Bank Marketing Campaign Analysis

- Ankit Beladiya

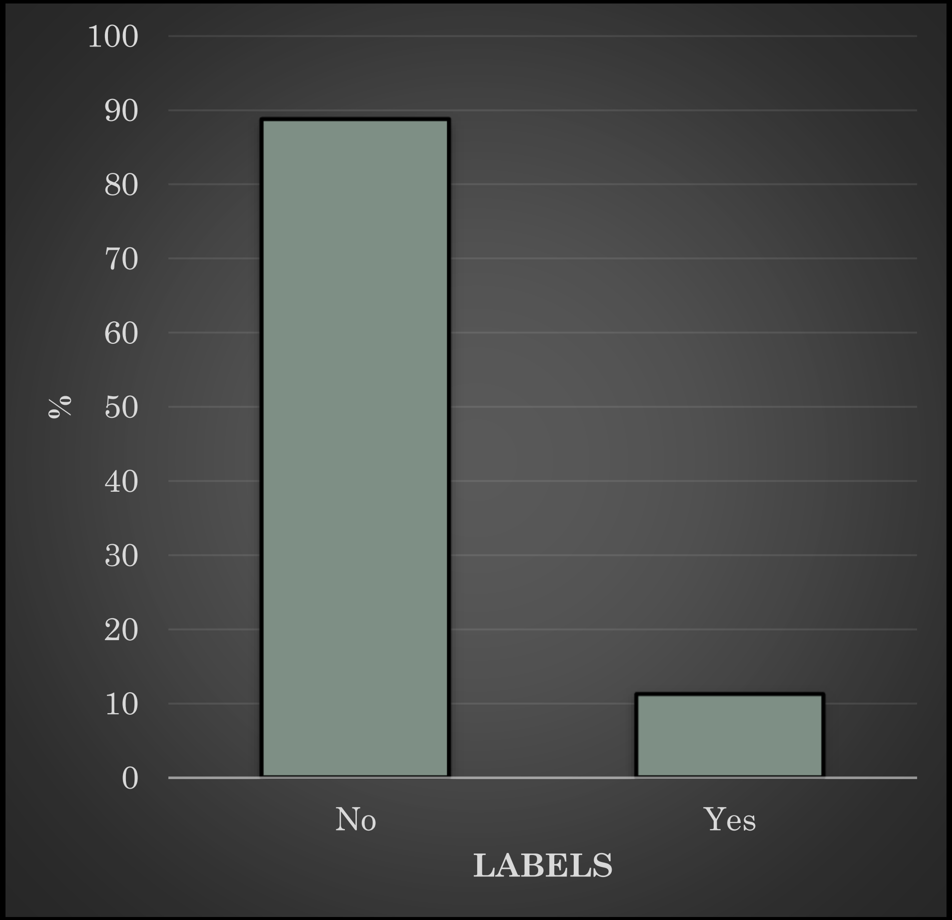
Table of Contents

Problem Definition

- Portuguese banking institution conducted direct marketing campaign (phone calls) to the potential customers to sell the bank product (bank term deposit).
- Our goal is to predict whether the client will subscribe the term deposit or not

Problem Definition – ML Language

- In machine learning terminology, this is a **binary class classification** problem.
- Furthermore, target class distribution is imbalanced(9:1). So, it is an **imbalanced class classification problem.**



Data Exploration and Processing

Data set has total 20 attributes, and 41188 records.

Dataset has 12 duplicate records that will be dropped.

Data Exploration and Processing – Attributes

- Campaign data has mainly 4 types of attributes
  - Client related information e.g. age, job, education, etc
  - Client’s campaign related information e.g. day, month and duration of last contact
  - Social and economic index of the client e.g. employment variation rate, etc
  - Client’s previous information e.g number of contacts performed, etc

Data Exploration and Processing - Categorical Features

- Dataset has 10 categorical features.

Name	Description
Job	Type of job
Marital	Marital status
Education	Highest Education
Default	Has credit in default?
Housing	Has housing loan?
Loan	Has personal loan?
Contact	Communication type
Month	Month of last contact
Day of week	Day of week of last contact
poutcome	Outcome of the previous marketing campaign

Data Exploration and Processing - Categorical Features

- Feature like ***education*** is ordinal feature. So, we can use ordinal encoder for encoding.

education	basic.4y
	basic.6y
	basic.9y
	high.school
	illiterate
	professional.course
	university.degree
	unknown

Data Exploration and Processing - Categorical Features

- Feature like ***Contact*** is binary features and binary encoder can be used for encoding.

contact	cellular
	telephone

Data Exploration and Processing - Categorical Features

contact	2
day_of_week	5
default	3
education	8
housing	3
job	12
loan	3
marital	4
month	10

- None of the remaining feature has high cardinality. Therefore these features can be converted into numeric using one-hot encoding.

Data Exploration and Processing - Numerical Features

- Dataset has 9 numerical (discrete) features.

Name	Description	Number of Distinct Values
Age	Age	78
Campaign	Number of contact performed during this campaign	42
Cons.conf.idx	Consumer confidence index	26
Cons.price.idx	Consumer price index	26
Emp.var.rate	Employment variation rate	10
Euribor3	Euribor 3 month rate	316
Nr.employed	Number of employees	11
Pdays	Days passed by after the client was last contactedfrom a previous campaign	27
Previous	contacts performed before this Campaign	8

Data Exploration and Processing - Numerical Features

	mean	std
age	40.024060	10.421250
campaign	2.567593	2.770014
cons.conf.idx	-40.502600	4.628198
cons.price.idx	93.575664	0.578840
emp.var.rate	0.081886	1.570960
euribor3m	3.621291	1.734447
nr.employed	5167.035911	72.251528
pdays	962.475454	186.910907
previous	0.172963	0.494901

- All the numerical features are not in same scale. So, standard scaling should be performed.

Data Exploration and Processing - Numerical Features

- Some of the features are highly corelated feature therefore these can be dropped.

	emp.var.rate	euribor3m	nr.employed
emp.var.rate	1.000000	0.972245	0.906970
euribor3m	0.972245	1.000000	0.945154
nr.employed	0.906970	0.945154	1.000000

Data Exploration and Processing - Sanity Check

- Duration Feature
  - It is duration of the call of last contact
  - This attribute highly affects the target e.g. if duration is 0 than output is No.
  - However, duration is not known before call is performed
  - Therefore, it should be dropped for modelling.

