**MODELLING CONSUMERS RESPONSE TO DIRECCT MALL MARKETING** 

TEACHER :SIR TEHSEEN JILLANI

GROUP MATES: nimra siddiqui (roll no 51)

Tooba Iftikhar( roll no 76

Maryam Shafiq(roll no 34)

Shumaila naz (roll no 67)

# CONTENTS

1. Introduction ----------------------------------------------------------------------------------------3
2. Business Understanding -------------------------------------------------------------------------3
3. Data Understanding ------------------------------------------------------------------------------4
   1. Data Frame ------------------------------------------------------------------------------------4
   2. Variables ---------------------------------------------------------------------------------------5
   3. Quick Data Insights --------------------------------------------------------------------------6
4. Data Preparation -----------------------------------------------------------------------------------7
   1. Data Cleaning and Transformation -------------------------------------------------------7
   2. STANDARIZATION (To Reduce Variability) ----------------------------------------------7
   3. Splitting Data in Testing and Training Sets ----------------------------------------------7
   4. EDA (Exploratory Data Analysis) -----------------------------------------------------------8
   5. Feature Engineering --------------------------------------------------------------------------9
      1. Spending Variable --------------------------------------------------------------------9
      2. Polynomials ----------------------------------------------------------------------------9
      3. Dummy Variables --------------------------------------------------------------------10
5. Modeling ---------------------------------------------------------------------------------------------10
   1. Logistics Regression --------------------------------------------------------------------------10
   2. K Nearest Neighbors (KNN) -----------------------------------------------------------------11
   3. Random Forest ---------------------------------------------------------------------------------12
   4. Cross Validation --------------------------------------------------------------------------------12
      1. Hyper Parameters ---------------------------------------------------------------------12
      2. Feature Selection and Class Imbalances -----------------------------------------13
6. Model Evaluation ------------------------------------------------------------------------------------13
   1. Confusion Metrics for Logit Random KNN ------------------------------------------------14
7. Deployment --------------------------------------------------------------------------------------------15
8. ROC -AUC -----------------------------------------------------------------------------------------------15
9. EDA (Figures) -------------------------------------------------------------------------------------------16

1.Introduction

This report is based on data from a clothing store chain that uses different types of marketing initiatives to promote sales. our task is to classify which customers will respond to direct mail marketing based on data collected for past customers.

To achieve this goal we are following the cross industry standard process because of its powerful practicality, its flexibility and its applicability when using analytics to solve thorny business issues.

## Cross-Industry Standard Process for Data Mining (CRISP-DM)

**The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a open standard process modal that specifies a six phase life cycle for data mining projects: 1. Business understanding.**

**2. Data understanding. 3. Data preparation. 4. Modelling. 5. Evaluation. 6. Deployment.**

## 2. BUSSINESS UNDERSTANDING

In direct email marketing, sending relevant information only to customers that might be interested will be the right strategy for any reasonable direct marketers. Well-focused customer targeting is need. Data mining techniques can be used for this purpose.

We are working on data from a clothing store and because most of the dealings of stuff is digital, direct mail seems old and boring. But direct mail is still a great way to reach your audience, grab their attention, and connect with them on a personal level. In 2016, The Data & Marketing Association reported that the direct mail customer response rate increased by 43% as compared to 2015.

Here are few reasons why direct mail marketing is still effective:

1. It puts the ROI in third behind email and social media marketing. Social media is ahead by only 1 percentage point, hence providing more bucks on your investment. (source: data and marketing association DMA).
2. Smart marketers not only rely on digital marketing but they actually combined direct mail and digital marketing to increase loyalty with their clients.
3. As fewer brands are using this strategy which may seems outdated your brand has a higher chance of standing out.
4. It gets undivided attention of your client and you can step up the game by sending coupons and special offer, most people will save it for future use.
5. It increases your brand awareness, you send an email, use Facebook Ads, or do any online marketing, your chance of reaching an older demographic isn’t good. So DM is for all age groups.

