Machine Learning Approach to Client Subscription Classification in Bank Marketing Campaign

Git Repository link: https://github.com/DMYomal/Bank_Marketing_ML.git

I. Introduction

The dataset originates from a direct marketing campaign conducted by a Portuguese banking institution, with the primary objective of promoting long-term deposits through telemarketing phone calls. The campaign involves human agents reaching out to a list of clients in an attempt to sell the bank's term deposit products. The outcome of each call is categorized as either a successful ("yes") or unsuccessful ("no") contact which is considered as the target variable (Subscribe).

Banks need to identify the right customers at the right time. The outcomes change with many external and internal factors and ML approaches are the best for identifying potential clients.

II. Data Overview

Each record in the dataset comprises key information, including the target output (contact outcome), telemarketing attributes (e.g., call duration), and client-related details (e.g., age). The dataset includes a combination of both categorical and numerical attributes, each providing insights into the factors that influence the success of the marketing campaign. The data set is unbalanced with 4648 success records and 36548 unsuccess records.

The independent attributes are as follows:

1. age: Age of the client
2. job: Type of job (e.g., "admin.", "blue-collar")
3. marital: Marital status (e.g "married", "single")
4. education: Educational background (e.g., "primary", "tertiary")
5. default: Whether the client has credit in default ("yes", "no")
6. balance: Average yearly balance in euros
7. housing: Whether the client has a housing loan ("yes", "no")
8. loan: Whether the client has a personal loan ("yes", "no")
9. contact: Communication type for contact (e.g. "telephone", "cellular")
10. day: Last contact day of the month
11. month: Last contact month of the year (e.g., "jan", "feb")
12. duration: Duration of the last contact in seconds
13. campaign: Number of contacts made during this campaign for the client
14. pdays: Number of days since the client was last contacted from a previous campaign (-1 indicates no previous contact)
15. previous: Number of contacts made before this campaign for the client
16. poutcome: Outcome of the previous marketing campaign (e.g., "success", "failure")

The dataset's attributes collectively offer valuable insights into customer characteristics, communication patterns, and historical interactions. The dataset consists of a combination of categorical and numerical variables, providing a comprehensive view of the factors contributing to the campaign's success.

Data Preprocessing

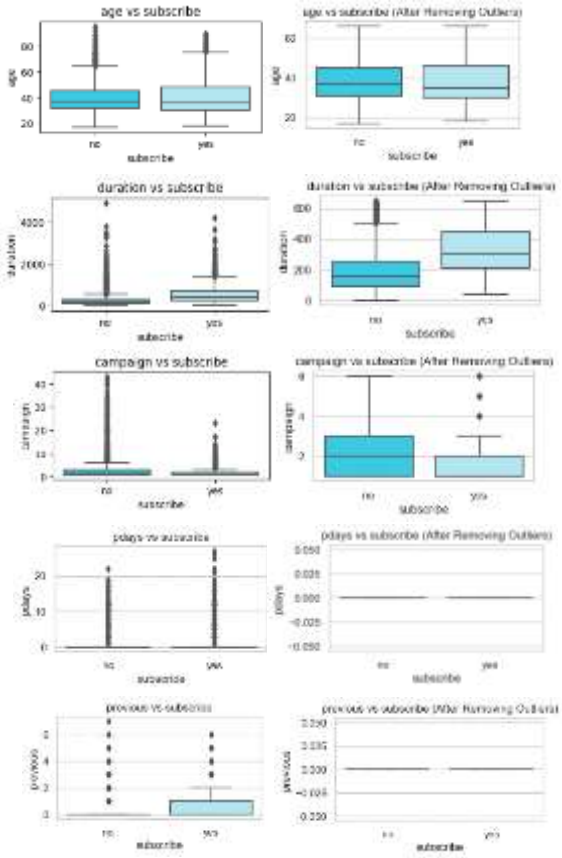
The dataset comprises 45,211 records and encompasses 17 attributes. Initial preprocessing involved scrutinizing each attribute for potential null values. Further examination revealed occurrences of "unknown" data within categorical attributes such as "job," "marital," "education," "loan," and "default." To prepare the data for modeling, categorical variables required encoding. Given the "unknown" category's contribution to dimensionality challenges, a preliminary action involved removing these entries.

