Bank Marketing Campaign

30th Aug 2022

Agenda

- Background
- Attributes Information
- Data Analysis and ML approaches
- Data Understanding
- Data cleaning
- Outlier Analysis
- Which Models we used to classify data?
- Evaluation Metric
- ROC Curve
- Results
- Next steps

Data Glacier
Your Deep Learning Partner

Background : Bank Management Campaign

Problem Statement:

- ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Problem description:

- One of the most common marketing strategy In Banking sector is direct marketing campaigns through phone calls ,it is a form of advertising that allows organizations to communicate directly with customers to offer their services based on the client’s existing bank profile .Here we will consider term deposit as a banking service .

Business Objective and understanding:

- The plan is to help ABC company to provide a short list of customer that are more likely to buy their product based on their bank details information such as loan. Marital status, account balance etc. This goal will be achievable by using a sophisticated machine learning algorithm capable of using a customer record to predict their future action in a blink of an eye to reduce the company’s time and resources.
- Objective: ABC Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more.
- The success criteria: for this business problem would be based on how much maximum number of customers we are able to predict who have subscribed to the product.

Data Glacier

Attributes Information

Input variables:

bank client data:

1 - age (numeric) 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no').
Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Data Glacier

Data Exploration and Attributes Information

Other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: 'yes', 'no')

Data Details:

No. of observations are 41188, used only one file with size 4.18 mb.

Data Glacier

Data Analysis and ML approaches

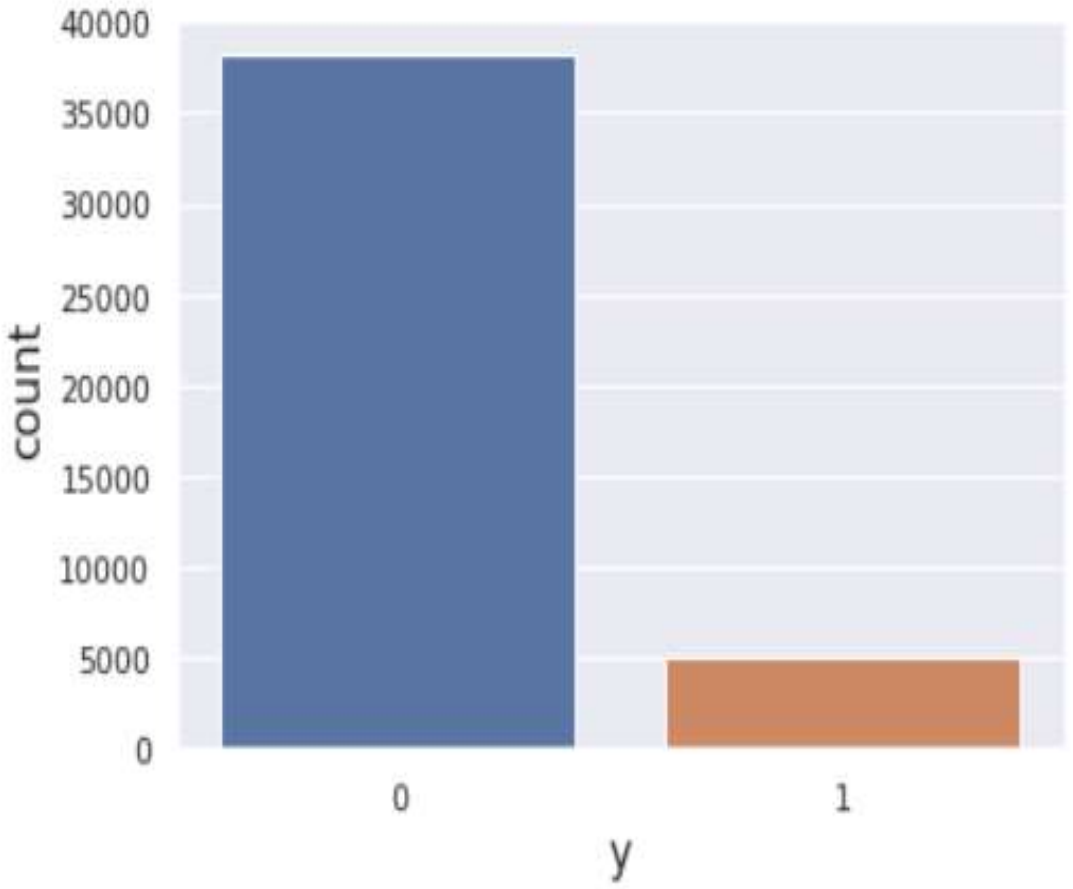
Approaches:

1. The data looks pretty clean. These approaches are basically where were the pain points and what we are trying to solve , The columns which has two values('yes' and 'no') and slightly imbalanced such as default, loan, y, has been converted to (1,0) numerical values. rest are continuous variable were binned so that outliers value are converted into count values.
2. While approaching data cleaning method dropped the outliers and ambiguous values, such as "others" and "unknown".
3. Skewness doesn't provides much insights in data, as values of columns are nearly zero apart from 'previous'. data seems symmetrical.
4. Data has been cleaned with flooring and clapping using interquintile range(IQR) Outliers are removed by dropping values that is below 25% and 75% percentile.
5. In Feature Engineering used undersampling method to delete the samples in majority class as There are huge difference between y=1 and y=0, So the ML algorithm will omit the smaller value, which may affect the performance of algorithm.
6. Imbalanced dataset is a common problem in data science; however some approaches have been used on the dataset such as over and undersampling methods as well as boosting algorithm (for traditional machine learning approach) so we have used XGBOOST and Catboost algorithms
7. We can also use booting trees like adaboost, xgboost and random forests that are more robust to imbalanced datasets.
8. Also a popular method for dealing with imbalanced datasets for machine learning models and deep learning models within the preprocessing phase is data augmentation.
9. We used logistic Regression model as it can provide better accuracy after providing model pipeline like min_max normalization.
10. Heterogeneous ensembling methods which combines several base model XGBoosting, Gradient Boosting, Logistic Regression and Catboost to produce final optimum solution by calculating their cross entropy loss.

