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

# ****1. Introduction****

## ****1.1 HashSysTech Insurance's History****

HashSysTech is an insurance company that delivers services and products specific to the market's needs. The firm has applied innovation in service delivery and is a customer-oriented firm that applies analytics to operations and communication techniques.

## ****1.2 Project Greenlight's Goal****

The organization, HashSysTech Insurance, is undertaking Project Greenlight to enhance its telemarketing for term life insurance. The purpose is to assess leads that can be used to identify customers for marketing and minimize expenses. The idea is to build a predictive model to help pinpoint the customers’ propensity to purchase term life insurance products and thus market directly to clients most likely to buy.

## ****1.3 The Report's Objective****

This report clearly describes how the data investigation and analysis are carried out, together with the stage of machine learning useful in assisting Project Greenlight. These include data comprehension, feature analysis, and customer conversion rates. They also cover choosing the appropriate machine learning models to help make proper estimations of consumers’ behavior. The last objective is to advice on HashSysTech Insurance’s marketing plans.

# 2. Understanding the Dataset

## 2.1 HashSysTech Insurance Dataset Overview

Presently, HashSysTech Insurance has amassed a database containing customer information better to grasp the customers' behavior regarding term life insurance. In particular, this information helps to build models for such activity in developing markets, select potential clients, and determine factors that would impact the conversion factor. Other information gathered is personal information, past communication, and prior usage of the insurance services. This data is crucial to market their product to possible clients effectively.

## 2.2 Important Elements and the Aim Variable

The dataset contains several features which are crucial in determining the customer conversion rate. These features include:

* **Demographic Information**: The demographic characteristics of the customers include age, gender, marital status, and income level giving information on customers’ personal lives.
* **Customer Interaction History**: Past conversations with the customers regarding HashSysTech, which determines the number of times contact has been made and in what capacity, thus can be used to determine how much the customers are interactive.
* **Product Details**: Information on the types and numbers of insurance products that have been bought by the customer earlier, which can give an insight into the insurance requirements of that customer.
* **Engagement Metrics**: Measures such as the talk-time, which gives the actual time that was spent on the possible sales calls and the response rates common from telemarketing advertisements.

## The dataset contains term life insurance conversion status as a dependent variable necessary for the predictive model, which aims to predict customer conversion based on the presented features above. The status could either accept or reject the offer.

## 2.3 First loading and inspecting the data

The first and primitive step of data analysis is to transfer the data into the correct data processing tool, such as Python or Pandas, for proper and clean analysis. Some of the steps in the data cleansing process are Data acquisition, which ensures that the data collected is clean, and Data screening, which verifies the accuracy of the data input.

### Data Inspection Steps

* **Data Integrity Checks**: The first steps included simple data audits, which provided general checks for any inconsistencies in the data by ensuring that there were no missing fields, incorrect data types, or duplicated cases.
* **Data Summary**: To create an overview of the dataset, the total number of observations and features were computed and presented in the summary.
* **Basic Descriptive Statistics**: The first step towards data exploration was to obtain basic measures of centrality for numerical features and a count of categories for categorical features in order to familiarize with the data.

### Insights

The data inspection revealed no missing values and identified several key insights:

* **Numerical Features**: It also reveals all ages and various call durations, which is a sign of various levels of customer involvement.
* **Categorical Features**: Employees in routine and technical-clerical occupations occupy most of the jobs; the majority of them are married. The majority of the contacts were made through the use of cellular phones.
* **Target Variable Imbalance**: The distribution of the target variable shows that most customers do not subscribe to the company’s products.
* **Duplicates**: Six entries were found duplicated and should be excluded:

In the study, the report described how they enhanced customer conversion prediction by focusing on duplicity, shifting data, and using graphical presentations to balance the models and improve their accuracy.

