**Capstone Project Report: Marketing Campaign Analysis**

**Table of Contents**

**Background**

**1.1 Introduction**

**1.2 Project Timeline**

**Business Understanding**

**2.1 Problem Statement**

**Data Understanding of Raw Data**

**3.1 Dataset Overview**

**3.2 Business Description**

**3.3 Hypothesis**

**3.4 Objective**

**Technologies Used**

**4.1 Models and Machine Learning Algorithms**

**4.2 Libraries**

**Data Exploration**

**5.1 Data Exploration Techniques**

**5.2 Descriptive Statistics and Data Visualization**

**5.3 Data Preparation (Data Processing and Exploratory Data Analysis)**

**5.4 Feature Engineering**

**5.5 Models**

**5.6 Challenges**

**Accomplishments**

**6.1 Results and Analysis**

**6.2 Steps**

**6.3 Model Evaluation**

**6.4 Conclusion**

**6.5 Future Recommendations**

**References**

**1.1 Introduction**

In this project, I was tasked to analyse a marketing campaign dataset and implement machine learning algorithms to predict customer responses. The project must follow a structured project management methodology, encompassing data exploration, preprocessing, and the application of classification models. This report details the entire project plan, business understanding, data understanding, data preparation, machine learning implementation, findings, conclusions, and future recommendations. The report should be presented in a clear and concise manner, and it should demonstrate your ability to use a project management methodology. The project management methodology should enable you to prioritise tasks and monitor the progress of the capstone project.

After a lot of searches, I decided to embark on focusing on a analysing the responses of customers to help businesses market their products. In this initiative, I am embarking on a data-driven exploration to elevate marketing strategies by putting advanced analytics. The aim of the analysis is to be sure of customer responding to marketing campaigns.

The world revolves around sales and good marketing aids great sales. I was able to find a comprehensive customer marketing response dataset in Kaggle. This dataset gives **2240** different customers basic information, their product purchasing preferences as well as their reactions to some marketing campaigns.

The Marketing response dataset is a large dataset with 2240 rows, 29 columns made up of 23 numerical and 3 categorical features. I used the CRISP DM methodology for this project.

**1.2 Timeline**

The project timeline was structured into distinct phases:

Data Exploration: Understand the dataset and identify key features.

Data Preprocessing: Handle missing values, outliers, and transform variables.

Machine Learning Implementation: Apply classification models to predict customer responses.

Evaluation: Assess model performance using metrics like accuracy and recall.

Documentation: Compile a comprehensive report.

**Business Understanding**

**2.1 Problem Statement**

I am trying to solve the problem of companies instead of spending money to market a new product to every customer in the company’s database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that segment.

**Data Understanding of Raw Data**

**3.1 Dataset Overview**

This section is about understanding the dataset. The first thing I did was read the dataset into the data frame. But to know more about the dataset, I use .head() to look at the first five rows of the data frame which will have 2240 observations and 29 features.

**3.2 Business Description**

**3.3 Hypothesis**

H0: There is no significant relationship between the predictor variables and target variable.

H1: There is a significant relationship between the predictor variables and target variable.

I am trying to predict the number of customer responses to marketing campaigns and to predict the customer response to certain products to see if they will do well in the market.

Advantages of using Random Forest algorithm:

1. It works for both types of problems including classification and regression
2. Works efficiently on large datasets
3. It provides higher level of accuracy while predicting the outcome

**3.4 Objective**

The main objective of this analysis was to predict customer responses to optimize future marketing strategies.

**Success Criteria/ indicators**

I found a correlation between product type and response, it shows that customers that buy groceries like fruits, meat, fish are responding more. This means food products will gain more marketing response. I used different machine learning models and algorithms to find the best predictions and results with higher accuracy.

So, I have also used random forest algorithm to predict the outcome. I have achieved 89% percent of accuracy through random forest classification.

**Technologies Used**

**4.1 Models And Machine Learning Algorithms**

I have used Crisp-Dm, Exploratory Data Analysis, Data Preparation. The machine learning models used are Cross Validations: Stratified K-fold, Random Forest, Support Vector Machine, Logistic Regression and Decision Tree.