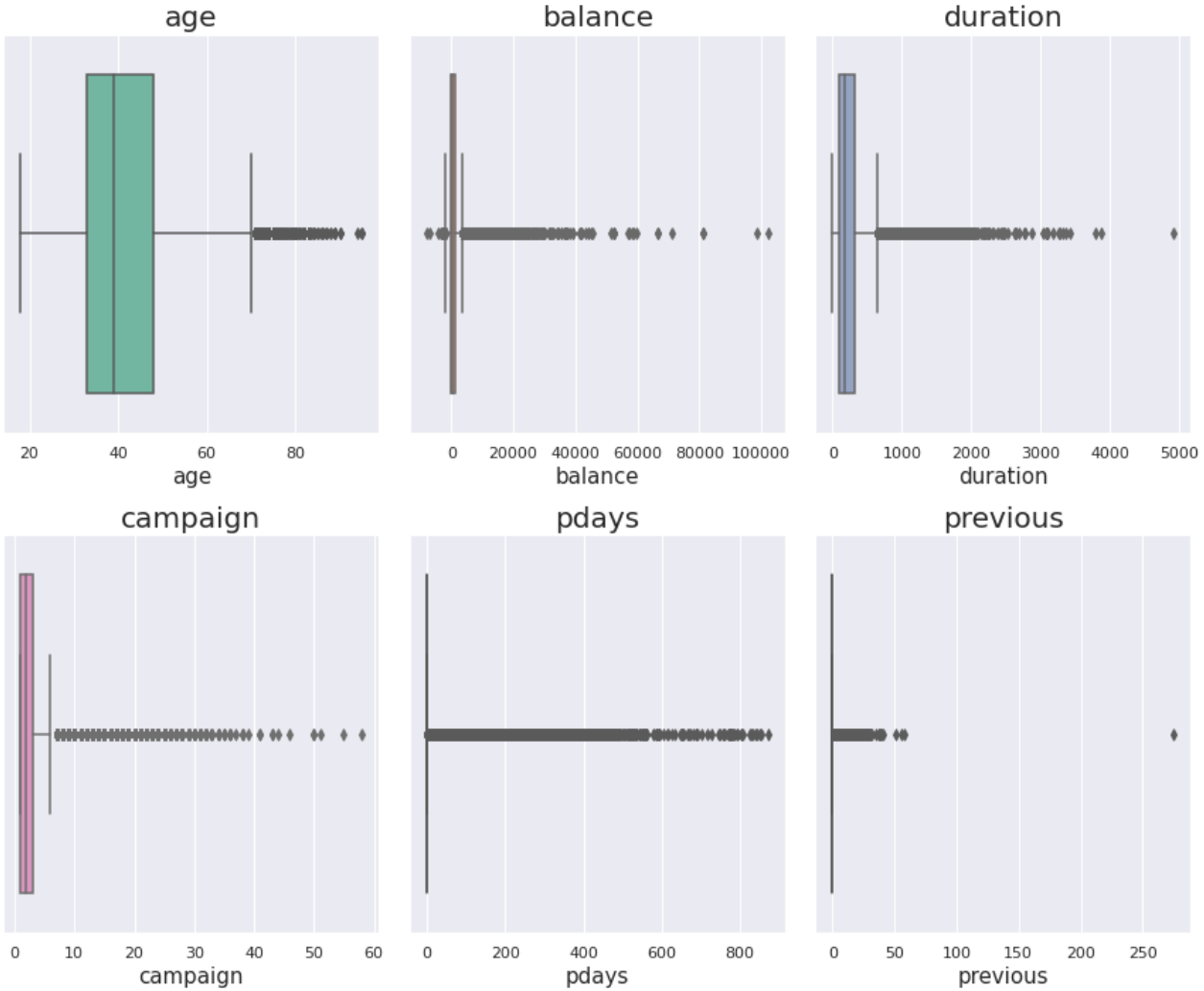
Data Glacier

Data Understanding

- There was a challenge in the data when we found the target was imbalanced so we used skewness and interquantile range to find outliers and remove them from dataset.
- As shown in the figure, term deposit “Y” is highly imbalanced and handled with undersampling technique.
- The majority class is 88.38% and minority class is 11.62%
- Now the challenge is to check which customers bank can take into consideration as convert them to take term deposite.