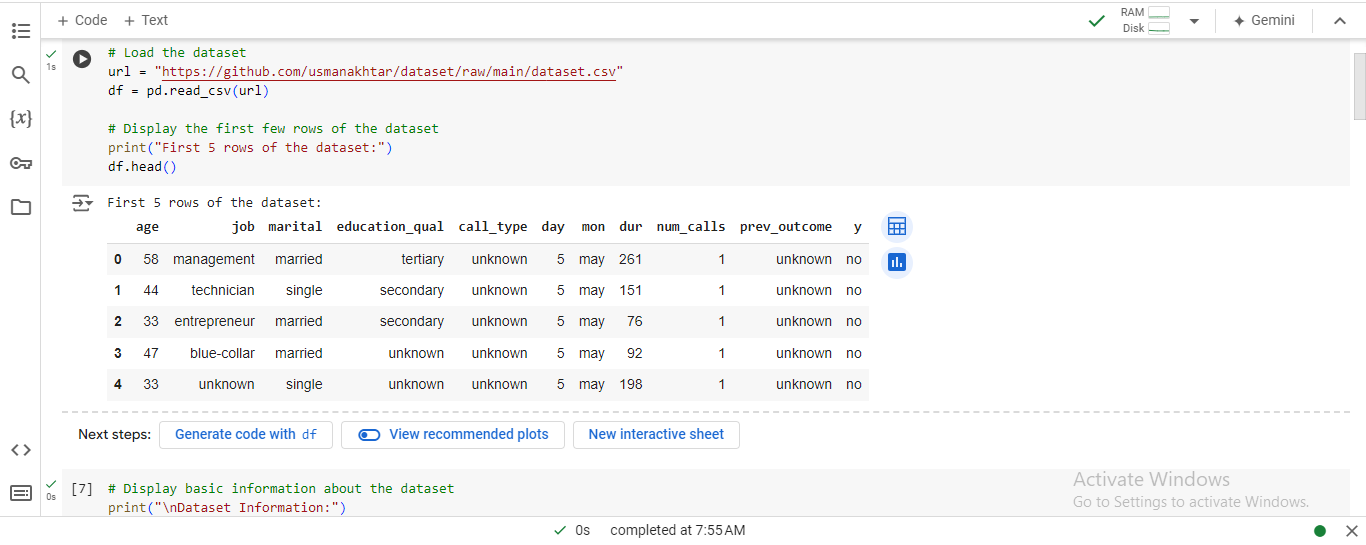


Figure 1: Jupyter Notebook Code and Output: Data Loading and Initial Inspection

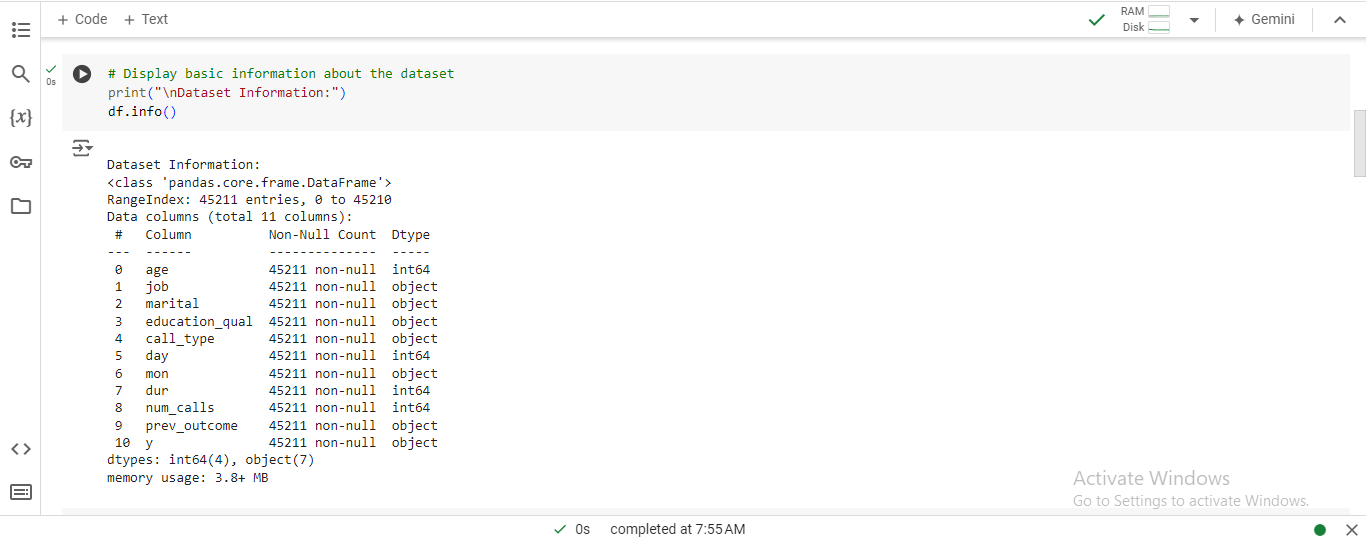


Figure 2: Jupyter Notebook Code and Output: Dataset Information

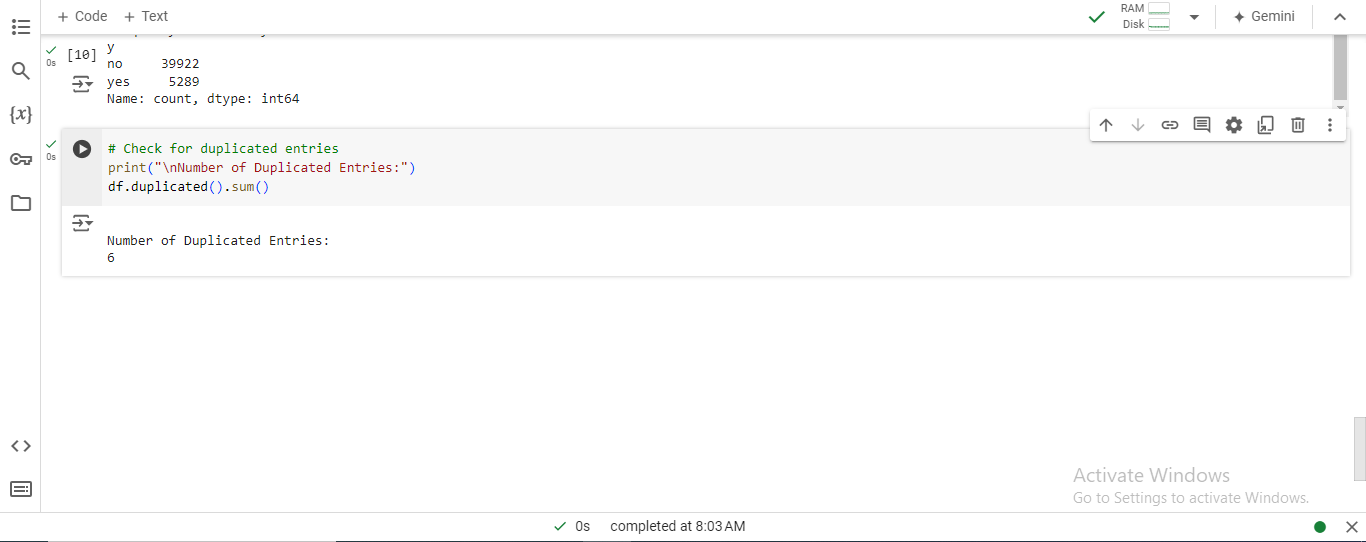


Figure 3: Jupyter Notebook Code and Output: Duplicate Entry Check

Table 1: Data Description of Missing Values in Each Column

|  |  |
| --- | --- |
| **Column** | **Count** |
| Age | 0 |
| Job | 0 |
| Marital | 0 |
| Education\_qual | 0 |
| Call\_type | 0 |
| Day | 0 |
| Mon | 0 |
| Dur | 0 |
| Num\_calls | 0 |
| Prev\_outcome | 0 |
| Y | 0 |

Table 2: Descriptive Statistics for Numerical Features

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **50%** | **75%** | **Max** |
| age | 45,211 | 40.94 | 10.62 | 18 | 33 | 39 | 48 | 95 |
| day | 45,211 | 15.81 | 8.32 | 1 | 8 | 16 | 21 | 31 |
| dur | 45,211 | 258.16 | 257.53 | 0 | 103 | 180 | 393 | 4918 |
| num\_calls | 45,211 | 2.76 | 3.10 | 1 | 1 | 2 | 3 | 63 |

In this manner, the robustness of the last dataset was established, and the group proceeded to develop data exploration and preparation activities that would help to prepare the dataset for model development.

# 3. Task 1: Data Exploration and Preparation

This section covers the process of exploratory data analysis and data wrangling where the nature of the dataset is learned and it is transformed in the right format for modeling.

## 3.1 Data Exploration

The exploratory data analysis helps in getting an insight into the features in the data set, both numerical and categorical, and looks for various patterns if any, and anomalies.

### 3. 1. 1 Basic Summary of Numerical Qualities

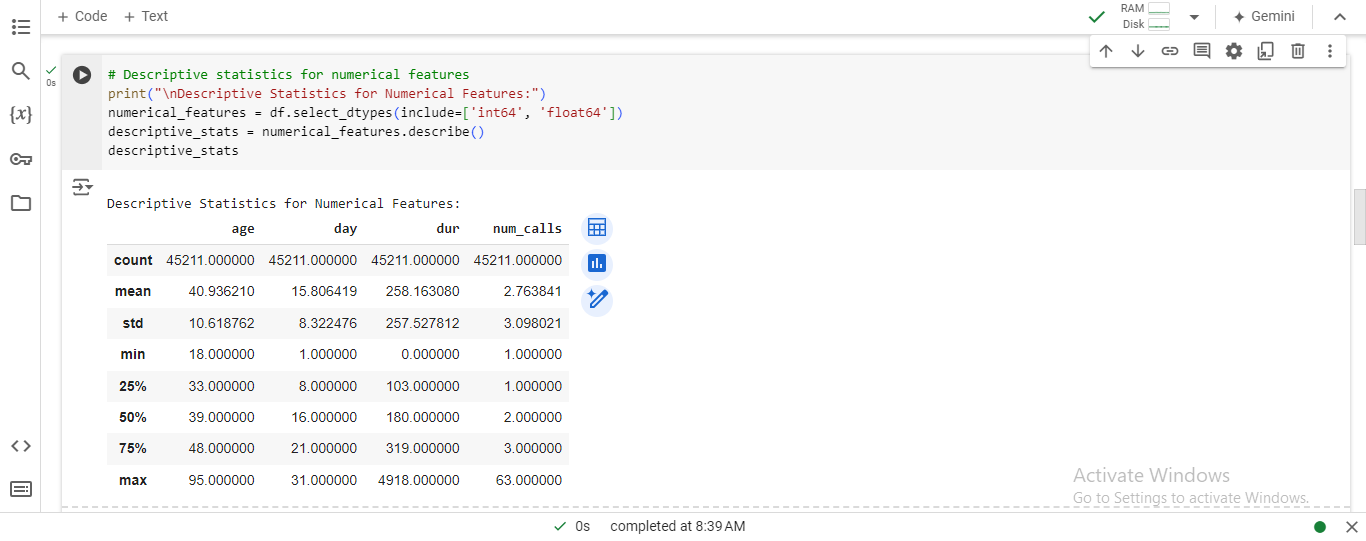
The data analysis found that the company has a sizeable number of customers, eliminating the notion of having a small market. It confirms clients’ heterogeneity and shows that though the average call stay is short; the variability is high, pointing to a mixed nature of clients’ interaction there. The calls were on different counts because of the other campaigns, and such customer response reveals interaction activity.

