BANK MARKETING CASE STUDY

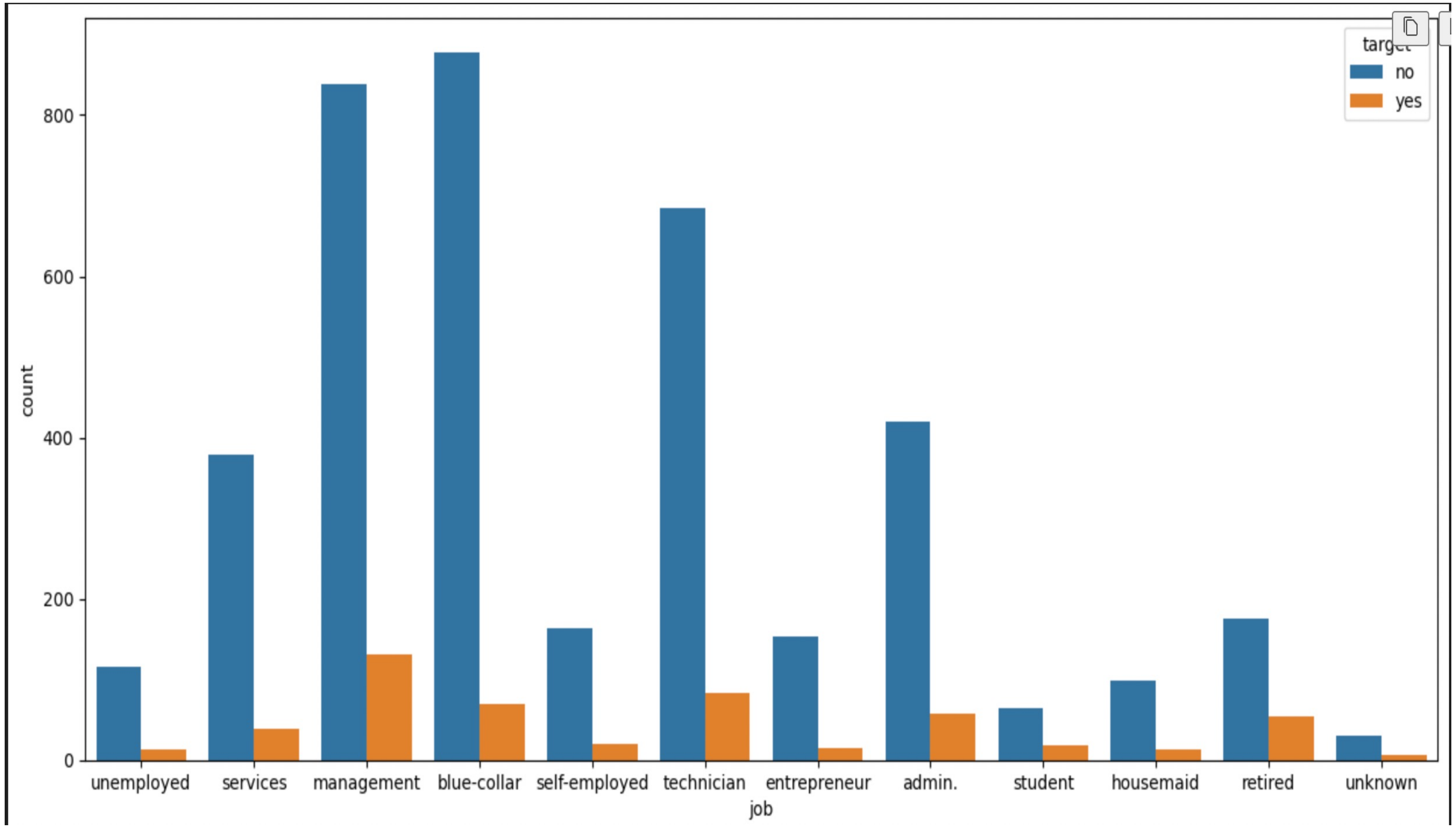
PROBLEM STATEMENT

- A Portuguese banking institution wants to predict whether a client will subscribe a term deposit after their direct marketing campaigns(phone calls)
- We need to build a ML Classification model which reduces the resource and time loss of the company and correctly predict a customer who will subscribe for a term deposit.

