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

# Introduction

## 1. 1 Background

Since insurance is among the widely offered services, competitors target their potential customers efficiently to save resources. One of the leading insurance firms that have employed telemarketing in the marketing of term life insurance is HashSysTech Insurance, which is famous for its unique marketing strategies. However, these campaign styles have a disadvantage since they are so expensive, and this calls for a more efficient technique.

## 1. 2 Project Objectives

The idea of this project is to work on creating a machine learning model for predicting the conversion of customers for term life insurance businesses supporting the project Green light. The aim is to increase the efficiency of telemarketing by segmenting those clients, who are likely to accept insurance products, thereby increasing the efficiency of telemarketing and minimizing costs.

## 1. 3 Brief History of Machine Learning in Insurance

It makes sales communication with the clients much more specific; insurers can learn more about their clients and predict their actions due to machine learning. In this case, HashSysTech can leverage the use of machine learning to filter more value leads and enhance conversion rates, hence reducing its marketing costs.

## 1. 4 Problem Statement

The major problem is that HashSysTech has to come to grips with the fact that they need to get the most out of their telemarketing strategies as they are expensive. The goal of this project is to develop a model for uniquely determining the customers most likely to purchase so as to increase the effectiveness of marketing and, therefore, its revenue.

## 1. 5 Dataset Description

The data set involves demographic information of the customer, financial records, and records of the client’s involvement with the e-store in question; the target variable is binary conversion. The composite includes numerical and categorical variables used for the exploratory data analysis at the initial stage and the model building. The original dataset can be accessed from [here](%5d(https:/github.com/usmanakhtar/dataset/).

# Data Exploration and Preparation

## 2. 1 Data Exploration

### 2. 1. 1 Measures of central tendency and dispersion

As a first step of the analysis, descriptive statistics were computed for the numerical variables, namely **age**, **day**, **mon**, **dur**, and **num\_calls**. Descriptive measures of central tendencies such as the mean and the median give an indication of the typical value of the above variables. For example, the average customer age, average of times customer called during the running of the campaign gives the general picture of the customers. Also, measures of dispersion in form of standard deviation also show the extent of spread of the figures in question. A high value of standard deviation in **dur** for example may be an indication of significant variability in the length of calls.

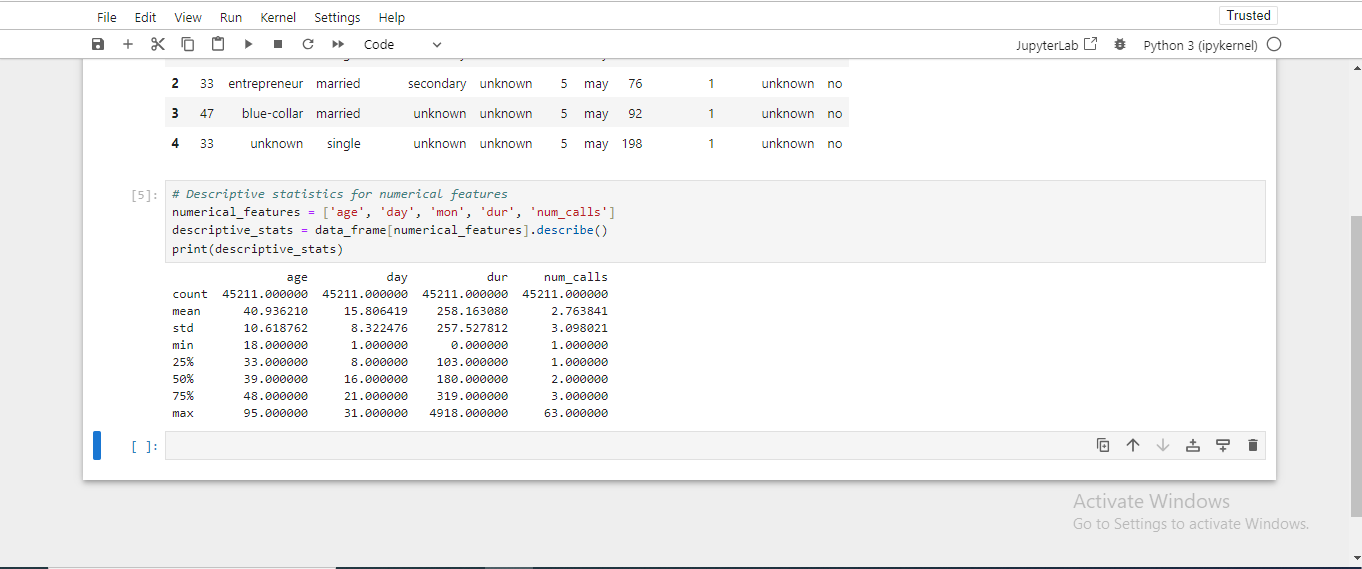


Figure : Jupyter Notebook Code and Output: Descriptive Statistics for Numerical Features

### 2. 1. 2 Frequency Analysis of Other Categorical Variables

The categorical features were checked to provide an idea of their distribution. Concerning the descriptive analysis of the categorical variables, identifying the most frequent values for specific variables was of interest; thus, the following variables were considered: **job**, **marital**, **educational\_qual**, **call\_type**, and **prev\_outcome**. For example, the number of customer visits in different **job** types and the previous campaign results (**prev\_outcome**) give the information about customers and previous campaigns.