**Building the Cost/Benefit Table**: Classification models are often evaluated on accuracy rates, error rates, false negative rates, and false positive rates. In business problems, company managers may require that model comparisons be made in terms of cost/benefit analysis. The goal of developing a classification model for predicting a customer response to direct mail marketing ensures the highest return for marketing efforts, ultimately increasing the profitability of the store.

* + True Negative - No sales are made, no costs are incurred. • True Positive - Promotion is mailed, customer responds and thus the benefit is profit less cost.
  + False Negative - Promotion is not mailed to a customer who would have responded. No cost is incurred, however the store loses a potential profit.
  + False Positive - Promotion is mailed to a customer who does not respond. Only the cost of sending promotion is incurred.

For direct mail marketing, it can be seen that a false positive is far less serious, with But it can be higher in other environments. For example, in direct marketing, a false positive may cost no more than a postcard, while in HIV testing, a false positive on the ELISA test will be more expensive, leading to second-level HIV testing. (On the other hand, of course, false negatives in HIV testing are very serious indeed, which is why the ELISA test allows a higher rate of false positives, to maintain the false negative rate as low as possible.)

|  |
| --- |
| **Outcome** **Classification Actual Response Rationale** |
| True Negative Nonresponse Non response No marketing costs, No |
|  |
| True Positive Response Response Profit - Marketing cost |
|  |
| False Negative Nonresponse Response profit loss, No marketing cost |
|  |
| False Positive Response Nonresponse Marketing cost, No profit |

## 3. DATA UNDERSTANDING

3.1. DATA FRAME: our data consists of 50 regressors(features) and a binary response variable

(i.e. the response we get on our direct mails from our clients). There are two type of variables categorical and numerical.

3.2. VARIABLES: Customer information that can be collected/used in selection of customers may include the followings in general (the specific description is given in following table);

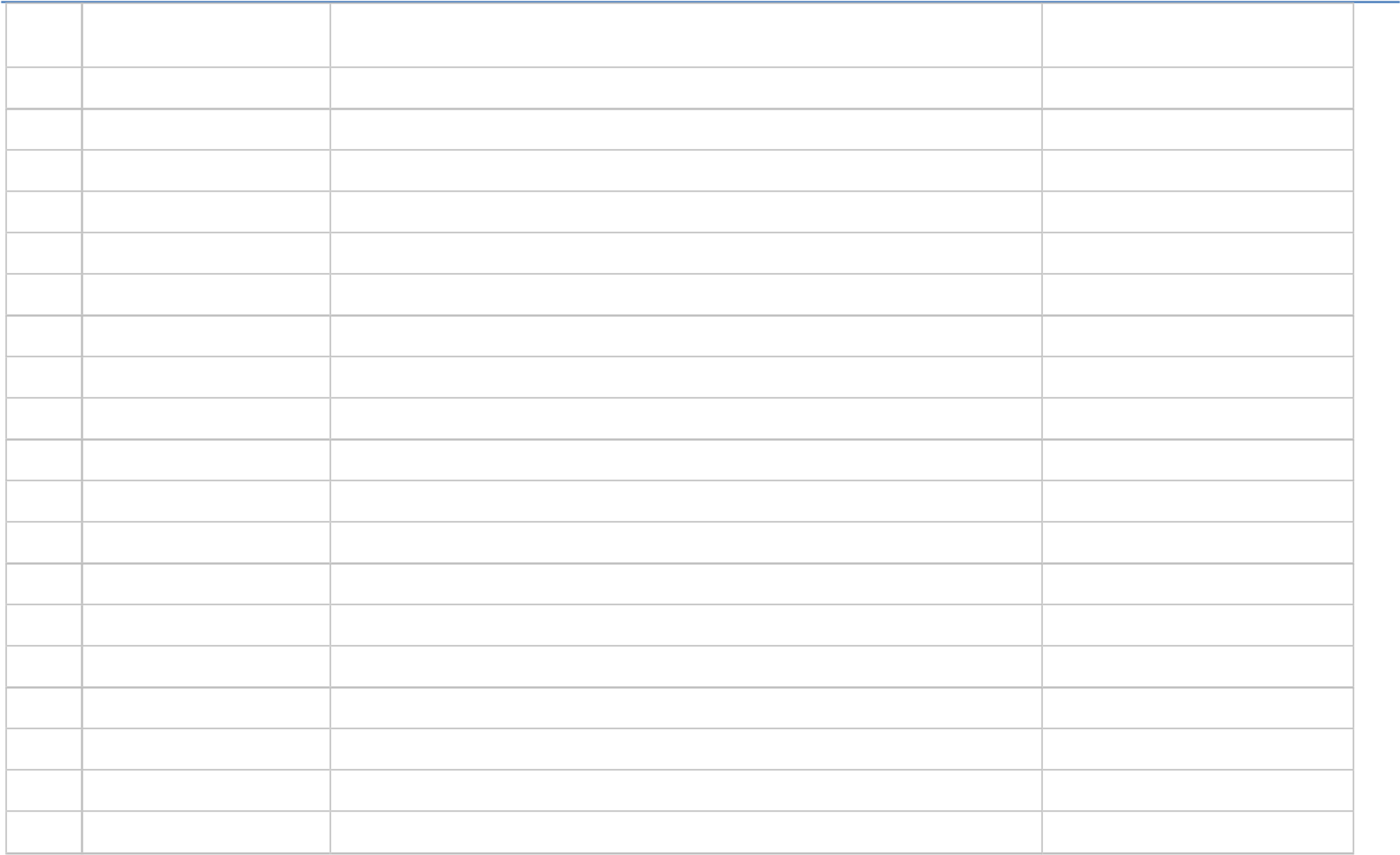
* Demographic variables describe characteristics of populations and include age, gender, race, education, occupation, income, religion, marital status, family size, children, home ownership, socioeconomic status, and so on.
* Geographic variables include various classification of geographic areas, for example, zip code, state, country, region, climate, population, and other geographical census data.
* Behavioral variables include product usage rate and end, brand royalty, benefit sought, decision making units, location of purchase, no. of purchase visits and so on.