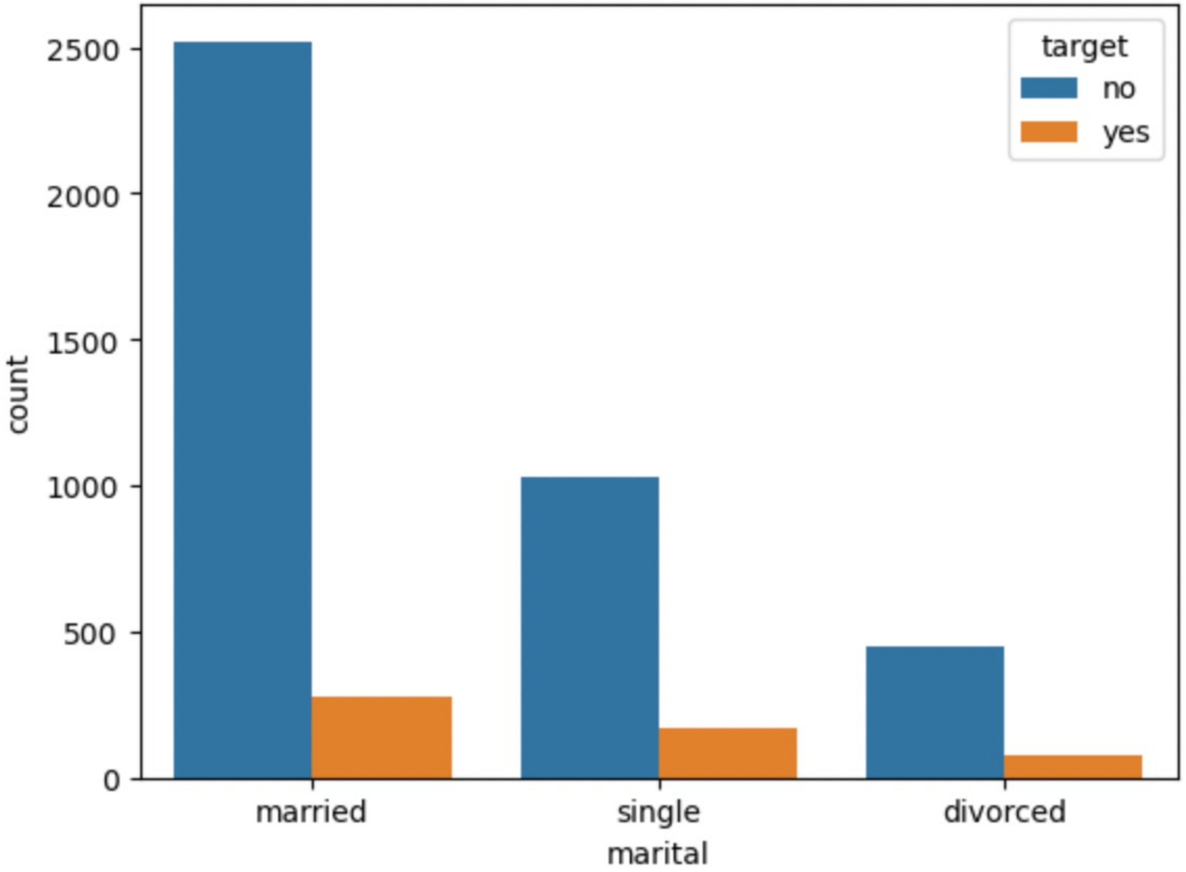
❑ STRATEGY

- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for subscribing
- Scaling features
- Prepare the data for model building
- Build a Logistic Regression, Random Forest and Decision Tree based models
- Handling Imbalanced data using Random Oversampling, Random Under-sampling, SMOTE and SMOTE+TOMEK approach
- Hyper-parameter tuning of the model
- Choosing the best approach among all data imbalance techniques from the Evaluation metrics obtained
- Fine tuning the model on the chosen approach
- Finalising the model with best performance metrics