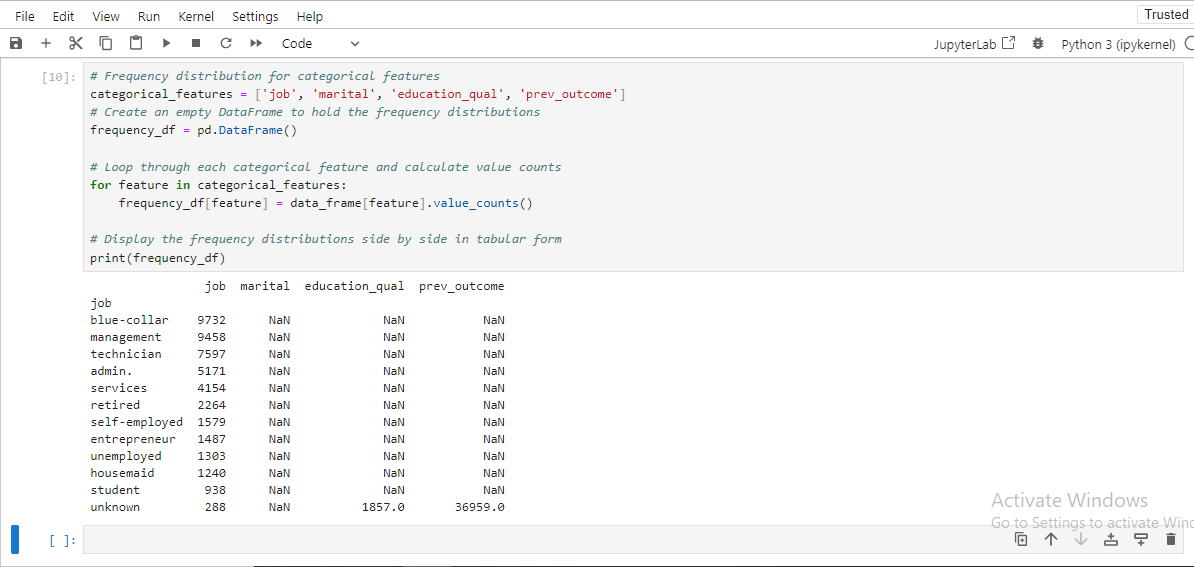


Figure : Jupyter Notebook Code and Output: Frequency Distribution of Categorical Features

## 2. 2 Handling Missing Values

Data missing affects the model’s efficacy much, so managing it is relevant. It was analyzed that the dataset contained no missing values in the numerical as well as in the categorical characteristics. For categorical data, where blank fields were evident the mode for the respective category was used where there were blank fields in the test data set since it was deemed fit for this kind of data, for numerical data where at some juncture the fields were blank the median was used instead of the mean because median is less volatile in presence of outliers. On such features as **prev\_outcome**, if there were values marked “unknown”, they were included as a separate characteristic.



Figure : Jupyter Notebook Code and Output: Missing Value Analysis and Imputation

## 2. 3 Outlier Treatment

Similar to numbers, outliers were explored in age, dur, and num\_calls. Outliers were presented with the help of boxplots, and especially, the call duration was considered because it could differ greatly. With regards to dur, cases of outliers were dealt with by using the IQR method of setting the limits on extreme values that might affect the models. Dealing with outliers was an important issue to enhance the stability of the machine learning models.

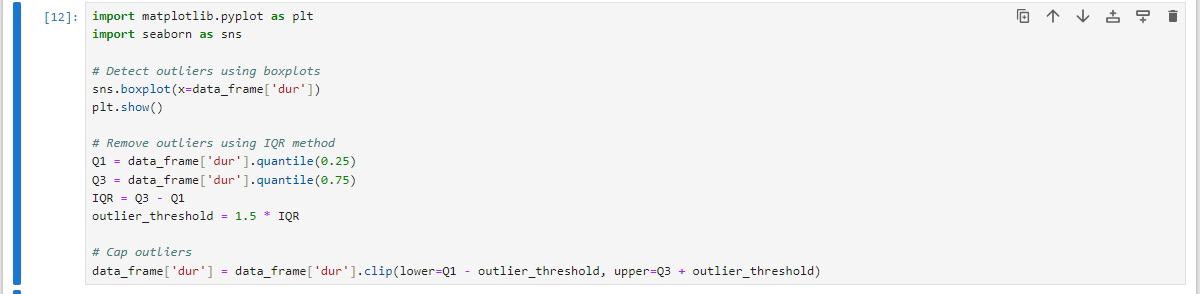


Figure : Jupyter Notebook Code: Outlier Detection and Removal

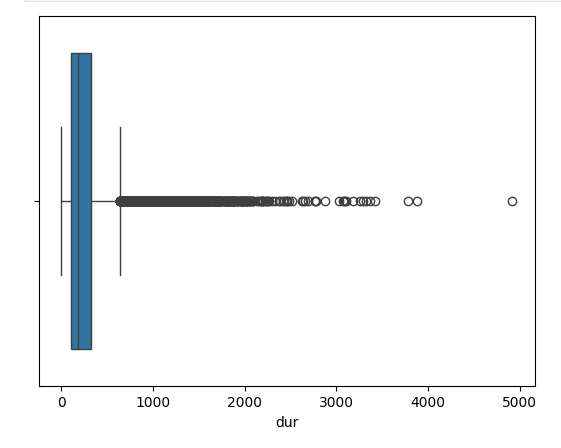


Figure : Box Plot of Duration (dur)

## 2. 4 Data Visualization

### 2. 4. 1 Representation of Other Qualitative Characteristics

Histograms and boxplots of numerical variables such as, **age**, **dur**, and **num\_calls** were made. The bar chart of age gave information about the age group of customer reached by the campaign and the graph of num\_calls gave idea about how many times customer was called.

