Marketing Campaign Prediction

By

Pandya Harsh Maheshkumar,

Reddy Charan Pulivendula,

Sheta Rushank Ghanshyam

hapandya@iu.edu , repuli@iu.edu , rsheta@iu.edu

Under the guidance of **Prof. Yuzhen Ye**

Abstract – Prediction is crucial in businesses as this may help increase the impact of marketing campaigns or advertisements by using targeted marketing strategies. Using predictive modeling techniques it is possible to assess the probability of a new user becoming a customer or someone who purchases a product. A heavy share of the company's revenue is invested back into marketing campaigns and advertisements to get good results, but making a data-informed decision here can help reduce the risk of marketing campaign failure. Nowadays, companies use targeted marketing to increase the return on this hefty investment. In this project, we have tried to generate value out of the historical marketing campaign data, to mine insights out of it, through data analysis and training a binary classification model aimed to provide good classification for the next product.

Keywords – Exploratory Data Analysis (EDA), Data Analysis, Univariate Analysis, Bivariate Analysis, Correlation Analysis, Classification, Logistic Regression, Decision Tree, Random Forest, Boosting, Oversampling.

1. Introduction –

On average a company spends 4% to 25% of its net revenue on Marketing campaigns and Advertisements. While some companies like Salesforce and Tableau spent around 50% of their revenue on marketing campaigns. Targeted marketing is used to get good returns from this investment which aims at identifying groups of customers who are highly likely to become future revenue-generating users Customer segmentation allows us to group customers based on demographics, geography, interests, user activity, etc. Marketing campaign analysis helps us to predict if the customer becomes our future customer. In order to make sure the Marketing campaign analysis makes use of a variety of alternative data including their education and gender who brought successful results in the previous campaign. We conduct exploratory data analysis on the open-source dataset and trained multiple models to asses which one is the best one. In the EDA phase, we conducted univariate, bivariate, and correlation analysis, also tried some feature transformation, feature generation to get valuable insights from the data. We used various classifier models ranging from logistic regression, and random forest and tried methodologies like oversampling, outlier handling, boosting, and bagging in modeling phases. Finally, we compared all the trained models using valid metrics and pre-processing steps.

2. Problem Statement -

To classify if a new customer will respond positively or negatively towards the next product based on responses taken while a previously conducted marketing campaign. Marketing campaign analysis allows companies to find customers which are highly likely to become future customers this can be done either by data analysis or classification models. We will use both, firstly we try various plots to see if we can generate some insights using data visualization then train classifier models to make predictions given the input attribute including customer demographics like age, income, marital status, account balance, gender, occupation, last campaign results, no of times contacted, etc.

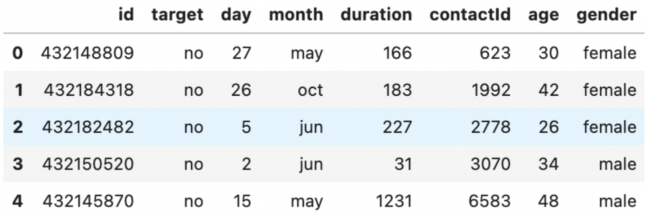
3. Methodology -

- Data Validation
- Handling missing values
- Exploratory Data Analysis
  1. Univariate Analysis
  2. Multivariate Analysis
  3. Correlation Analysis
- More Data Analysis
- Data Pre-process
- Modelling
- More Modelling(using oversampling)

4. Data validation -

This includes steps like Checking for missing values and understanding the five-number summary and statistical summary for each attribute.

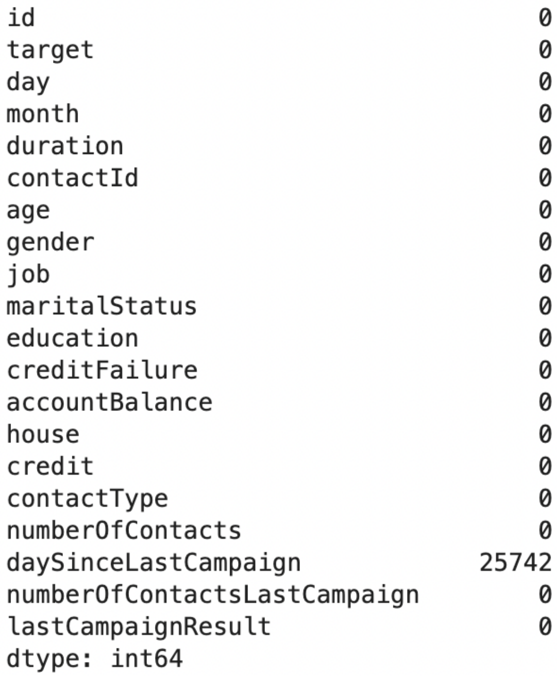
Training Data contains attributes like:



	id	target	day	month	duration	contactId	age	gender
0	432148809	no	27	may	166	623	30	female
1	432184318	no	26	oct	183	1992	42	female
2	432182482	no	5	jun	227	2778	26	female
3	432150520	no	2	jun	31	3070	34	male
4	432145870	no	15	may	1231	6583	48	male

Fig 4.1

- Target – ‘1’ refers positive result and ‘0’ refers to negative result of customer purchasing the product
- Duration – Call duration time of the customer
- Age – age of the customer
- Gender – {male, female} for the customer
- Job – the current job of the customer
- Marital Status – {married, single, divorced}
- Education – the highest ed. Degree of the customer
- Credit Failure – provides the credit score status of the customer



```
id                              0
target                          0
day                             0
month                           0
duration                        0
contactId                       0
age                             0
gender                          0
job                             0
maritalStatus                   0
education                       0
creditFailure                   0
accountBalance                  0
house                           0
credit                          0
contactType                     0
numberOfContacts                0
daySinceLastCampaign        25742
numberOfContactsLastCampaign    0
lastCampaignResult              0
dtype: int64
```

Fig 4.2

Checking for Null or missing values we found out that the attribute 'Day Since Last Campaign' has most of the values null, this may be because most of the participants in the current campaign might not had participated in the previous campaign. We decided to replace null values with -1 here which would suggest that this customer has not participated in the previous campaign. Also, by checking for duplicate values we found that there were no duplicated rows in the data frame.

Further classified attributes into numeric or categorical since classification is needed before data visualization which will help in selecting the type of plot to use for an attribute. Day, Duration, Contact Id, Age, Account Balance, No of contacts, day since last campaign, no of contacts last campaigns are set as numeric whereas columns like gender, target, credit failure, job, etc will be used as categorical ones. Out of a total of 19 columns, we have 10 categorical attributes, 8 numeric attributes, 1 binary attribute which is a target.

5. Univariate Analysis -

To visualize the distributions of attributes we have removed the outliers in data using Inter Quantile Range(IQR) method for outlier handling. Followed by density and scatter plots for each attribute. After removing outliers we are left with 28k instances from 31k(with outliers).