This information can be extremely useful for marketing purposes.

* Past business history includes various business statistics on customers, especially purchasing patterns. This provides essential business indicators and therefore is very important information.
* Rates and ratio metric like RFM marketing oriented variables for example Customers purchased *recently* tend to buy again. Customers purchased *frequently* tend to buy again. Customers purchased most *monetary* values tend to buy again.

VARIABLES DESCRIPTION.

**no. Code Meaning Data type**

1. HHKEY Customer ID, Unique Categorical
2. ZIP\_CODE zip code Categorical
3. REC number of days between (since) purchases Integer
4. FRE Number of purchase visits Integer
5. MON Total net sales Float
6. AVRG: av amount spent per visit Float
7. PC\_CALC20 brand of choice (encrypted) Categorical
8. PSWEATERS proportion spent on sweaters Float 9. PKNIT\_TOPS proportion spent on knit tops Float 10. PKNIT\_DRES proportion spent on knit dress Float 11. PBLOUSES proportion spent on blouses Float 12. PJACKETS proportion spent on jackets Float 13. PCAR\_PNTS proportion spent on career pants Float 14. PCAS\_PNTS proportion spent on casual pants Float

15. PSHIRTS proportion spent on shirts Float 16. PDRESSES proportion spent on dresses Float

1. PSUITS proportion spent on suits Float
2. POUTERWEAR proportion spent on outerwear Float
3. PJEWELRY proportion spent on jewelry Float