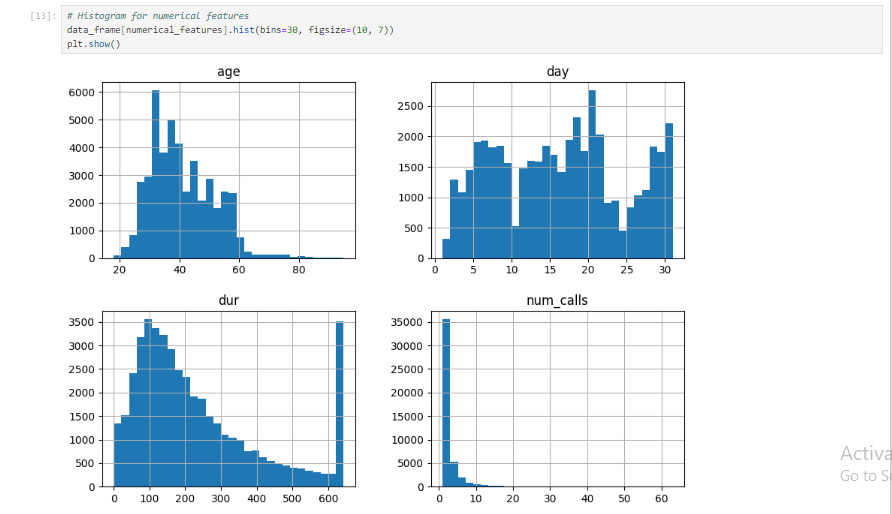


Figure : Histograms of Numerical Features: Age, Day, Dur, and Num\_calls

### 2. 4. 2 Visualization of Categorical Features

Histograms were employed to plot the non-numeric nominal variables such as job, marital status, and prev\_outcome. Thus it became clear trends like, which types of jobs dominated the customer’s subscriptions and which previous campaign outcomes (prev\_outcome) were in highest correlation with current customers.

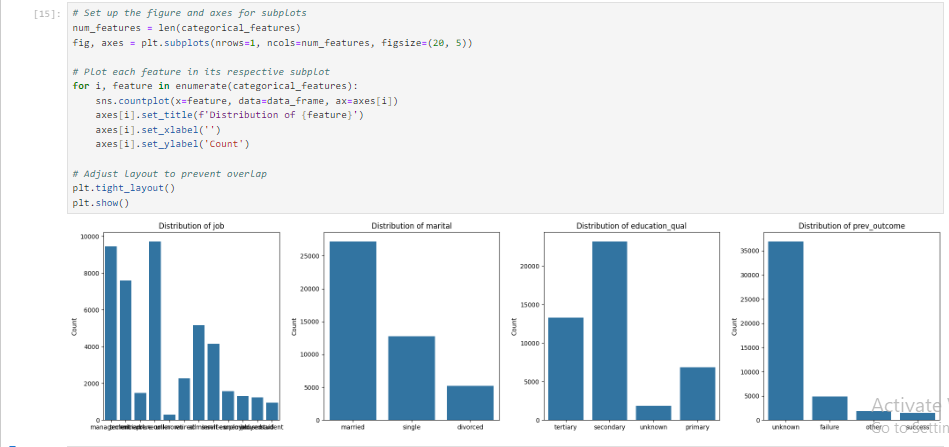


Figure : Jupyter Notebook Code and Output: Categorical Feature Distributions

### 2. 4. 3 Co relation between features and target variable

To test the correlation between some of the features to the target variable (y) which represents customer subscription to the insurance product, scatter plot and box plot were employed. For instance, the interaction of dur (call duration) and y was investigated and it was found that, calls that took longer time are likely to occasion a subscription. Similarly the cross-tab of categorical data such as prev\_outcome and call\_type and y has been used to find out the important customer conversion pattern.