Exploratory Data Analysis (EDA)

The EDA has been mainly divided into two sections which is to identify the relationship between:

a) Numerical Variables and Target
b) Categorical Variables and Target

a) **Numerical Variables and the Target**



There are 5 numerical variables in the data set and each

Figure 1 - Numerical Vs Target Variable Relationship

relationship is depicted in below box plots.

According to the box plot the number of outliers is as below

-age: 457
-duration: 2174
-campaign: 1675
-pdays: 1296
-previous: 4652

The **interquartile range (IQR)** method has been used to clean the outliers.

IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
pper_bound = Q3 + 1.5 * IQR

After the removing the outlier shape of the data set is (22218, 16).

The box plots of 'pday' vs 'subscribe' and 'previous' vs 'subscribe' indicate that a majority of customers were approached by the bank for the first time. Before outlier removal, approximately 96% of the 'pdays' column and 82% of the 'previous' column contained zero values. Following outlier removal, both columns can be omitted from the dataset as their values have become uniform.

The heatmap illustrates the relationships among numerical variables. The displayed values reveal that there is no significant correlation between these numerical variables, indicating the absence of multicollinearity effects.

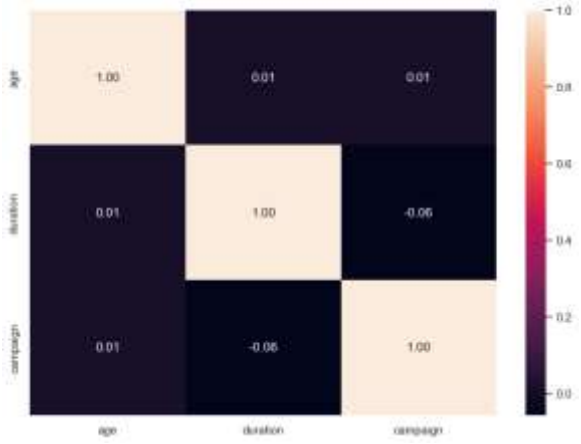


Figure 2 - Heat Map

Below bar chart depict the relationship between categorical variables and the target variables.