I used different supervised that may often be used in Regression and Classification problems.

**4.2 Libraries**

There are different libraries used to perform this task and modeling of algorithms. These includes; Pandas, Numpy, Seaborn, Matplotlib, scipy, sklearn.model, etc.

**Accomplishment**

**5.1 Data Exploration**

The data for Market Customer response has been used for analysis, having 2240 observations and 29 variables. There were 26 numerical variables and 3 categorical variables, these includes; ID, Year Birth, Income, Kidhome, Teenhome, Recency, MntWines, MntFruits, MntMeatProducts, MntFishProducts, NumWebVisitsMonth, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, AcceptedCmp2,Complain, Z\_CostContact,Z\_Revenue, Response.

**Source**

The data has been taken from an online source that is Kaggle <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>.

**5.2 Descriptive Statistics and Data visualization**

I have gone through the overview of the dataset using head to view the first five rows to know what types of data and have general oversight view of the data. Descriptive statistics were used to understand the distribution of variables, identify missing values, and observe outliers.

Table 1: General overview of the first five rows

A screenshot of a computer

Description automatically generated

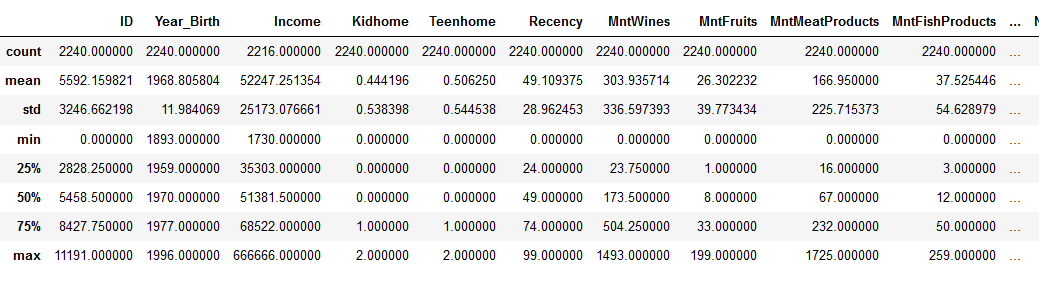
What I did here is read the dataset into data frame to have the oversight view of the dataset.

Table 2: Table below shows the full data types, attributes of the dataset

A screenshot of a computer

Description automatically generated

Table 3 Table below shows the results of statistical analysis of the numerical features



Findings:

1. Average age of the customers respondent according to the Year 2023 is 53 years old.
2. Average income of the customer is 51381(0000).
3. Average customer that responded does not have kid or teen at home.

The snippet comprises statistical and exploratory analysis because we can start exploring missing values and outliers from looking at this table above.

**5.3 Data Preparation (Data Pre-Processing and Exploratory Data analysis)**

From the oversight view of our statistical analysis, I noticed there are some missing values and there was a possibility of outliers in some variables. I have used different steps in data preparation such as cleaning the data, resizing of data, and feature engineering. In the section, I will use different steps to check the actual missing values, outliers, resizing of data, feature engineering, general cleaning of the data including treating outliers and treating the missing values so that our dataset will be ready for model testing to give us reliable predictions and accuracy.

Since my prediction is on how customers will respond to the marketing products and looking at the data dictionary there are some features that are not necessary for the analysis and model testing. In this section, we dropped the features we believed were not necessary to complete our regression prediction problem i.e. we initially concentrated on the features we believed important for completing the regression problem. It is easier to choose the features we require than drop the features we don't require. This reduces the number of features from 29 columns to only 19 columns. Now, we have looked for missing values in our dataset and remove those values. There were 24 missing values in our Income variable as shown in the below table and graph.

Figure 1: Table of missing values

A white background with black text

Description automatically generated

As is seen above in the table, there are 24 missing values in our Income variable which was treated using the average to replace and fill the missing values. The reason for using the mean as replacement was because of the distribution of the income value. To know more about the distribution, the graph below shows how it was distributed.

