Interpretable Machine Learning for Predicting Term Deposit

John Mary Karanzi
Department of Computer Science, CoCIS
Makerere University
karanzi.johnmary@students.mak.ac.ug

Linda Kellen Ayebale
Department of Computer Science, CoCIS
Makerere University
lindakellen9@gmail.com

***Abstract*—This paper presents an application of machine learning techniques to predict client subscription to term deposits based on historical marketing data. Using the UCI Bank Marketing dataset [1], we explore supervised learning methods, specifically Logistic Regression and Random Forest, to address class imbalance and improve predictive accuracy. The Random Forest model achieved superior performance with an accuracy of 92.3%. To mitigate class imbalance, the NearMiss-2 algorithm [2] was employed, effectively balancing the dataset by undersampling the majority class. Interpretability techniques, such as SHAP [3] and LIME [4], were utilized to ensure model transparency and actionable insights. This study highlights the role of machine learning in optimizing marketing strategies, reducing costs, and enhancing customer satisfaction. Future work focuses on advanced ensemble techniques, cost-sensitive learning, and real-time prediction systems to further refine marketing campaign effectiveness.**

***Index Terms*—Term deposit, Machine Learning, Explainable Artificial Intelligence (XAI), NearMiss-2**

I. Introduction

Machine learning (ML) has revolutionized marketing analytics by enabling businesses to make data-driven decisions. In banking, targeted marketing campaigns require accurate prediction of client responses to optimize resources and increase success rates. This study addresses the challenge of predicting whether a client will subscribe to a term deposit, a critical aspect of marketing in financial institutions.

Key methods include:

- **Term Deposit:** This could also be referred to as a fixed deposit in some countries, is a financial product offered by banks where an individual deposits a sum of money for a fixed period at a predetermined interest rate.
- **Machine Learning:** Essential for predictive modeling in binary classification tasks.
- **NearMiss-2 Algorithm:** A supervised undersampling technique for addressing class imbalance [2].
- **Explainable AI:** An AI that is easy for people to comprehend and analyze the acts it predicts is explainable[4].

By leveraging these techniques, this work contributes to optimizing resource allocation in marketing campaigns, ultimately benefiting the banking community.

II. Background and Motivation

A. Selecting a Template

Machine Learning: Focused on automating prediction tasks, ML is pivotal in modern marketing strategies for data segmentation and behavioral analysis [5].

NearMiss-2 Algorithm: Enhances data balance by selecting majority class samples farthest from the minority class, preserving decision boundaries and improving model robustness [2].

Explainable Artificial Intelligence (XAI): Focuses on making AI systems' decision-making processes transparent and interpretable to humans[6]. As AI models, particularly deep learning networks, become more complex, their "black box" nature poses challenges in understanding how specific inputs lead to certain outputs. XAI addresses this by providing insights into the internal workings of AI models, enhancing trust and facilitating informed decision-making. .

Marketing Analytics: Utilizes data-driven insights to optimize marketing strategies and customer targeting [7].

III. Literature Review (Existing Works)

Machine learning techniques have been extensively applied in the banking sector to enhance marketing strategies and customer segmentation. Numerous studies have explored various supervised, semi-supervised and unsupervised learning methods to address these challenges.

1) Supervised Learning: Supervised learning involves training models on labeled data to predict outcomes. Algorithms like Random Forests and Support Vector Machines (SVM) have been employed to predict customer responses to marketing campaigns. Tang and Zhu [8] demonstrated that a Random Forest model achieved an accuracy of 92%, outperforming SVM, which attained 87%.

2) Semi-Supervised Learning: Semi-supervised learning combines labeled and unlabeled data, making it valuable in scenarios where labeling is costly or impractical [9].

3) Unsupervised Learning: Unsupervised learning uncovers hidden patterns in unlabeled data. Clustering algorithms like K-Means have been used for customer segmentation, enabling institutions to tailor marketing strategies [10].

A. Research Gaps

Despite all the advancements, a number of gaps persist in the application of machine learning techniques to bank marketing

According to the literature reviewed, the complex models, especially ensemble often operate as "black boxes", making it difficult to interprete their predictions and gain actionable insights[3]

Identifying the most influential features remains challenging. Overlooking criticall features can lead to suboptimal model performance and misinformed marketing decisions.

