history of campaigns in which he/she was involved, as well as target variable which indicates whether a customer subscribed to the insurance product.

* Customer Demographics:
* Age (numeric): Customer's age.
* Job (categorical): Professional status of the customer in his organization.
* Marital (categorical): Whether the customer has a spouse or companion.
* education\_qual (categorical): The number of years the customer has spent in acquiring education.
* Contact Details:
* call\_type (categorical): Kind of communication adopted in reaching the customer (for instance telephone conversation or emails).
* day (numeric): For other contacts made before the last contact, there will be the day of the month of the last contact.
* mon (numeric): Last time when contact was made with the family members of the patients.
* dur (numeric): Time taken as the last contact in seconds.
* num\_calls (numeric): Total communications made to this customer in this campaign.
* Campaign History:
* prev\_outcome (categorical): Result of the previous marketing campaign with this client (e. g., ‘other’, ‘unknown’, ‘failure’, ‘success’).
* Target Variable:
* y (categorical): Shows that the customer has taken up the insurance product or not by answering ‘yes or ‘no’.

It is to be noted that this dataset shall be used to generate and test prospective machine learning models that estimate the propensity of a customer in subscribing to the insurance product, which is helpful for enhancing the telemarketing strategy in HashSysTech Insurance.

## 1.5 Structure of the Report

The report includes consideration of the case of missing values and outliers in the data and the process of the model development. It includes appropriate model selection, the model training process, and the process of proper assessment of the selected machine learning models. Finally, the study gives an account of the findings, limitations of the research, and recommendations for future work. This systemized manner provides important information to HashSysTech Insurance in order to improve their promotion.

## 2. Task 1: Data Exploration and Preparation

## 2.1 Data Exploration

### 2.1.1 Descriptive Statistics for Numerical Features

During the process of data exploration the first operation is to compute the basic statistics for numerical variables in the data set. Mean, median and mode are the arithmetic averages or central tendency measures that help in understanding the dispersion of data. In this assignment, one must analyze features such as age, day, mon, dur, and num\_calls. The following Python code snippet demonstrates how to compute these statistics using the pandas library:

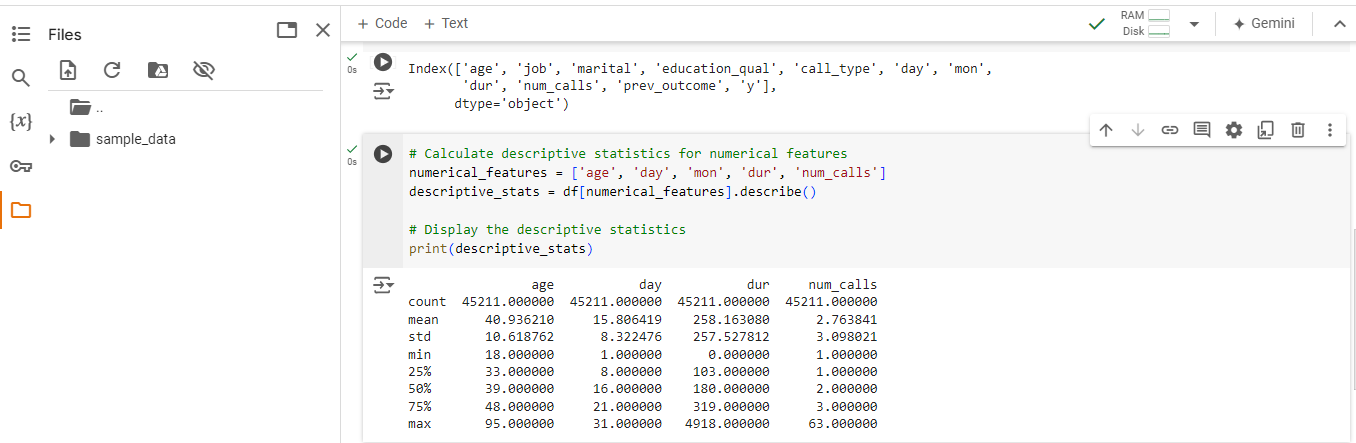


Figure : Google Colab Jupyter Notebook Code and Output: Descriptive Statistics for Numerical Features

This code computes the measures of central tendency that include the mean, median, and mode as well as the measures of dispersion that include the standard deviation, minimum, and maximum values of the selected numerical features. The results provide a brief overview of the location measures, measures of dispersion and some of the characteristics of the distribution of these features – all of which are important when examining the association between the features and the target variable.