|  |  |  |  |
| --- | --- | --- | --- |
| 20. | PFASHION | proportion spent on fashion | Float |
| 21. | PLEGWEAR | proportion spent on legwear | Float |
| 22. | PCOLLSPND | proportion spent on collectibles | Float |
| 24. | AMSPEND | Amount spent at the AM franchise | Float |
| 25. | PSSPEND | Amount spent at the PS franchise | Float |
| 26. | CCSPEND | Amount spent at the CC franchise | Float |
| 27. | AXSPEND | Amount spent at the AX franchise | Float |
| 28. | TMONSPEND | Amount spend in last three months | Float |
| 29. | OMONSPEND | Amount spend in last one months | Float |
| 30. | SMONSPEND | Amount spend in last six months | Float |
| 31. | PREVPD | Amount spend the same period last year (SPLY) | Float |
| 32. | GMP | Gross margin percentage | Float |
| 33. | PROMOS | number of marketing promotions on file | Integer |
| 34. | DAYS | number of days the customer has been on file | Integer |
| 35. | FREDAYS | number of days between purchases | Integer |
| 36. | MARKDOWN | markdown percentage on customer purchases | Float |
| 37. | CLASSES | number of different product classes purchased | Integer |
| 38. | COUPONS | number of coupons used by the customer | Integer |
| 39. | STYLES | total number of individual items purchased by customer | Integer |
| 40. | STORES | number of stores at which the customer shopped | integer |
| 41. | STORELOY | maybe store location | categorical |
| 42. | VALPHON: | if valid phone is given | categorical/ true/false |
| 43. | CC\_CARD | if is credit card user | categorical T/F |
| 44. | WEB | if is web shopper | categorical T/F |
| 45. | MAILED | Number of promotions mailed past year | integer |
| 46. | RESPONDED | Number of promotions responded to past year | integer |
| 47. | RESPONSERATE: | Promotion response rate for last year | float |
| 48. | HI | Product uniformity (low=diversity) | Categorical |
| 48. | LTFREDAY: | Lifetime average time between visits | Float |
| 49. | CLUSTYPE: | Micro vision Lifestyle Cluster Type | Categorical |
| 50. | PERCRET: | Percent of returns | Float |
| 51. | RESP: | flag responded to promotion - target variable | categorical True/False |

3.3. QUICK DATA INSIGHTS :

* + - The descriptive statistics shows our variables are numerical i.e. floats / integers and categorical. we must find a way to address this problem so we can incorporate all categorical variables into our model. one common thing is most of them having median closer to 1st quartile.
    - There are no missing values to incorporate.
    - Thebox plots and histogram plots of all features are indicating the two main problems; outliers and skewness in our features so we have to transform and clean our data before modelling.
    - It is important to know what is the proportion of responders to the direct mail marketing promotion? The computed stats show that only 4762 of the 28,799 customers, or 16.54%, responded to last year’s marketing campaign (1 indicates response, 0 indicates nonresponse.) Since the proportion of responders is so small, we may decide to apply balancing to the data prior to modeling.

After knowing the main problems in this section we are moving towards data preparation phase where we state our strategies to overcome these problems.

## 4. DATA PREPRATION

We prepare our data in following steps as follow:

**4.1 DATA CLEANING AND TRANSFORMATION:**

Initially we drop the 1st two features customer ID as it is unique and encrypted, it can contain no information that is helpful for our task of predicting which customers are most likely to respond to the direct mail marketing promotion. The zip code actually represents categorization of the client data base by geographic locality. However, for the present problem, we set this field aside and concentrate on the remaining variables.

We separate the 31 numeric variables which are positively-skewed having many outliers and apply transformations to achieve normality or symmetry also to tackle outliers in order to avoid overfitting.

It is important to note that we use natural log transformation (1 plus) for variables which contained positive as well as zero values.

The reason is many data mining methods and models, such as principal components analysis and logistic regression (which we are going to use), function best when the variables are normally distributed or at least symmetric. As Doane and Seward (2011) proposed, it is appropriate to transform data which has a skewness > 0.5.

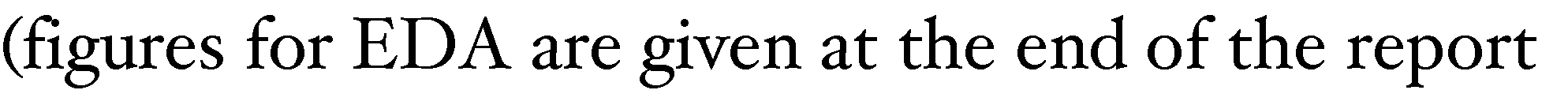
**4.2 STANDARIZATION TO REDUCE VARIABILITY:**

in order to have less variability among the S.DEV of numerical variables we choose to use Zscore scaling (mean=0, Standard deviation=1) this is done by subtracting the mean of the variable and dividing by the standard deviation.