Data Exploration and Processing - Sanity Check

- Some categorical features has missing value (unknown). Possible solutions are as follow
  1. Impute missing category with most repeating category.
  2. Define missing category as its own category and let model handle it.

	missing value count
job	330
marital	80
education	1731
default	8597
housing	990
loan	990

Data Exploration and Processing - Sanity Check

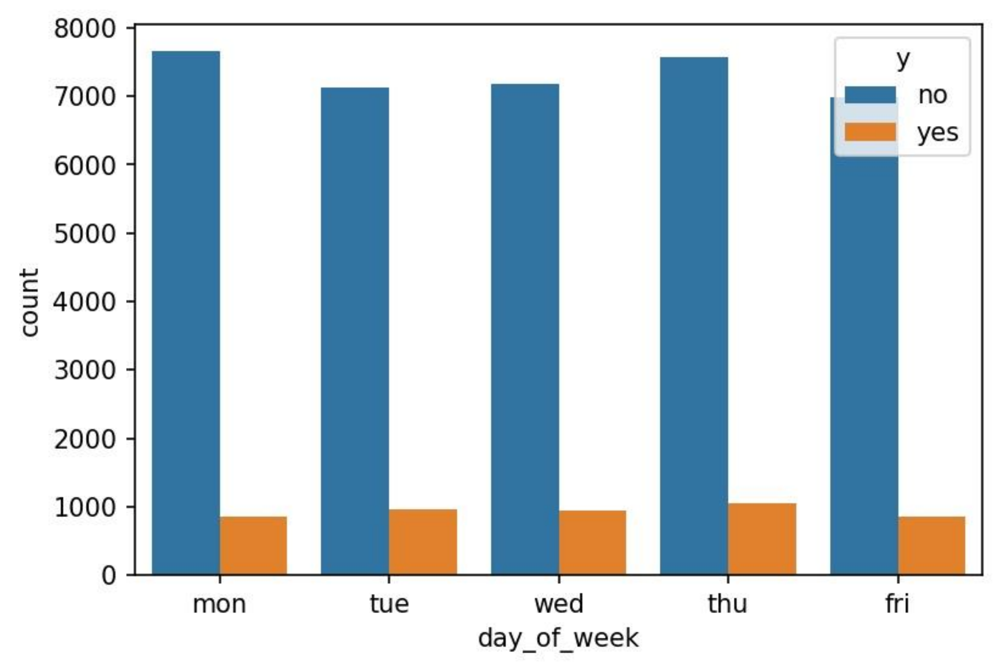
```
data['pdays'].value_counts()
```

```
0          15
1          26
2          61
3         439
4         118
5          46
6         412
7          60
8          18
9          64
10         52
11         28
12         58
13         36
14         20
15         24
16         11
17          8
18          7
19          3
20          1
21          2
22          3
25          1
26          1
27          1
999     39673
```

- Pdays (days passed by after the client was last contacted) feature has outlier (i.e. 999 means client was never contacted)
- Therefore, we can replace 999 by value 0.

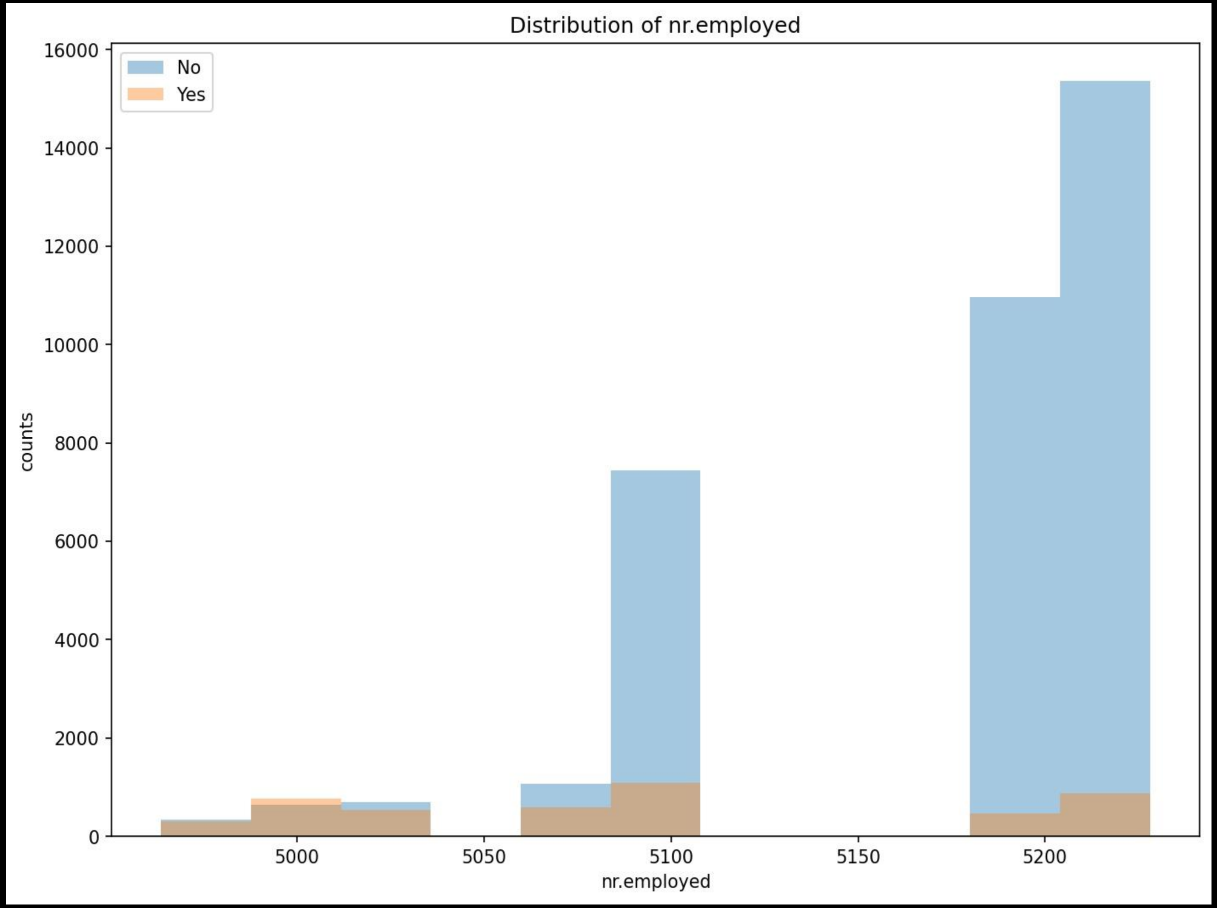
Data Exploration and Processing – Features Distribution

- Dataset contains many features which do not have much information to separate target classes. For example, day of week feature has same number of counts for target classes across all its categories.
- This type of features can be drop as it do not add any value.