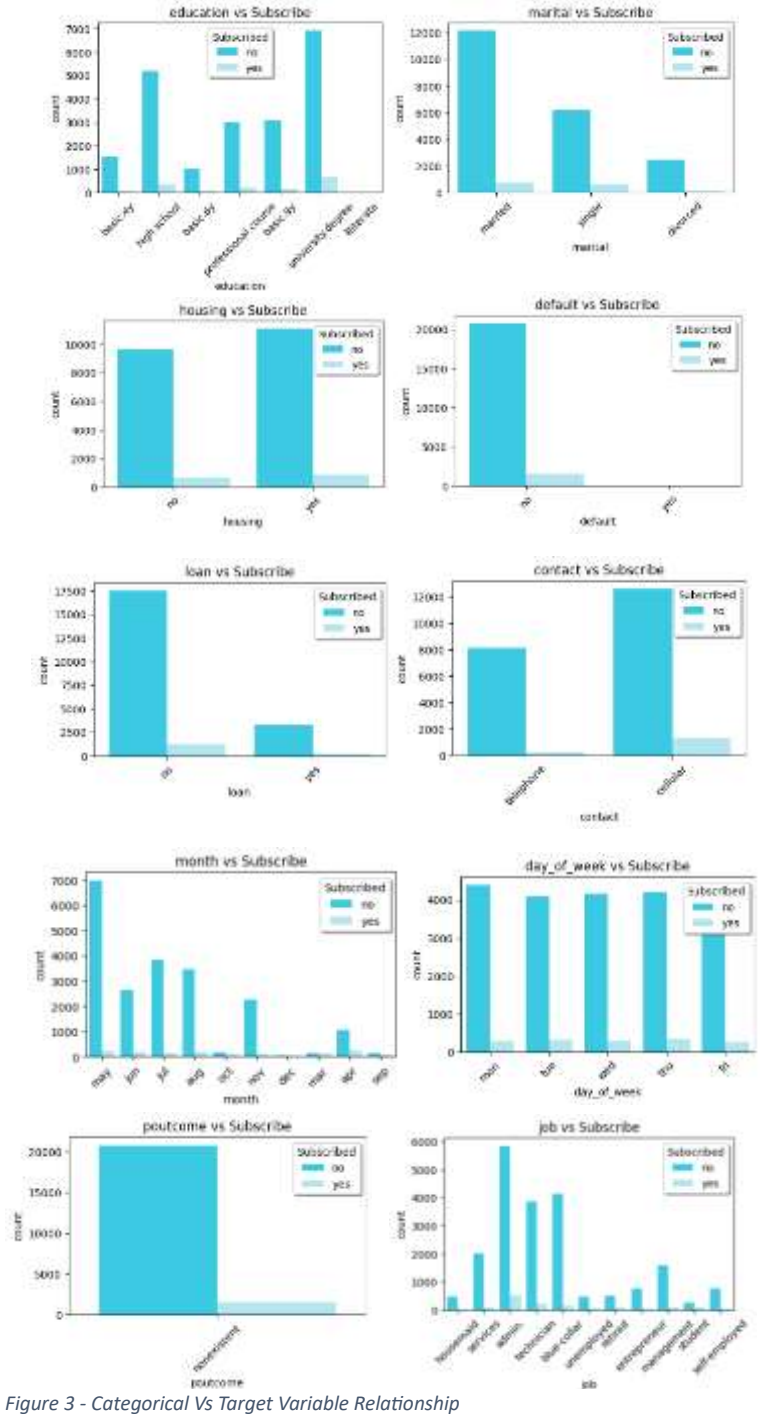


Figure 3 - Categorical Vs Target Variable Relationship

Through the exploration of the interplay between categorical and numerical variables and their relationship with the target variable, we can discern intricate patterns. As a result, the classification approach becomes more inclined towards multivariate analysis.

b) **Categorical Variables and the Target**

III. Model Fitting

Preparation of the data set for model fitting

The categorical variable columns' unique categories labels are as below.

-job: 11
-marital: 3
-education: 7
-default: 2
-housing: 2
-loan: 2
-outcome: 1
-contact: 2
-month: 10
-day_of_week : 5

Here the categorical variables have been encored before the model fitting. Here **One-Hot encoding** has been used. This can be considered as a **feature engineering** step.

The data set has been splited as 80% training and 20% for testing with a random state of 42.

Model Selection

A supervised learning model is appropriate due to its training and testing on historical data. For imbalanced binary classification, techniques such as **Random Forest Classifier, Gradient Boosting Algorithms, Support Vector Machines (SVM),** and **Logistic Regression** were employed.

To address data imbalance, class weights were applied to the minor class based on the class proportions of 'Yes' and 'No'.

1. **Random Forest with Class Weights**: Random Forest exhibits resilience in dealing with imbalanced data. By assigning higher weights to the minority class during training (via the class_weight parameter), the model can prioritize accurate classification of the minority class.
2. **Gradient Boosting Algorithms** (XGBoost, LightGBM, CatBoost): These algorithms inherently incorporate mechanisms to handle imbalanced data. They offer the flexibility to assign class weights or utilize specific loss functions that penalize misclassifications of the minority class more heavily.
3. **Support Vector Machines (SVM)** with Class Weights: SVMs can be potent when coupled with class weights, emphasizing precise classification of the minority class.
4. **Logistic Regression with Class Weights**: Logistic regression, though straightforward, can prove effective. Adjusting class weights allows for achieving a more balanced and accurate outcome in the model's predictions.