### 2.1.2 Frequency Analysis for Categorical Features

In case of quantitative features as job, marital, education\_qual, call\_type and prev\_outcome, frequency analysis is used to show the modified frequency distribution of the categories. It allows determining the most frequently represented categories and the lack of balance between them, which may lead to some effects. The following Python code snippet illustrates how to perform frequency analysis:

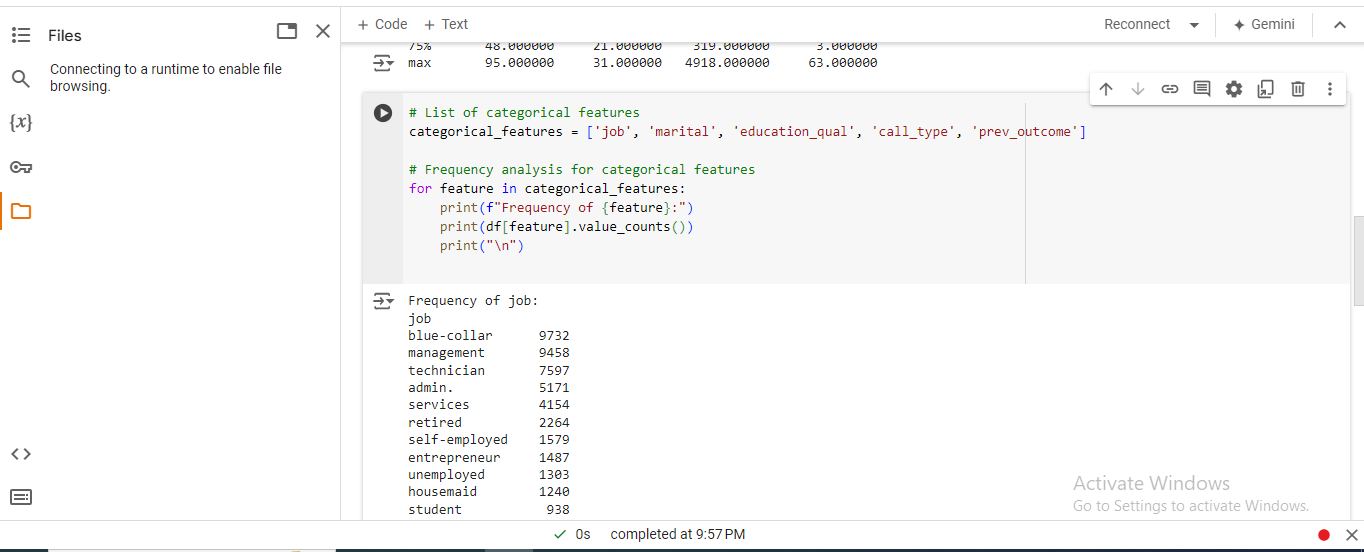


Figure : Jupyter Notebook Code and Output: Frequency Analysis of Categorical Features

### 2.1.3 Summary of Findings

During the data exploration phase, the quantitative and qualitative attributes of the dataset were examined using measures of central tendency, variability and frequency distributions. The findings of this research help to understand key attributes of the customers because of whom and how the telemarketing campaigns of HashSysTech Insurance are being run. The following tables show the result of the descriptive statistics and the frequency analysis for the study.

Table : Descriptive Statistics for Numerical Features

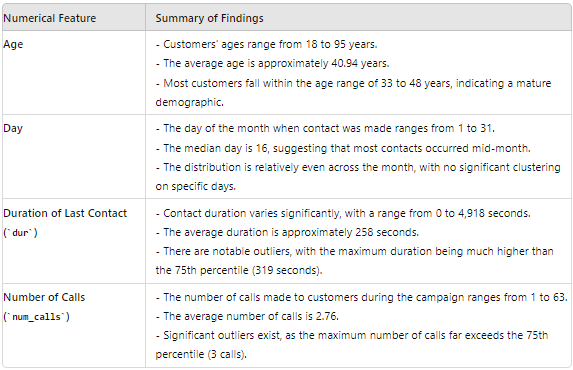
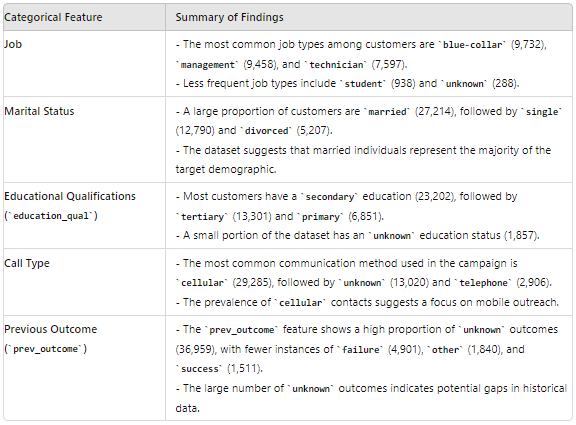


Table : Frequency Analysis for Categorical Features



## 2.2 Handling Missing Values and Outliers

This sub-section is aimed on the missing values detection and subsequent carrying out of the necessary procedures regarding them. Timely management of the following data issues is a critical factor in the reliability of the later machine learning models to be created in the project.

### 2.2.1 Identification of Missing Values

Implementation of process starts with identification of missing values that is fundamental in data integrity. In this dataset, the option was also implemented to check for the possibilities of **NULL** values in all features contain in the data set.