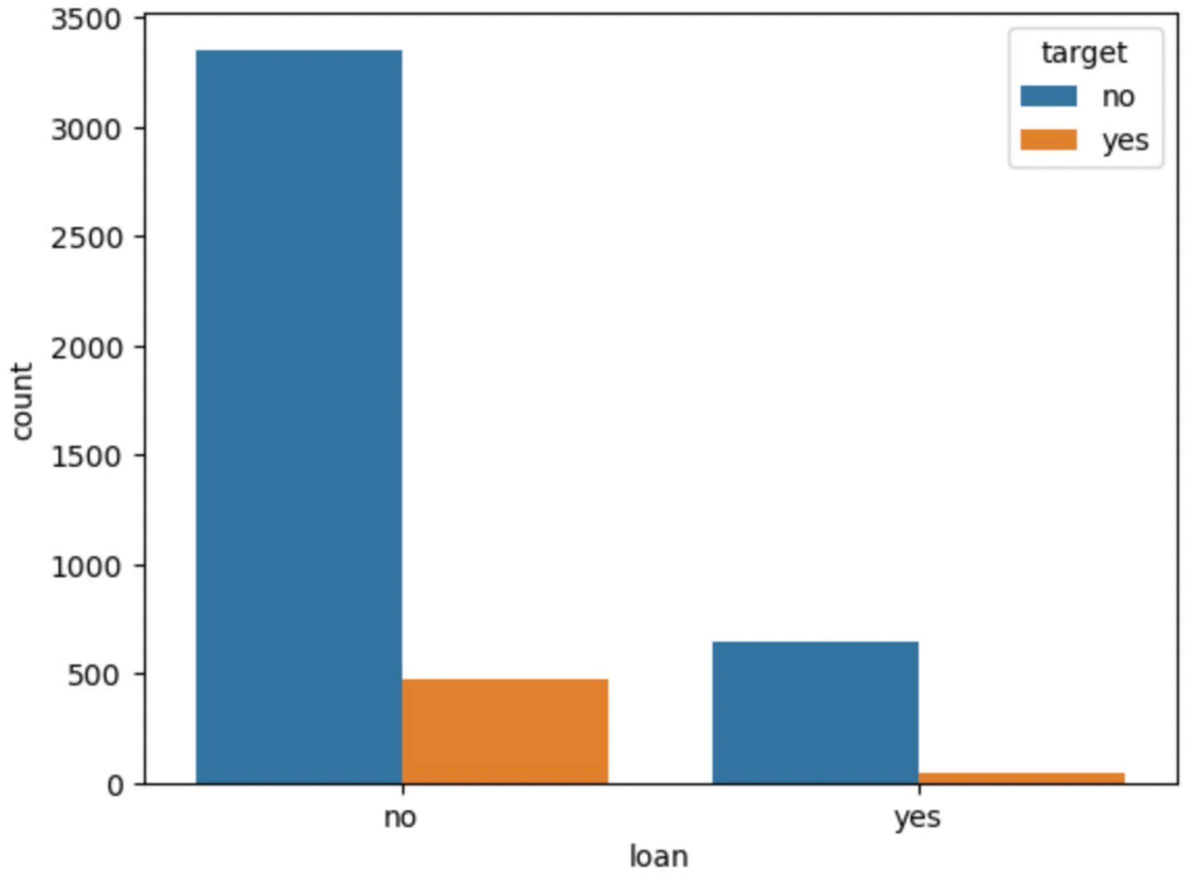
❑ EXPLORATORY DATA ANALYSIS



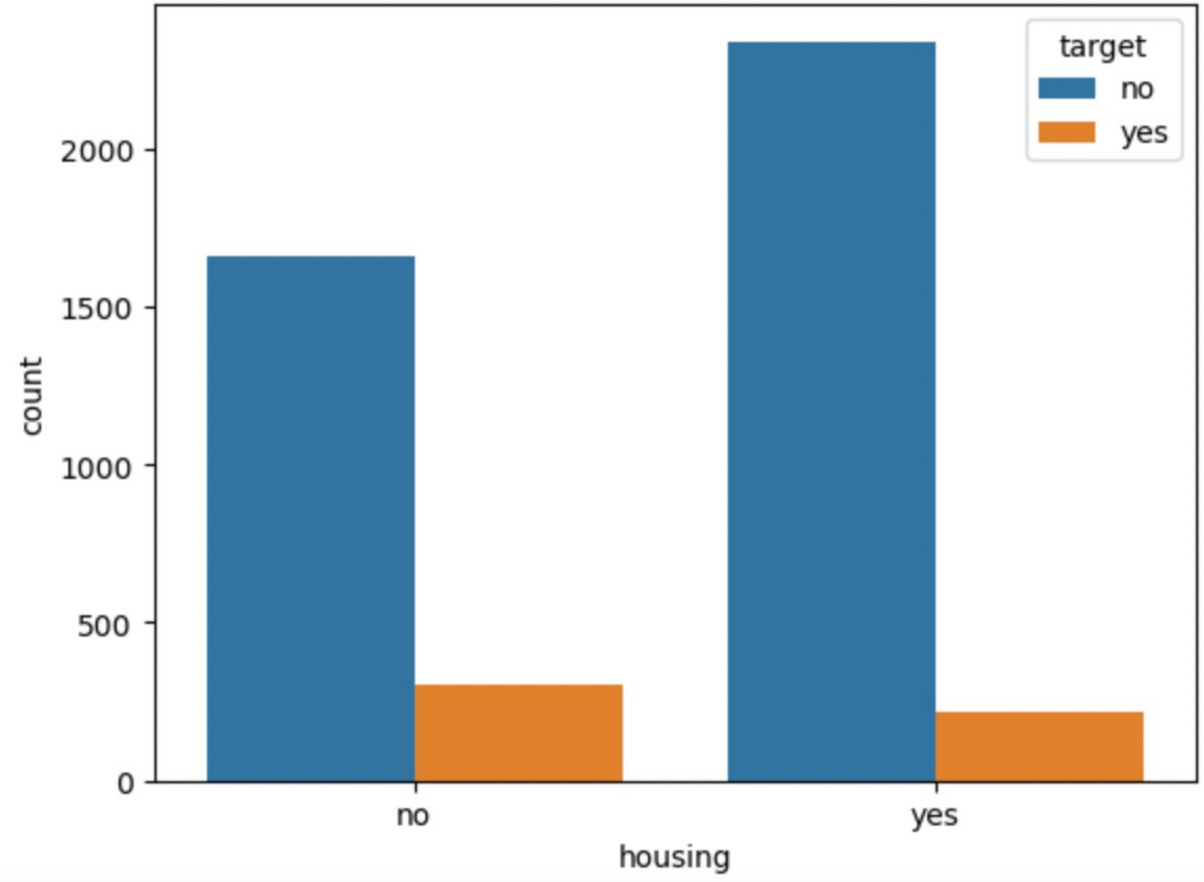
- ## Job vs Target Variable
- We see that clients from management, blue collar, technician, admin, retired background are more likely to subscribe to a term deposit.



