Final Report

Bank Marketing Case Study

11/30/2023

Group: Profit
Jinghao Shen

Agenda

- Problem Statement
- Data Exploration
- EDA Summary
- Model Selection
- Model Building
- Results
- Conclusion

Problem Statement

- ABC Bank wants to sell it's 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).

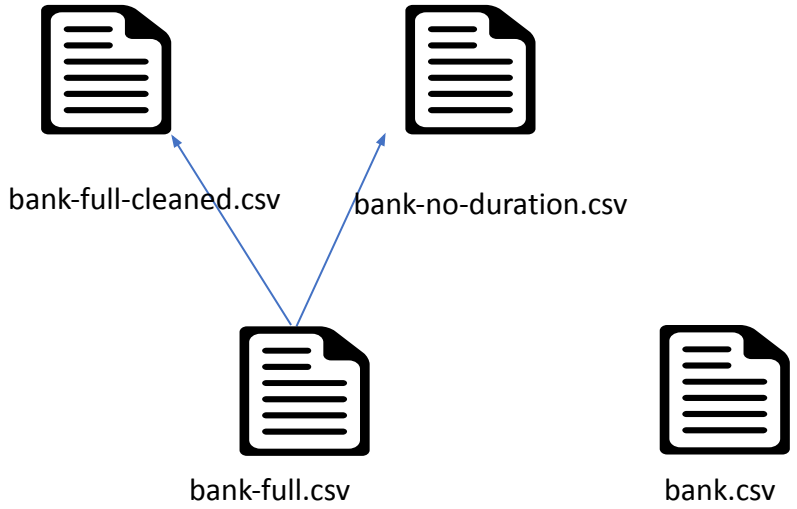
- Object: Develop an ML model to shortlist customer whose chances of buying the product is more so that their marketing channel can focus only to those customers whose chances of buying the product is more.

Data Exploration

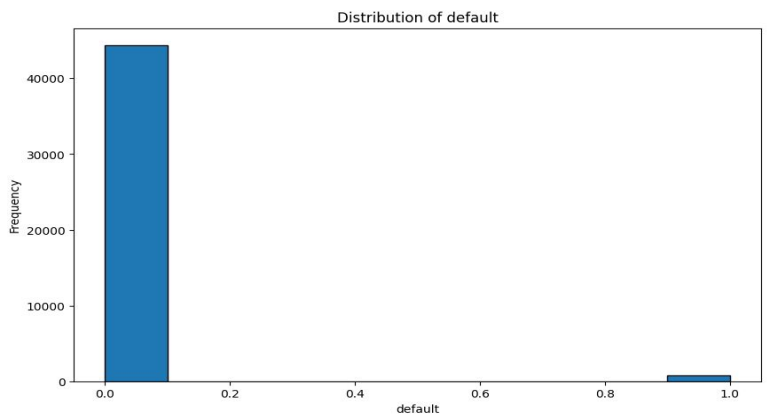
- Features: 16+1
- Time period of data: 2012
- Total number of observations: 45211

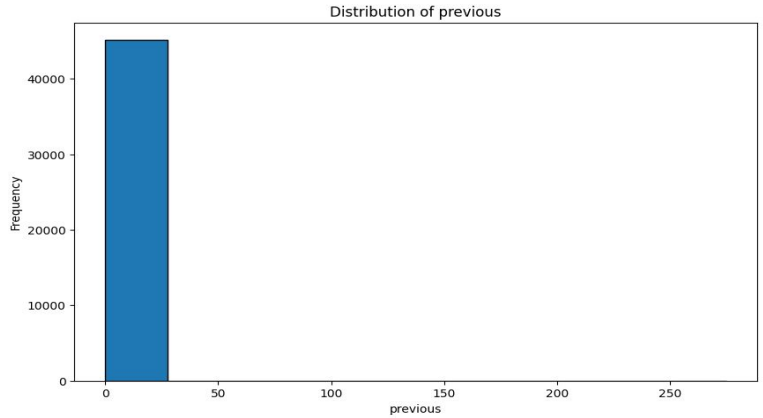
Transformations:

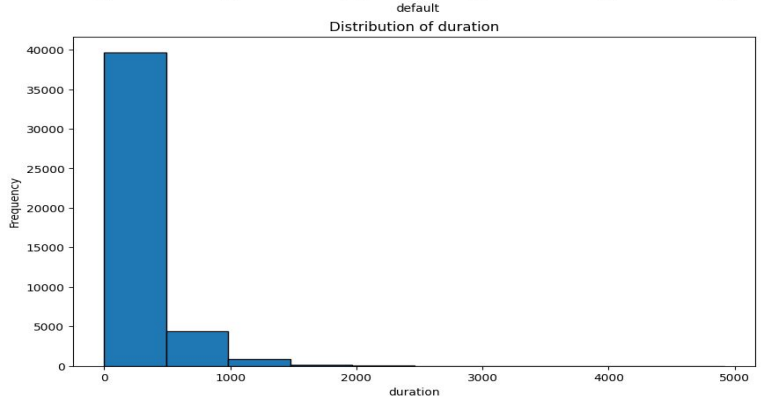
1. bank.csv with 10% of the sample randomly selected from bank-full.csv
2. bank-full.csv is transformed to bank-full-cleaned.csv that all non-numerical attributes is turned into numerical values.
3. bank-no-duration.csv removes ‘duration’ attribute from bank-full-cleaned.csv