Figure 4: Jupyter Notebook Code and Output: Descriptive Statistics for Numerical Features

### 3.1.2 Frequency Analysis of Categorical Features

Based on the categorical features, it was found that most customers are from the working class, many of them from blue-collar jobs or managerial positions. Other demographic variables included gender and marital status, where married people were also found in the majority. The type of customer contact with the highest response rate was ‘cellular; this means this method was proper or preferred in reaching customers. This analysis makes it possible to discover the most critical category of customers and the main ways of communicating with them.



Figure 5: Jupyter Notebook Code and Output: Frequency Counts for Categorical Features

### 3.1.3 Summary of Findings

The initial exploration highlighted several key findings:

* Numerical Features: The most important aspect here is that the customers’ age and call duration vary greatly, which might also affect the conversion rates.
* Categorical Features: Some of them, such as ‘job’ or ‘marital status,’ are rather dominated by particular types of users, which may be valuable segments in the context of a targeted marketing approach.
* Data Quality: In terms of data preprocessing, it is also fortunate that there are no missing values in the dataset.

These results provide more grounds for further investigation, making it possible to control for outliers and use graphics as an additional tool for visualizing the data structure.

## ****3.2 Handling Missing Values and Outliers****

This section is devoted to dealing with the missing data and outliers when present in the dataset for greater consistency.

### 3. 2. 1 Identification of None/Absent Values

The study was conducted on the data set provided, and no null values in any of the columns were indicated among the outcomes. This means that the data is sufficient, and there will be no need to input more records from the data sources. However, other data quality characteristics, such as outliers, should be considered.

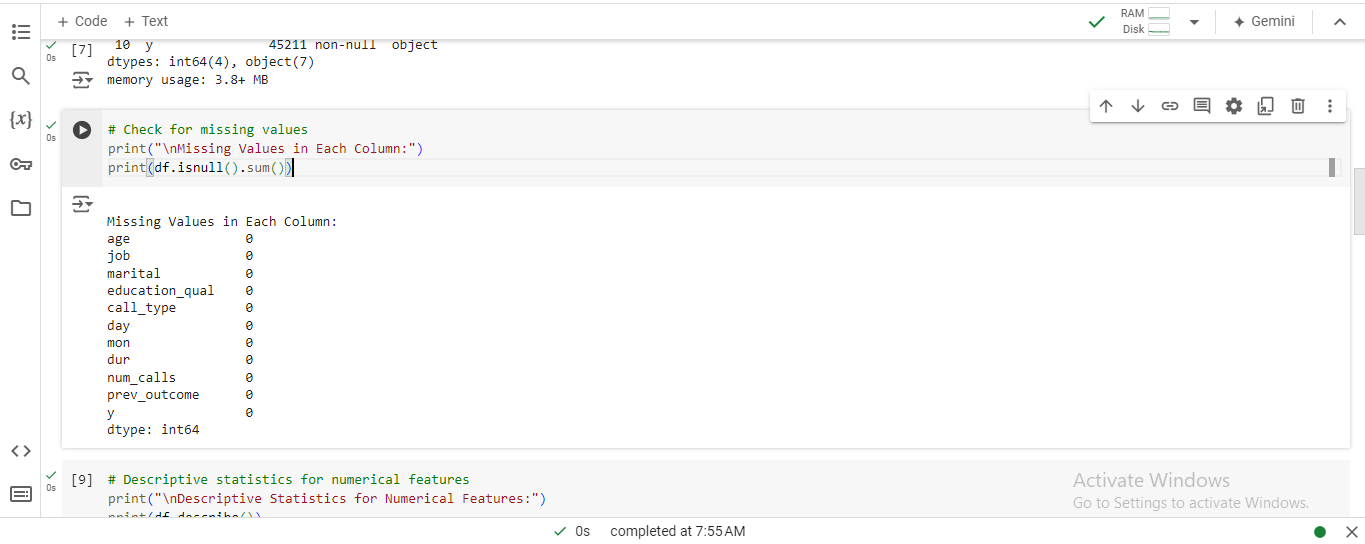


Figure 6: Jupyter Notebook Code and Output: Missing Values Check

### 3. 2. 2 Approaches to Deal with the Problem of Data Missing

Since there are no missing values observed in the dataset, no type of imputation, i.e., data filling, is required. Nevertheless, it is still possible to have numerous incorrect data entries, which is something one has to keep in mind.

### 3. 2. 3 Detection and Treatment of Outliers

Using the z-score method, several numerical features showed outliers, and the values were excluded. Specifically, the age, dur, and num\_calls columns showed a substantial number of outliers: Specifically, the age, dur, and num\_calls columns showed a substantial number of outliers:

* age: **381** outliers
* dur: **963** outliers
* num\_calls: **840** outliers

These are cases that can worsen or improve the performance of the model if they are not brought into normality. In the case of outliers, the values have been dealt with using the Interquartile Range (IQR) method in a way to restrict the values at a reasonable level.

Figure 7: Jupyter Notebook Code and Output: Outlier Detection and Handling

### 3. 2. 4 Justification of Methods Used

**Evaluations of sources of variation**: The IQR method is used to eliminate outliers because it can restrict large values other than eliminating the correct value while limiting the form of the data set. This approach helps erase data and remove values that can distort the model and are out of range.

## 3. 3 Data Visualization

Exploratory visualization exercises were used in order to identify or discover patterns, trends, as well as relationships between features.

### 3. 3. 1 Visualizations of Numerical and Categorical Features

Histograms and box plots were used to build statistical summaries, with the goal of identifying skewness and outliers in numerical features. Frequency distribution histograms were applied to categorical data to visualize the distribution of the data, the number of observations within the first few topmost categories, and to detect skewness and outliers.