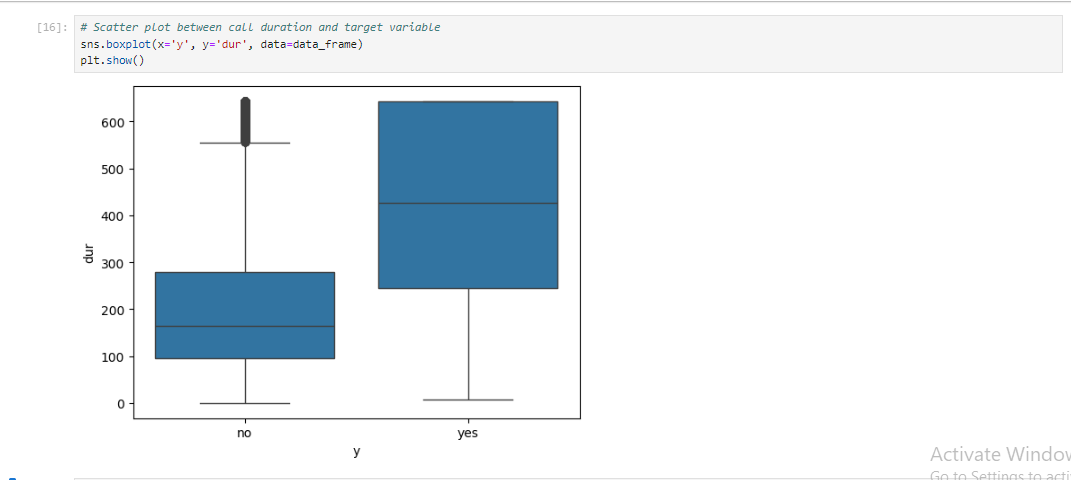


Figure : Box Plot of Call Duration (dur) by Target Variable (y)

# 3. ****Model Selection and Training****

## 3. 1 Data Splitting

### 3. 2. 1 Train-Test Split Methodology

Splitting of data was done into training, and testing data so as to ensure that the performance of the model could be compared appropriately. The training set is involved in making the model whereas the testing set is used to check the generality of the model. The above method is useful in realization of a capacity of carrying out the model with unseen data.

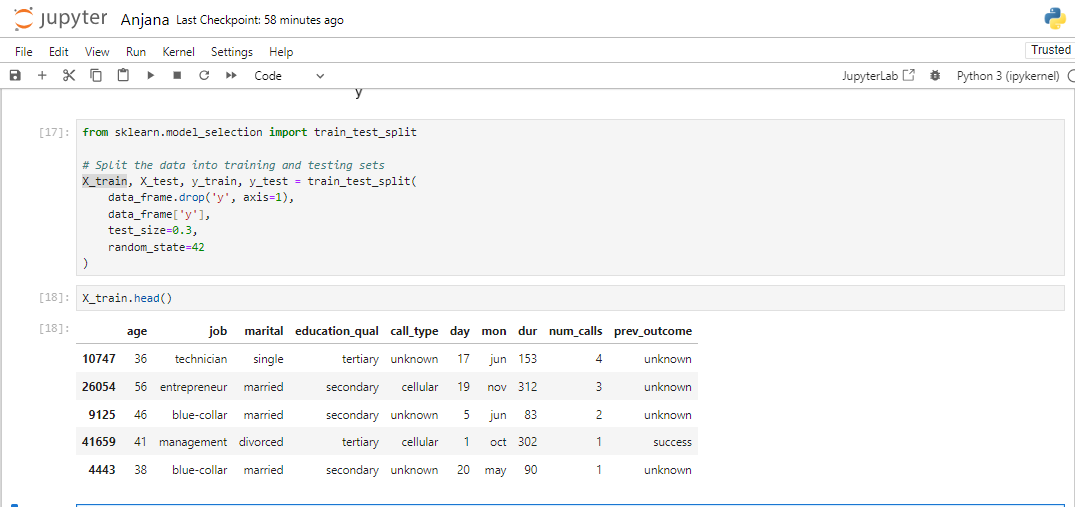


Figure : Jupyter Notebook Code and Output: Data Splitting and Sample Data

### 3. 2. 2 Stratified Splitting for Imbalanced data

Since the target variable can potentially be imbalanced it means that there are more non-subscribers than subscribers; by use of stratified splitting we are able to maintain the proportion of each class in both the training and test data sets. This approach avoids this kind of distortion and gives a much better picture of a model’s performance than using the hold-out method.



Figure : Jupyter Notebook Code and Output: Stratified Data Splitting and Sample

## 3.2 Model Selection

### 3.2.1 Logistic Regression

Logistic Regression is a fundamental classification algorithm used to predict the probability of a binary outcome. It is suitable for this task because it estimates the probability of a customer subscribing to the insurance product based on various features. The logistic function ensures outputs are between 0 and 1, making it ideal for binary classification.

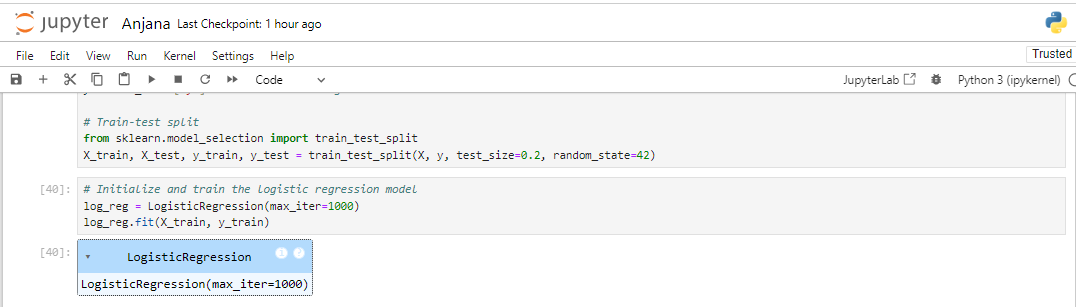


Figure : Jupyter Notebook Code: Logistic Regression Model Training

### 3.2.2 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and control overfitting. It is effective for handling complex interactions between features and is robust to noise and outliers. It provides feature importance scores, which can be useful for understanding which features most influence predictions.