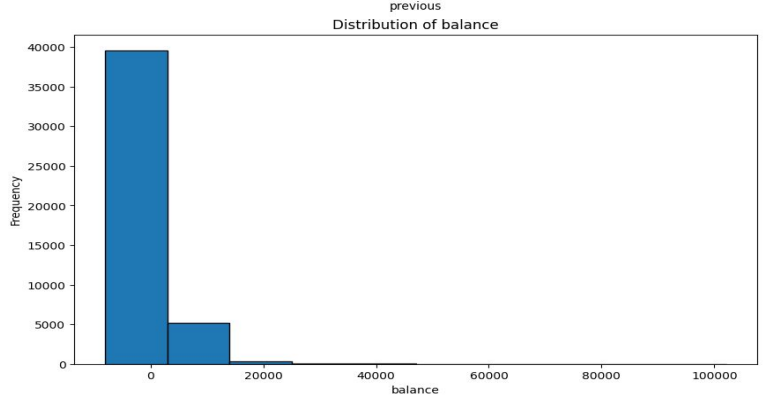


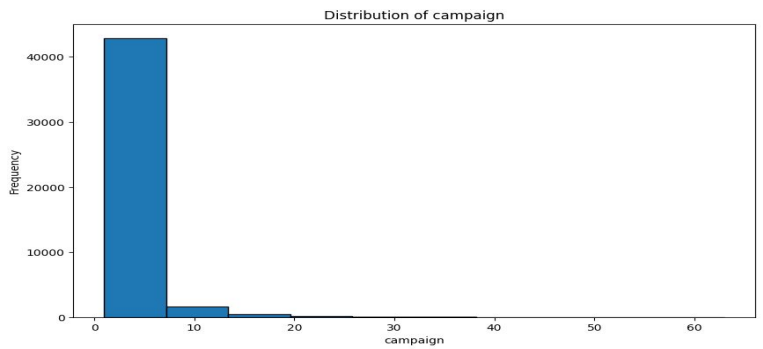
Attributes Analysis

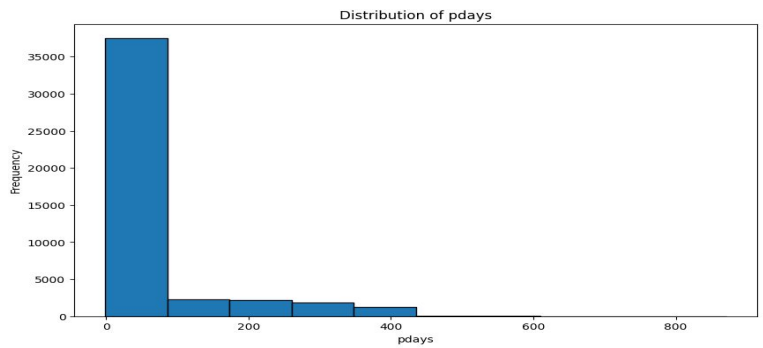






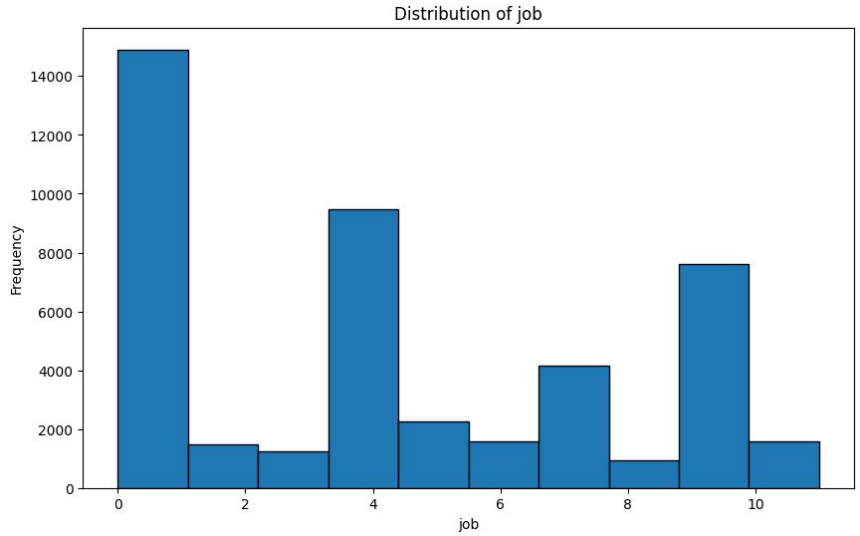


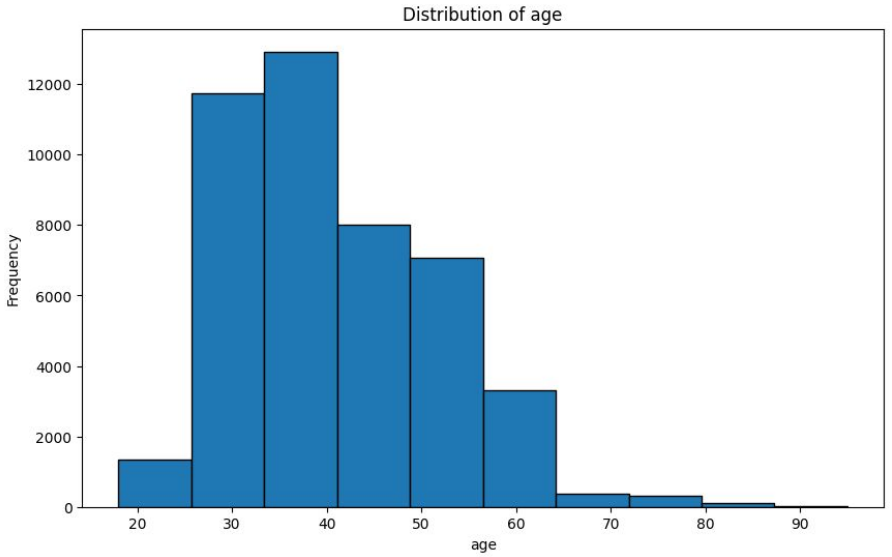


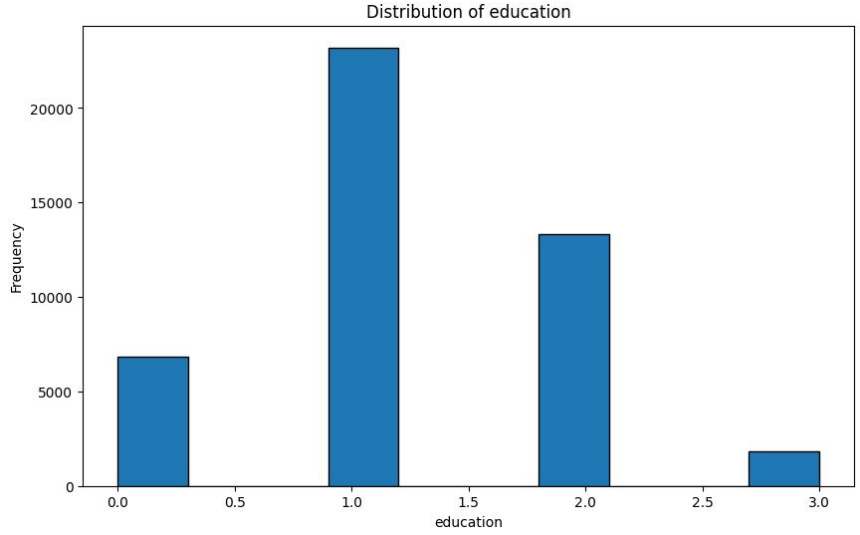


- We can observe long-tail distribution in several attributes.
  - Number of instances in head classes are way more than instances in tail classes.
- Missing values are replaced as ‘unknown’.