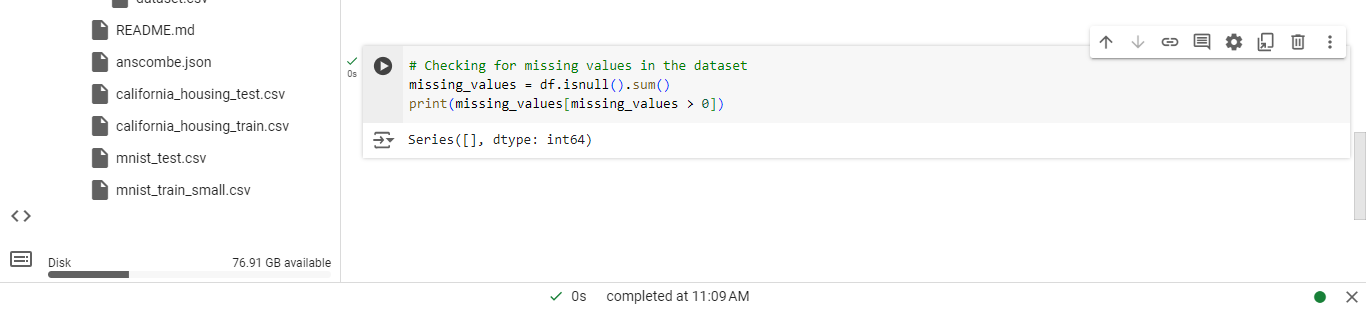


Figure : Jupyter Notebook Code and Output: Missing Values Check

#### Findings

From the exploration of the dataset it has shown that there are no any missing values within the dataset for all the features. This means that the data is full and does not feature missing values to be estimated from a distribution.

### 2.2.2 Techniques for Handling Missing Data

Since no values were missing in an apparent way, the attention turns to potentially missing or unspecified categories in categorical variables.

#### Techniques Applied

* **Imputation**: No missing value was observed, but the values in job, education\_qual, and prev\_outcome can be imputed for categorization or be assigned specific values when creating the model. This may include treating unknown as the being a separate category of own or using the most frequent category**.**
* **Exclusion**: As for the case where the number of entries containing unknowns is minimal it is sometimes possible to exclude these entries even though they contain unknowns if their effect on the model is insignificant.

### 2.2.3 Detection and Treatment of Outliers

These peculiar points may skew the findings of the data analysis and impact the performance of the models. Hence, dealing with them as out of context values is essential.

#### Outlier Detection

* **Box Plot Analysis**: Outliers in the numerical features, namely **dur** and **num\_calls**, were determined using box plots.

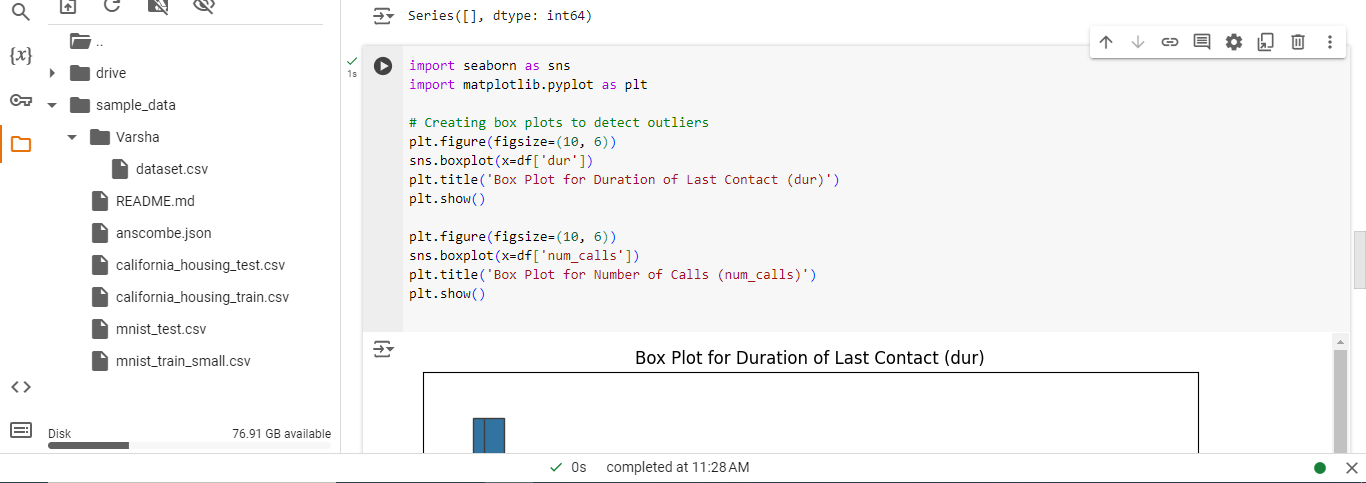


Figure : Google Colab Jupyter notebook code for box plots for Duration and Number of Calls