**4.3 SPLITTING THE DATA INTO TRAINING AND TESTING SET:**

We randomly apportioned our data into testing and training sets with 80-20 split.

**4.4 EDA (EXPLORATORY DATA ANALYSIS):**



)

**It refers to the critical process of performing initial investigations on data so as to discover patterns to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.** Moving forward we try to visualize our training data graphically in the context of the response variable to know whether the customer responded or not.

Multivariate analysis:

To avoid information overlapping and multicollinearity b/w predictors and to find which predictors have strong relation with response variable we use correlation heatmap. the matrix for all predictors indicates the high correlation between ’classes’, ’coupons’, ’styles’ and ’stores’, ’mailed’ and ’responded’, as well as ’tomonspend’, ’omonspend’ and ’smonspend’.

whereas we further build a matrix b/w response variable and predictors which are highly correlated with response variable. The predictors ’fre’, ’classes’, ’styles’, ’responded’, ’mon’, ’coupons’, ’stores’, ’responserate’ and ’smonspend’ are all moderately correlated with the response which implies that a no. of predictors are required in order to accurately classify the response.

Bivariate analysis:

We take predictors for bivariate analysis for better understanding. A **violin plot** combines elements of a box plot and density plot, allowing the visualization of both the distribution of the data and its probability density. This is significant in revealing the modality of a distribution,for example ‘FRE’ that for both a response and no response, the common values are right skewed towards fewer purchase visits. However notably the mean and upper quartile range for a response is higher than no response, indicating a higher number and greater spread of average visits. In case of ‘CLASSES’ the number of product classes is more evenly distributed for response, indicating that people who buy different classes of goods tend to respond more.

We use **histogram** to get more information about the data concentration. One interesting thing is most of data in numeric features falls in 1st quartile for ex ‘Total net sales’ and ‘visits’. Also we can see some franchise having fewer visits by customer whereas some are frequently visited.

Another interesting pattern is the **bar chart** of CLUSTER TYPE: it’s the market segmentation of customer into 50 parts define by Claritas Demographics. The 5 most common lifestyle cluster types in our data set are:

1. Cluster 10: Home Sweet Home—families, medium-high income and education, managers/professionals, technical/sales
2. Cluster 1: Upper Crust—metropolitan families, very high income and education, homeowners, manager/professionals
3. Cluster 4: Midlife Success—families, very high education, high income, managers/professionals, technical/sales
4. Cluster 16: Country Home Families—large families, rural areas, medium education, medium income, precision/crafts.
5. Cluster 8: Movers and Shakers—singles, couples, students, and recent graduates, high education and income, managers/professionals, technical/sales.

It can be seen cluster 10 and 1 is the most prevailing type and its associated with high education and income, which make sense.

Lastly, we regress all key predictors individually on response variable using logistic regression to decide whether it’s statistically significant or not. based on these plots, we can say that the predictors with a higher correlation to the response will have clear difference between the distribution of the response and no response classes as it follows the typical sigmoid curve for example the predictor “CLASSES”.

**4.5. FEATURE ENGINEERING:**

To improve the classification performance of the models, feature engineering was used to create additional predictors from the data provided. The creation of additional predictors was informed by the exploratory data analysis conducted at the beginning of this report.

* + 1. **Spending Variables**

The data contains three cumulative spending variables which were of interest, which represented the amount spent by a customer in the last month, three months and six months, the amount spent by a customer in the previous month is also going to be contained within the last three and six months. Hence, this is redundant data, which is being triple counted within the model building process. We drop the variable for past three and past six month and try to fix this as follow:

|  |  |  |
| --- | --- | --- |
| PREVIOUS VARIABLE | NEW VARIABLE | FEATURE ENGINEERING |
| TMONSPEND | Month23 | TMONSPEND-OMONSPEND |
| SMONSPEND | Month456 | SMONSPEND-TMONSPEND |

Due to error in data entry we notice that we are having some negative values which is not the intuitive, so we just replace the negative values with zero to avoid NaN issues after taking a log or square root transformation.