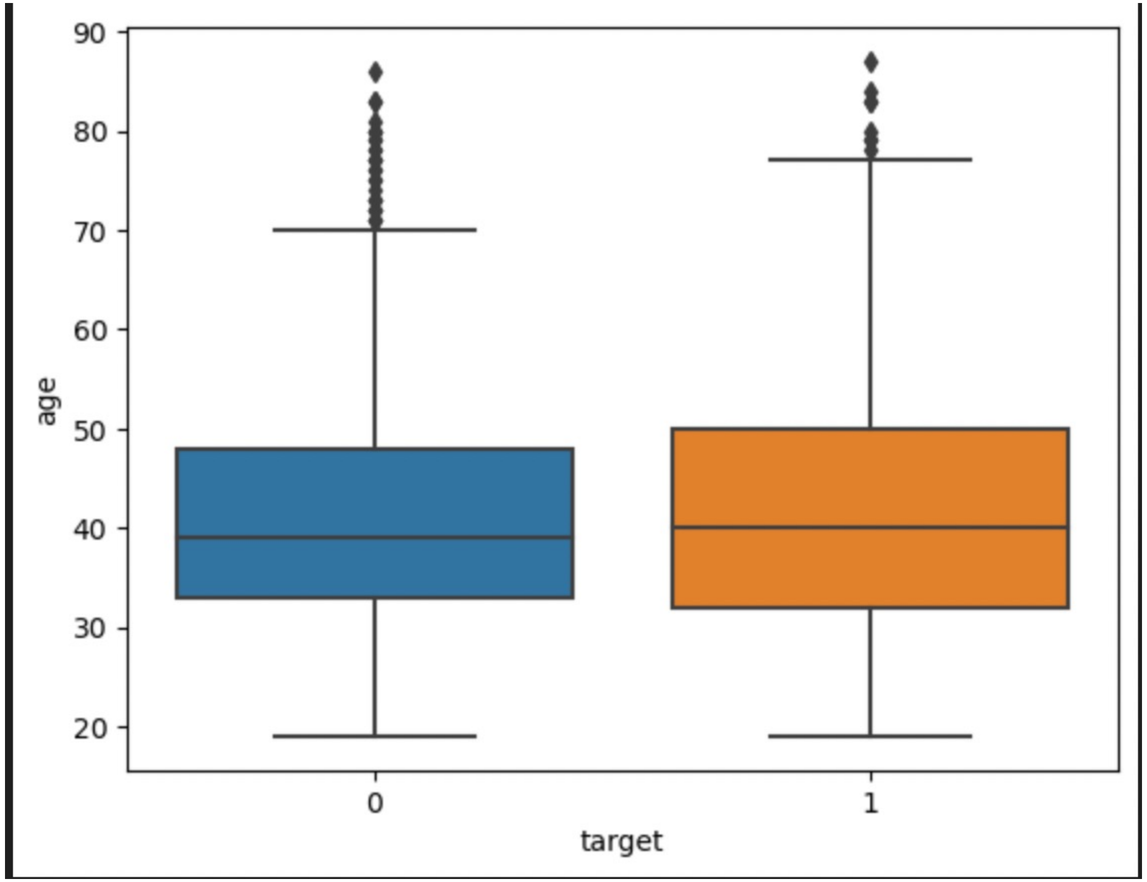
- Marital Status Vs Target Variable
  - We see that married clients are more likely to take a term deposit