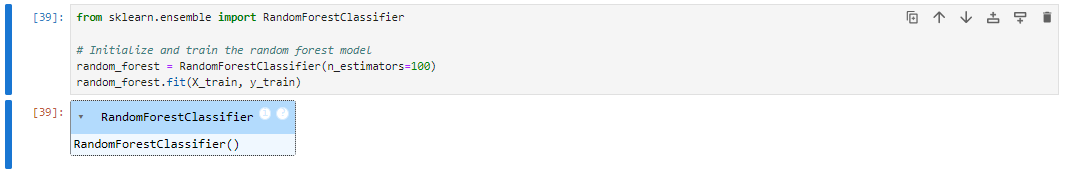


Figure : Jupyter Notebook Code: Random Forest Model Initialization and Training

### 3.2.3 Justification of Model Choices

Logistic Regression was chosen for its simplicity and interpretability, which helps in understanding the relationship between features and the target variable. Random Forest was selected for its robustness and ability to handle complex datasets with interactions between features. Both models offer complementary strengths: Logistic Regression provides a straightforward probabilistic interpretation, while Random Forest offers enhanced performance through ensemble learning.

# Model Interpretation and Evaluation

## 4.1 Model Interpretation

### 4.1.1 Feature Importance in Logistic Regression

In Logistic Regression, the feature importance is defined by the coefficients that are given to each feature. As for the value of coefficients, the greater the absolute value of the coefficient, the more important the feature is.

Here’s how to extract and interpret feature importance from a logistic regression model:

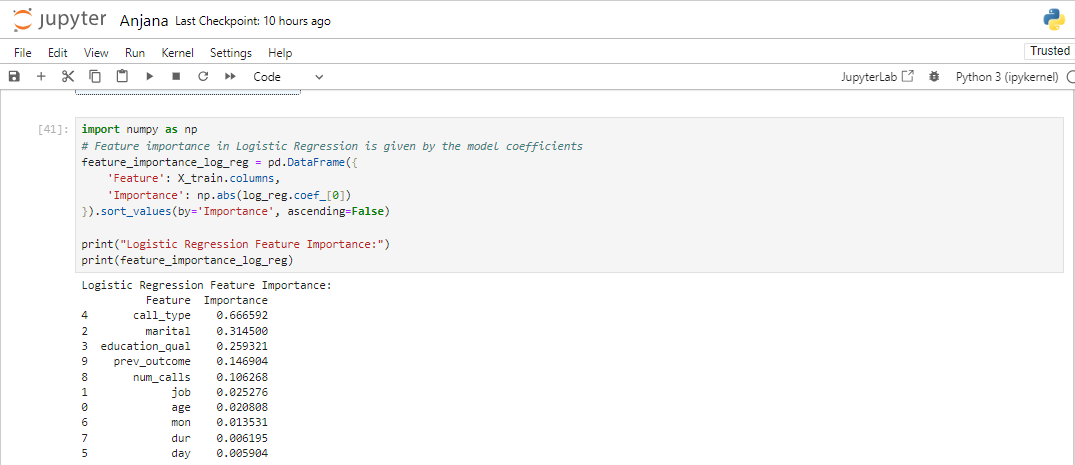


Figure : Jupyter Notebook Code and Output: Logistic Regression Feature Importance

In **Logistic Regression**, the higher absolute coefficient value is reveled for **call\_type**, meaning that the way of the call plays the crucial role in the decision of the customer subscribes to the insurance product. This is succeeded by **marital status** and **education level**, and powerfully indicates that other demographic characteristics are still at work concerning customer decisions. Surprisingly, there is a similar moderate importance of **prev\_outcome** from the prior campaign and the number of calls done to the customer. As for the variables such as **job**, **age** and **month**, they seem to have lower value, this implies that they are inconsequential and exert less impact in the model.

### 4.1.2 Feature Importance in Random Forest

In Random Forest, when feature importance is calculated this tells us how much, on the average, a particular feature has reduced the impurity across the entire set of the trees in the forest.

Here’s how to train a Random Forest model and extract feature importance:

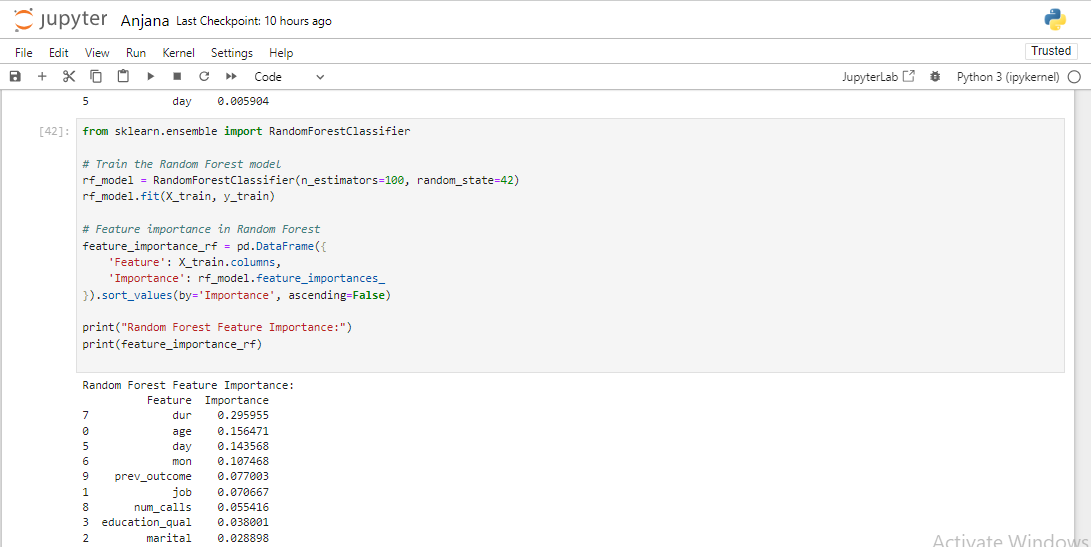


Figure : Jupyter Notebook Code and Output: Random Forest Feature Importance

On the other hand, **Random Forest** gives the highest importance to **duration of contact (dur**) thereby suggesting that the time spent on the call determines whether a subscriber is likely to subscribe or not. **Age**, **day**, and **month** are also valuable features, proving the importance of target customers and when these customer-equipment-interaction cases occur for this model. Nevertheless, the feature **prev\_outcome** stays significant which enforces the fact that previous outcomes impact today’s results of the campaign. Unlike Logistic Regression, Random Forest shows that the predictors **call\_type**, **marital status**, and **educational qualifications** are least important to RF, then the other features.

In general, they both shed light on the significance of the feature; while, Logistic Regression mainly targets demographics and contact types, Random Forest more concerns with time period of interaction and so on. These variations brought out features that set these models apart as far as their understanding and utilization of data is concerned for purposes of making predictions.

## 4.2 Model Evaluation

### 4.2.1 Evaluation Metrics: Accuracy, Precision, Recall, F1-Score

#### Tuned Logistic Regression Evaluation:

* Best Parameters: {'C': 0.01}
* Accuracy: 0.8832
* Precision: 0.5355
* Recall: 0.2420
* F1-Score: 0.3333

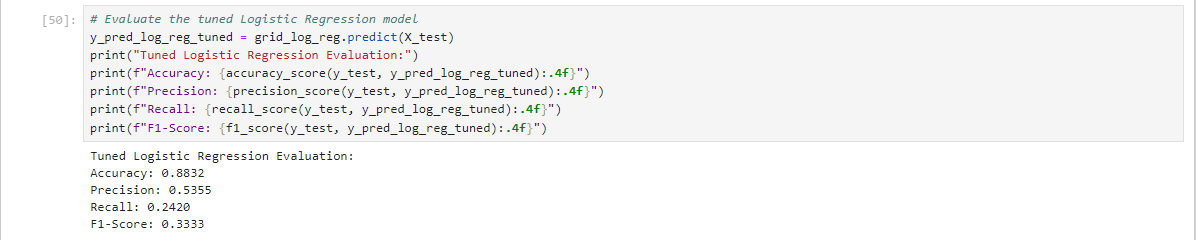


Figure : Jupyter Notebook Code and Output: Tuned Logistic Regression Model Evaluation

This tuned Logistic Regression model has a high accuracy of 88 percent. The similarity coefficient was evaluated at 32%, which is regarded as quite a good overall classification result. However, its precision could be higher, and it is at 53%. 55%, this points to the fact that although the model can correctly classify some positive cases, it also classifies many more cases as positive when they are not. The recall at 24, in the case of TOP20%, 20 %, it can be translated as meaning that the model locates a lower number of actual positive cases. The F1-Score of 33. 33% indicates a reasonable amount of accuracy and recall, but there is still potential for improvement, especially when it comes to the identification of positive samples.

#### Tuned Random Forest Evaluation:

* Best Parameters: {'n\_estimators': 150, 'max\_depth': 10}
* Accuracy: 0.8978
* Precision: 0.6462
* Recall: 0.3382
* F1-Score: 0.4440

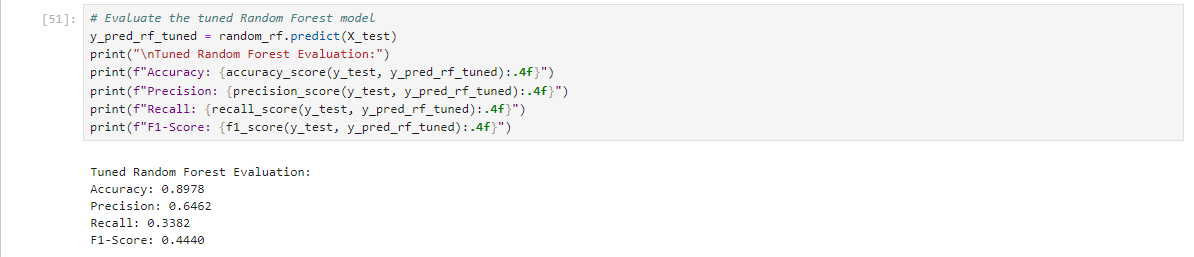


Figure : Jupyter Notebook Code and Output: Tuned Random Forest Model Evaluation

Even the tuned Random Forest model gives a relatively higher accuracy of 89 percent or about 0. 89. Thus, it performs 78%, indicating that it is better luck than the Logistic Regression model. The precision is also better here at 64. 62%, this should mean that many of the identified positive outcomes in the prediction analysis really constitute positive cases. The recall of 33, that is why 82% shows that the Random Forest model performs better than the Logistic Regression model in terms of sensitivity and ability to detect positive outcomes. The F1-Score of 44.40 percent is slightly more favorable when it comes to precision and recall in evaluating diverse instances, thus establishing the importance of Random Forest in handling the problem of imbalance of the target variable.