After the one hot encoding, the independent variable attributes increased to 48 and this may reason for to curse of dimensionality. Hence, principal component analysis was done with 5 components.

IV. Model Evaluation

The models are evaluated **using precision, recall, F1 score, model accuracy**, and **confusion matrix**. The summary of the test results is as below.

Table 1 - Classification Report Summary (without PCA)

	Accuracy	Subscribe	precision	recall	f1-score
Random Forest	93.61%	0	0.94	0.99	0.97
		1	0.58	0.14	0.22
Extreme Gradient Boosting Algorithms	89.63%	0	0.97	0.91	0.94
		1	0.35	0.67	0.46
Support Vector Machines (SVM)	85.62%	0	0.98	0.87	0.92
		1	0.27	0.71	0.39
Logistic Regression	82.22%	0	0.98	0.82	0.90
		1	0.25	0.82	0.38

Table 2 - Confusion Metrix Summary (without PCA)

		Negative	Positive
Random Forest	False	4120	29
	True	255	40
Extreme Gradient Boosting Algorithms	False	3786	363
	True	98	197
Support Vector Machines (SVM)	False	3597	552
	True	87	208
Logistic Regression	False	3411	738
	True	52	243

Random Forest:

- High Accuracy (93.61%): Random Forest performs well in correctly classifying instances overall.
- Class 0 (No subscription): High precision (0.94) and recall (0.99) indicate that the model can identify non-subscribers accurately.
- Class 1 (Subscription): Low precision (0.58) and recall (0.14) suggest that the model has difficulty identifying subscribers, leading to a lower F1-score (0.22).
- Confusion Matrix: A high number of false negatives (FN) for class 1 indicates that the model misses a significant portion of actual subscribers.

Extreme Gradient Boosting Algorithms:

- Good Accuracy (89.63%): Gradient Boosting provides a balanced performance between precision and recall.
- Class 0: High precision (0.97) and recall (0.91) indicate accurate identification of non-subscribers.
- Class 1: Moderate precision (0.35) and recall (0.67) suggest better performance in identifying subscribers than Random Forest.
- Confusion Matrix: Still a significant number of false negatives (FN) for class 1, indicating room for improvement.

Support Vector Machines (SVM):

- Moderate Accuracy (85.62%): SVM provides a trade-off between accuracy, precision, and recall.
- Class 0: High precision (0.98) and recall (0.87) show accurate identification of non-subscribers.
- Class 1: Low precision (0.27) and moderate recall (0.71) suggest challenges in correctly identifying subscribers.
- Confusion Matrix: Still a notable number of false negatives (FN) for class 1.

Logistic Regression:

- Moderate Accuracy (82.22%): Logistic Regression provides relatively balanced performance.
- Class 0: High precision (0.98) and moderate recall (0.82) indicate accurate identification of non-subscribers.
- Class 1: Low precision (0.25) and recall (0.82) suggest difficulty in identifying subscribers.
- Confusion Matrix: Moderate number of false negatives (FN) for class 1.

According to the results, Random Forest has the highest accuracy (93.61%), but accuracy alone might not be the best metric for imbalanced datasets.

Considering the business case, where correctly identifying subscribers (class 1) is essential, Gradient Boosting Algorithms seem to be performing relatively better than the other models. It provides a more balanced trade-off between precision and recall for both classes.