Figure 2: Income Distribution

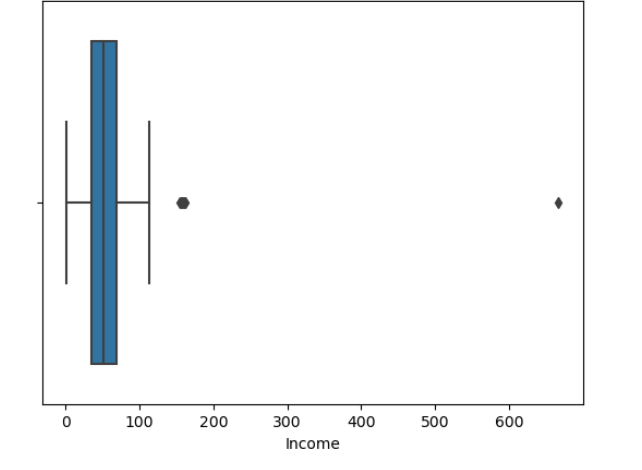
A graph of income distribution

Description automatically generatedAs seen above, the observation through the graph, the distribution is a bimodal distribution. And also the missing value is relatively small percent of the variable, it is only 1% and this reason why I didn’t drop it but rather replace it with mean ( average of the income distribution).

Figure 3: The box plot below shows the outlier noticed in Age and Income distribution; these two variables are of core importance to our analysis.

**A graph with a bar and numbers

Description automatically generated with medium confidence**



There are outliers in Income and Age as seen from the graphs above. These outliers will be treated by removing the rows where it occurred and that is removing any row that is greater than 200k income and where Age is above 100.

I will go ahead to visualize the Univariate, Bivariate and Multivariate Analysis of the variables to see how they are distributed.

A graph showing a number of income and age

Description automatically generated with medium confidence

There are outliers in Income and Age. We will treat the outliers by zscores and rectifying the outliers by capping to make sure our values do not go too low or too high to give us more accurate results in our analysis going forward.

Figure: Graph of Age and Income after outlier treated.

A diagram of a graph

Description automatically generated with medium confidence

**Flowchart of Data Preparation and Modelling**



Descriptive Statistics through

Distributions of Income and age

Top 5 Age Groups by purchase

Top 5 Age Groups by response

Correlation analysis

Feature Selection- Initial removal/dropping of the unwanted columns to avoid noise and reduce the complexity of the EDA process

Hot-Encoding categorical variables Like Education etc.

Smote Analysis and Resampling dataset.

Feature Engineering

Dataset check- understanding the raw data

Data Dictionary- check for quantitative and qualitative data

Analyze Features relevance/significance in regards to the project objectives

Modelling and Algorithm Defining the X and Y features

Splitting the dataset into the test and training set Creating and fitting the regressor to the training set

Calculating the accuracy of the training and the test set.

Hyper parameter tuning

Under sampling unbalance dataset and model testing after Under sampling

Figure 4: The table below shows the Age Distribution of customers.

A graph of age distribution

Description automatically generated

From the Distribution above, the age of the customers is mainly centered in their 40s or 70s, the young people under 30s and elderly above 80s are very few. These people are at their middle ages or old ages, so their family condition should be further taken care of.

Figure 5: Pie chart below shows the Marital and Education distribution of customers.

A pie chart with different colored circles

Description automatically generated

A pie chart with different colored circles

Description automatically generated

The pictures from the above graphs show a quick look of the customer distribution, we could see the most our customer (64%) is in relationships (Married (38.64%) or Together (25.85%)) and most (97%) are at least bachelor’s degrees with having (PhD (21.69%), Master (25.54%), Bachelor (50.36%))).

Figure 6: Histogram to show the customers with kids and teens at home. Want to show their distribution.

A graph of a bar

Description automatically generated with medium confidence

A graph of a number of blue bars

Description automatically generated with medium confidence

The visualization above shows that most customers have 1 kid/Teen or 0 kid/Teen at home, very few have 2 kids/Teens, and no one have kids/Teens over 2.

Figure 7: The histogram below shows the general response of the former survey which is being used for this model training and testing. The table below shows will show if the dataset is balance or imbalance.