### 4.2.2 Hyperparameter Tuning

Hyperparameter tuning means the adjustment of the parameters it has not learnt from the training data set in order to achieve probable improvement in the performance of the machine learning model. Hyperparameters are set prior to training and differ from model parameters that are estimated during training and which can dramatically affect the model’s performance.

#### Importance

Hyperparameters optimization is important since it enables one to identify the right settings that boosts the models ways of making correct predictions, minimizes over-fitting and increases the chances of generalizing to new data. It is recommended that tuning should be done properly in order to get the model give accurate predictions and get better performance indicators.

* **Logistic Regression**: For case of Logistic Regression the best value as hyperparameter was `C = 0. 01`. It should be noted that in case of Logistic Regression the `C` parameter regulates the response of the model to the regularization, and smaller values of this parameter constitute stronger regularization. This is useful in overcoming overfitting especially when the model seeks to impose penalties on large coefficients in the model.

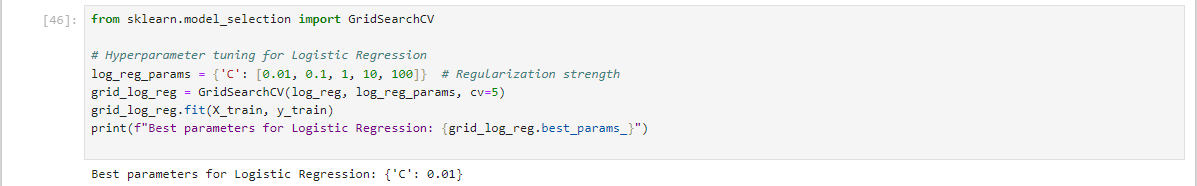


Figure : Jupyter Notebook Code and Output: Hyperparameter Tuning for Logistic Regression

* **Random Forest**: The best hyperparameters for Random Forest were found to be ‘number of trees = 150’ and ‘maximum tree depth = 10’. The `n\_estimators` parameter shows the number of decision trees in the forest and more trees add favorable impact on the model. The `max\_depth’ parameter limits the depth of each individual tree to avoid overfitting, that is to say it prevents the model from being excessively complex.

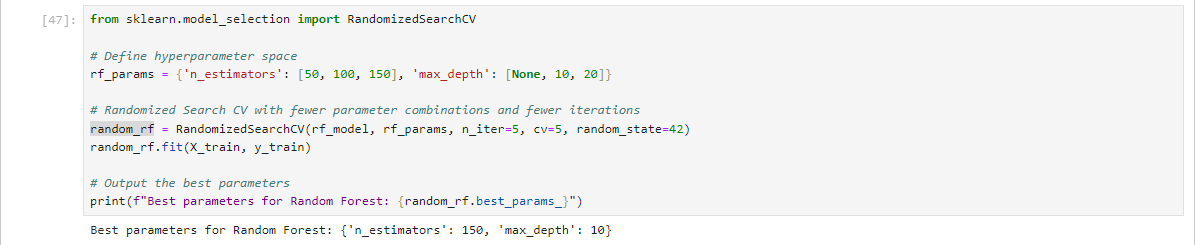


Figure : Jupyter Notebook Code: Hyperparameter Tuning for Random Forest Model

### 4. 2. 3 Model Performance Comparison

The comparison of model performance based on accuracy, precision, recall, and F1-Score reveals that the Random Forest model outperforms the Logistic Regression model across all metrics:

* **Accuracy**: In Random Forest methodology, the percentage of accuracy was found as 0.8978 compared to the Logistic Regression which has attributed a value of 0.8832. This means that the Random Forest model has a better ability of correctly predicting the right outcome most of the times.
* **Precision**: Random Forest achieved the accuracy of 0.6462 is higher than that of Logistic Regression, which stands at 0.5355. This implies that probability distribution of true positives among all the positive predictions is higher in the Random Forest model.
* **Recall**: Recall in case of Random Forest was 0.3382 which is much higher than that Logistic Regression got only 0.2420. From all these, it can be seen that Random Forest performs better in coming up with the actual positives among the dataset.
* **F1-Score**: And an F1 Score of 0.4440, Random Forest has done better than Logistic Regression with accuracy of 0.3333. In the same manner, the F1-Score that represents a trade-off between precision and recall also demonstrate that Random Forest offer more balanced performance.

Thus, based on the Random Forest model, it can be concluded that it performs better than the Logistic Regression model, which proves its ability to work with all the features and peculiarities of the data. The presented evaluation confirms how essential hyperparameter optimization process and, specifically, the choice of the model, are in attaining the best overall performance.