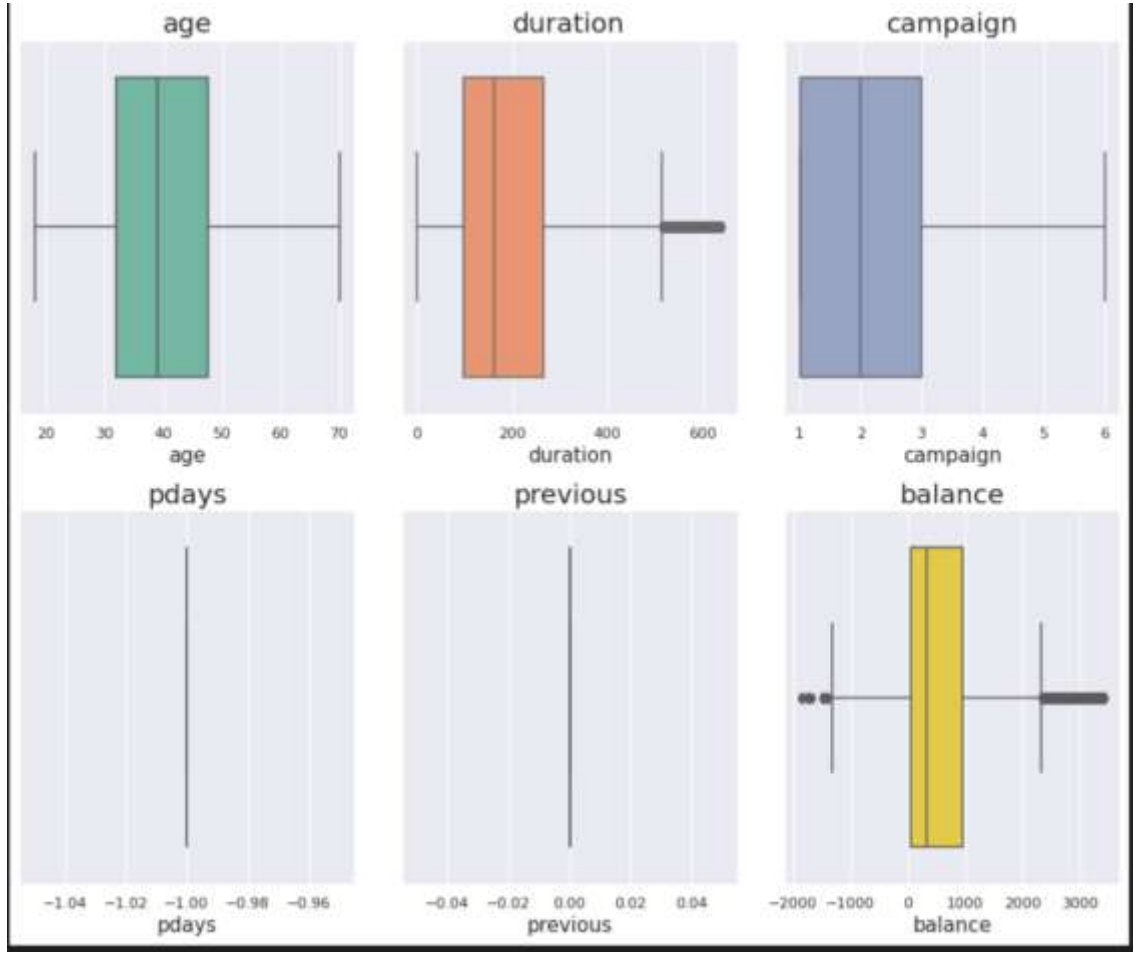


Data cleaning



- While dropping outliers are found some equally spaced values that are selected from the cumulative distribution function of a random variable in a way that divides the data set into equal parts.
- The difference between a good and an average machine learning model is often its ability to clean data. One of the biggest challenges in data cleaning is the identification and treatment of outliers.
- So used interquantile range and skewness to detect and remove outliers.

Outlier Analysis



The difference between a good and an average machine learning model is often its ability to clean data. One of the biggest challenges in data cleaning is the identification and treatment of outliers. In simple terms, outliers are observations that are significantly different from other data points. Even the best machine learning algorithms will underperform if outliers are not cleaned from the data because outliers can adversely affect the training process of a machine learning algorithm, resulting in a loss of accuracy

Which Models we used to classify data?

Models for Boosting:

- Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. We've used '**gradient boosting classifier**'. We experiment with two main hyperparameters now - learning_rate (shrinkage) and subsample . By adjusting the learning rate to less than 1, we can regularize the model. there's a trade-off between learning_rate and n_estimators - the higher the learning rate, the lesser trees the model needs (and thus we usually tune only one of them). Also, by subsampling (setting subsample to less than 1), we can have the individual models built on random subsamples of size subsample .We got accuracy of 90%.
- Next we used "**Cat Boosting classifier ".** CatBoost is a recently open-sourced machine learning algorithm from Yandex. It yields state-of-the-art results without extensive data training typically required by other machine learning methods, and Provides powerful out-of-the-box support for the more descriptive data formats that accompany many business problems. "CatBoost" name comes from two words "Category" and "Boosting". We got the accuracy of 91%.

Models for Ensemble:

- In Our model we used first "**XGBoost classification**". XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) where we got the accuracy of 90%.

Data Glacier

Which Models we used to classify data?

Models for Linear:

- Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). We got the accuracy of 88%.
- Next model we used is K-Nearest neighbour where we got an accuracy of 88%. The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm , when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

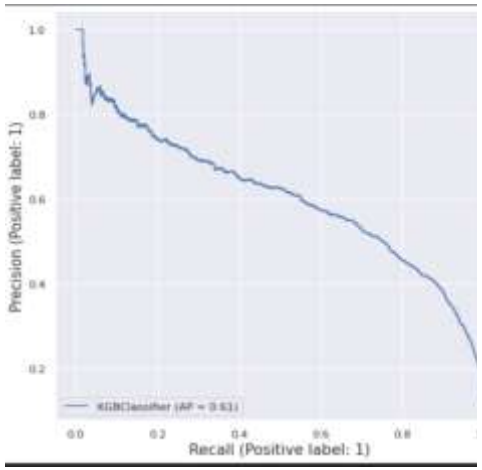
Models for stacking:

- Voting classifier is considered to get final accuracy of all the algorithm, All the models have been stacked , this brings diversity in the output thus it is called heterogeneous ensembling , we got the accuracy 90%
  .