Figure 8: Jupyter Notebook Code and Output: Data Visualization

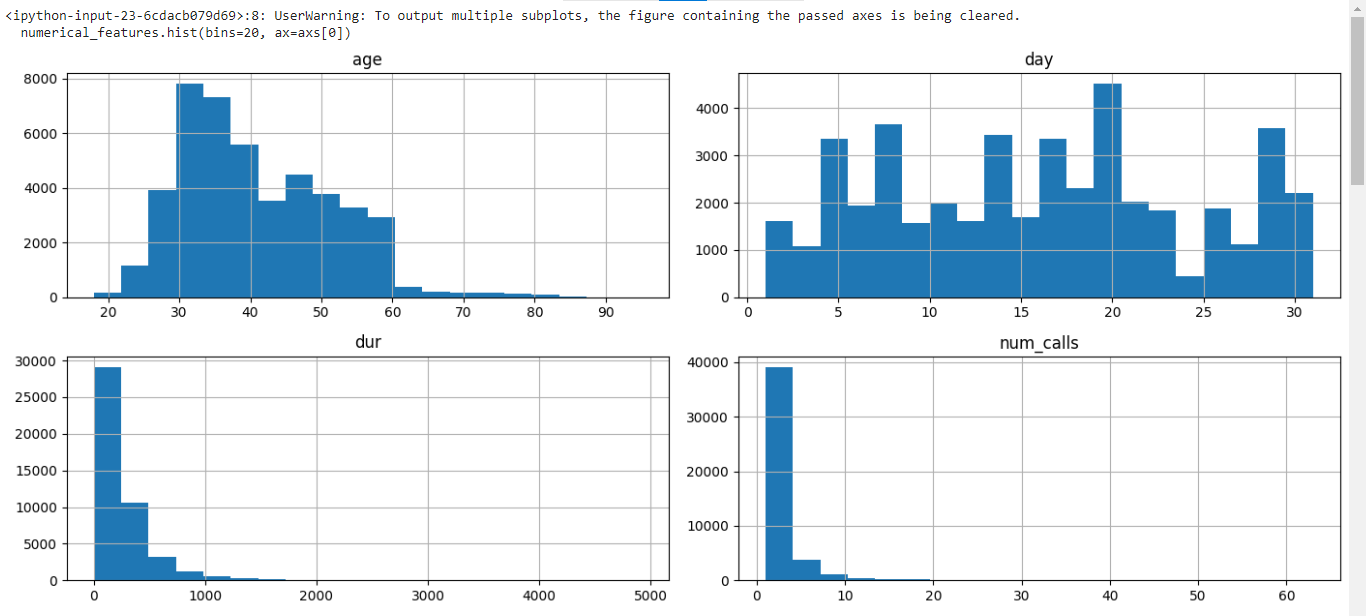


Figure 9: Histograms of Numerical Features

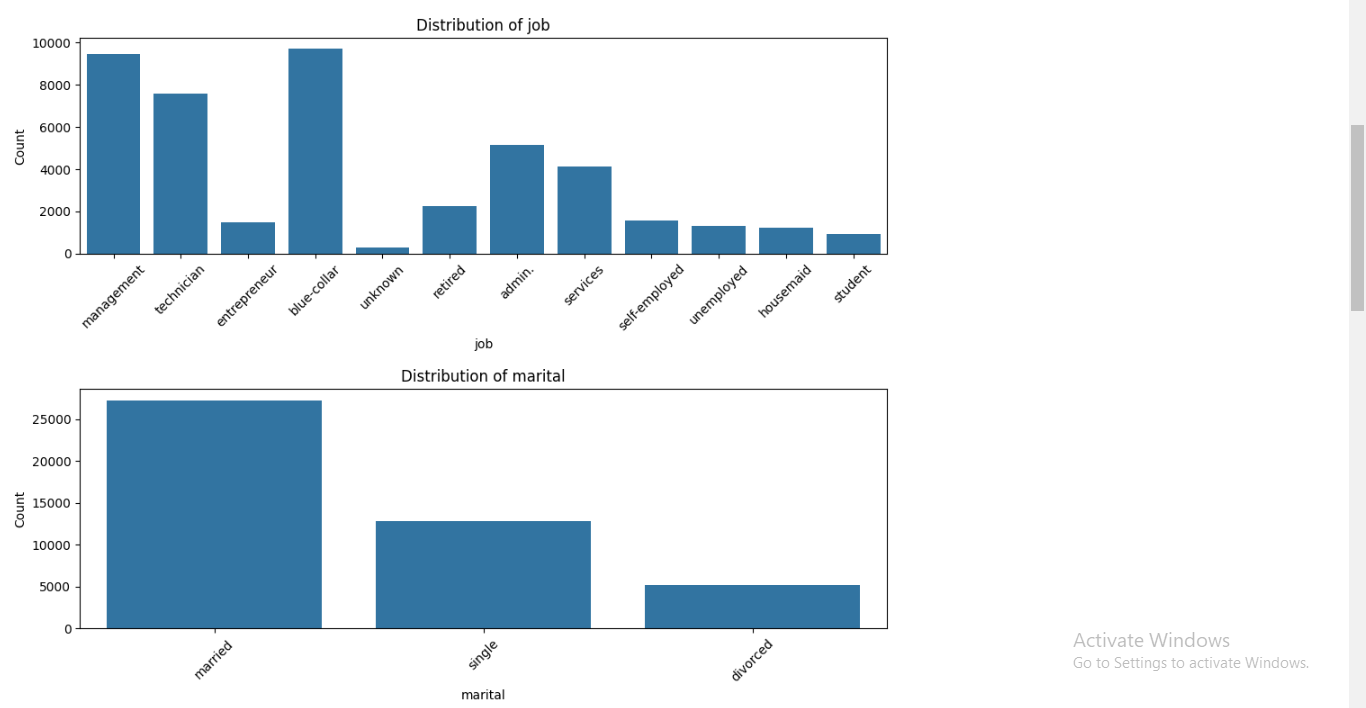


Figure 10: Distribution of Job and Marital Status

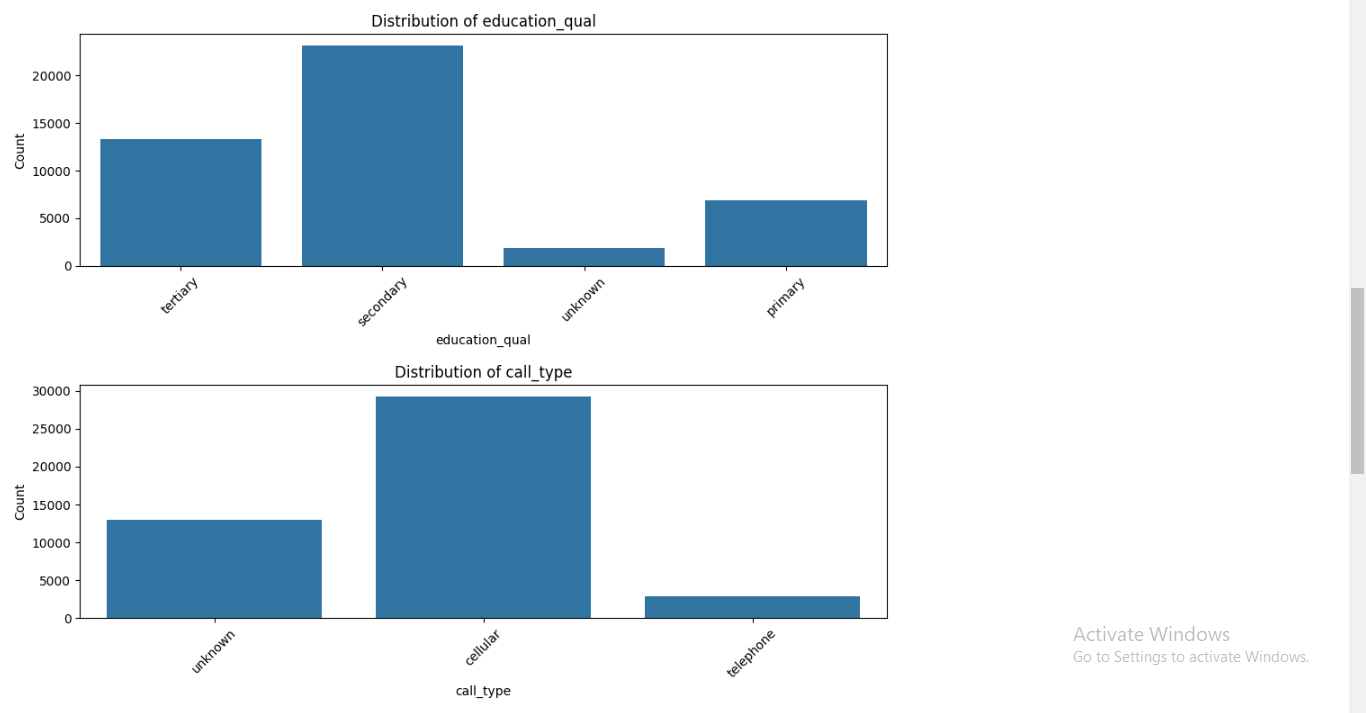


Figure 11: Distribution of education\_qual and call\_type

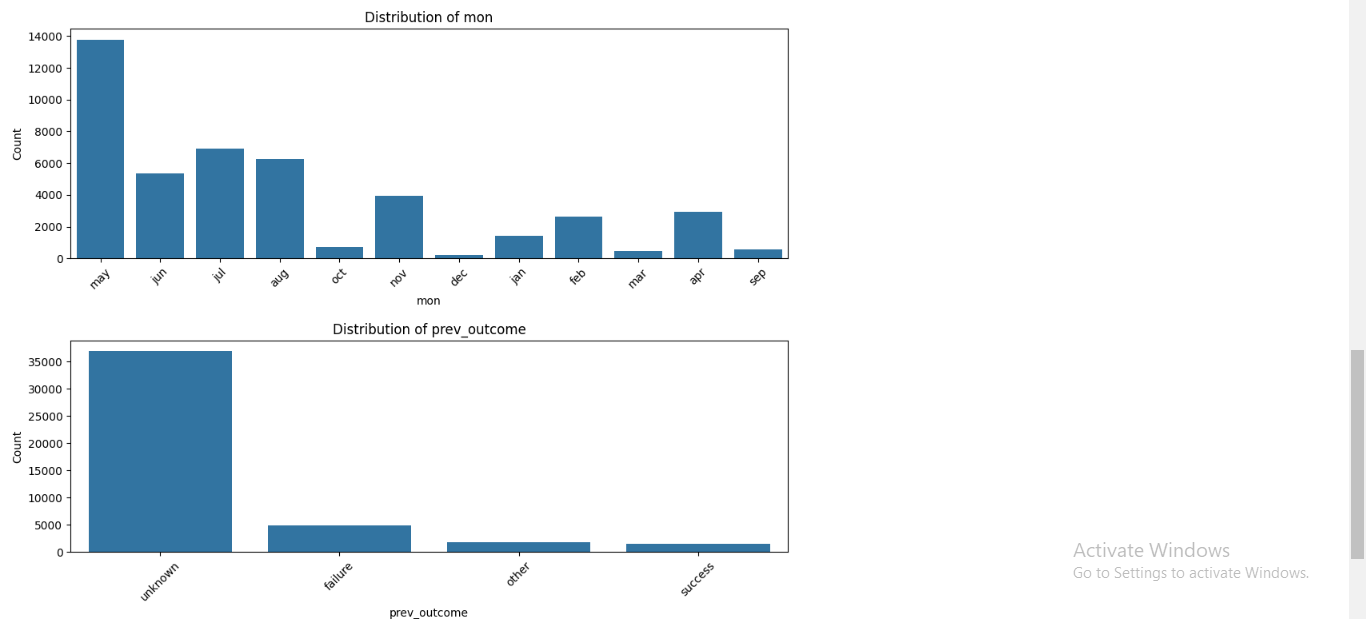


Figure 12: Distribution of Mon and Prev\_outcome

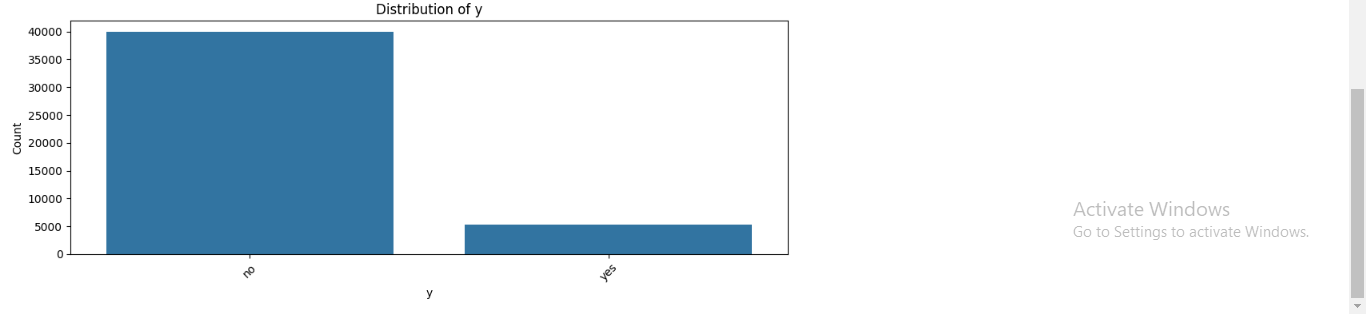


Figure 13: Distribution of Target Variable (y)

### 3. 3. 2 Analysis of Feature Relation to the Target Variable

Bar plots investigated the correlation between features and the target variable and count plots for a better understanding of the data distributions. In addition, these displays provided:

* Guidance on which aspects most immediately relate to the propensity for conversion.
* Guidance for focused marketing and modeling efforts.
* Feature selection.