Data Exploration and Processing – Features Distribution

- Feature like nr.employed separates target well. Therefore, this kind of feature will get more importance.

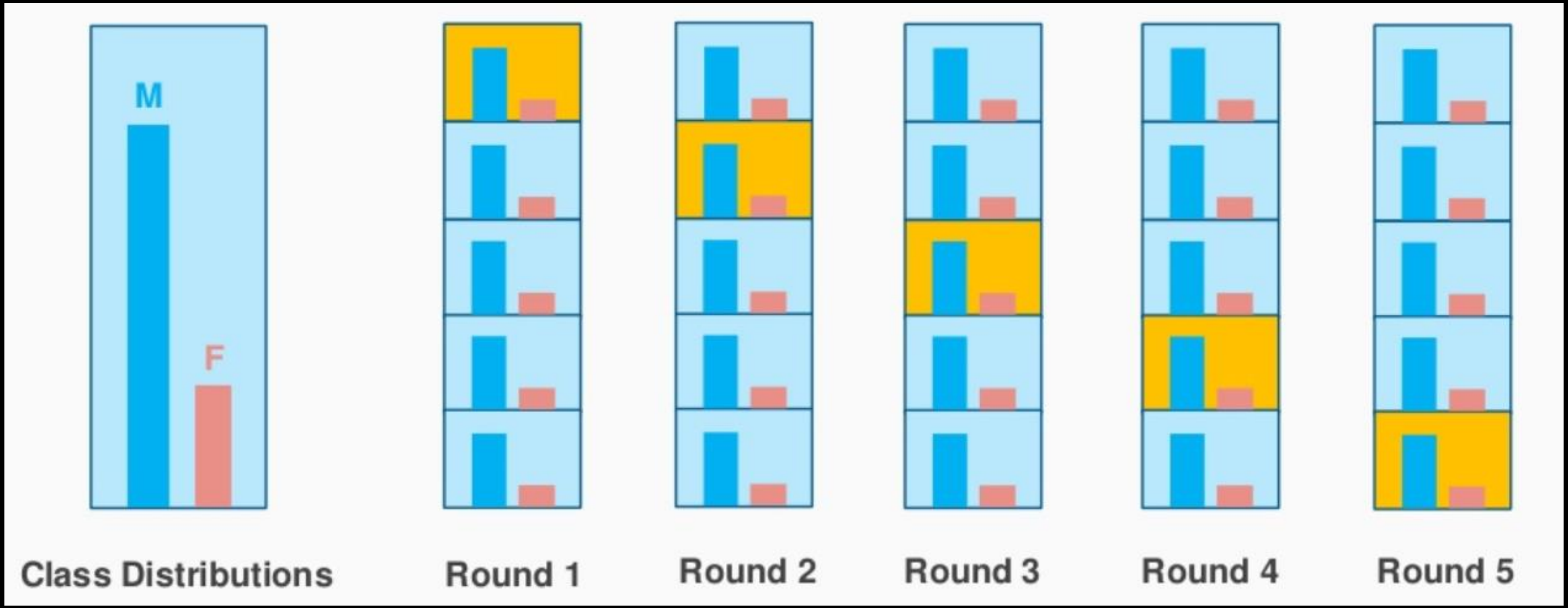


Modeling – Handling imbalanced classes for model training

- To handle imbalanced classes for model training there are mainly 2 options.
  1. Penalize Model **:-** This method imposes an additional cost on the model for making classification mistakes on the minority class during training. These penalties can bias the model to pay more attention to the minority class.
  2. Resample the dataset :- This method suggest to under sample majority classes and/or over sample minority classes(using synthetic data generation).

Modeling – Cross Validation

- As we already know that we are dealing with imbalanced class classification, we should use stratified cross validation to unsure that we have same classes distribution in training and testing set to validate model performance.



Modeling – Metrics to Measure Performance

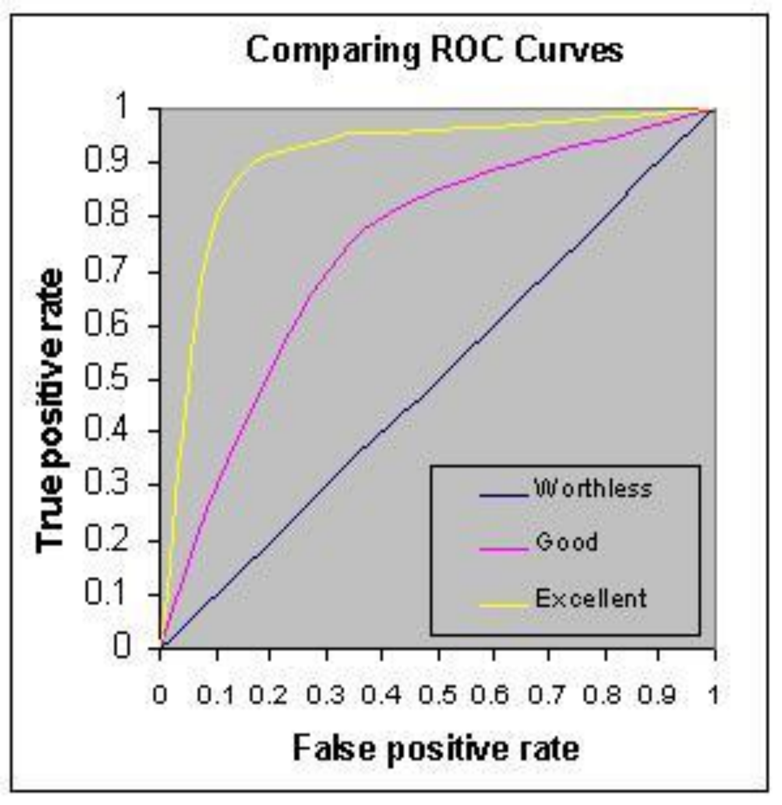
- As our problem is imbalanced class classification, we can not use accuracy to measure the performance of the model.
- For this kind of problem, we can use one of the following metrics
  - Area under ROC curve (AUC)
  - Area under PR curve (AUC-PR)
  - Recall (TPR)
  - Precision
  - F1-Score,
  - Etc…

Modeling – Right Metric to Answer Right Question

- Selection of the right metrics depends on business problem.
- For example, let's say bank has 1 million clients but bank can only contact 1000 clients everyday. Therefore, bank would like to prioritize call first to only those customers, who would most likely to subscribe the term.
- In this case, we should select precision as a metrics.