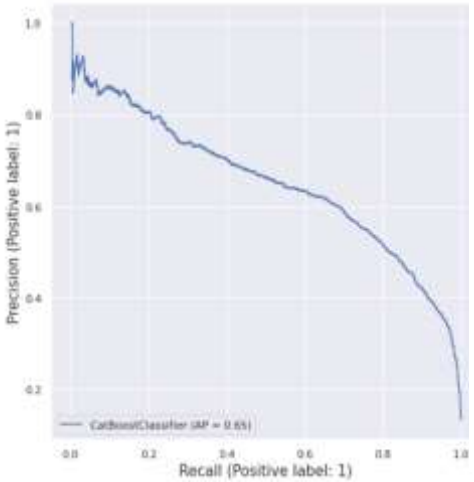
Data Glacier

Evaluation Metric

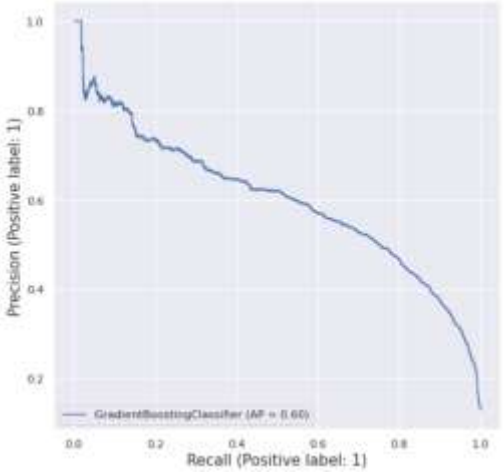
The precision-recall curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).



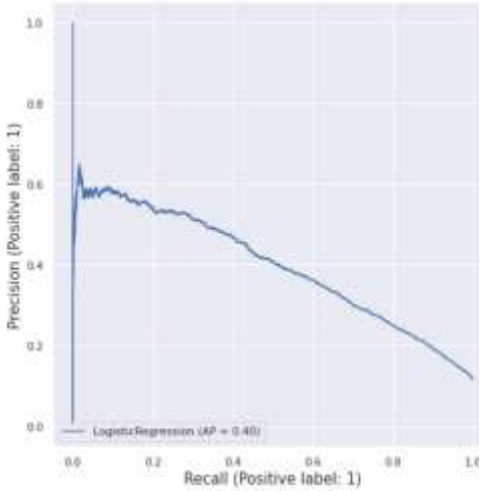
XGB classifier



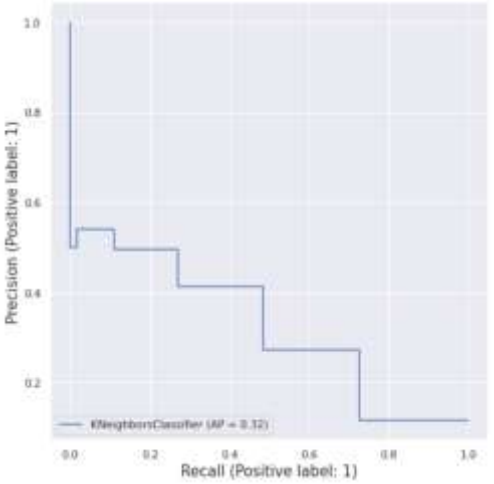
CatBoost classifier



Gradient Boosting classifier



Logistic Regression



KNN Classifier

ROC Curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

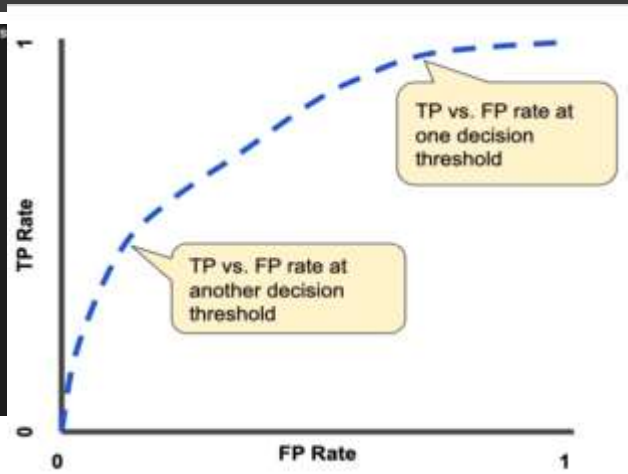
* True Positive Rate
* False Positive Rate

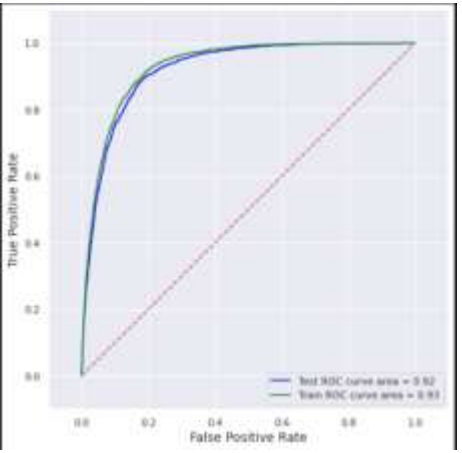
True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