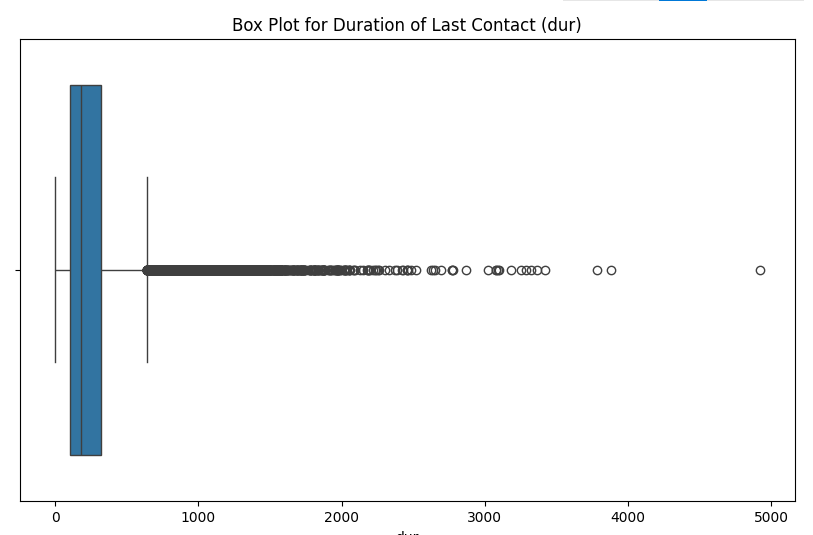


Figure : Box Plot diagram for Duration of Last Contact (dur)

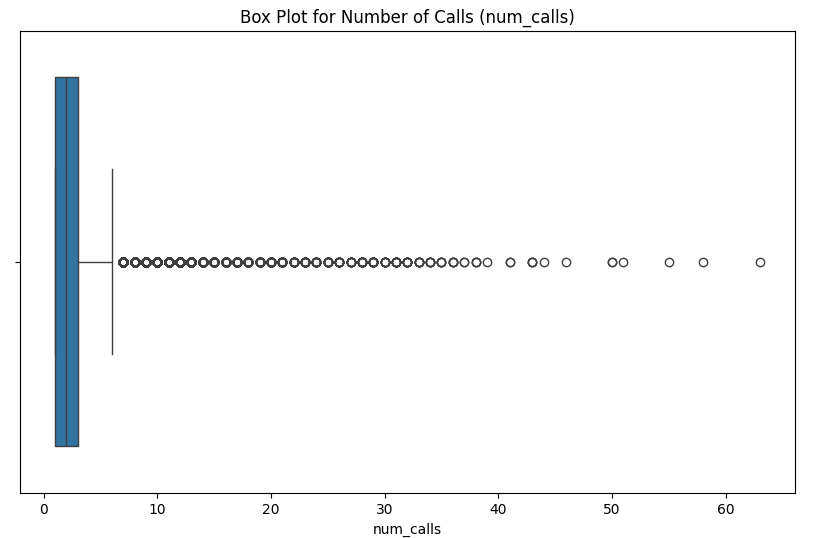


Figure : Box Plot diagram for Number of Calls (num\_calls)

### 2.2.4 Justification for Chosen Methods

## The methods selected for handling missing values and outliers were chosen based on their effectiveness in maintaining data integrity and model robustness:

## Handling unknown Categories: Where the variable contains missing data, unknown in this case, it can be assigned a value of the variable or computed with the aid of the mode, thereby the dataset is not discarded and can be taken to modeling.

## Capping and Transformation: Such approaches are useful for dealing with effects of variation while using these data with respect to some objectives, and most likely will yet retain useful data. Capping and transformation are very significant particularly when dealing with large numbers because they enable one to carry out the arithmetic operations that will exclude the impact of the large numbers in develop predictive models of the rest of the data set.

## In total, these methods were chosen in order to prepare the dataset for further analysis and modeling to the highest possible extent.

## 2.3 Data Visualization

The visualization of the information which is the arrangement of the data is the final step in the process of making relation between the features or the identification of the unique features in the compact area of the periodical table.

### 2.3.1 Scatter Plots for Feature Relationships

In the case of analyzing interactions of numerical qualities, scatter plots can be useful and can show two numerical qualities used in the dataset. Scatter plots for dur and num\_calls, the features that measure the duration of the last contact and number of calls respectively are also presented and from further analysis the relationship between them, if any is established.