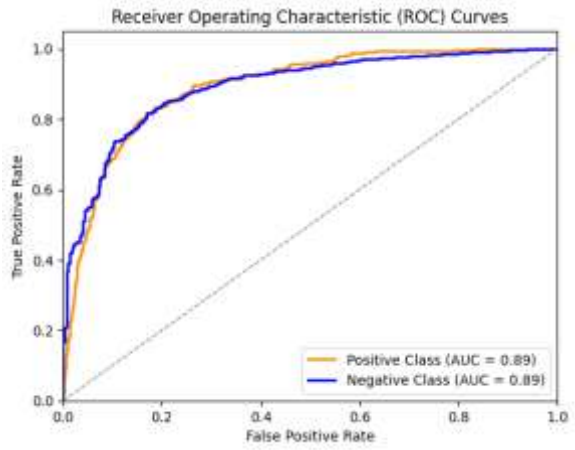


Figure 4 - ROC Curve for Extreme Gradient Boosting Algorithms

According to the ROC curve, both the positive classes and the negative classes are 0.89 which is greater than 0.5. Also, the AUC (the area under the curve) is greater. Hence, the gradient boost model can be considered a good model.

Table 3 - Classification Report Summary (with PCA)

	Accuracy		precision	recall	f1-score
Random Forest	93.31%	0	0.94	1	0.97
		1	0.47	0.06	0.10
Extreme Gradient Boosting Algorithms	87.08%	0	0.96	0.90	0.93
		1	0.26	0.53	0.35
Support Vector Machines (SVM)	74.84%	0	0.98	0.74	0.85
		1	0.19	0.82	0.30
Logistic Regression	75.13%	0	0.97	0.75	0.85
		1	0.17	0.71	0.28

Table 4 - Confusion Matrix Summary (with PCA)

		Negative	Positive
Random Forest	False	4130	19
	True	278	17
Extreme Gradient Boosting Algorithms	False	3714	435
	True	139	156
SVM	False	3084	1065
	True	53	242
Logistic Regression	False	3129	1020
	True	85	210

Similarly, to the previous case, Gradient Boosting Algorithms exhibit comparatively superior performance compared to other models. It achieves a more equitable balance between precision and recall for both classes. Notably, in the absence of PCA, the Gradient Boosting model performs even better. Among the models considered, Random Forest, SVM, and Logistic Regression demonstrate lower performance levels compared to XGBoost.

In order to address the class imbalance issue, a second approach involves applying SMOTE (Synthetic Minority Over-sampling Technique) to the XGBoost model for oversampling the target class. This aims to enhance the model's predictive capabilities further.

Table 5 - Classification Model for final XGBoost Model

	precision	recall	F1-Score
No: 0	0.98	0.82	0.90
Yes: 1	0.25	0.82	0.38

Table 6 - Confusion Matix for final XGBoost Model

	Negative	Positive
False	3411	738
True	52	243

Accuracy: The accuracy of the model is 82.22%. It indicates the overall correctness of the predictions made by the model.

Class 0 (No) Evaluation:

- Precision: The precision for class 0 is 0.98. This high value suggests that when the model predicts a client won't subscribe to the term deposit, it is often correct.

- Recall: The recall for class 0 is 0.82. This indicates that the model can correctly identify a good portion of actual non-subscribers.
- F1-score: The F1-score for class 0 is 0.90. This is the harmonic mean of precision and recall and provides a balance between the two metrics.

Class 1 (Yes) Evaluation:

- Precision: The precision for class 1 is 0.25. This indicates that when the model predicts a client will subscribe to the term deposit, it is less accurate and may have false positives.
- Recall: The recall for class 1 is 0.82. This means that the model can correctly identify a significant portion of actual subscribers.
- F1-score: The F1-score for class 1 is 0.38. This indicates the balance between precision and recall for class 1.

Interpretation:

- The model's performance in identifying non-subscribers (class 0) is good, as indicated by high precision and recall values.
- However, the performance in identifying subscribers (class 1) is not as satisfactory, with a lower precision and a higher number of False Negatives.
- This suggests that while the model is good at identifying non-subscribers,

Overall, the model's performance after SMOTE resampling seems to be an improvement compared to the initial model.