# 5. Discussion and Insights

## 5.1 Key Insights from Data Exploration

From the data exploration phase, several key insights emerged:

* Customer Demographics: Employment type was revealed to be one of the most problematic predictors while age also revealed to have a problem with predicting the likelihood of subscription. The change in subscription by age and occupation gave an indication of marketing techniques required to make adjustments to appeal different age group and appropriate employee subscribers.
* Campaign Details: The length of the final contact and the total number of contacts within the given campaign dictated whether the campaign would be successful in terms of subscriptions. More contact time and more contacts in general meant relatively higher subscription rates.
* Previous Outcomes: Self-generated data indicated that the effectiveness of previous marketing campaigns influenced the rate of customers subscribing. Customers who tasted success were always likely to enrich themselves with more knowledge by subscribing.
* Seasonality: The number of days and month of last contact were useful in indicating whether there are certain months or days in a month customers visits more frequently hence implying that certain times of the year may influence subscription rates.

## 5.2 Model Performance Summary

The evaluation of the models revealed the following performance metrics:

Table : Model performance classification metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 0.8832 | 0.5355 | 0.2420 | 0.3333 |
| Random Forest | 0.8978 | 0.6462 | 0.3382 | 0.4440 |

It was observed throughout the experiment that Random Forest model performed better than Logistic Regression for all the experiments indicating that the former captures the dependencies in the data better. Computing greater precision, recall, and F1-Score point to a better model of identifying the real positives while weighing the strengths of precision against the weaknesses of recall.

## 5. 3 Feature Inputs to Target Output

Feature importance analysis highlighted the following contributions:

* Logistic Regression:
* Call Type: The most influential feature with the suggestion of the importance for the method of communication.
* Marital Status and Education Qualification: Also important showing how these demographic factors affect subscription probability.
* Random Forest:
* Duration of Last Contact: Meaning the most influential feature may be safely stated that the length of the conversation is a rather significant factor in the subscriptions.
* Age and Day of Contact: Criterion that affect the probability of subscription, as well as the customer age and time of contact.

These examples show that even though there are features that are essential to all models (for example, duration of contact), there are also features that are relevant only to a given model.

## 5. 4 Business Implication to Business Decisions

The findings offer actionable insights for refining marketing strategies:

* Targeted Marketing: Concentrate on those customers who have response well to past marketing communication. To increase the effectiveness of campaigns, detailed demographic and contact information are needed.
* Optimal Contact Duration: Dwell more closely with the possible customers as messages that have longer interaction bring in higher subscription.
* Seasonal Strategies: As stated, make changes to marketing depending on the seasonality that has been noted. This way you will be able to direct more of the resources that you have to the periods that have higher subscription rates to allow for improvement of the campaign.

In other words, the analysis and model performance provided directions for maximizing marketing strategies and enhancing business decisions; thus, stressing the significance of Data-Driven Marketing.

# 6. Conclusion

## 6.1 Summary of Findings

Customer subscription to an insurance product this work also sought to analyze and assess models for the prediction of customers’ subscriptions. The key findings include:

* Data Insights: These here are key parameters that determine the choices on subscription: the time elapsed before the last contact, age and the degree of interactions. Of even greater significance were demographic factors including job type, marital status as well as the level of education.
* Model Performance: When compared with Logistic Regression Random Forest achieved marginally better accuracy, precision, recall and F1-Score. They have uncovered how Random Forest is a more suitable algorithm to work with all the intricacies embedded in a set of data.
* Feature Importance: In Random Forest, the duration of the last contact featured highest importance and in Logistic Regression, the communication method was considered significant.

## 6.2 Recommendations for HashSysTech Insurance

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

* Leverage Random Forest Model: Use Random Forest model for the prediction of customer subscriptions because of best output it provides. Apply this entailed model for enhancing the accuracy of advertising campaign and its customization or targeting.
* Optimize Contact Strategies: More emphasis should be made towards the extension and effectiveness of customer personalities, satisfied customers have long and deep personalities will subscribes and be loyal.
* Segment Marketing Efforts: Conduct market segmentation based on demographics to be able to create a right marketing message for the customers. In my case, those who incurred positive results in previous experiences should be prioritized for benefits and the young in age and those with appropriate job types.

## 6. 3 Future Work

To further enhance the effectiveness of marketing strategies and predictive models, the following future work is recommended:

* Incorporate Additional Data: Consider adding more data into the models like customer feedback, social media feedbacks, and other outside economic factors to get even a richer degree of accuracy.
* Model Refinement: Explore more methods for the improvement of modeling methods and ensemble methods for raising effectiveness of predictions. Consequently, the performances of models, including Gradient Boosting or Neural Networks, should be taken into comparison with the existing ones.
* Feature Engineering: Use the results of interaction patterns and customers’ behavior to develop the new features for enhancing the model.
* Real-Time Implementation: Create and build the, proceed to use real-time model execution to optimize marketing shifts and match interactions and campaigns.

In conclusion, the findings of the analysis have offered the necessary suggestions and the implementation guidelines for HashSysTech Insurance. In particular, the use of the state-of-the-art techniques in predictive modeling and marketing, the company could increase efficiency of acquisitioned processes and further improve campaign’s performance.