Attributes Analysis (client data)





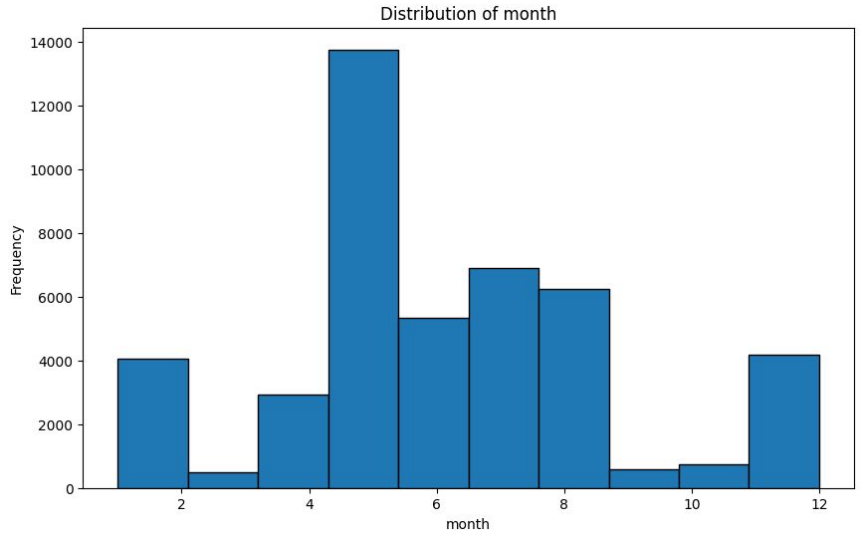


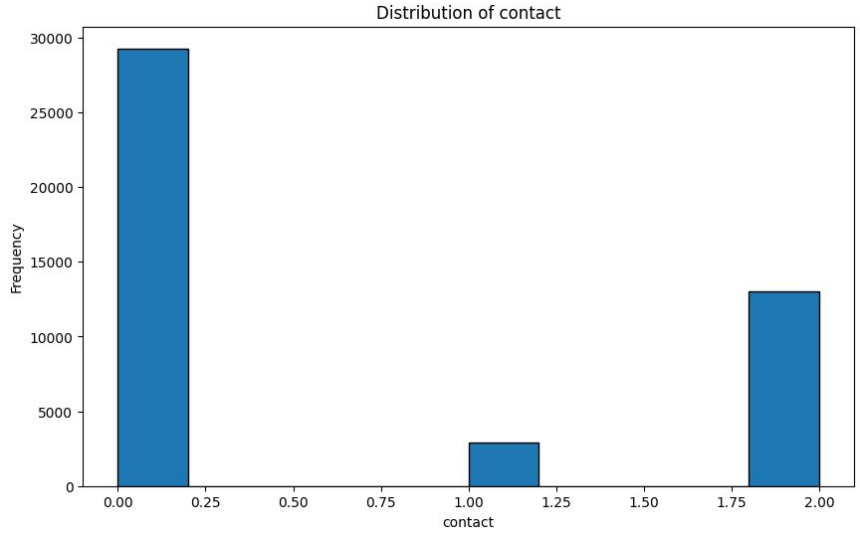
- Most client age falls within the range 30 to 40
- Top three client jobs are: blue-collar, management, and technician.
- Most client have an education level higher than secondary school.

Assuming: education:

- secondary: 1
- tertiary: 2
- primary: 0
- unknown: 3

Attributes Analysis (contact)





Most contact happened in May.

Less contact in March and September.

Cellular Phone is the most used contact approach.

Assuming:

Cellular: 0

Telephone: 1

Unknown: 2

EDA Summary

Through previous EDA, we have observed and concluded that

- Data set: We do observe the presentation of long-tail distribution in several attributes, which may lead to bad model performance. Samples in head classes are over-represented and samples in tail classes are under-represented.
- Clients Age: We have most clients at age 32. They have the highest loan percentage among all ages and they have the best chance to subscribe the product.
- Contact: Number of contacts during previous campaigns has negative correlation with number of contacts during current campaign. We also found that longer contact duration won't necessary mean higher chance of subscription.
- Previous Outcome: Clients who subscribed during previous campaigns have higher chance to subscribe in this campaign.