The research also shows that marketing datasets often exhibit imbalanced classes, with a majority of non-responders and a minority of responders. This imbalance can bias models towards the majority class, reducing the effectiveness of marketing strategies.[3]

B. Summary of Contributions

This paper addresses the aforementioned challenges by:

- **Implementing Ensemble Learning:** Utilizing Random Forest algorithms to handle data imbalance and enhance predictive accuracy. [2].
- **Feature Importance Analysis:** Employing techniques to identify and prioritize influential features, thereby improving model interpretability and decision-making. [4].
- **Model Interpretability:** Applying model-agnostic interpretability methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), to elucidate model predictions and ensure transparency.
- **Adaptive Modeling:** Designing models capable of updating with new data to reflect changing customer behaviors, ensuring sustained relevance and accuracy.

Due to the outcome class imbalance, the research involved experimenting with various undersampling methods/algorithms. The NearMiss-2 is what we used in this research. NearMiss-2 is a supervised undersampling technique for handling class imbalance in datasets. It enhances data balance by selecting majority class samples that are farthest from the minority class samples. It helps removing noise points, preserving decision boundaries which in turn increases the robustness of classification models.

IV. Methodology

1) Problem Statement: The primary objective is to predict whether a client will subscribe to a term deposit following a marketing campaign and explain how the model reached the prediction. Accurate predictions enable banks to target potential customers more effectively, optimizing marketing resources and enhancing campaign success rates.

2) Significance: Improving prediction accuracy in marketing campaigns leads to:

- Increased subscription rates.
- Reduced marketing costs.
- Enhanced customer satisfaction through personalized marketing.

3) Scope: The study focuses on applying supervised learning techniques to historical marketing data to develop predictive models. It also incorporates interpretability methods to ensure the models provide actionable insights.

4) Proposed AI Approach: For our data collection, we utilized the UCI Bank Marketing dataset, which includes client information and past campaign outcomes. While processing the data:

- No null values were found.
- Categorical features were encoded, and continuous features were normalized.
- Undersampling was performed on the negative class of the outcome.

In model development, supervised learning models, specifically Random Forest and Logistic Regression, were trained to predict client subscription. Model results were assessed using metrics like accuracy, precision, recall, and F1-score. Finally, SHAP and LIME were applied to interpret model predictions and understand feature contributions.

5) AI Evaluation Framework: Confusion Matrix Analysis was applied to examine true positives, false positives, true negatives, and false negatives to understand model performance.

A. Dataset Description

The data is related to directed marketing campaigns (phone calls) of a Portuguese banking institution. The campaigns were based on phone calls, often requiring more than one contact with the same client to assess whether the product (bank term deposit) would be subscribed to.

Our dataset comprises 45,211 rows and 17 columns (features). The data did not have any missing or null values.

TABLE I
FEATURE DESCRIPTIONS IN THE DATASET

Feature Name	Description
`age`	Age of the client
`job`	Job type of the client (e.g., admin, technician, etc.)
`marital`	Marital status of the client (e.g., single, married, divorced)
`education`	Education level of the client
`default`	Credit default status (yes or no)
`balance`	Balance in the client's account
`housing`	Housing loan status (yes or no)
`loan`	Personal loan status (yes or no)
`contact`	Type of communication used (e.g., cellular, telephone)
`day`	Day of the month when the client was last contacted
`month`	Month when the client was last contacted
`duration`	Duration of the last contact in seconds
`campaign`	Number of contacts performed during this campaign
`pdays`	Number of days since the client was last contacted
`previous`	Number of contacts performed before this campaign
`poutcome`	Outcome of the previous campaign (e.g., success, failure, unknown)
`y`	Target variable: has the client subscribed to a term deposit? (yes or no)

This dataset enables us to develop models to predict client subscription behaviour allowing banks to: identify potential product (term deposit) customers, tailor marketing strategies to specific client segements and optimizes resource allocation for bank marketing campaigns.

V. Data Preparation and Exploratory Data Analysis

The data had several categorical features and a few continuous features. Categorical features include job, marital status, education level, default, housing loan, personal loan, contact type, poutcome and the y which is what we want to predict. The continuous fields are age, balance, day contacted, duration, campaign, pdays and previous.

```
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
```

Fig. 1. Categorical Features encoded.

The categorial features were each plotted to show their effect on the outcome y.

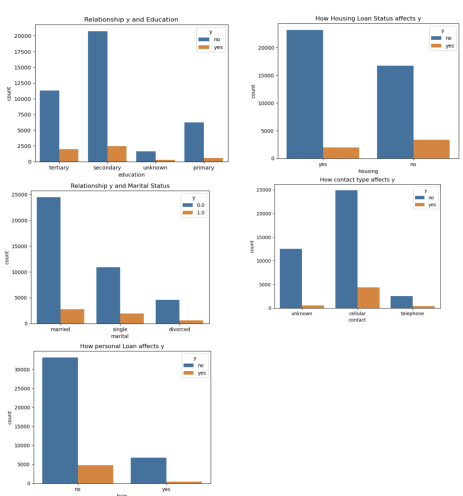

Fig. 2. Categorical Features encoded.