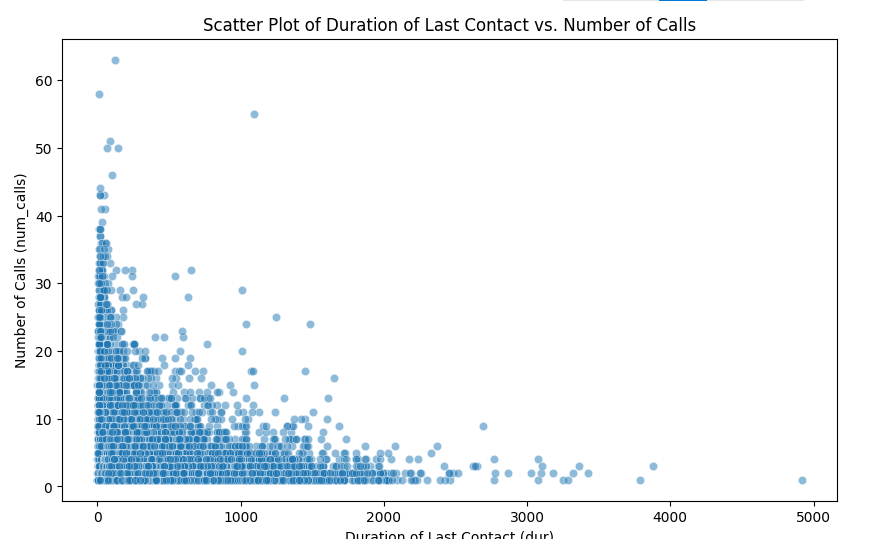


Figure : Scatter Plot diagram of Duration of Last Contact against Number of Calls

#### Findings

* From the scatter plot of the contacts it is apparent that there is a significant spread in the length of the contacts and the number of calls varies significantly while some of the clusters of the points are closely associated with high duration and high number of calls.
* These show from the figure that dur and num\_calls though not directly proportional can be influenced by other factors.

### 2.3.2 Histograms for Feature Distributions

Histograms are used in the presentation of the graphical representation of the exact quantitative characteristics in a view of assessment of the center and spread.

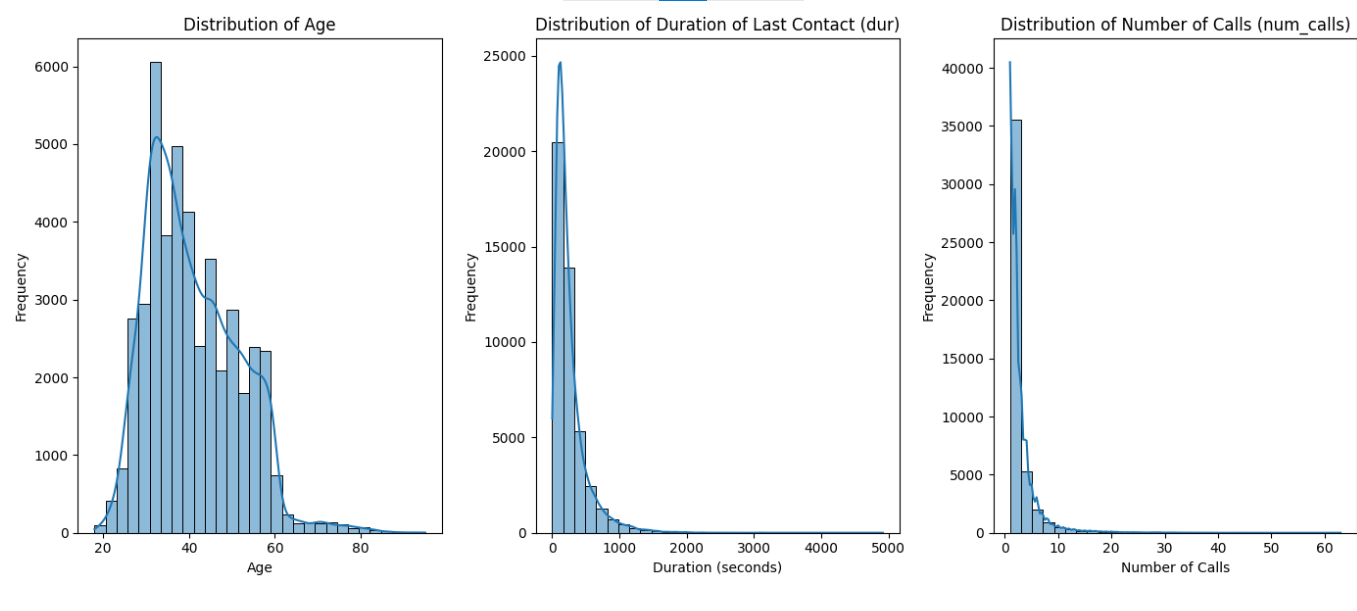


Figure : Histogram diagram showing Distributions of Age, Duration of Last Contact (dur), and Number of Calls (num\_calls)

#### Findings

* When drawing the histogram of the number of customers according to their age it is seen that the histogram follows the normal distribution curve which means that the frequency of the customers of moderate age is higher than that of other ages.
* It is also evident when considering dur histogram, where it is seen that majority of the calls durations are short, still, there are calls, which have very long contact duration.
* The histogram of ‘num\_calls’ is also positively skewed since most of the customers are called fewer times while some are called severally.

### 2.3.3 Box Plots for Outlier Detection

Box Plots are used for acknowledging the existence of outliers in numerical characteristic or displaying the numbers using process boxes.