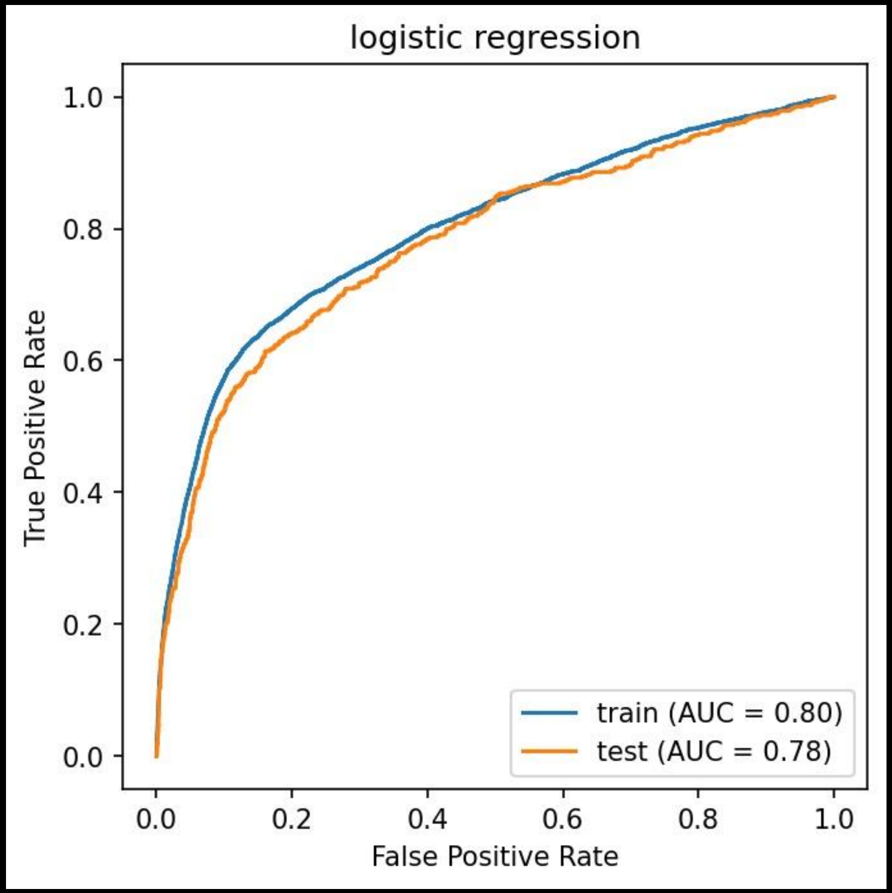
$$Precision = \frac{T_p}{T_p + F_p}$$

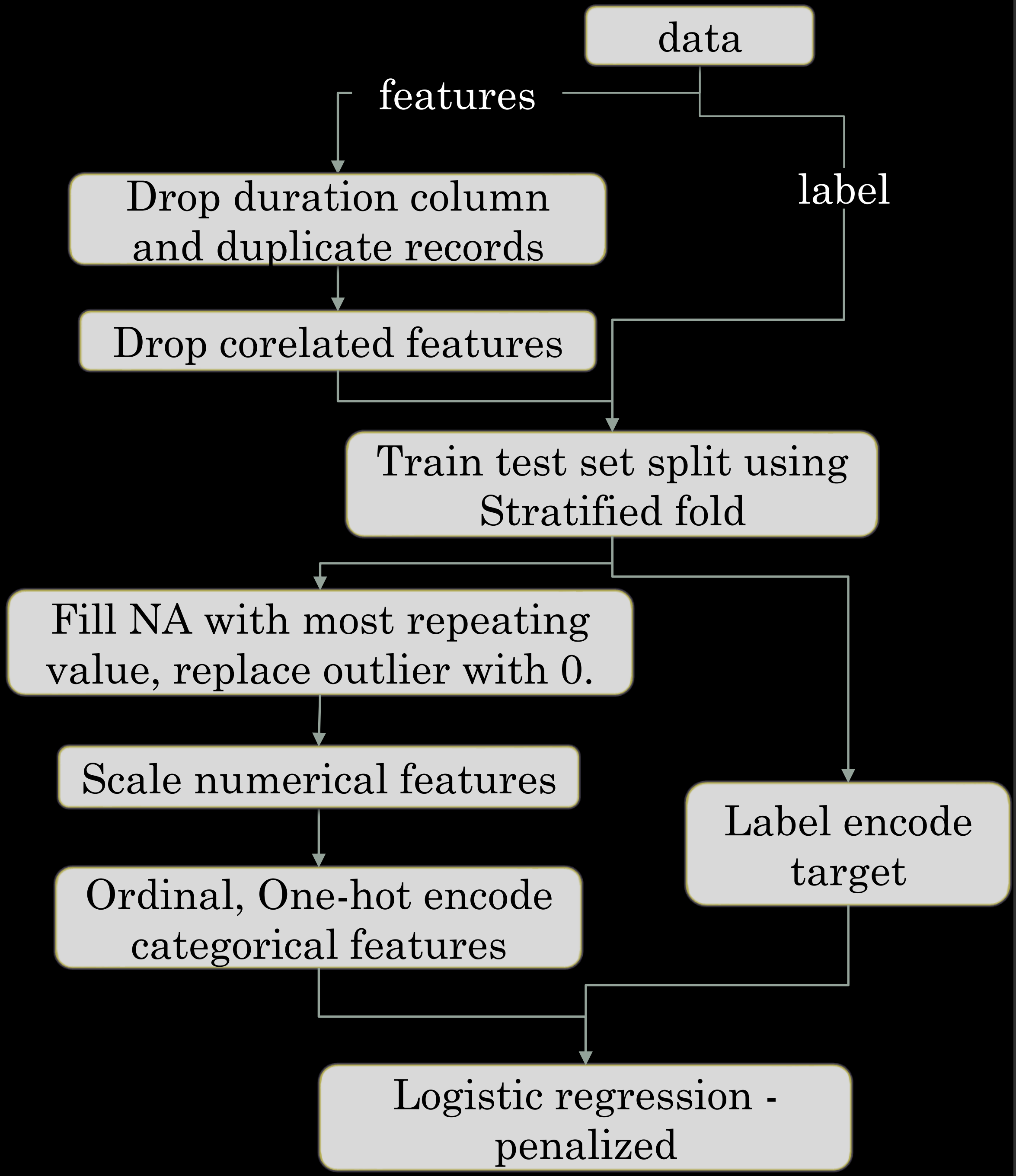
Modeling – Selection of Metric



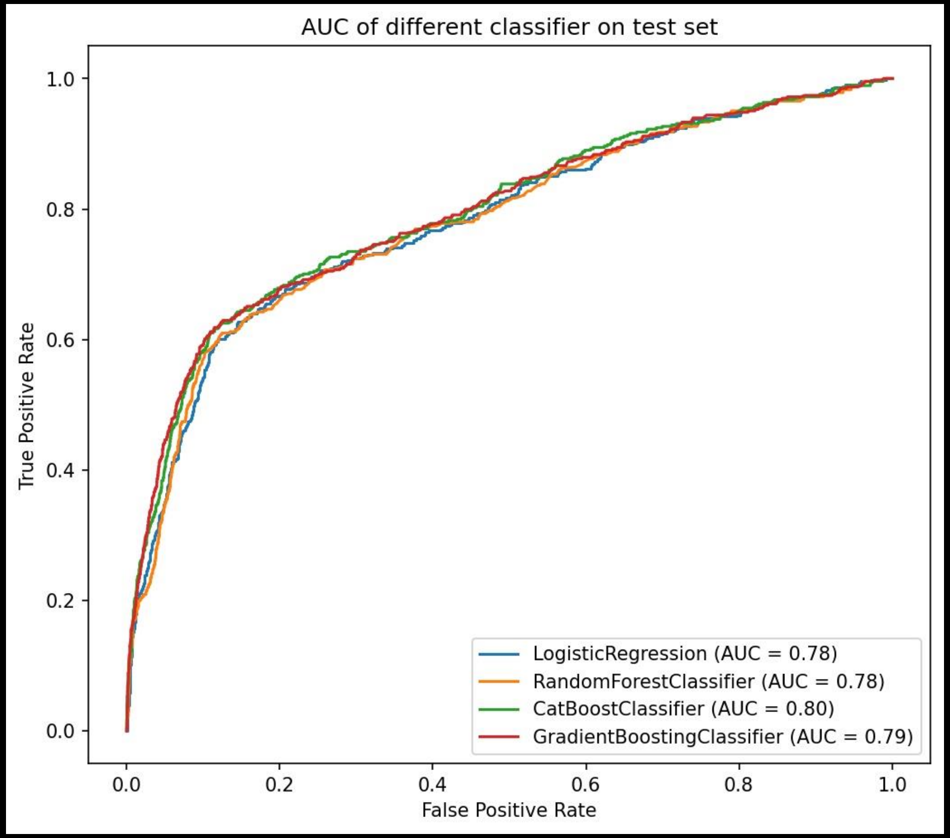
- For this problem, we want to find clients who will be receptive to the marketing campaign. Therefore, we should optimize over model to give higher **recall(TPR)**
- However, any model that is predicting only positive class will give higher recall. So, we should also account for false positive rate(FPR).
- Therefore, area under ROC curve (AUC) is a good metrics to calculate model's goodness to separate both classes for different probability threshold.
- After selecting best model, we can tune probability threshold for higher TPR and lower FPR.