In summary, figure above shows how each categorical feature affects the outcome. In education, clients with tertiary level have a higher likelihood of subscribing. For both housing and personal loan status, clients without these loans are more likely to subscribe to the deposit. Clients contacted via cellular phone have the highest likelihood of subscribing. For the marriage status, single clients have a higher probability of subscribing that the married or divorced. The continuous data fields were normalized and a heat map generated in order to visualize their correlation.

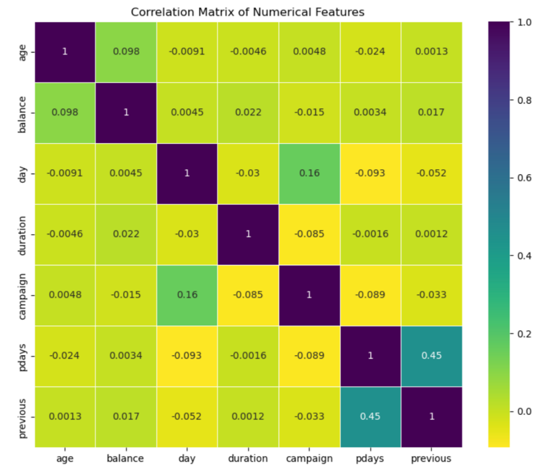

Fig. 3. Categorical Features encoded.

It's observed that most variables have weak correlation with each other. The highest correlation is between pdays and previous (0.45), which implies moderate positive relationship. As pdays increases, previous tends to increase as well. Variables such as duration and campaign, or day and balance, have near-zero correlation indicating almost no linear relationship. On training on further assessment of the data, we discovered there was a class imbalance between the negative class and positive class of the outcome. The negative class had more numbers than the positive class as shown in figure below:

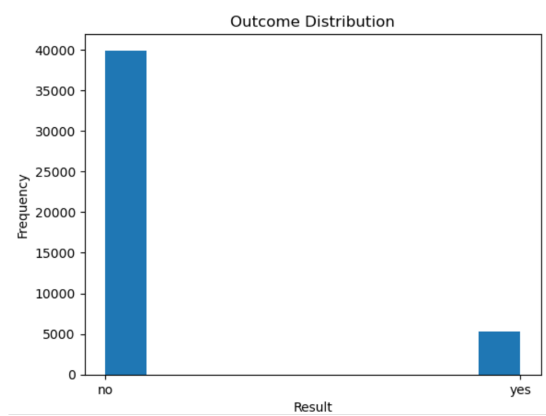

Fig. 4. Categorical Features encoded.

We experimented with a number of undersampling algorithms including random undersampling, NearMiss-1,NearMiss-2, NearMiss-3, Tomek Links, ENN, OSS and NCR but settled for the NearMiss-2 algorithm to undersample the negative class in out dataset. This is because its results produced the best accuracy, precision, recall and F1 score when both models were trained among all the other algorithms. The NearMiss-2 reduced the negative class to the same number of records as the positive one, 5,289

A. ML model selection and optimization.

We used the Logistic regression and random forest modals for this task. Logistic regression: It enables prediction of the probability of a binary outcome. It is simple and interpretable making it ideal for our dataset that is a binary classification. The model uses logit function to transform predicted values into probabilities. Random Forest: This is a collection of decision trees. It makes predictions by majority voting for classification ML analysis. RF creates multiple trees by bootstrapping and each tree is trained on a subset of the data. It is powerful for handling complex, non-linear relationships in the datasets. It is robust to outliers and provides feature importance, which can assist in interpreting which features are most predictive.

In order to use the categorical feature in modal training, we did some hot encoding for those that had 2 categorical values, then label encoding for those which had more than 2 categorical values.

[31]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	4	1	2	0.0	2143	1.0	0.0	2	5	8	261	1	-1	0	3	0.0
1	44	9	2	1	0.0	29	1.0	0.0	2	5	8	151	1	-1	0	3	0.0
2	33	2	1	1	0.0	2	1.0	1.0	2	5	8	76	1	-1	0	3	0.0
3	47	1	1	3	0.0	1506	1.0	0.0	2	5	8	92	1	-1	0	3	0.0
4	33	11	2	3	0.0	1	0.0	0.0	2	5	8	198	1	-1	0	3	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45206	51	9	1	2	0.0	825	0.0	0.0	0	17	9	977	3	-1	0	3	1.0
45207	71	5	0	0	0.0	1729	0.0	0.0	0	17	9	456	2	-1	0	3	1.0
45208	72	5	1	1	0.0	5715	0.0	0.0	0	17	9	1127	5	184	3	2	1.0
45209	57	1	1	1	0.0	668	0.0	0.0	1	17	9	508	4	-1	0	3	0.0
45210	37	2	1	1	0.0	2971	0.0	0.0	0	17	9	361	2	188	11	1	0.0

45211 rows × 17 columns

Fig. 5. Categorical Features encoded.