A graph with a blue rectangle

Description automatically generated

According to Response from the graph above, this is an unbalanced dataset, over 80% customers say no to the last campaign. So, we will do a smote analysis to be sure of an accurate prediction. There are systematic algorithms that you can use to generate synthetic samples. The most popular of such algorithms is called SMOTE or the Synthetic Minority Over-sampling Technique. This technique is used to treat the imbalance data to check further the accuracy of the prediction if there will be much difference.

Figure 8 : The table below shows the response based on Marital and Education Status of the customers.

A graph of a number of students

Description automatically generated

A graph with blue and orange bars

Description automatically generated

From the table above, I noticed that campaign acceptance rate in high education groups (Master and PhD) are higher than that in low education groups.

Single people tend to say yes to the campaign more.

Campaign acceptance rate in high education groups(Master and PhD) are higher than that in low education groups.

Single people tend to say yes to the campaign more

Married people has the highest percent of not responding to the campaign.

Figure 9: Going further to visualize the customer response based on how many kids and teens they’ve got at home.

A graph of a number of objects

Description automatically generated with medium confidence

A graph with blue and orange bars

Description automatically generated

From the visualization above, Customers with no kids and no teens at home are more likely to accept the offer in this campaign.

Figure : Top 5 Age Groups that made the total purchase.

A graph of a group of people

Description automatically generated with medium confidence

The Top 5 age group that purchased the products are 48,52,53,54 and 73 years old respectively.

The customers with the of 73 years has the highest total purchases.

The reason for doing this analysis is to ascertain if there is a correlation between the age group that purchase the product and the ageo group that responded to the campaign.

Figure : Top 5 Age Groups that Responded to the Campaign

A graph of different colored squares

Description automatically generated

The Top 5 age groups that responded to the campaign are 41,51,52,53 and 73 years old respectively.

The customers with the of 73 years has the highest response.

The reason for doing this analysis is to ascertain if there is a correlation between the age group that purchase the product and the ageo group that responded to the campaign.

From the above graph, we can see that the age group of 41 and 52 years old do responded to the campaign but are not among the top 5 age group that has the highest purchases.

**Correlation**

I will be using pair plot that is commonly used to find the correlations between different variables to see any specific relationship among variables.

Figure 10: Checking correlation using Pair plot.

A graph with red and blue squares

Description automatically generated

A screenshot of a graph

Description automatically generated

From the plot above, we could see that these numerical variables do not have clear linear trend between each other.

While Income might have log relationships with these products purchasing amount (say, Mntfruit, MntSweetProducts and etc.).

**Heat map**

The below visualization with aims to give the better understanding of the dataset using different color coding in high volume of locations with most data matters.

A colorful squares with numbers

Description automatically generated with medium confidenceFigure 11: Heat map among all numerical variables

**5.4 Feature Engineering / Feature Selection**

In order to improve the predictions of the result, I feature engineering analysis by extracting the information from birth year. I did the following in the feature engineering analysis. I replace the “Birth Year’’ with a ‘’Age’’ by subtracting the current Birth Year from 2023 the current right now. This was done for better model compatibility.

I selected relevant variables and created a new data frame for it called ‘’df\_mkt\_new’’.

After researching each of the features, we will now drop the features that appear to reference the same features, are also detailed in other features or may not be meaningful for our initial EDA analysis. This will reduce the number of our features from 29 to 19 which make our dataset easier to analyse.

**Machine Learning**

The objective is to see if customers will buy a new product when launched. If they will respond Yes to the product.

**5.5 Models**

I have used three different machine learning algorithms including logistic regression, decision tree and random forest algorithm. The models have focused on to resolve the regression and classification prediction of featured variable. I have done different steps to make the data useful. Such as removing unnecessary variables and making the dummy variables of symptoms using Marital status of customer. I also replaced categorical variables to numerical values, I did that on the Education variable and assigned 0, 1, 2, 3 to each Educational level respectively.

**Preparation of data for machine learning algorithms**

After dropping off some variables that are not necessarily for our machine learning algorithms, I proceeded to check our response (y) feature to see if the data is balance or not.