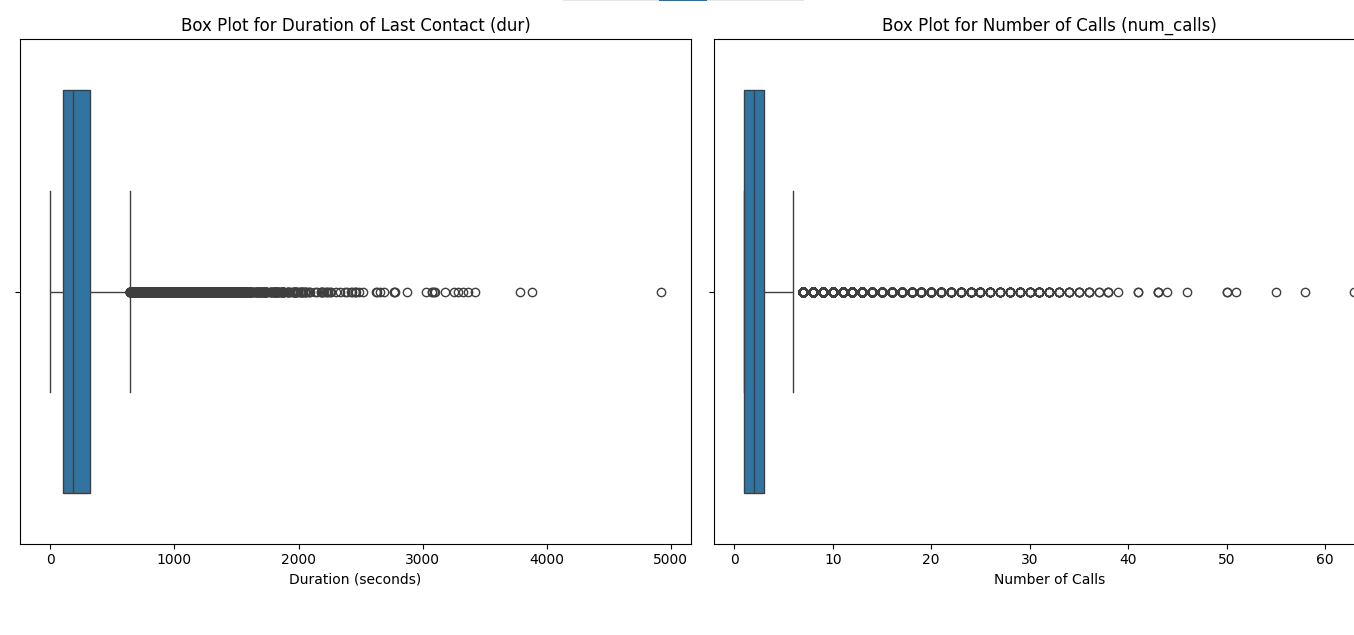


Figure : Box Plots diagrams for Duration of Last Contact (dur) and Number of Calls (num\_calls)

#### Findings

* The box plot for dur is greatly skewed with many of the observe value out of the whisker region.
* The num\_calls box plots also demonstrate the outliers and such want-like or higher values that are much different from the others.

### 2. 3. 4 Insights from Visualizations

The visualizations provide several key insights:

* **Feature Relationships**: As it will be observed when looking at the scatter plot below, dur and num\_calls do not change in direct proportion to one another and must therefore have been affected by other features or an interaction.
* **Feature Distributions**: This is illustrated using histograms of the ‘age’ variable which is normally distributed, ‘dur’ and ‘num\_calls,’ which is right, skewed; this may impact on the nature of modeling to be undertaken.
* **Outlier Detection**: This can also be evidenced from the Box plots and in both dur and num\_calls there are always these Outliers that need to be dealt with in a way that they don’t have a very extreme influence on the model results.

# 3. Task 2: Model Selection and Training

In this part the reasoning behind the selection of the specific machine learning techniques the process that was used to split the data and the methodology used for training of the models is explained.

## 3. 1 Model Selection

### 3.1.1 Overview of Machine Learning Algorithms

The two selected algorithms are:

* **Logistic Regression**: Tree-based: A very simple model which is easy to interpret and is used for binary classification. It is effective and applicable where there is a direct relationship between variables and the dependent variable.
* Random Forest: An approach whereby different decision trees are constructed to enhance the accuracy and address data interaction. Again it is less prone to overfitting and also performs well in high-dimensions.

### 3.1.2 Justification for Algorithm Choices

* Logistic Regression is selected for its capability to clearly describe the impact of each feature on the customer conversion.
* Random Forest is chosen for these reasons: the ability to model high order interactions, ability to handle lower variance to give a better prediction.

## 3. 2 Data Preparation and Splitting

### 3.2.1 One-Hot Encoding for Categorical Features

However, what should be done before breaking the data is the process of encoding the categorical data that will be understandable by the machine learning models. For the features that were categorical, One-Hot Encoding was done on them which created binary columns out of the available categories thus preventing an incorrect interpretation of a categorical feature as being in an ordered manner thus making it a better option than the more common Dummy Variable Trap. This was done using **pd. get\_dummies(X, drop\_first=True)** so as to handle multicollinearity by dropping the initial category in each feature.