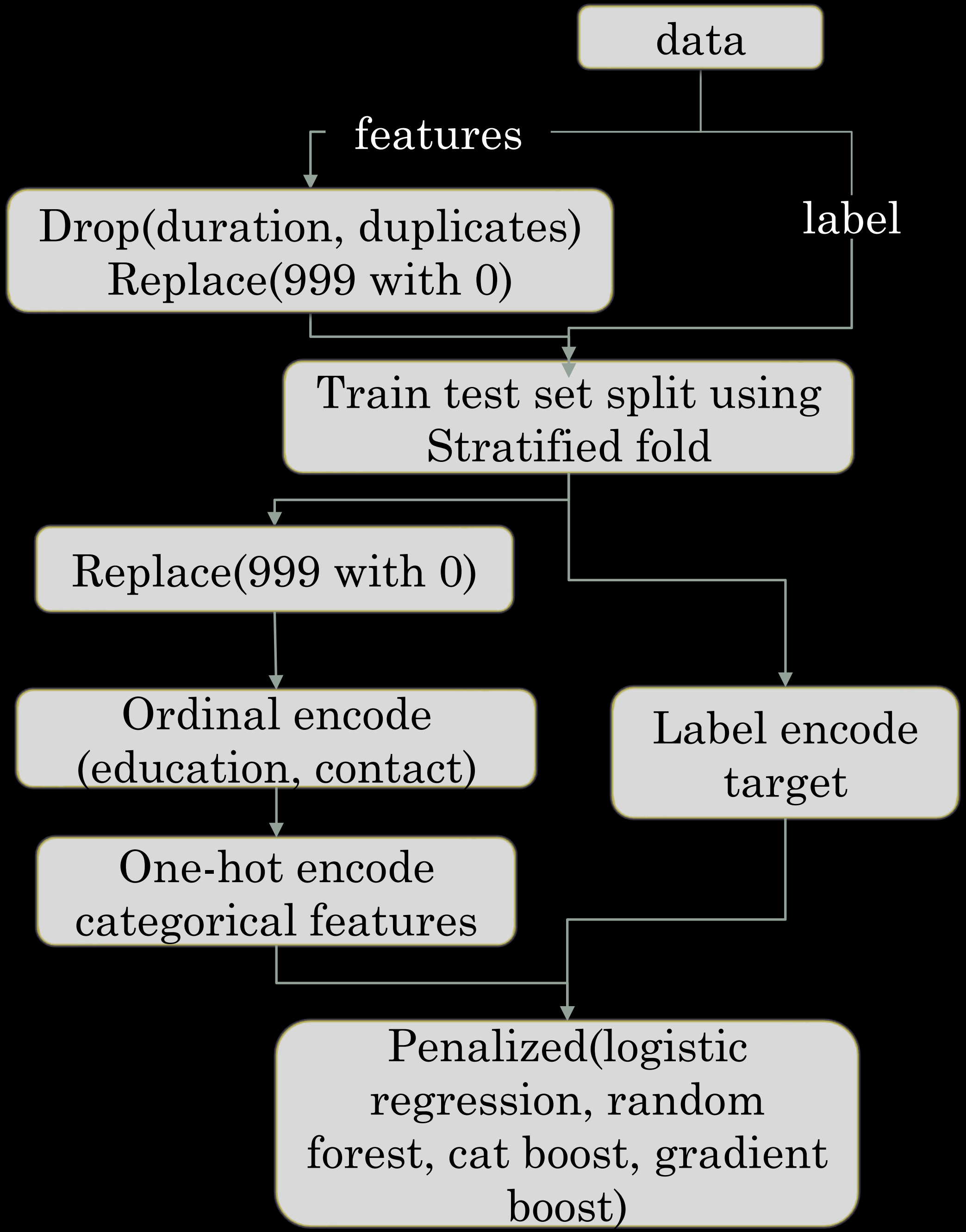
Modeling - Baseline



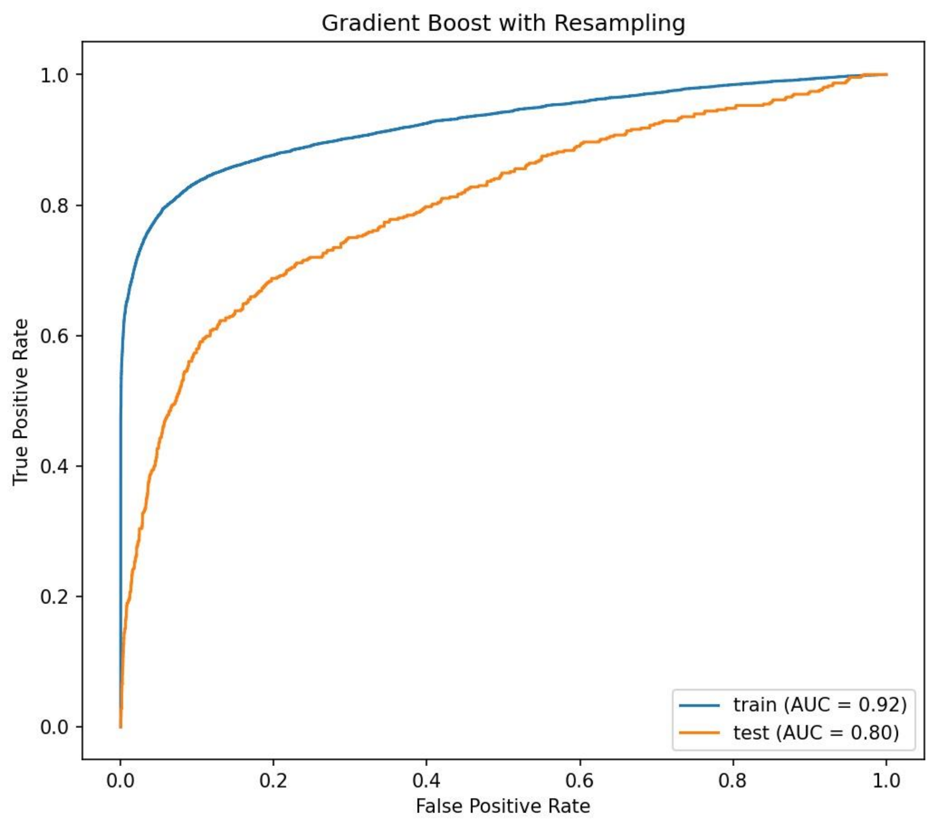


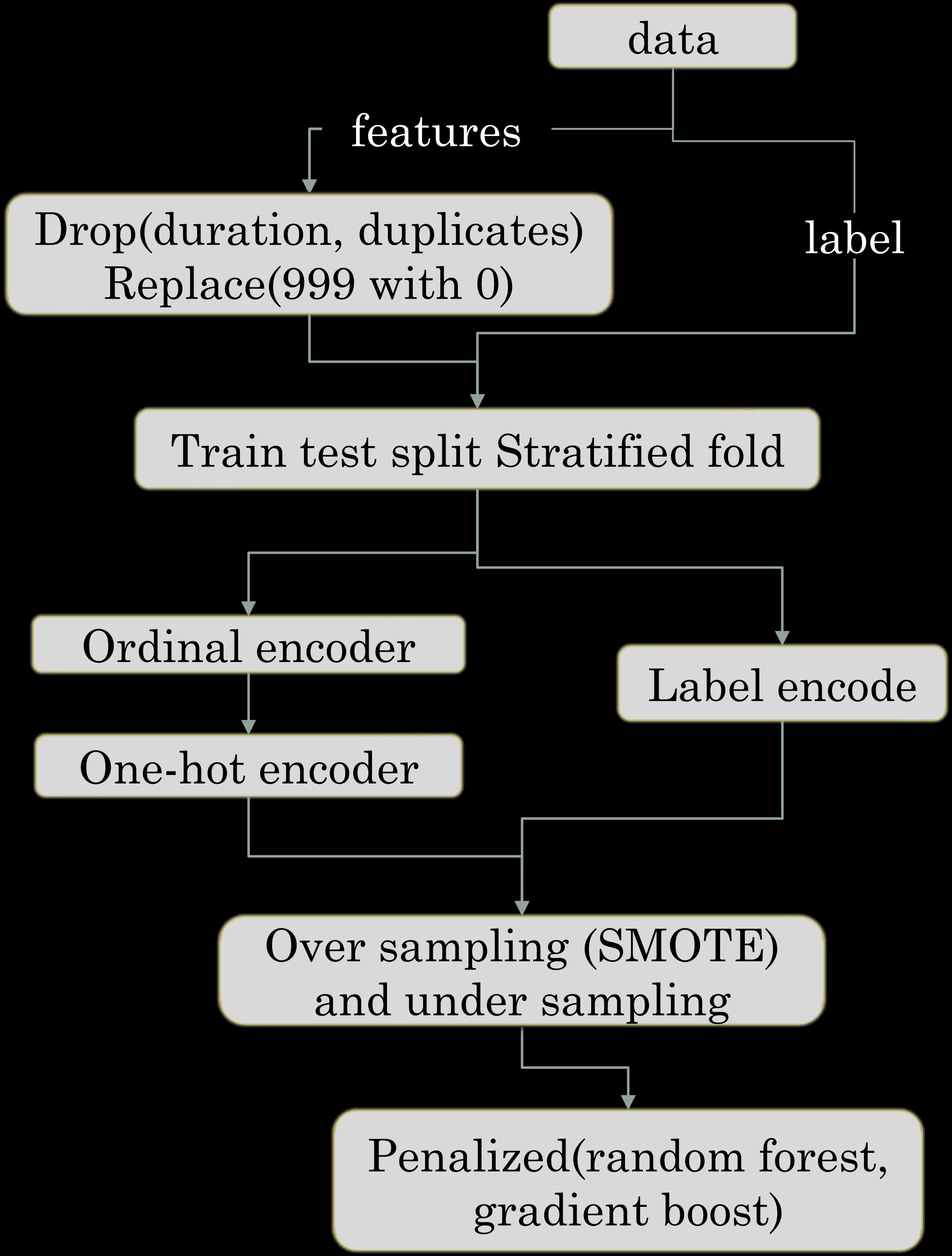
Modeling – Improving Performance





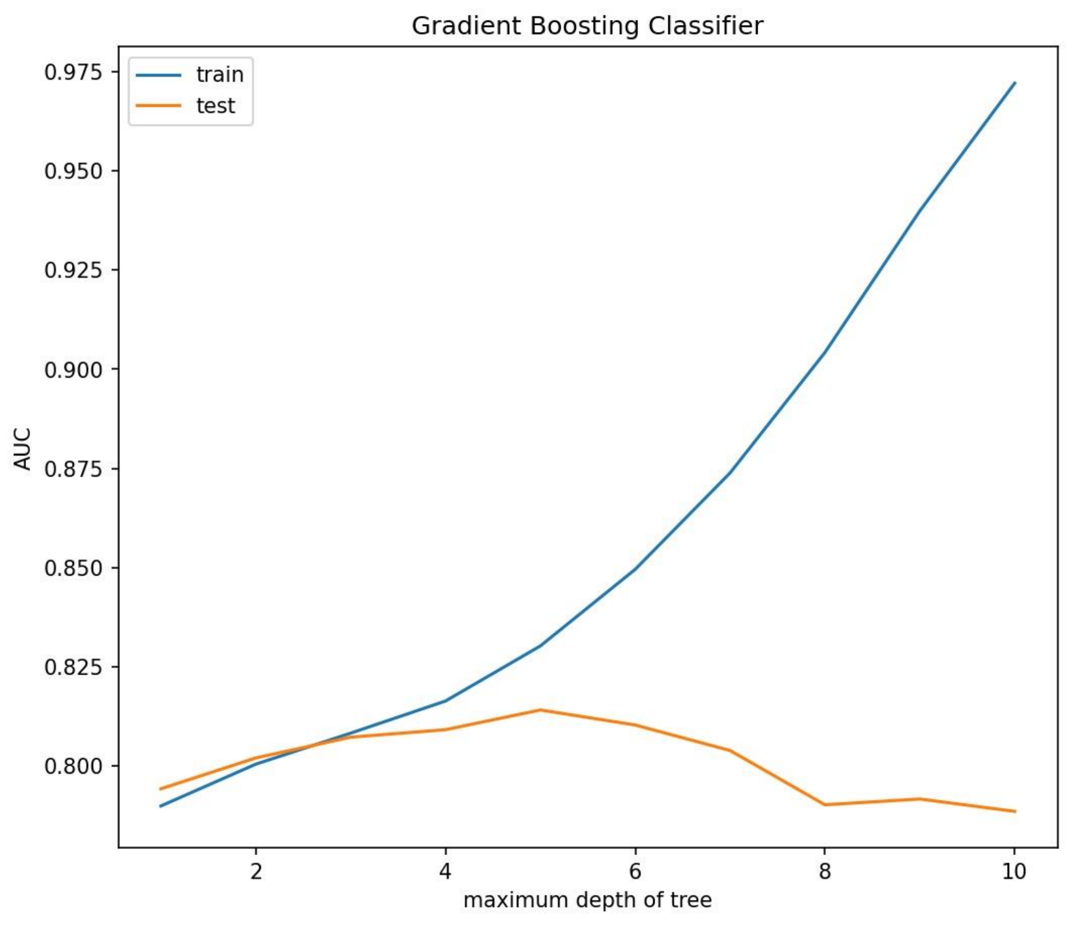
Modeling - SMOTE





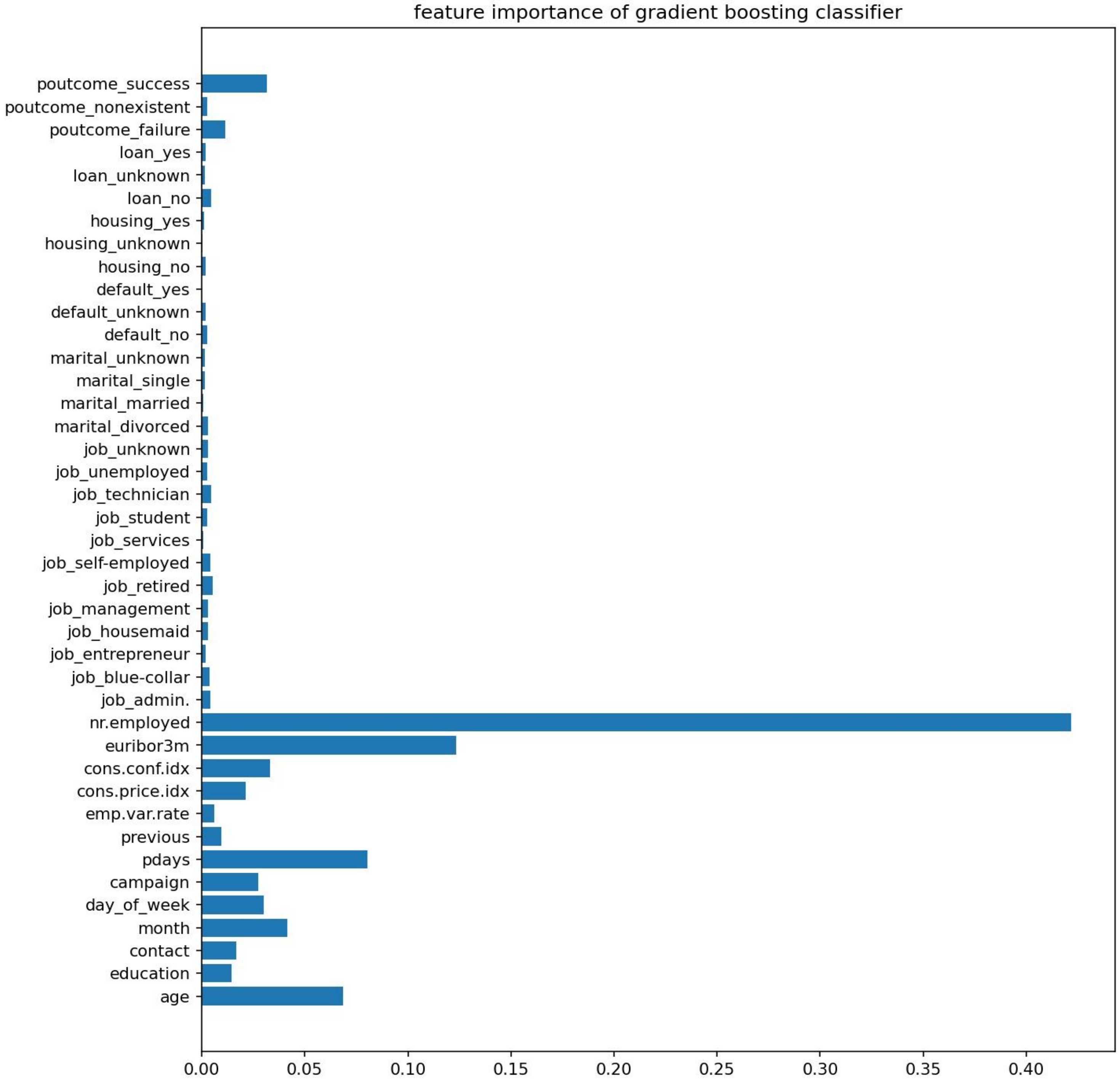
Modeling – Handling Overfitting

- For producing baseline, regularization was used to prevent overfitting.
- Trees were pruned using maximum depth of tree parameter to reduce overfitting in gradient boosting and random forest classifier.



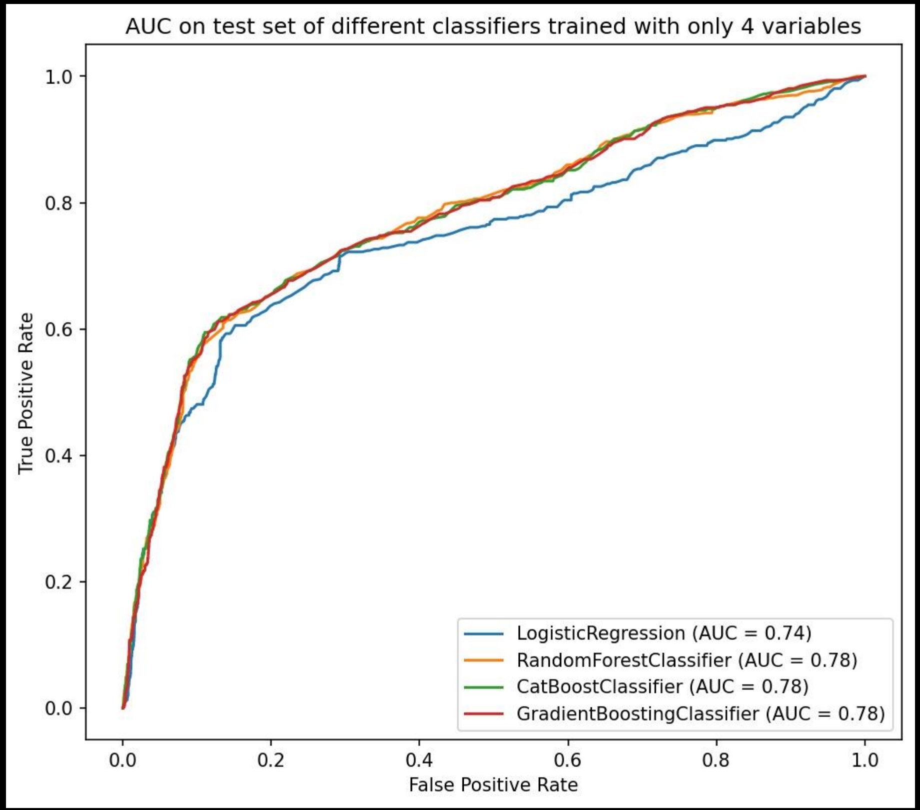
Modeling - feature importance

- This graph shows feature importance learned by Gradient boosting model.