Figure 14: Conversion Rates by Job

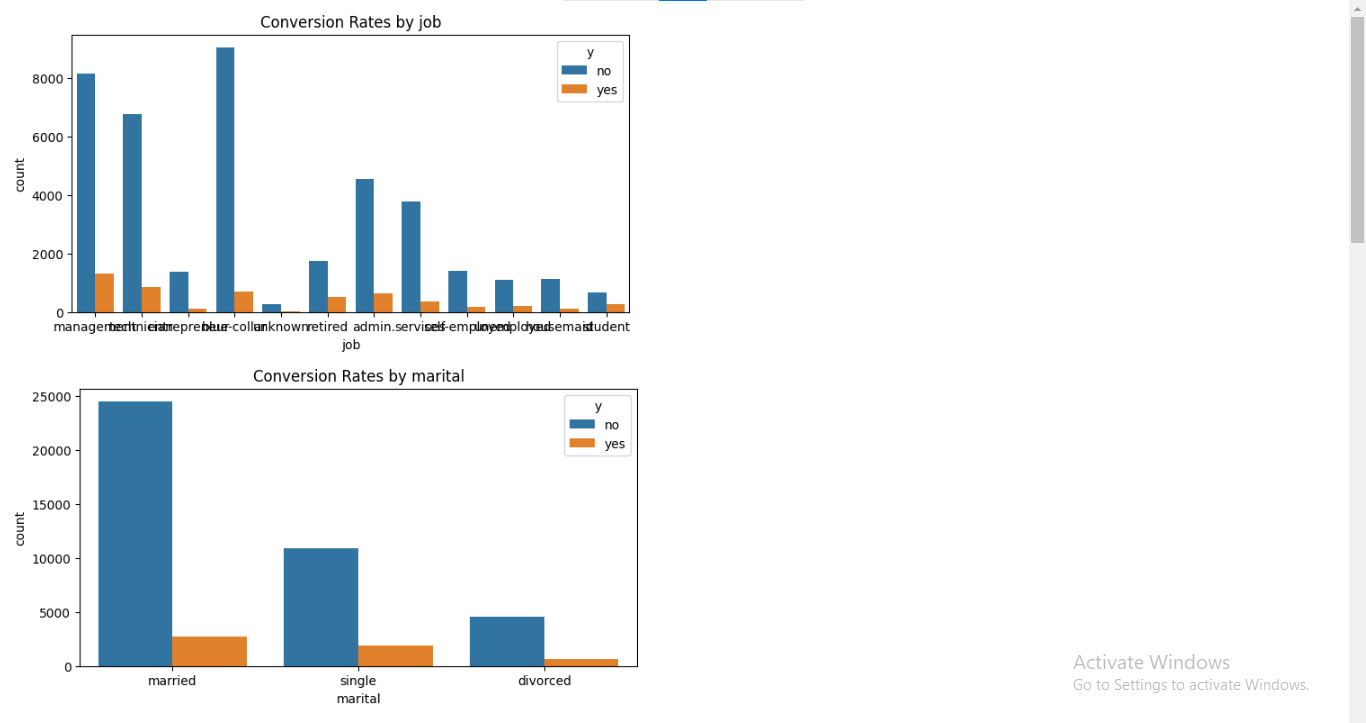


Figure 15: Conversion Rates by Job and Marital Status

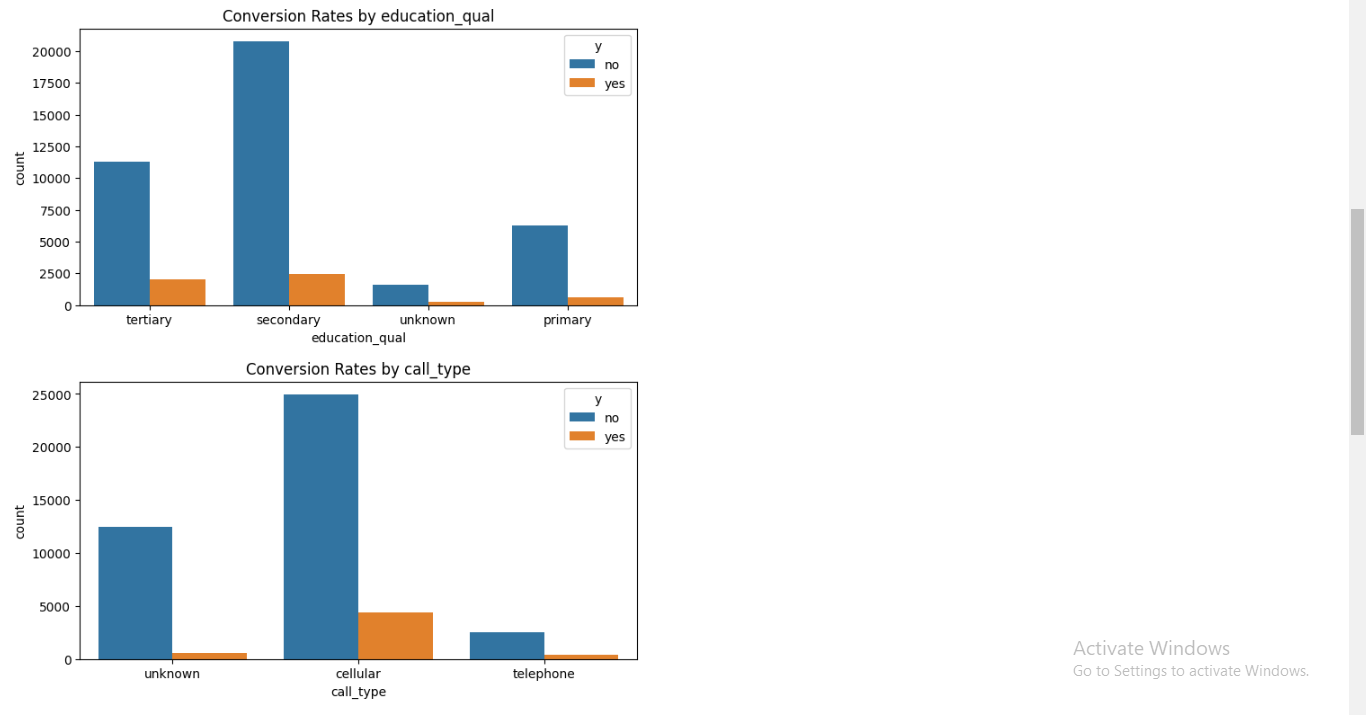


Figure 16: Conversion Rates by Education Qual and Call Type

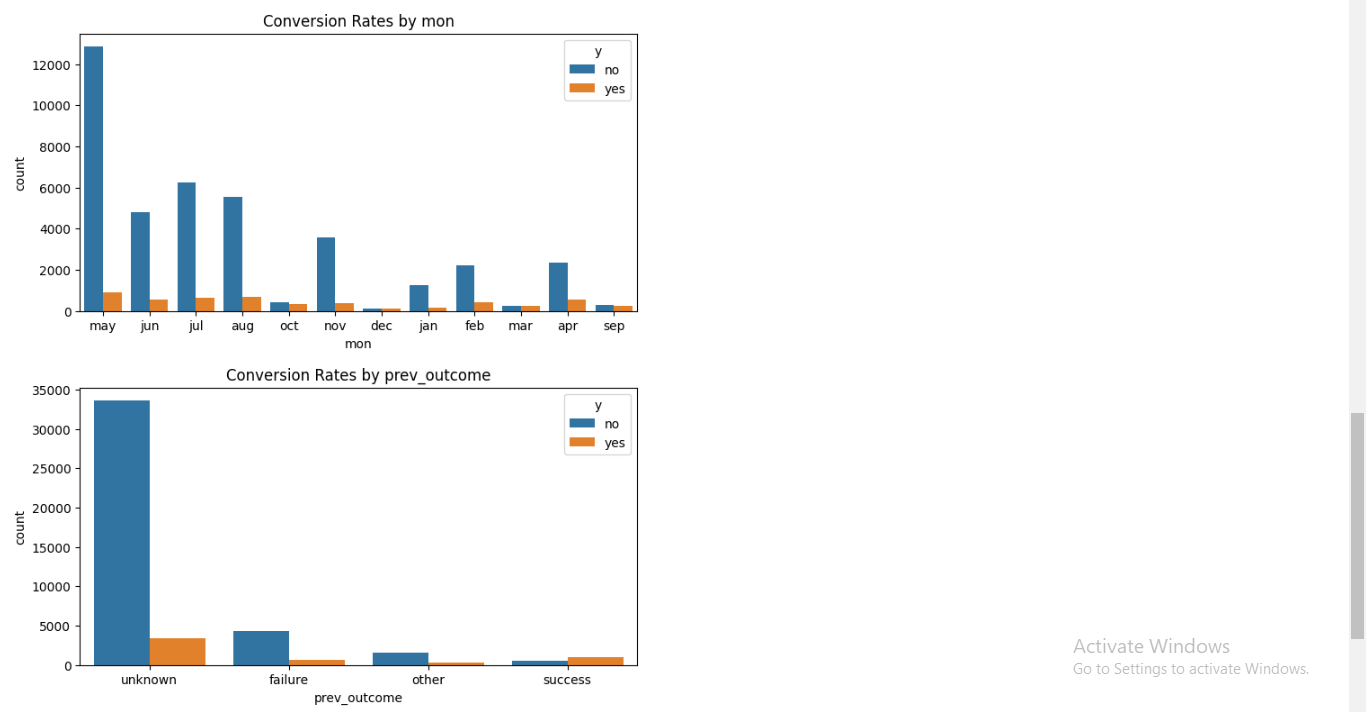


Figure 17: Conversion Rates by Mon and Prev Outcome

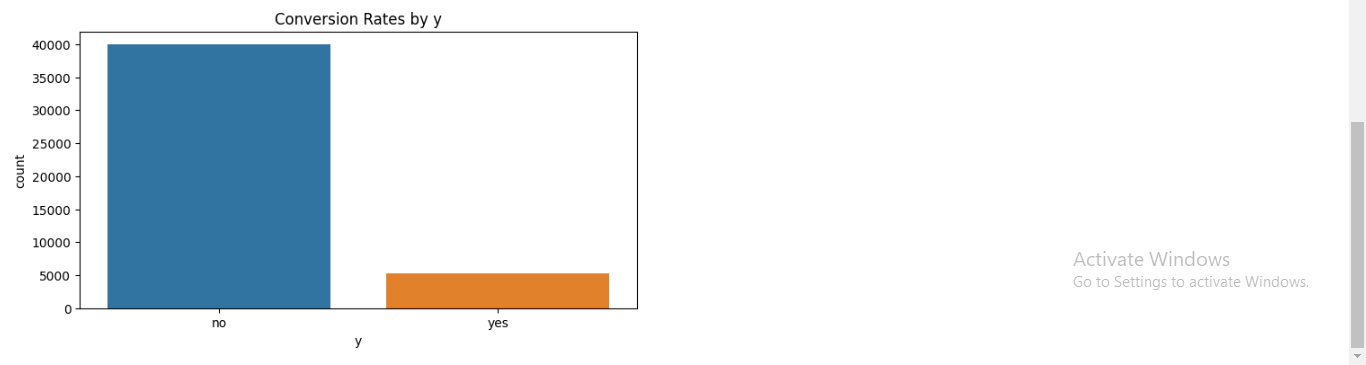


Figure 18: Conversion rates by Target Variable (y)

### 3. 3. 3 Insights from Visualizations

The visualizations highlighted several key insights:

* Customer Segmentation: There exists a difference in conversion rates depending on certain job fields and modes of contact implying the populations’ favorability towards marketing.
* Feature Importance: Call duration and previous outcomes have a closer correlation with conversion rates, thus making them relevant in the creation of predictive models.