* + 1. **Polynomials**

The predictors with non linear relationship with response are transformed using higher order polynomials in order to get the desire linear pattern, for example:

|  |  |  |
| --- | --- | --- |
| PREVIOUS VARIABLE | NEW VARIABLE | FEATURE ENGINEERING |
| RESPONDED | RESPONDED-sq | Taking sq-root of variable  RESPONDED |

* + 1. **Dummy variables**

now we transform all the categorical variables into dummy variables. They can be thought of as numeric stand-ins for qualitative facts in a regression model, sorting data into mutually exclusive categories (such as VALPHON 1= if the phone no. is valid and 0= otherwise).

## 5. MODELING

The model selection we used here is based on intuition that these models are considered to be standard for classification problems and often yield reliable results with minimal assumptions. it includes both parametric and non-parametric models as follow:

### 5.1. Logistic Regression

logistic regression is a type of regression model where the response variable is categorical. In this case, the categorical response variable is the consumer’s response. In the logit model, the probability p that the observed value of your response variable y takes the value 1 is:



Using maximum likelihood estimates, we get the linear equation between the regressors and response as:



usually easier to interpret (you can interpret it as the log odds) and because their corresponding cumulative distribution function is a closed form expression. we are trying to predict the probability this means that the fitted values of our logit model (i.e. the probabilities) must fall between 0 and 1.

**ASSUMTIONS:**

* the dependent variable is binary.
* For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
* Only the meaningful variables should be included.
* The model should have little or no multicollinearity.
* The independent variables are linearly related to the log odds.
* Logistic regression requires quite large sample sizes.

### 5.2. K Nearest Neighbors (KNN)

KNN classifier is a non-parametric approach in which the output is a class membership. An object is classified by a plurality vote of its neighbors with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. It’s an approximation of the Bayes classifier;



Run the algorithms several times with different values of k and choose the one which minimize the error without loosing accuracy, keep in mind that if you decreases the value of k predictions become less stable inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging.

**Assumptions:**

* KNN assumes that the data is in a feature space. More exactly, the data points are in a metric space. The data can be scalars or possibly even multidimensional vectors.

* Each of the training data consists of a set of vectors and class label associated with each vector. It will be either + or – (for positive or negative classes). But KNN, can work equally well with arbitrary number of classes.

### 5.3. RANDOM FOREST

It is also a non-parametric approach. The random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features. The final predictions of the random forest are made by averaging the predictions of each individual tree.

Select a random sample of features and observations (with replacement) from the entire data. For each feature (node) you'll test different thresholds and see which gives you the best split criterion (generally entropy, gini, information gain) Keep the feature and its threshold that makes the best split and repeat for each other feature Stop after a you completely reach a single leaf node or a stopping criterion (max.depth size or min.samples\_leaf).

**Assumptions:**

* It also assumes that data is in a metric space.
* Also Each of the training data consists of a set of vectors and class label associated with each classes.

5.4 CROSS VALIDATION

Cross validation (CV) is one of the techniques used to test the effectiveness of a machine learning models, it is also a re-sampling procedure used to evaluate a model. Our goal is to finding the model which yields the largest expected profit based on the validation data.

**K-FOLD CV:** It is a popular and easy to understand, it generally results in a less biased model compare to other methods. Because it ensures that every observation from the original dataset has the chance of appearing in training and test set. We randomly split the data in to k folds (generally we take k=5 or 10). Train the data on k-1 folds and validate it on the kth fold. Repeat this process until every K-fold serve as the test set. Then take the average of your recorded scores. Because our response is binary we are using STRATIFIED K FOLDS as it maintains the percentage of samples for each class in every fold. So that model gets equally distributed data for training/test sets.

* + 1. HYPER-PARAMETERS

One important thing in increasing the effectiveness of models that we choose here is the selection of hyper-parameter. By this ten-fold (k=10) cross validation method the candidate hyper-parameter was fit to the k-1 folds and tested on the kth fold. For each of the k folds the precision was calculated. Finally, the value for the hyper-parameter was selected based on the average precision on each of the k folds

* + 1. FEATURE SELECTION AND CLASS IMBALANCE:

To increase model accuracy we try to find most relevant features that has greater impact on response variable. Here are the three methods used:  Sequential feature selector: a family of greedy search algorithms that are used to reduce an initial d-dimensional feature space to a k-dimensional feature subspace where k < d.