By normalising the value\_counts on our target feature and printing and plotting the results, we can see that our target feature is imbalanced with 80% of the values customers say no to the last campaign. So we will do a smote analysis to be sure of an accurate prediction. To address this issue, we can use a re-balancing solution called SMOTE further below. I also use under sampling also. I want to see which one works better then I can go with the Machine Algorithm with more accurate result.

A graph with a rectangle and a rectangle with text

Description automatically generated

I did not encode my Target variable as my Target variable is already in numeric values as 0,1. 0 no respond and 1 as customers that respond respectively.

I split my dataset into training and testing datasets.

* from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.25, random\_state = 4)

It is important to note that for classification models, it is recommended to use a test\_size of 0.25.

I carried out three Machine Learning Models using Random Forest, Logistic Regression and SVM without smote analysis to balance the unbalance data, I did this to know how the machine learning models will perform and to see if there will be bias in their predictions.

I did not use Hyper parameter tuning when I first carried out the analysis, but I will use Grid searchCV and also Stratified K-fold Cross validation going further in this analysis.

**Hyper parameter tuning:**I used GridSearchCV for my parameter tuning and also define a pipeline for this tuning . Here is an example of one hyper paramet tuning used on one of the machine learning algorithm;

**A screenshot of a computer code

Description automatically generated**

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

The parameter tuning was performed on all the Machine learning algorithms.

**Cross- Validation: Stratified K-fold**

I used stratified k-fold cross validation to compare the prediction outcome of each machine learning algorithm, we need to choose the best models to perform on my training and testing dataset. I tried to further use Support vector machine model but it took forever for it to train the dataset, so I deleted the SVM and stick to using only the three machine learning algorithms (Random Forest, Logistic Regression and Decision Tree).

For us to choose the model to work with, we must first perform cross-validation which will give us the best model outcome. It shuffles the dataset randomly. It splits the dataset into a certain number of groups i.e. k number of groups. For each of these unique groups, it then takes that group as a test dataset (a holdout) and uses the remaining groups as the dataset to train on. It then fits the model on the training dataset and evaluates it on the testing dataset. It then retains the evaluation score and discards the model. It does this multiple times to see which is the best underlying generalisation.

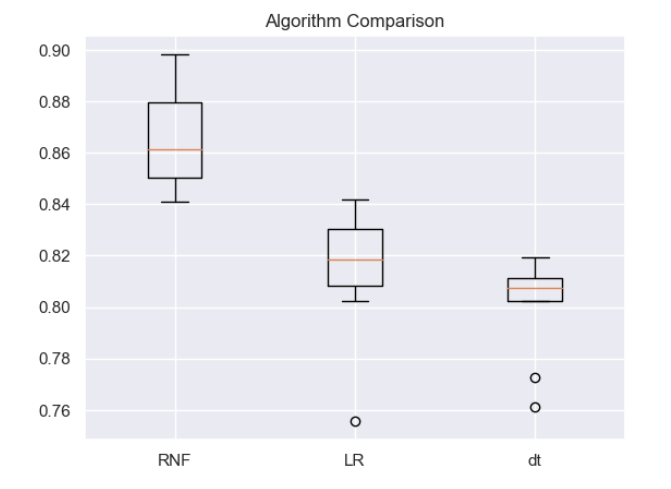
After I carried out my cross validation on the models, these are my outcomes below;

RNF: 0.866930 (0.019516)

LR: 0.815928 (0.023857)

dt: 0.801217 (0.018160)

Looking at the cross-validation outcome above, we can see that random forest performed the best with a 87% accuracy score and Decision Tree performed worst at 80%. I will also visualise the algorithm comparison to look at how they perform graphically to see if there is any outlier in each other the model using the interquantile range. It is important to note that accuracy scores should be clarified with an analysis of the dependent feature in the classification reports for each of the models as the dataset is imbalanced. However, using cross-validation is useful to help us decide on what machine learning models to use below.