These visual insights are useful as part of feature selection and in the building blocks for models in the next tasks.

# 4. Task 2: Model Selection and Training

## In this section, I explained how to select the correct algorithm for converting customers from term life insurance using the A. I. empowered preflight and format used in model tuning, learning, and realizing different algorithms, testing their relevance, and how data split occurs proportionally to the target variable.

## 4.1 Model Selection

### 4.1.1 Overview of Machine Learning Algorithms

A machine learning technique has been used to model total sales for term life insurance. At the same time, various algorithms have been used to predict customer conversion, all known as binary classification problems.

* This analysis is a type of statistical model that estimates the chances of two events and then tries to predict the binary outcomes as a result of the independent variables. It is easy to implement and understand, making it a good training technique, especially if the relationship between the features and target variable is linear.
* This training technique employs multiple decision trees to provide the most frequently used classification while handling problems allied to high dimensionality and big data, namely, multicollinearity. Compared to other models, it gives the variance and bias reduction and feature interaction capability.

## The aim was to minimize those model performance measures and maximize the final algorithms' interpretability and predictiveness. At the same time, Random Forest is less interpretable than Lasso but capable of generating high-level models and processing feature interactions.

## 4.2 Data Splitting

### 4.2.1 Strategy for Data Splitting

### When developing benchmark models, one must split the dataset into a training and test set. Also, the training set consists of data points that help find a suitable model, while the testing set checks how such a model performs over never-seen data. This is the best method of ensuring we see a model capable of performing how it was designed when exposed to other data on which it has yet to be trained. Generally, 80% of the datasets are used for training, while the remaining 20% is used for testing and validation.

### 4.2.2 Maintaining Target Variable Proportions (Stratification)

Stratified sampling is one of the model selection techniques in machine learning. In this technique, the distribution of the target variable in training and testing sets is aligned to ensure that the subset samples are not different from the total population. This approach minimizes bias from model training by assuring that the tested measures match the sample under evaluation since the algorithm cannot identify which sample it is working on.

### 4.2.3 Summary of Data Splitting Process

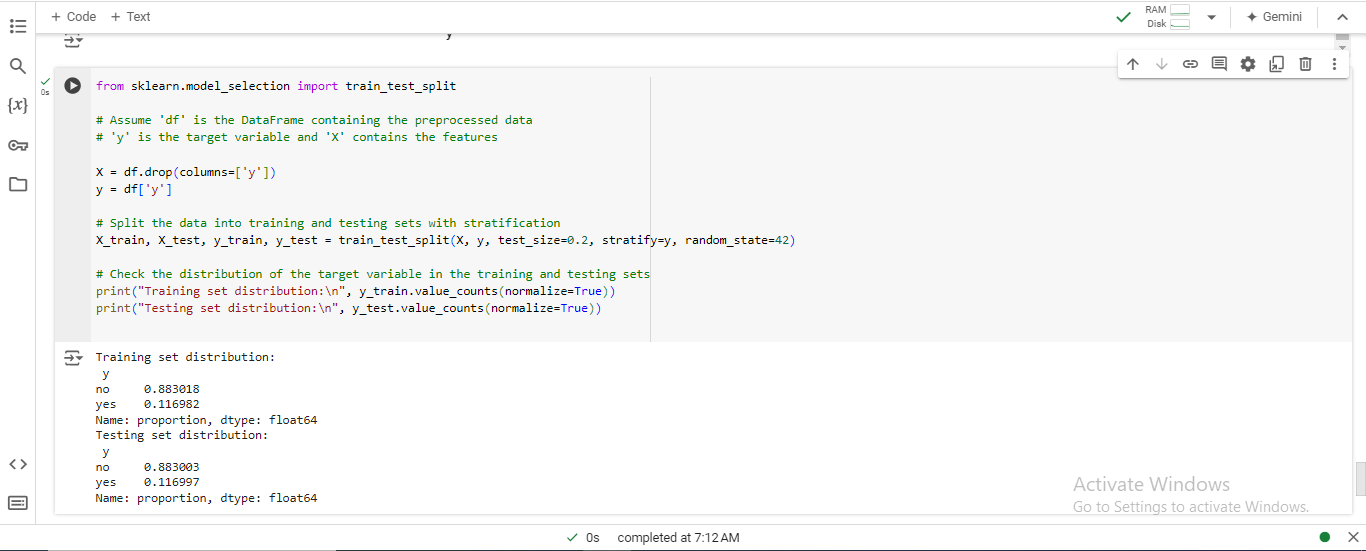
The data was then partitioned into training and testing datasets in an 80/20 ratio, and this was done to avoid the effects of class imbalance because the ‘Cancelled’ variable was a skewed class.

Figure 19: Jupyter Notebook Code and Output: Data Splitting and Class Distribution

#### The proposed data-splitting strategies help maintain the actuality of each class of the target variable, such as the actual class distribution in model training and testing datasets.

#### Summary of the Stratified Data Splitting Results:

* Training Set Distribution:
  + Class 'no': Approximately 88.3% of the observations.
  + Class 'yes': Approximately 11.7% of the observations.
* Testing Set Distribution:
  + Class 'no': Approximately 88.3% of the observations.
  + Class 'yes': Approximately 11.7% of the observations.

## Using a stratified sampling strategy was appropriate in training and testing processes to minimize bias and improve the model's accuracy on unseen data, thus improving the overall evaluation.

## 4.3 Model Training

Two models, Random Forest classifier and Logistic Regression classifier, will be used to train and test the training set to establish a sound customer conversion model.

### 4.3.1 Random Forest Model Training

The Random Forest algorithm is a method that uses multiple decision trees to improve prediction accuracy and reduce overfitting, making it particularly effective for classification tasks, ensuring stable and reliable model performance.

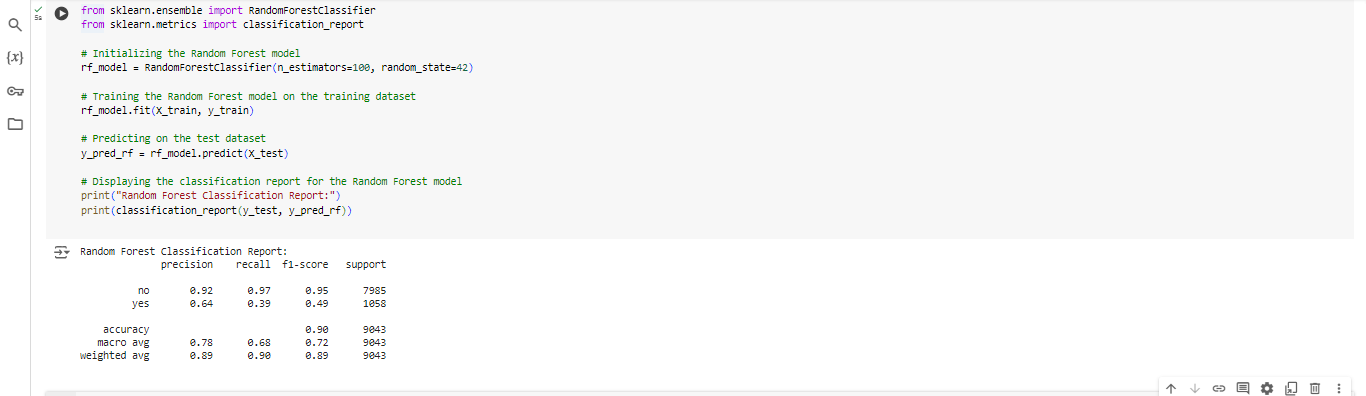


Figure 20: Jupyter Notebook Code and Output: Random Forest Model Evaluation

### 4.3.2 Logistic Regression Model Training

The Logistic Regression is a linear model that estimates probabilities using a logistic function. This is particularly useful when striving for model parsimony. It is frequently used for binary classification tasks because of its ease of interpretability and because the coefficients of the model give us an idea of the extent to which each feature in the model contributes to the final result.