* Random forest classifier.
* Gradient boost classifier.

Another important thing is that our data has more points of ‘NO RESPONSE’ So it may give bias results to overcome this problem we have to resample Our data by different techniques, we use:

* Unsampled
* Smote
* Down sampled

SMOTE gives us best balancing with 0.78 F1 score.

## 6. MODEL EVALUATION

Generally, classification algorithms are examined on how accurately it classifies the observation in right class but it is not appropriate to use the accuracy of prediction alone as the measure of model performance. This is because the penalties for misclassification are not equal, as defined by the **cost-benefit matrix** in Section 2. Following are 5 metrics that we are using in this case:

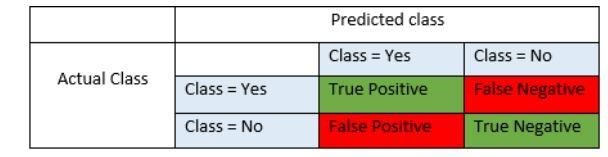
* Accuracy: simply the proportion of rightly classified points.
* Precision: it’s the percentage of result which are relevant
* Recall: it’s the percentage of total relevant results correctly classified by your algorithm.
* F1 score: it’s the is the weighted average of Precision and Recall. In our case the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | AUCCURACY | PRECISION | RECALL | F1 SCORE |
| RANDOM  FOREST | 0.61 | 0.78 | 0.58 | 0.63 |
| GRADIENT BOOST | 0.36 | 0.66 | 0.52 | 0.58 |
| LOGIT | 0.49 | 0.71 | 0.82 | 0.75 |

From above results we see that random forest gives us more accuracy as compare to the other models.

**6.1 CONFUSION METRICS FOR LOGIT, RANDOM FOREST, KNN CLASSIFIER**:

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.



False positives and false negatives, these values occur when your actual class contradicts with the predicted class. Here we know that we have to reduce the no. of false negative and false positive and we must find balance between them.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MODEL | TN | TP | FP | FN | ERROR RATE |
| LOGISTIC | 19 | 3428 | 684 | 174 | 51% |
| GRADIENT BOOST | 88 | 2159 | 615 | 1451 | 64% |
| RANDOM FORESTS | 458 | 2035 | 245 | 1567 | 39% |

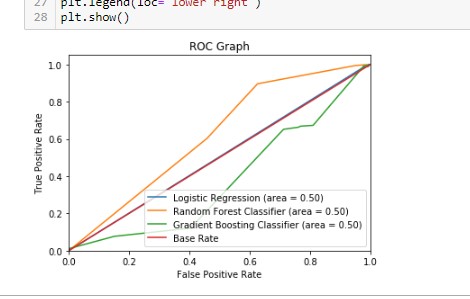
Based on above metrics the management will decide whether or not the cost of sending mail to the wrong customers (FN) is greater than the loss of profit it will face by not sending mail to their potential customers (FP). We can see that gradient boost performs the worst and in logistic model despite of its lower error rate, the values contained within the loss matrix penalize misclassification heavily for false negatives of which the logistic regression model contains substantially more than the random forest model model. Therefore, of these two models, the random forest model which has an expected profit per customer is the more preferred model.

## 7. DEPLOYMENT

We have to set a benchmark for our business, what if we send the mail to our entire list? How much profit we made? Based on past experience if very fewer customers are going to respond.

In case, the profit we earned on that is **>** cost incurred (due to sending mail to entire list) then we should select the model which enable us selecting the list that ***exceeds*** the profit of sending mail to entire list. Clearly, the company would like to maximize the true positive result and can afford to mis-classify customers provided that it is a false positive due to its low cost.

We can see model sensitivity is equal in all model. In this case we select the **Random forest model** with 39% error rate while this value is still relatively large, the low misclassification rate for false negatives ensures that the overall cost to the expected profit is minimal.



EDA (FIGURES)