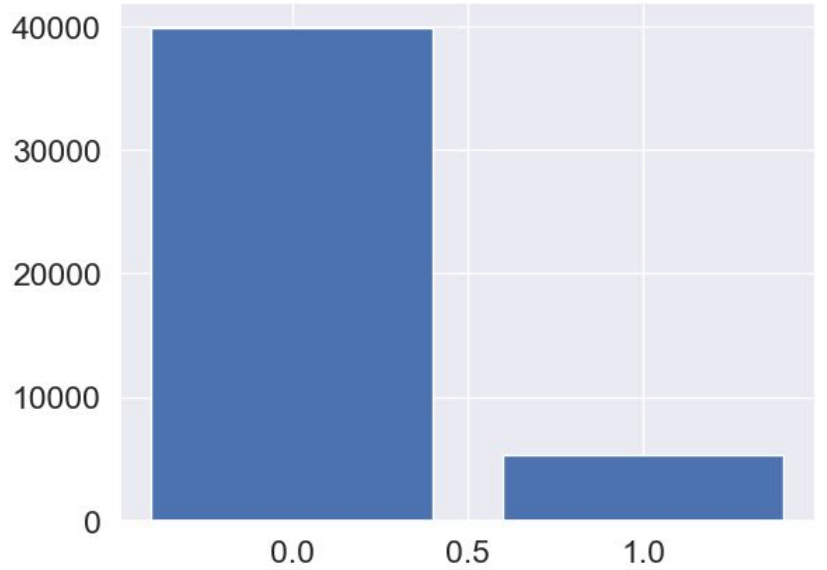
Model Selection

We propose two approaches:

1. Using logistic regression model with class balanced weight.
2. Oversampling the tail class. Then apply logistic regression model.

Both approaches can effectively robust against long tail distribution observed in data set.

Two models are trained and tested with transformed dataset bank-full-cleaned.csv and bank-no-duration.csv.



Distribution of the outcome (y)

Class=0, n=39922 (88.302%)
Class=1, n=5289 (11.698%)

Approach 1: Class balanced weight

Class weights: We can train the algorithm to take the imbalance distribution into account by giving different weights to both the majority and minority classes. The purpose of that is to penalize the misclassification by setting the minority class a higher weight and reduce weight for the majority class.

Weight is calculated as:

$$W_j = n\_samples / (n\_classes * n\_samples_j)$$

Where n_samples is the total sample size, n_classes is the total classes number, and $n\_samples_j$ is the total number of sample in j class.

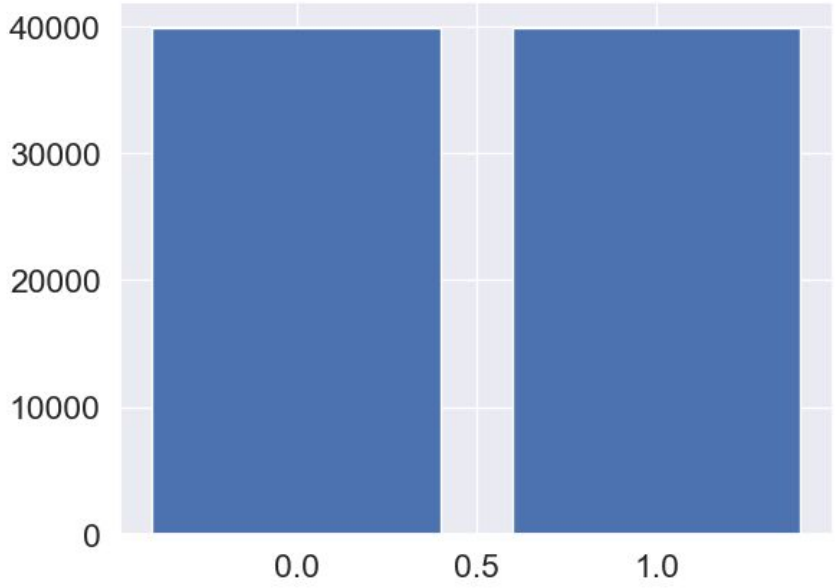
For our example,

Class weights: {0: 0.566241671258955, 1: 4.274059368500661}.