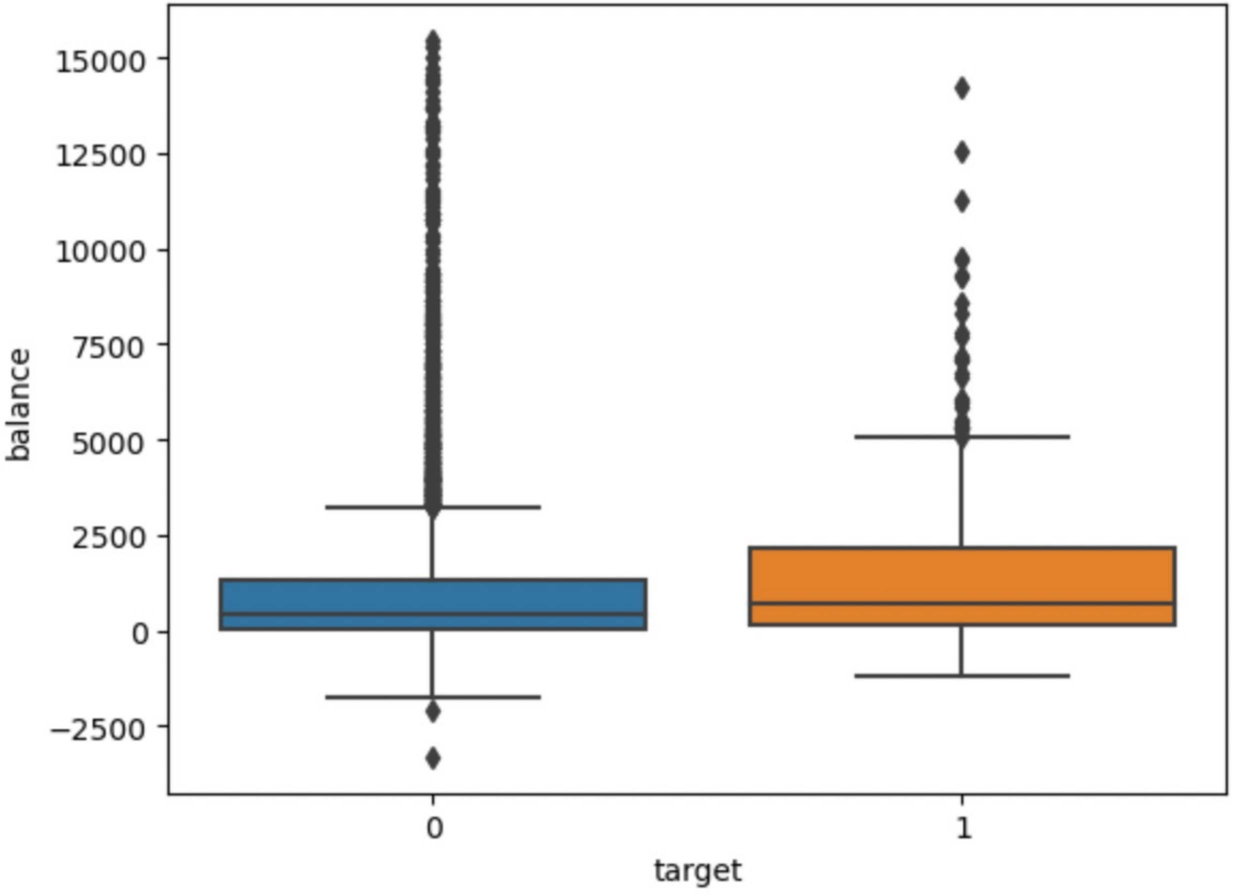
- Loan Vs Target Variable
- Clients who have not taken any loan might take a term deposit than the ones who have taken loan