Model comparisons

**Machine Learning algorithms After SMOTE analysis/ Under Sampling**

Previously, in my former analysis I did not use the correct smote analysis and I did not also try under sampling to see how best my Machine could perform, but now I was able to improve on that and the unbalanced data was balanced through these methods. I perform model training on my unbalance dataset to see the results, random forest performed best with 96% accuracy but there might be bias due to unbalance dataset and it perform correctly on the no response than response, so I decided to treat my unbalance dataset with smote and under sampling. After I used the two methods of treating unbalance data, the results for the smote analysis came out more better with the models and the predictions are more accuracy, so I will choose to use smote analysis for treating the unbalance data for this model training.

Show casing the reason for choosing Random Forest prediction after Smote analysis.

1. **Random Forest:**

Random forest is a great model which is suitable for both regression and classification. I set my n\_estimator, criterion parameters myself when I previously worked on the model training, but I improved on it by using GridsearchCV hyper parameter tuning for choosing the best parameters.

The steps followed was;

* Setting hyper parameter grid for datapoint training
* Fitting the GridsearchCV object to the training data
* Creating random forest classifier with the best parameters
* To predict the category that a new datapoint will belong to, the best parameters was used to predict the class that the datapoint should belong to and assign that datapoint to that class.

A screenshot of a computer

Description automatically generated

From the above classification result, the model achieved an accuracy of approximately 89.96% and a recall of around 40.79%. we can see the relatively low recall and high precision score which indicates that the dependent class wasn't correctly recognised most of the time i.e., a larger number of false negatives or small positives. This indicates that the model correctly predicted about 89.96% of all instances and identified approximately 40.79% of the positive cases correctly (customer responses). The precision for the positive class is 74%, suggesting that when the model predicts a positive response, it is correct around 74% of the time. The classification report and confusion matrix show that the model performed well in identifying instances with no response (class 0).

But, it struggled to correctly identify positive responses (class 1), as evidenced by the lower recall and precision values for class 1. However, the overall performance of the model is relatively good, as indicated by the weighted average F1-score of 0.89.

I also printed visualised the confusion matrix and classification report which makes a comparison between the y\_test and y\_pred and allows me to visualise the performance of the algorithm and to analyse the dependent feature in more detail. The below confusion matrix shows that of the 482 who did not respond to the campaign marketing, the algorithm predicted that 11 would respond. Of the 76 customers who responded, the algorithm predicted that 31 would respond and 45 won’t respond.

A graph showing a heatmap

Description automatically generated

1. **Decision Tree**

This model is a graph in the shape of a tree, a sequential diagram that shows all the potential decision options and their associated results. Decision Tree is supervised classification method. Decision Tree is a collection of decision nodes, connected by branches, extending downward from root node to terminating leaf nodes. Beginning with root node, attributes tested at decision nodes, and each possible outcome results in branch. Each branch leads to decision node or leaf node. Starting from the root of a tree, every internal node represents what a decision is made based on; each branch of a node represents how a choice may lead to the next nodes; and finally, each terminal node, the leaf, represents an outcome yielded.

In this case of predicting whether customer responded or not ‘Yes’ or ‘No’ response to marketing campaign.

All records enter root node, and CART evaluates possible binary splits.

Decision Tree works more perfectly on target variable that are categorical, this might be why decision tree was the worst performing model in this analysis because our target variable is in binary which is 0 and 1.

In the below classification report and confusion matrix report which makes a comparison between the y\_test and y\_pred and allows me to visualise the performance of the algorithm and to analyse the dependent feature in more detail. The below confusion matrix shows that of the 482 who did not respond to the campaign marketing, the algorithm predicted that 48 would respond. Of the 76 customers who responded, the algorithm predicted that 41 would respond and 36 won’t respond.

A screenshot of a computer

Description automatically generated

A graph with a green square

Description automatically generated

From the above results, model achieved an accuracy of approximately 85.13% and a recall of around 53.95%. This model performed relatively well in identifying positive responses compared to the other models, as evidenced by its higher recall value. However, the precision for class 1 is lower at 46%, indicating that the model's positive predictions are correct only about 46% of the time. The classification report and confusion matrix show that the model has a balanced performance in predicting both classes, but it still struggles with precision for class 1, this results maybe due to the target variable not been a categorical variable but a binary values.