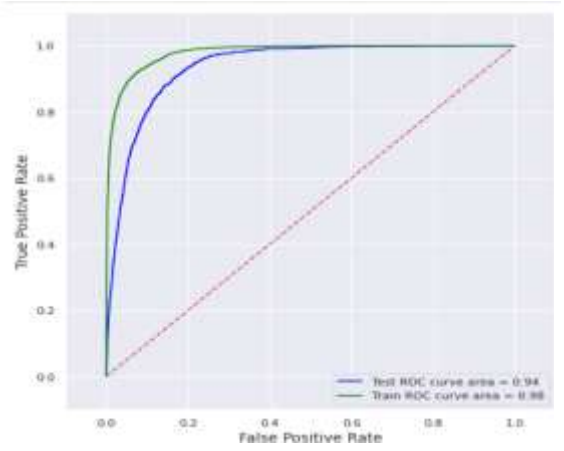
False Positive Rate (FPR) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

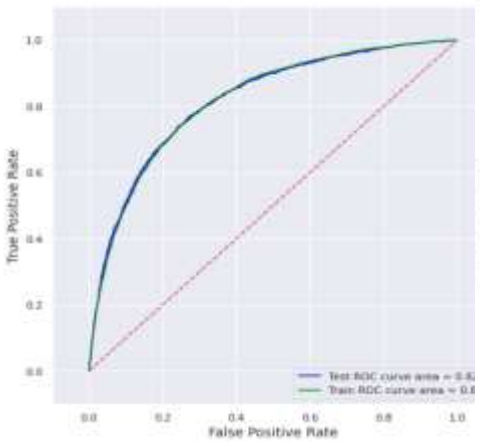




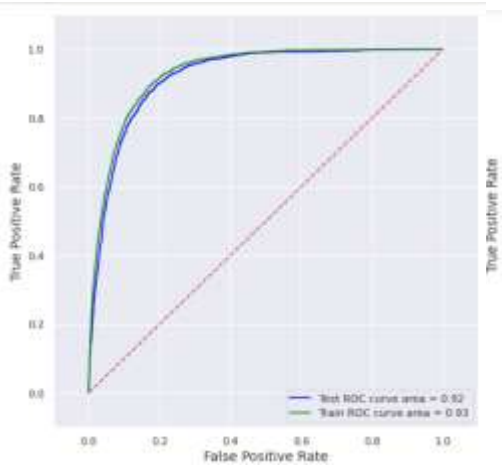
XGB classifier



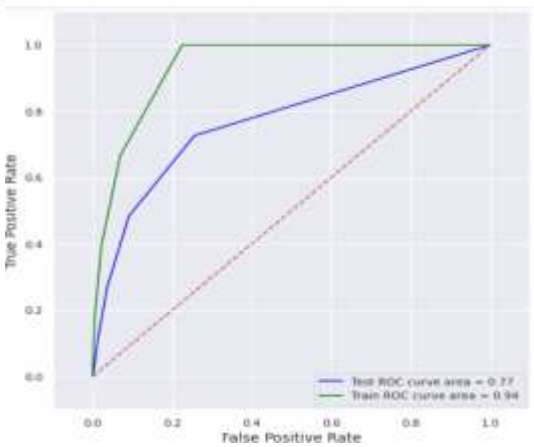
CatBoost classifier



Logistic Regression



Gradient Boosting Classifier



KNN Classifier

Data Glacier

Results

	Accuracy	Precision	Recall	F1 -Score
XGB classifier	90%	0.92	0.97	0.95
CatBoost classifier	91%	0.93	0.97	0.95
Logistic Regression	89%	0.90	0.98	0.94
Gradient Boosting Classifier	90%	0.92	0.97	0.95
KNN Classifier	88%	0.91	0.96	0.94
Voting Classifier	90%			

The best Model is a Catboost classifier providing the accuracy of 91%.

Data Glacier

Next Steps

Next steps to do for marketing team:

- Collaborate with economic Experts
- Be a fast mover, Capture the customers before competiters capture the chance.
- Target relatively old age people
- Convey peace of mind, self investment, steady income source as value of preposition.
- Try to engage customer have long calls.
- Prefer telephone over mobile calls
- Prioritize those customers who were part of previous marketing campaign.

Thank You