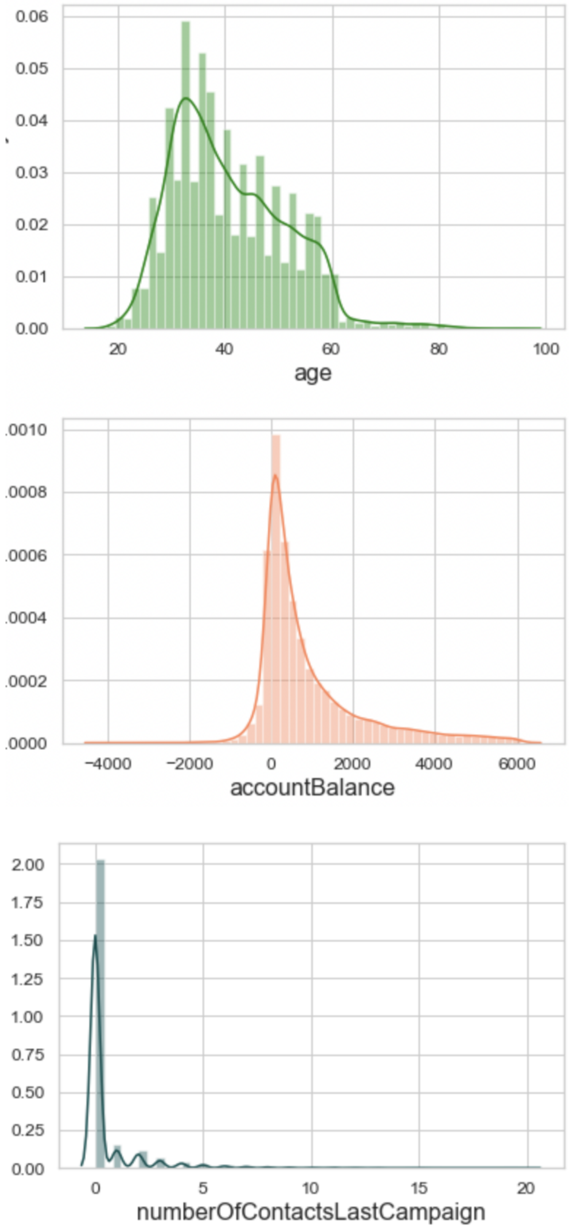


Fig 5.1

Some insights that we got from the univariate analysis were as follows: Density plots are looking better after removing outliers and we can analyze the results better. There are comparatively fewer instances for the third and fourth weeks of the month and more in the first and second weeks(day attribute). The mode age is around 34 as visualized form the above plot. Number of times a person was contacted in previous campaign ranges form 1 to 9. Most of the participants in this marketing campaign are contacted after 180 days or 6 months from the previous campaign. Account balance for most of the participants is close to zero and positive but not zero since mode account balance close to and greater than zero.

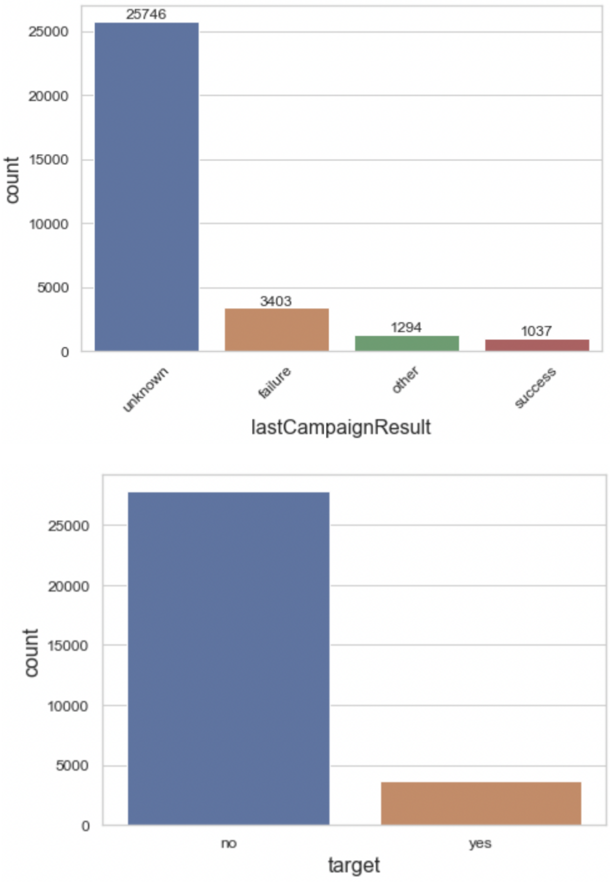


Fig 5.3

For further analysis, we have generated a new feature combining two features the target and last campaign result. The categories for new feature are as follows: Double Positive Customers: are the customers who accepted both previous and the current campaigns. New Positive Customers: are the customers who did not participate in the previous campaign and accepted the current campaigns. Negative to Positive Customers: are the customers who refused the previous campaign but turned to positive answer in the current campaigns. Positive to Negative Customers: are the customers who accepted the previous campaign but turned to negative answer in the current campaigns. New Negative Customers: are the customers who did not participate in the previous campaign and refused the current campaigns. Double Negative Customers: these are the customers who refused both previous and current campaigns. The ones not falling into any of the categories are classified as Unknown.