Approach 2: Oversampling

Oversampling: We copy or synthesizing new samples of the minority class so that the number of samples from minority class matches the number of samples from the majority class. The purpose of that is to emphasize the under-represented class.

We are using SMOTE (Synthetic Minority Over-sampling Technique) in our case to sample the data.

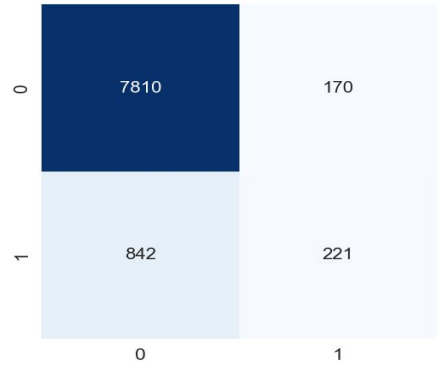


Outcome distribution after resampling

Class=0, n=39922 (50.000%)
Class=1, n=39922 (50.000%)

Model Results (Benchmark)

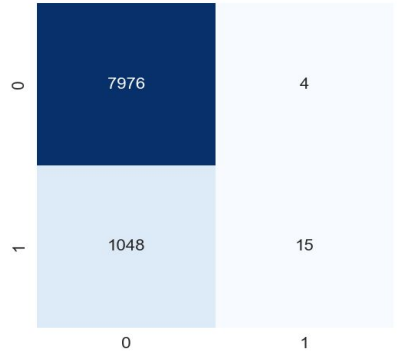
Model accuracy: To check the performance of the model, we will be using the f1 score as the metric, not accuracy. We will get high accuracy score because the model is biased towards the majority class. However, it will not serve any value to our problem statement. Therefore, we will be using f1 score as f1 score is the harmonic mean of precision and recall.



The confusion matrix for logistic regression without weight on bank-full-cleaned data.
F1-score: 0.3039889958734525

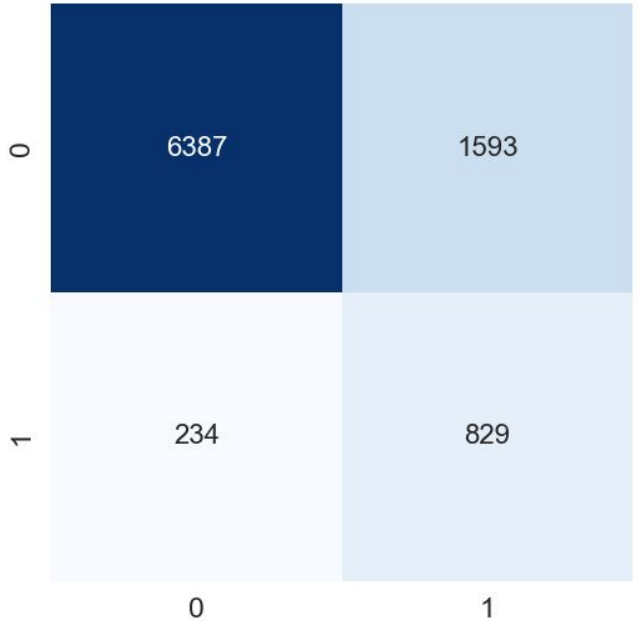
We got the F1-score only as 0.303 for a simple logistic regression model on data with duration.

We got the F1-score only as 0.027 for a simple logistic regression model on data without duration.



The confusion matrix for logistic regression without weight on bank-no-duration data.
F1-score: 0.027726432532347505

Model Results (with duration)