1. **Logistic Regression**

This model uses regression analysis which Predict the value of a dependent variable based on the value of at least one independent variable. A statistical technique for forecasting binary classes is logistic regression. The result or goal variable has a binary nature. There are just two conceivable classes when something is dichotomous. Logistic Regression outcome is discrete (not continuous) or categorical. We can apply the LogisticRegression models to the forge dataset, and visualize the decision boundary as found by the linear models.

The steps followed was;

* Setting hyper parameter grid for datapoint training
* Fitting the GridsearchCV object to the training data
* Creating random forest classifier with the best parameters
* To predict the category that a new datapoint will belong to, the best parameters was used to predict the class that the datapoint should belong to and assign that datapoint to that class.

In the below classification report and confusion matrix report which makes a comparison between the y\_test and y\_pred and allows me to visualise the performance of the algorithm and to analyse the dependent feature in more detail. The below confusion matrix shows that of the 482 who did not respond to the campaign marketing, the algorithm predicted that 18 would respond. Of the 76 customers who responded, the algorithm predicted that 30 would respond and 46 won’t respond.

Also looking at the accuracy, recall, F-1 score and precision, we can confirm that the Logistic Regression model achieved an accuracy of approximately 88.53% and a recall of about 39.47%. Like the Random Forest model, it performed better in identifying instances with no response (class 0) compared to positive responses (class 1). The precision for class 1 is 62%, indicating that the model correctly identifies positive responses around 62% of the time. Although the accuracy is slightly lower compared to the Random Forest model, the Logistic Regression model still demonstrates respectable performance in predicting customer responses.

A screenshot of a computer

Description automatically generated

A graph showing a heatmap

Description automatically generated

**5.6 Challenges Encountered**

I encountered different challenges while starting my research because the first thing is how to choose the right dataset that will have enough observations and attributes that are necessary and relevant to the prediction. After a series of research and finding a comprehensive dataset, the second challenge was getting more knowledge about the market campaign because business knowledge is the fundamental key to any model training analysis.

There are not many challenges in cleaning and treating missing data but there was a challenge to know which categorical variable to encode and treat as dummies. Also, finding the perfect model that may give greater accuracy is challenging.

I also encountered the problem of unbalanced data and which method will be the best to treat the unbalance data, I had to go through using two different methods and choosing from the best accurate prediction with less bias.

All the challenges were overcome by using the right libraries, deep searching to know the right models and best model for classification.

Removed all unnecessary variables, renamed and feature engineering to model compatibility.

**Accomplishments**

**6.1 Results and Analysis**

I used three models to ascertain the best possible accuracy outcome. I used logistic regression, Decision Tree and Random Forest. Random Forest gave the overall accuracy score of 89% accuracy after smote analysis done on Random Forest, the result for cross validation is 87% accuracy, making Random Forest the best performed model, the confusion matrix predicted more better.

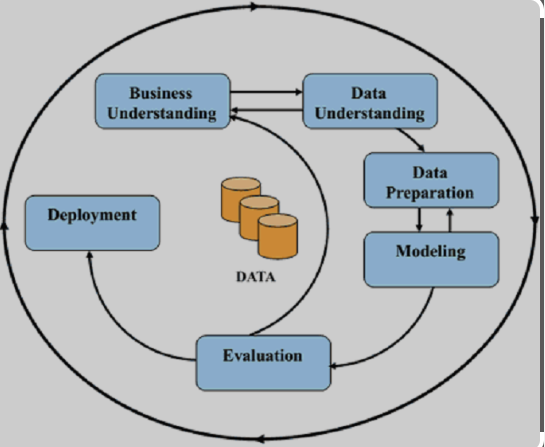
**6.2 Steps**

I followed the ML algorithm steps. The first thing I did was to divide the testing and training dataset. I separated out the target variable (y) and the independent variables (X), then convert them into arrays using values as ML algorithms require arrays (Menduni et al., 2022). I am looking to predict the response of our customers to our marketing products to see if they will purchase it or not.

Firstly, I imported the machine learning models libraries. Then I created the classifier variable and use n estimators to set the number of parameters i.e., it's the number of trees we want to add. I work with 300 decision trees and apply these as my parameter for the model. We then fit the classifier onto our X\_train and y\_train variables previously.