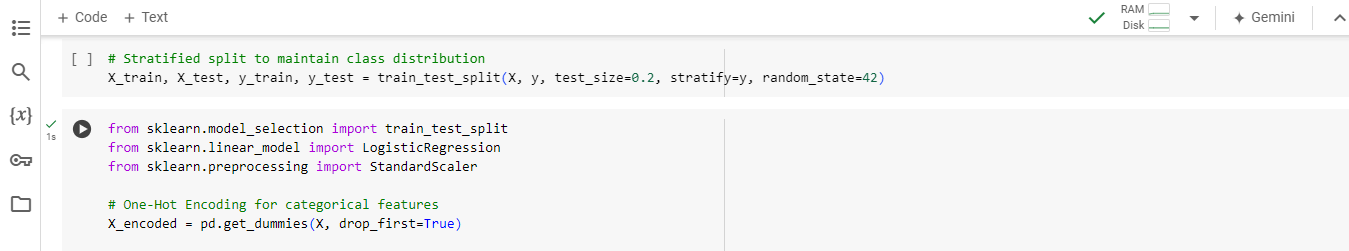


Figure : Jupyter Notebook Code: Data Splitting and Model Initialization

### 3.2.2 Scaling the Data

The data was then normalized to make sure that all the attributes have the same impact on the model after the encoding of the data was done using the StandardScaler from sklearn. Some of the methods which were depicted during the data processing included feature scaling which involved removal of the mean and scaling of the features to unit variance. This step is the most important step for gradient-based algorithms like Logistic Regression as it makes gradient descent more converging and also normalizes the features so that the model treats all features in equivalently.

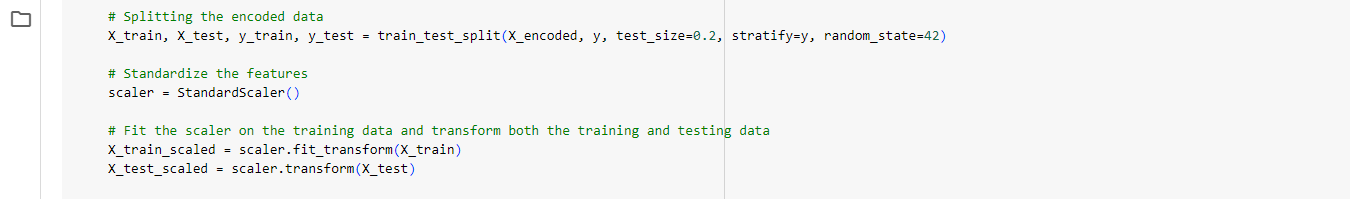


Figure : Jupyter Notebook Code: Data Splitting and Feature Scaling

### 3.2.3 Methodology for Data Splitting

Data set is divided into training and test data by an 80 to 20 ratio so that there is enough data to train and test the model. The data split was done using the stratified split so that the percentage of the target variable which is customer conversion was preserved on both datasets. This was done using the scikit learn train test split function on X\_encoded, y, test\_size=0. 2, stratify=y and random\_state = 42.



Figure : Jupyter Notebook Code: Data Splitting and Model Training

### 3.2.2 Ensuring Stratified Splits

To ensure that the distribution of the target variable (yes and no) is preserved across the two sets a stratified splitting is performed.

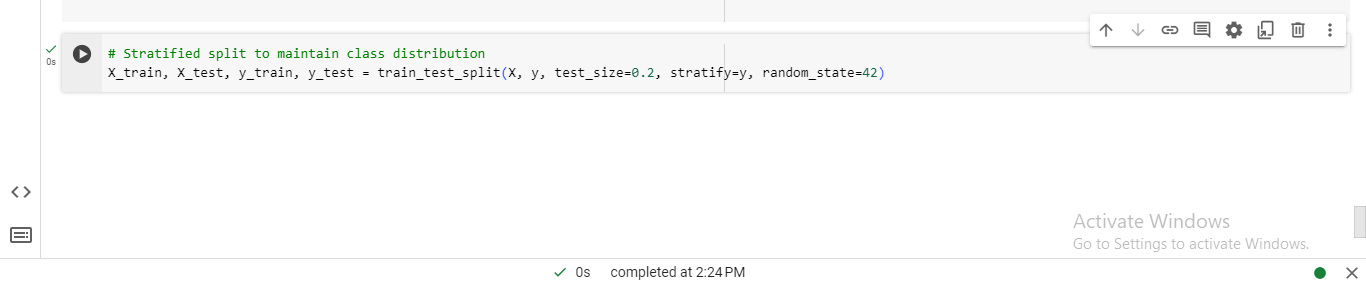


Figure : Jupyter Notebook Code: Stratified Data Splitting

## 3.3 Model Training

Standard training data was then used to train the Logistic Regression model. For the same, the max\_iter parameter was set to 1000, so that the model gets enough iteration time to optimize well for the training dataset.