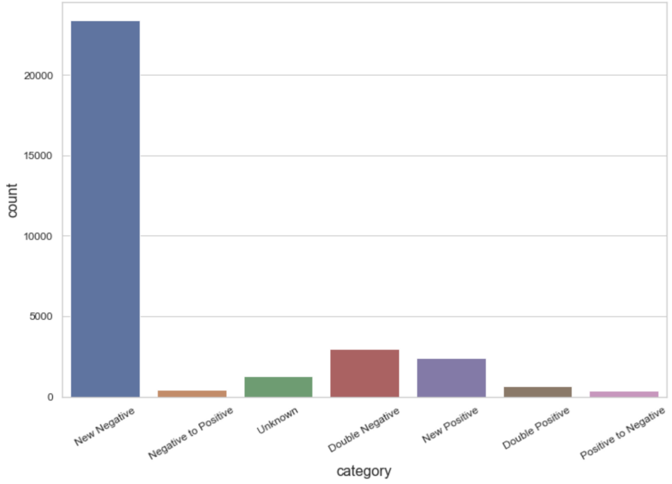


Fig 5.4

The top three categories are New negative, Double Negative and New Positive.

6. Bivariate Analysis -

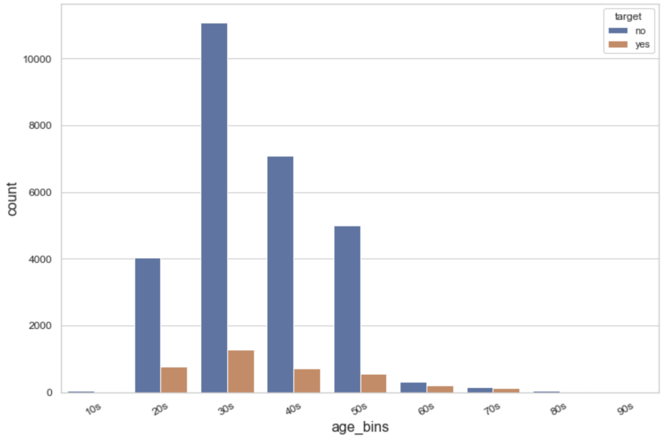


Fig 6.1

Using binning for Age attribute we can infer that people in thier 20s, 60s and 70s are more likely to respond more positively than negatively when compared to all other groups, i.e success rate is more in age group of 20-29, 60-69 and 70-79.