The results in this model showed a high accuracy of 88 percent. In the evaluation of the model, we have used Crisp DM model also as shown in figure. So, we must re-evaluate to see what else can be done to improve and make our prediction accurate classifying/ predicting the model.

Figure 12: Crisp DM model



I already revalued and I did not set the parameters myself but I used GridsearchCV to set the parameters when I revalued, and our best performing model is Random Forest. There might be need to still revalue because of the low recall in all the models, and also there might be a connection between the variables and predictions. We might need to create more new features and re select which features that are important and necessary for our prediction i.e prediction indicators.

**6.3 Model Evaluation**

In this paper, I have focused to analyze the Customer dataset which explains the customers response to marketing dataset.

Through our project I gained a better understanding of the data by utilizing the Data Dictionary and Research, we adjusted our approach to our EDA and Data Preparation by dropping unrelated features, managing null values, Feature Engineering – Binning Features and creating relational visualizations for better understanding of the variables and how they could help our analysis better. I also added hyper parameter tuning, used Cross Validation on 3 models to determine the best one, used SMOTE technique to imbalanced data, used proper metrics system for classification like precision and finally we used recall and F1 to fully interpret the performance of the model.

**6.4 Conclusion**

In this report, I have focused on analyzing the marketing response dataset of customers after marketing campaigns were carried out. I have reviewed CRISP-DM framework and applied its 6 phases to assist us in completing the objectives. The 6 stages we have used were: Business understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. The project successfully applied project management methodology to analyze a marketing campaign dataset. Key findings included the effectiveness of SMOTE in addressing imbalanced data and the superior performance of the Random Forest model over Logistic Regression. The insights gained can guide future marketing strategies to target potential customers more effectively. Using the Random Forrest classifier, I have been able to predict with 89% accuracy that customers are likely to response to the marketing campaigns in future but however, we may need to revisit the imbalanced data problem that has impacted the lower-class performance on the recall so as to have a high recall also. We may also look to improve the feature selection.

**6.5 Future Recommendations**

Feature Importance: Conduct a deeper analysis of feature importance to identify key factors influencing customer responses.

Dynamic Model: Implement a dynamic model that can adapt to changing customer behaviors over time.

Model revaluation: Recommending the evaluation of the CRISP DM model which means we may have to re-visit some of the earlier steps to obtain confidence in our predictions/ most importantly the imbalanced data must be revisited because it really impacted the performance of lower-class.

.

Github link:  
  
https://github.com/horlaralexie/Strategic-Thinking-3

**References**

Kenan Guan .(2021) ‘Customer\_Response\_Predict & Personality\_Analysis’ Available at: [https://www.kaggle.com/code/cainanguan/customer-response-predict- personality-analysis](https://www.kaggle.com/code/cainanguan/customer-response-predict-%20%20%20personality-analysis)

[Accessed: 10 December 2023]

Menduni, G., Zifarelli, A., Sampaolo, A., Patimisco, P., Giglio, M., Amoroso, N., Wu, H., Dong, L.,

Bellotti, R. and Spagnolo, V., 2022. High-concentration methane and ethane

QEPAS detection employing partial least squares regression to filter out energy

relaxation dependence on gas matrix composition. *Photoacoustics*, *26*, p.100349.

Geron, A.(2019) Hands-n Machine Learning with Scikit-Learn,Keras and TensorFlow,

United State of America: O’Reilly Media, Inc.

The imbalanced-learn developers ed., (n.d.). SMOTE. [online]

imbalanced- learn.org.

Available at: [https://imbalanced-](https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html) [learn.org/stable/references/generated/imblearn.over\_sampling.SMOTE.html](https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html)

[Accessed 04 May. 2024].

scikit-learn developers ed., (n.d. c). sklearn.ensemble.RandomForestClassifier. [online] Available at: [https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

[learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

[Accessed 02 May. 2024].

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python. O'Reilly Media, Inc.

Zaki, M. J., & Meira, W. Jr. (2020). Data Mining And Machine Learning:

Fundamental Concepts And Algorithms. Cambridge CB2 8BS, United Kingdom.