- As we can see that some features has vey low importance. So, dropping those features will not hurt performance of model.



Modeling – Feature Pruning

- As many features has very low feature importance dropping those features will not hurt performance if the model.
- Model trained with only 4 features (age, cons.price.idx, cons.conf.idx, nr.employed) can compete with model trained with all features.



Modeling – Probability Threshold for Higher TPR for lower Gradient Boost

Threshold = 0.5

Threshold = 0.11

	default
TPR	0.237069
FPR	0.016694
Threshold	0.5

	tuned
TPR	0.618534
FPR	0.114669
Threshold	0.119277

Modeling – Metrics summary

	Random forest		Random forest prune	
	default	tuned	default	tuned
TPR	0.706897	0.609914	0.609914	0.614224
FPR	0.259168	0.124521	0.136836	0.139026
Threshold	0.5	0.574769	0.5	0.477028

	Gradient Boosting		Gradient Boost prune	
	default	tuned	default	tuned
TPR	0.237069	0.618534	0.193966	0.612069
FPR	0.016694	0.114669	0.021346	0.1289
Threshold	0.5	0.119277	0.5	0.119949

	Cat Boost		Cat Boost prune	
	default	tuned	default	tuned
TPR	0.215517	0.609914	0.157328	0.618534
FPR	0.012863	0.108374	0.015326	0.133005
Threshold	0.5	0.14586	0.5	0.130592

Conclusion

- To conclude, as all three models give same recall, we can choose **Random Forest Classifier** in collaboration with feature pruning and best probability threshold to achieve higher recall while keeping false positive rate lower.