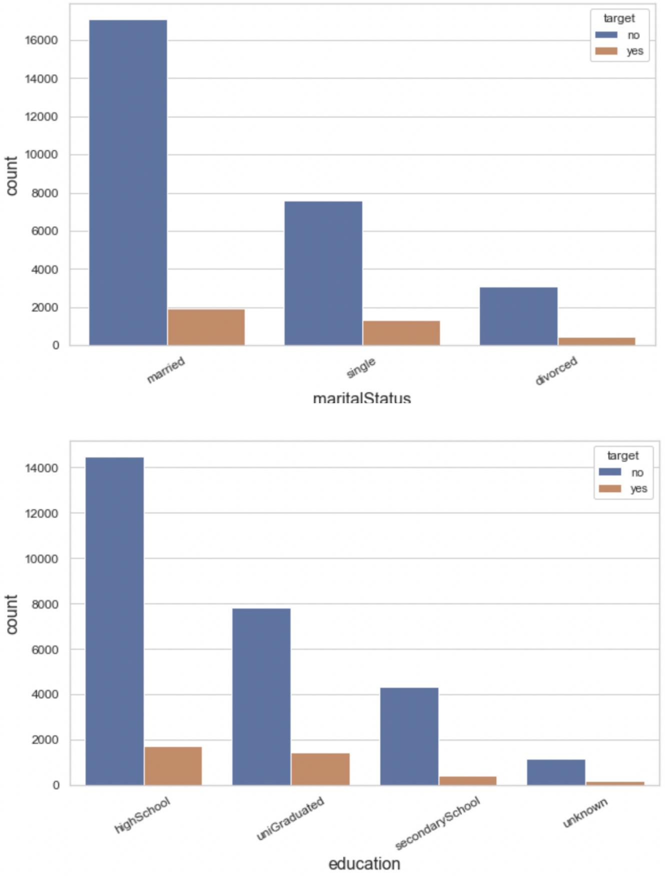


Fig 6.6

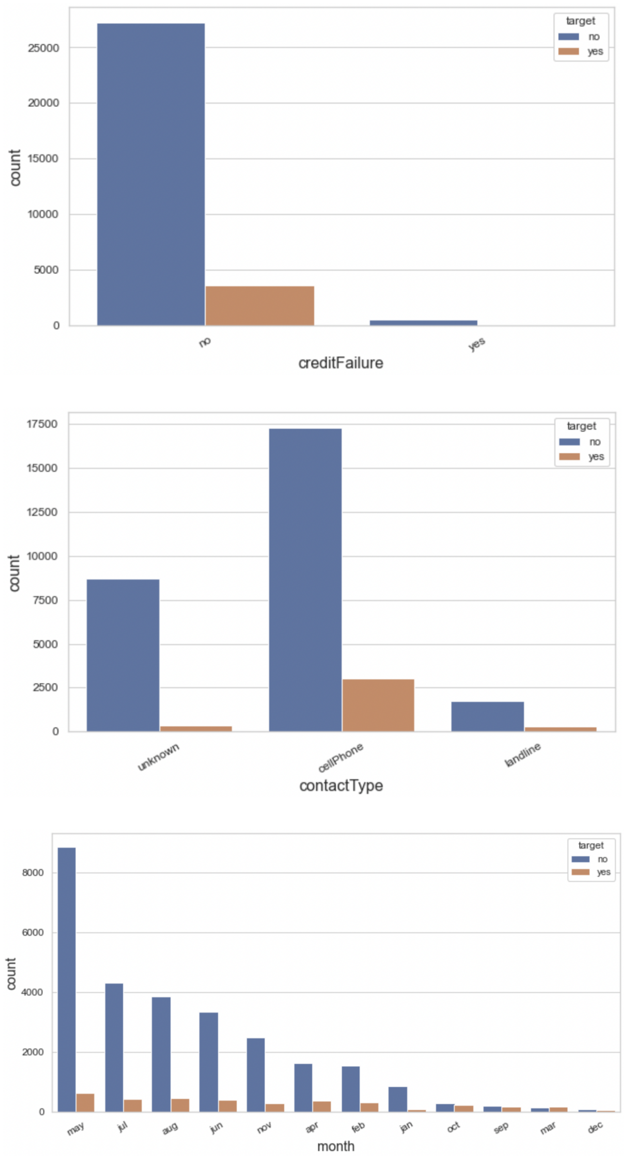


Fig 6.2

Some inference from the bivariate analysis is that From the above plot we can say that we get highest success rate during months of october, september, december. March seems to be the best month because the number of positive responses is higher than negative ones. Lowest success rate of marketing campaign is seem in month of may. CellPhone is the best medium, is suggested by the above analysis. There are no instances where someone has defaulted in credit and responded postitively, this suggest that campaign should target people who have positive credit. University graduates seem to be the best performing class on target while highschool seems to have lowest success rate.

7. Correlation Analysis -

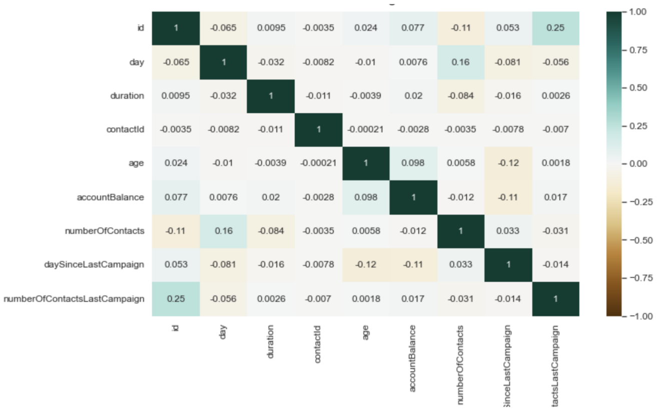
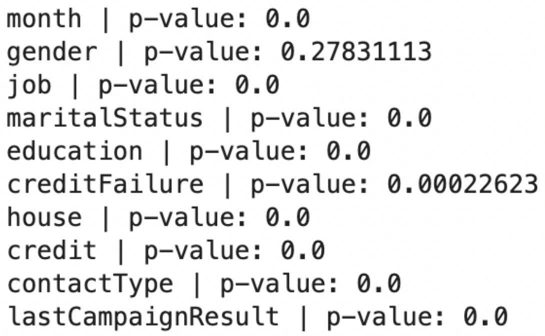


Fig 7.1

Duration has the highest positive correlation with target. Account balance, No of contacts last campaign, age also have significant positive correlation. No of contacts and day since last campaign has lowest negative correlation.



```
month | p-value: 0.0
gender | p-value: 0.27831113
job | p-value: 0.0
maritalStatus | p-value: 0.0
education | p-value: 0.0
creditFailure | p-value: 0.00022623
house | p-value: 0.0
credit | p-value: 0.0
contactType | p-value: 0.0
lastCampaignResult | p-value: 0.0
```

Fig 7.2

Using Chi-Squared test to test the correlation between categorical variables and the target variable. Here p-value is the Probability of H0 being True, H0 or null hypothesis assumes correlation, hence p-value < 0.05 suggests no correlation.