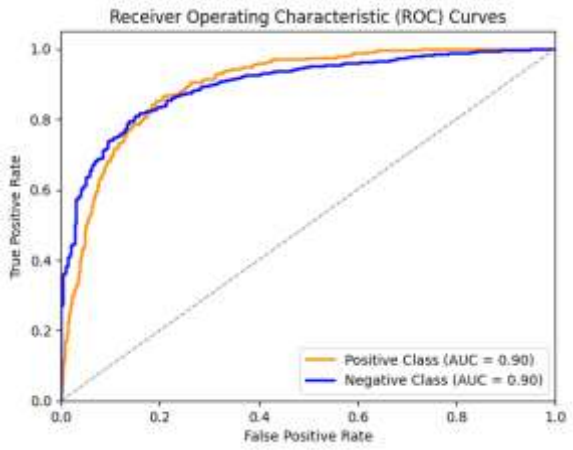


Figure 5 - ROC for final XGBoost Model

V. Conclusion

The classification goal of the study is to predict if the client will subscribe to a term deposit according to the features of the client and the marketing campaign.

1. After the EDA, it was identified that,
    a. The data set is imbalanced and after cleansing 7.2% of clients subscribed to the term deposit.
    b. Most of the clients are approached by the marketing team for the first time.
2. PCA on encoded data did not significantly improve the models by reducing the curse of dimensionality.
3. Extreme gradient boosts classifier performed better than the other three models such as random forest, SVM and logistic regression and SMOT can be used to improve the model accuracy.
4. The developed model is better for identifying a good portion of non-subscribers than subscribers.

VI. Limitation and Future Improvements.

The final model is good to identify the non-subscribers but there's room for improvement in accurately identifying subscribers. There is still a need to balance the trade-off between precision and recall, especially for identifying subscribers. Further adjustments to the model or using different techniques might be necessary to achieve a more balanced performance between the two classes.

The **hyperparameter techniques** such as **Grid Search, Random Search, Hyperopt,** etc. for better accuracy of the model.

VII. Reference

1.Asare-Frempong, J. and Jayabalan, M. (2017) Predicting customer response to bank direct telemarketing campaign. 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T) [online]. Available from: https://ieeexplore.ieee.org/document/8215961.

2.Kim, K.-H., Lee, C.-S., Jo, S.-M. and Cho, S.-B. (2015) Predicting the success of bank telemarketing using deep convolutional neural network IEEE Xplore. 1 November 2015 [online]. pp. 314–317. Available from: https://ieeexplore.ieee.org/abstract/document/7492828 [Accessed 28 April 2020].

3.Lahmiri, S. (2017) A two-step system for direct bank telemarketing outcome classification. Intelligent Systems in Accounting, Finance and Management [online]. 24 (1), pp. 49–55.

4.Moro, S., Cortez, P. and Rita, P. (2014) A data-driven approach to predict the success of bank telemarketing. Decision Support Systems [online]. 62, pp. 22–31.

5.Palaniappan, S., Mustapha, A., Mohd Foozy, C.F. and Atan, R. (2017) Customer Profiling using Classification Approach for Bank Telemarketing. JOIV : International Journal on Informatics Visualization [online]. 1 (4–2), p. 214. [Accessed 25 November 2019].

6.Tekouabou, S.C.K., Cherif, W. and Silkan, H. (2019) A data modeling approach for classification problems. Proceedings of the 2nd International Conference on Networking, Information Systems & Security - NISS19 [online].

7.Tékouabou, S.C.K., Gherghina, Ş.C., Toulni, H., Neves Mata, P., Mata, M.N. and Martins, J.M. (2022) A Machine Learning Framework towards Bank Telemarketing Prediction. Journal of Risk and Financial Management [online]. 15 (6), p. 269.

8.Vajiramedhin, C. and Suebsing, A. (2014) Feature selection with data balancing for prediction of bank telemarketing. Applied Mathematical Sciences [online]. 8, pp. 5667–5672.