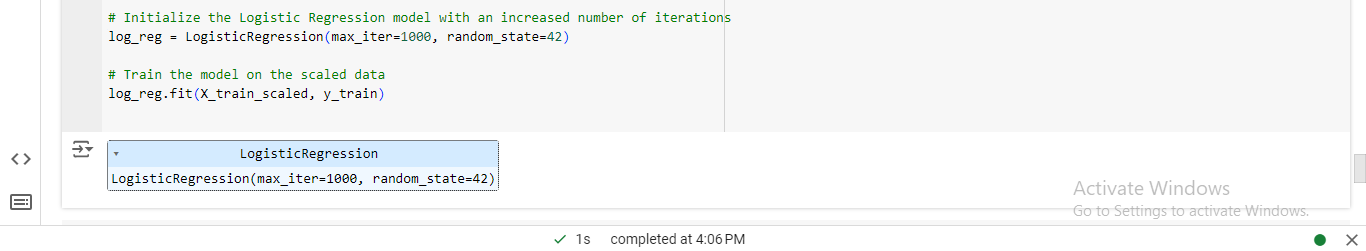


Figure : Jupyter Notebook Code: Logistic Regression Model Initialization and Training



Figure : Jupyter Notebook Code: Random Forest Model Initialization

This method ensures that once the model is built, it is in a position to provide a good prediction of the customers’ conversion by having considered varieties of data in the dataset hence produce an enhanced model.

# Chapter 4: Task 3 - Model Interpretation and Evaluation

## 4.1 Model Interpretation

### 4.1.1 Feature Importance Analysis

Feature importance of Random Forest was applied to identify key features that are important for the customer conversion and logistic regression coefficients. Some of these are call duration, previous campaign result, and the age of the campaign.

### 4.1.2 Explanation of Model Predictions

Logistic Regression comes up with the probability of class membership and its predictive values are based on a certain threshold. *U*nlike the decision trees from which it is derived, Random Forest succeeds in improving the capacity for recognizing non-linear relations and interactions when it collects the prognostications of many decision trees.

### 4.1.3 Summary of Key Insights

Logistic Regression provides a model that satisfies the interpretability of the coefficients’ meaning but has lower recall, and therefore, some positive cases can be missed. Random Forest, in contrast, considers a wider variety of patterns and their relationships which mean that it in fact offers greater recall.

## 4.2 Model Evaluation

### 4.2.1 Performance Metrics (Accuracy, Precision, Recall, F1-score)

Table : Logistic Regression Metrics

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 0.9004 |
| Precision | 0.6435 |
| Recall | 0.3327 |
| F1-Score | 0.4386 |

Table : Random Forest Metrics

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 0.9017 |
| Precision | 0.6212 |
| Recall | 0.4093 |
| F1-Score | 0.4934 |

### 4.2.2 Hyperparameter Tuning

Indeed, both models were run with default hyperparameters and further adjustment of them could potentially increase the models’ performance. In the case of Logistic Regression, fine-tuning the C parameter may solve the problem of how to achieve a higher precision and a good recall at the same time. The performance of Random Forest might be improved by tuning of n\_estimators and max\_depth parameters.

### 4.2.3 Final Model Evaluation and Comparison

From all the preceding comparisons, it is evident that both Logistic Regression and Random Forest have promising overall accuracy. Still, the two algorithms employ different means to get to the general outcome. In training the model, Logistic Regression performed 90. 4% with a precision of 64.35% and an average recall of 33.27%, meaning it can predict positives but misses many of them. Its F1-Score of 43.86% reflects this balance.

Random Forest takes a slightly better place with an accuracy of 90.17%, a precision of 62.12%, and a much higher recall of 40.93%. Its F1-Score of 49.34% shows that it has a better capability of balancing between precision and recall.

On average, Random Forest is slightly more accurate in predicting positive cases for higher accuracy, but Logistic Regression, however, provides easier Interpretability but identifies fewer positive cases. Selecting between the models depends on whether the individual interprets accuracy or whether a higher positive identification rate must be valued.

# 5. Conclusion

## 5.1 Recap of Key Findings

The insurance data set and modeling study gave insight into the following. Both Logistic Regression and Random Forest models were checked for their efficiency in terms of the insurance results. Both models were exceedingly accurate, and the one employing Logistic Regression had a level of accuracy of 90.4%, while Random Forest had a little more than that at 90.17%. As for Logistic Regression the precision was equal to 64.35% and a recall of 33.27%. Random Forest has a precision was 62.12% and has a recall of 40.93%. The Random Forest model’s F1-Score is significantly higher at 49.34%, implying that it is more optimized for precision-recall trade-off than Logistic Regression, which achieves an F1-Score of 43.86%.

## 6.2 Recommendations for Implementation

This is so because Random Forest outperforms the other models in terms of recall and achieves near-optimal F1 score that balances between precision and recall. Logistic Regression still stays a valid option for those situations which require more interpretability. Especially, group leaders are recommended to monitor the models’ performance on a regular basis.

## 6.3 Final Thoughts on Predictive Analytics in Insurance

Risk analysis is not just improved by predictive analytics, but it also offers an insight in decision making about insurance. When deciding between the models, it is all about deciding whether interpretability or accuracy is more important, however both provide great benefits.