8. Modelling - Classification

Classification is the task of predicting a class label given the set of inputs, if the model is trained to predict a class from two class labels, then it is referred as a binary classification, if the number of classes are greater than two then it is referred to as a multi-class classification problem.

We performed modeling on the following combinations:

- Balanced Dataset
- Imbalanced Dataset
- Dataset with outliers
- Dataset without outliers

This is a binary classification problem, hence we require a model which is a binary classifier. The models we used are as follows:

- Logistic regression
- RandomForest classifier
- AdaBoost classifier
- Decision Tree

Before providing data to the model, we have done various preprocessing steps to the data

- Normalization: We applied normalization to the entire dataset using the StandardScaler() method in sklearn.
- Encoding: To handle the categorical variables in the dataset we have applied one-hot encoding to the categorical variables.
- Oversampling/ Undersampling: The dataset is imbalanced hence we have applied undersampling using RandomUnderSampler() and oversampling using SMOTE.

Logistic Regression(LR):

Logistic Regression is an algorithm used for multi-class classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). During the final stage of prediction it uses a sigmoid function which predicts the probability of the outcome.

RandomForest Classifier(RF):

RandomForest Classifier follows Ensemble modeling wherein multiple models are trained and the output is the majority of the underlying models. It is basically a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It is also known as bagging classifier.

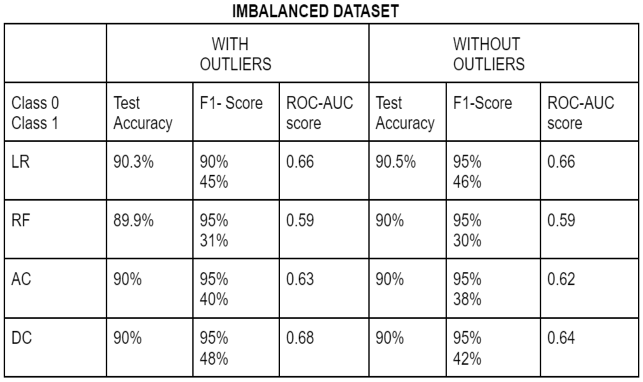
AdaBoost Classifier (AC):

Boosting is an ensemble modeling technique since. It has been a prevalent technique for tackling binary classification problems. These algorithms improve the prediction power by converting a number of weak learners to strong learners.

Decision Tree Classifier(DC):

A decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of examples from one class. Each non-leaf node of the tree contains a split point that is a test on one or more attributes and determines how the data is partitioned.

The performance of the models considering different combinations of the dataset is as follows:

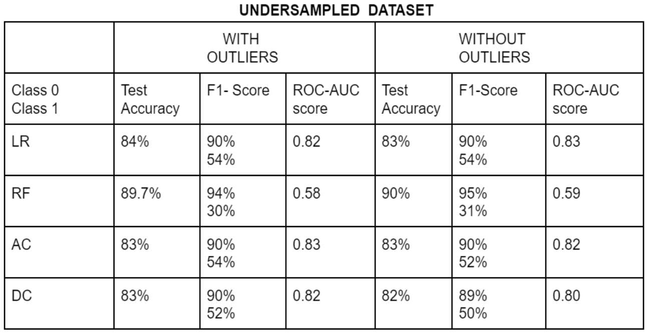


IMBALANCED DATASET

	WITH OUTLIERS			WITHOUT OUTLIERS		
Class 0 Class 1	Test Accuracy	F1- Score	ROC-AUC score	Test Accuracy	F1-Score	ROC-AUC score
LR	90.3%	90% 45%	0.66	90.5%	95% 46%	0.66
RF	89.9%	95% 31%	0.59	90%	95% 30%	0.59
AC	90%	95% 40%	0.63	90%	95% 38%	0.62
DC	90%	95% 48%	0.68	90%	95% 42%	0.64