- Housing Loan Vs Target Variable
  - Similarly, if a client has not taken any housing loan are more likely to subscribe for a term deposit



- **Age Vs Target Variable**
- We see that people having slightly more median age is likely to subscribe for term deposit



- Balance Vs Target Variable
  - Minimum balance and upper quartile is also more for people who opt for term deposit

MODEL BUILDING

- Splitting into train and test set
- Scale variables in train set
- Build Logistics Regression, Random Forest and Decision Tree Based model without resampling the data
- Further with hyper parameter tuning built Logistic Regression and Random Forest model after resampling the data using random oversampling, random under- sampling, SMOTE, SMOTE+TOMEK
- As we want the precision to be high, we will be taking random oversampling in the final model
- Fine tuned the model with random oversampled data
- Metrics such as accuracy, precision, recall, F1 score and AUC was used to determine the model performance

❑ PERFORMANCE METRICS

Without applying hyper parameter tuning and sampling techniques :

- Logistic Regression Performance :
  - Accuracy – 0.89
  - F1 Score – 0.22
  - Recall – 0.13
  - Precision – 0.68
  - AUC – 0.70
- Random Forest Performance :
  - Accuracy – 0.89
  - F1 Score – 0.2
  - Recall – 0.12
  - Precision – 0.63
  - AUC – 0.71
- Decision Tree Performance :
  - Accuracy – 0.81
  - F1 Score – 0.21
  - Recall – 0.21
  - Precision – 0.20
  - AUC – 0.55

❑ PERFORMANCE METRICS

After applying hyper parameter tuning and sampling techniques :

- We applied different sampling techniques and found random over sampling to give highest precision and with highest AUC as 0.71
  - Accuracy – 0.88
  - F1 Score – 0.31
  - Recall – 0.24
  - Precision - 0.43
  - AUC – 0.71

❑ PERFORMANCE METRICS

Further we checked performance of each parameters and did the fine tuning of the model

- Accuracy – 0.88
- F1 Score – 0.27
- Recall – 0.19
- Precision - 0.475
- AUC – 0.72

We see that the Area under curve(AUC) increased by 1%.

❑ LIMITATIONS AND FUTURE SCOPE

- Further in the next campaign we should target married clients
- Clients with jobs in management, blue collar, admin
- Clients who do not have high loans taken
- The top 3 features in prediction – balance, age, month
- In future we need more samples of clients who have taken a term deposit so model can be trained in different patterns of client behavior who will subscribe for a term deposit