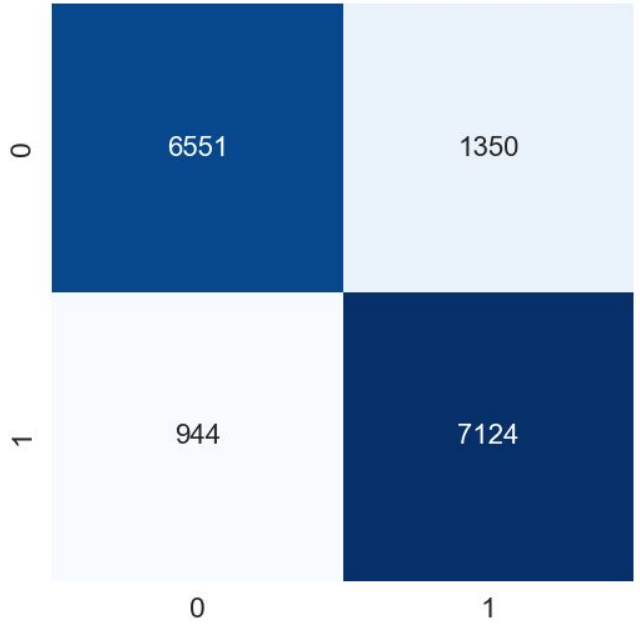


Approach 1: Class balanced weight

- The f1 score for the model is:

  0.4757532281205165

- 56.5% increase in accuracy



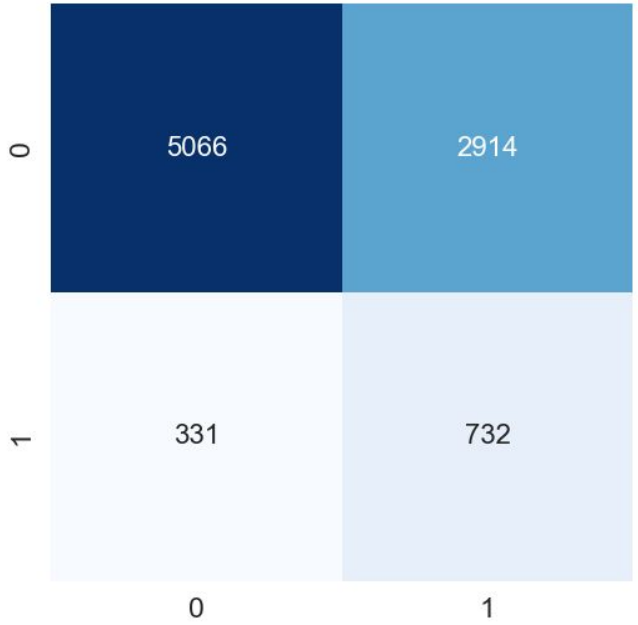
Approach 2: Oversampling

- The f1 score for the model is:

  0.8613226937492444

- 183.34% increase in accuracy

Model Results (without duration)

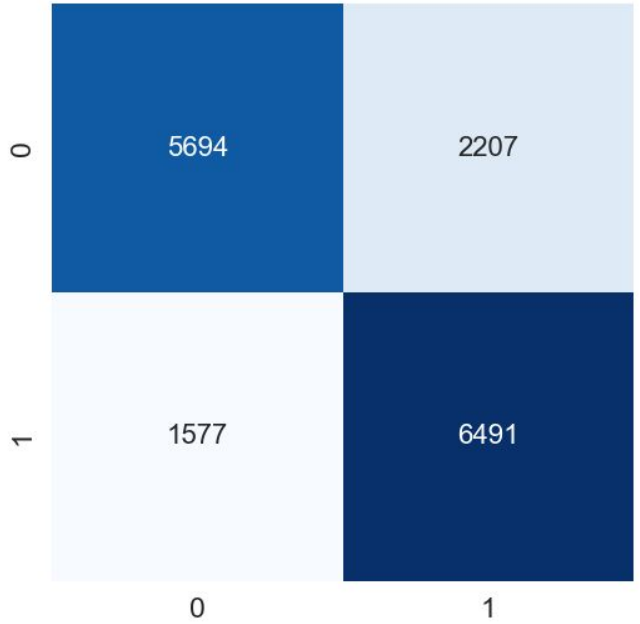


Approach 1: Class balanced weight

- The f1 score for the model is:

  0.3108940327033404

- 1021.29% increase in accuracy



Approach 2: Oversampling

- The f1 score for the model is:

  0.7758579599618686

- 2698.26% increase in accuracy

Conclusion

- For both approaches, we can observe a huge increase in F1 score compare to the benchmark result.
- Models with full bank information can perform better than models with data without ‘duration’ feature. We think the reason is that: Although ‘duration’ feature can be difficult to explain, it does have a relationship with the outcome and can contribute to the model performance.
- There is huge increase in accuracy for data without ‘duration’ feature since the benchmark is really close to 0.
- Oversampling provides a better result in both dataset than class balanced weighting.

Overall, we will recommend the Oversampling model with duration feature.

Thank You