Fig 8.1

Logistic Regression model works the best for the Imbalanced Dataset with outliers as well as without outliers with the Accuracy of around 90 % , F1-score of around 90% and 45 % for class 0 and class 1 respectively, ROC-AUC score of 0.66 as seen in above figure

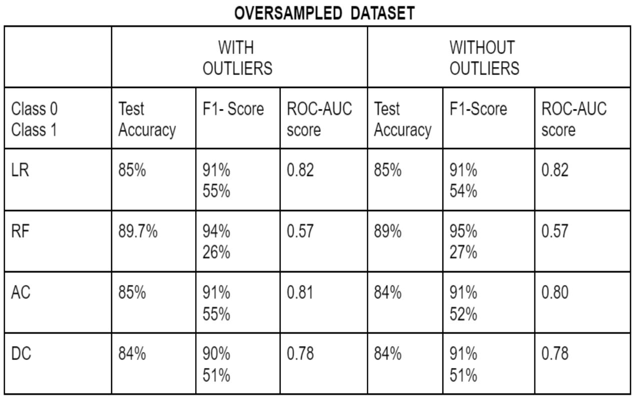


UNDERSAMPLED DATASET

	WITH OUTLIERS			WITHOUT OUTLIERS		
Class 0 Class 1	Test Accuracy	F1- Score	ROC-AUC score	Test Accuracy	F1-Score	ROC-AUC score
LR	84%	90% 54%	0.82	83%	90% 54%	0.83
RF	89.7%	94% 30%	0.58	90%	95% 31%	0.59
AC	83%	90% 54%	0.83	83%	90% 52%	0.82
DC	83%	90% 52%	0.82	82%	89% 50%	0.80

Fig 8.2

The Random Forest and AdaBoost classifier shows considerable good performance for the undersampled dataset without outliers with the Accuracy of around 90 % and 83% , F1-score of around 95%, 90% and 31%, 52% for class 0 and class 1 respectively, ROC-AUC score of 0.59, 0.82 for RandomForest and AdaBoost respectively as seen in above figure



OVERSAMPLED DATASET

	WITH OUTLIERS			WITHOUT OUTLIERS		
Class 0 Class 1	Test Accuracy	F1- Score	ROC-AUC score	Test Accuracy	F1-Score	ROC-AUC score
LR	85%	91% 55%	0.82	85%	91% 54%	0.82
RF	89.7%	94% 26%	0.57	89%	95% 27%	0.57
AC	85%	91% 55%	0.81	84%	91% 52%	0.80
DC	84%	90% 51%	0.78	84%	91% 51%	0.78

Fig 8.3

RandomForest classifier works the best for the Oversampled Dataset with both outliers and without outliers with the Accuracy of around 89 % , F1-score of around 95% and 27 % for class 0 and class 1 respectively, ROC-AUC score of 0.57. However AdaBoost has a satisfactory ROC-AUC of around 0.80 and F1-score of around 91% and 52% as seen above.

9. Conclusion and Future Scope:

- From the observation of the results of different models, we can infer that AdaBoost can provide satisfactory results considering the imbalance issue in the dataset.
- One of the problems with the above trained model is that it can only be applied to a product to which responses of users are collected. Means this model is highly training data specific for example, a model trained on customer responses for a cosmetic product can behave differently when predicting results of diverging categories like bank or property.
- Merging more relevant data from different sources can help build a better performing classifier which can possibly generalize the model

10. References –

1. https://towardsdatascience.com/how-to-predict-the-success-of-your-marketing-campaign-579fbb153a97

2. https://www.kaggle.com/code/seananguyen/marketing-campaign-analysis-python#3.-Data-Visualizations

3. https://www.kaggle.com/code/khanimar/bi-marketing-campaign-eda-analysis-prediction/data

4. https://waypointmc.com/blog/analyzing-marketing-results#:~:text=What%20is%20Marketing%20Analysis%3F,improve%20future%20conversions%20or%20sale

5. https://www.researchgate.net/figure/Adaboost-RF-algorithm-flowchart_fig1_333499498

6. https://towardsdatascience.com/comparison-of-the-logistic-regression-decision-tree-and-random-forest-models-to-predict-red-wine-313d012d6953

-