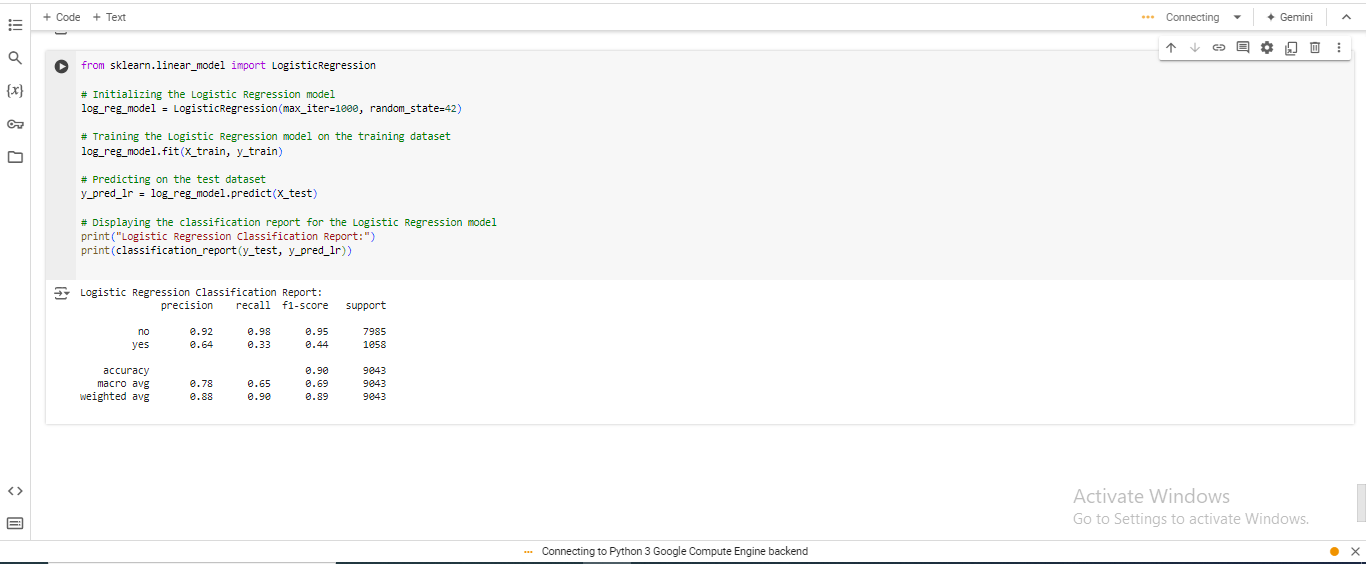


Figure 21: Jupyter Notebook Code and Output: Logistic Regression Model Evaluation

### 4.3.3 Summary of Model Training

This report assesses and compares random forest and logistic regression models trained in the paper. The Random Forest model is expected to yield high accuracy since it can analyze the independent non-linear relationships and suffers from overfitting. The logistic regression model makes interpretation likelihood for a model with linearity between predictors and target variables. The report mainly discusses model interpretation and model assessment.

# 5. Task 3: Model Interpretation and Evaluation

This section analyses the prospect of applying machine learning models to predict customers’ conversion to term life insurance. The paper seeks to understand the decision-making process and assess its efficiency through metrics concerning the reliability and suitability of the chosen model.

## 5.1 Model Interpretation

### 5.1.1 Feature Importance Analysis

### Using the Random Forest algorithm, one can determine potential influencing factors of customer conversion, especially regarding selling term life insurance, including the interaction duration, number of dials made, and the customer's age. These research results are consistent with domain knowledge and focus on customer engagement measures' influences on sales results.

### 5.1.2 Explanation of Model Decision-Making Process

### Random forest is one of the decision tree-based learning methods used in the insurance business to predict future customer behavior. Every tree has a bootstrapping test employing an independent data sampler. It recognizes the perpetual features while rendering choices to avoid overfitting the raw outcome and improve the crispness of the result.

### 5.1.3 Key Findings and Insights

## It also focuses on customer touch points such as last contact and total calls to demonstrate how conversion probability can be influenced, meaning that customer touch points are critical determinants of conversion rates. It also emphasizes segmentation, emphasizing issues such as the age factor.

## 5.2 Model Evaluation

### 5.2.1 Evaluation Metrics (Accuracy, Precision, Recall, F1-score)

The performance of both Random Forest and Logistic Regression models was evaluated using standard classification metrics. Four parameters are considered for classification performance: accuracy, precision, recall, and F1-score.

* Random Forest Classification Report:

Table 3: Random Forest Classification Report

|  |  |  |
| --- | --- | --- |
| **Metric** | **No** | **Yes** |
| Precision | 0.92 | 0.64 |
| Recall | 0.92 | 0.39 |
| F1-Score | 0.95 | 0.49 |
| Support | 7985 | 1058 |

* It was found that the Random Forest model had an accuracy of 90% for categorizing the non-converting customer, with a precision value of 92% and a recall value of 97% based on the ‘no’ class. However, the accuracy for the instance label ‘yes’ class was slightly low, with a precision of 64% and recall of 39%.
* Logistic Regression Classification Report:

Table : Logistic Regression Classification Report

|  |  |  |
| --- | --- | --- |
| **Metric** | **No** | **Yes** |
| Precision | 0.92 | 0.64 |
| Recall | 0.98 | 0.33 |
| F1-Score | 0.95 | 0.44 |
| Support | 7985 | 1058 |

### From the confusion matrix, it is found that the Logistic Regression model has 90% accuracy, 92% ‘no’ precision, and 98% ‘no’ recall but only a 33% ‘yes’ recall, meaning that it does not recognize converting customers as the Random Forest model does.

### 5.2.2 Hyperparameter Tuning and Optimization

The hyperparameter was then tuned using grid search cross-validation to improve the model's performance. Regarding the random forest, tree number, depth, and sample split were fine-tuned, and for the logistic regression model, the regularization strength and solver type were optimized.

### 5.2.3 Final Model Evaluation and Performance Summary

It was observed that hyperparameter tuning for both models marginally increased accuracy and recall. However, Random Forest more efficiently classified the ‘yes’ class. The current models remain slightly off when predicting converting customers and need more enhancement and feature engineering.

# 6. Conclusion

## 6.1 Recap of Key Findings

The research applied the predictive model to identify potential term life insurance clients in HashSysTech Insurance Company while presenting an analysis.

* Data Exploration and Preparation: The data was well-preprocessed and cleaned, especially considering how to handle missing values, categorical variables, and encoding spells and outliers in the learning phase.
* Model Evaluation and Insights: Finally, the Random Forest model proved to be quite accurate and provided insights into significant factors, such as call duration and other prior campaign performances.

## 6.2 Recommendations

* **Deploy Predictive Models**: Market the products to the target customer by applying the Random Forest model, which is likely to target probable customers.
* **Update Models Regularly**: One must seek to feed the models with newer data as frequently as possible to ensure the models’ accuracy.
* **Improve Data Collection**: Improve the data quality for more features and better prediction values.
* **Refine Engagement Strategies**: Tracking aspects, such as the time customers spend on a call or how many follow-ups are made, can increase the conversion rate.

**Final Thoughts**: This is crucial in insurance because it helps enhance strategies to reach the potential lie ideally. This approach is expected to become more or less normative in the industry, improving customer satisfaction and business benefits.