I used all the features to train my modals and below is how I split my data for each of the modals.

```
#split data into training and testing set.
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)
print("y_train shape:", y_train.shape)
print("y_test shape:", y_test.shape)
```

```
X_train shape: (36168, 16)
X_test shape: (9043, 16)
y_train shape: (36168,)
y_test shape: (9043,)
```

Fig. 6. Data Distribution for Model training.

My data is split into 80% training and 20We did not use any hyperparameters.

B. ML model selection Accountability.

AI accountability is the assurance that AI is functioning properly throughout its entire lifecycle. Accountability within AI systems is not merely an extension of traditional accountability concepts but a distinct challenge dure to the autonomous and often opaque nature of AI decision-making processes. As AI becomes more integrated in our day to day lives, the imperative for robust accountability is crucial.[2] For our dataset we used model agnostic approaches SHAP and LIME on our model results to add the accountability or explainability. Lime creates a local approximation of the model around the specific prediction instance.it goes ahead to slightly modify feature values and observes how the model's predictions change and then base on this to determine which features most influenced the prediction of that data record or instance.

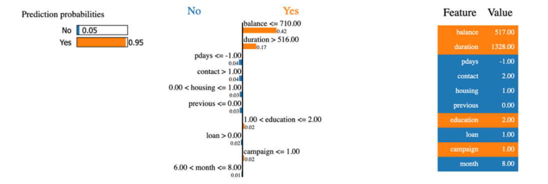

Fig. 7. Lime Positive Instance.

The figure above shows an instance from the test set. A lower balance (`balance` $\leq$ 710.00 [42%]) is a strong indicator for predicting 'Yes,' and longer call durations (`duration` $>$ 516.00 [17%]) also strongly support predicting 'Yes.' Therefore, for this particular instance, the model is highly confident in predicting 'Yes.' Features like `pdays` and `contact` contribute weakly towards 'No.'

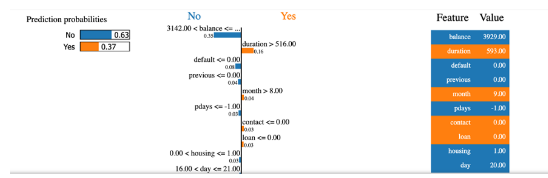

Fig. 8. Lime Negative Instance

Figure 2 shows another instance with a confidence of 63% predicting 'No' and a confidence of 37% predicting 'Yes.' It shows that a high balance (`3929.00`) strongly supports the 'No' prediction. Not having credit in default and not being contacted before also contribute to 'No,' although for the previous contact it's a slight contribution. It is also observed that a high call duration (`593.00` seconds) predicts a high likelihood of 'Yes.' The month being September (represented as `9`) also contributes to 'Yes' as well as the contact type being cellular.

The two examples provide a clear understanding of how individual features, as well as their combinations, influence the model's prediction toward either 'Yes' or 'No.' They highlight the relative importance of specific features (e.g., `balance`, `duration`, and contact history) in determining the outcome, showcasing how the model interprets and weighs these factors for different instances.

VI. Results and Discussion

TABLE II
Performance Metrics for Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	Sensitivity
Random Forest	0.923	0.951	0.884	0.916	0.884
Logistic Regression	0.886	0.886	0.876	0.882	0.876

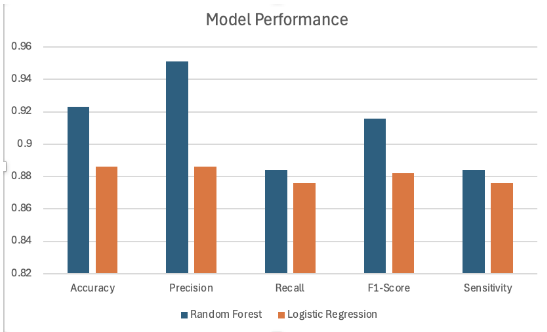

Fig. 9. Model Metrics.

The Random Forest model achieved an accuracy of 92.3%, surpassing Logistic Regression in all evaluation metrics. SHAP and LIME provided transparency by identifying influential features, such as balance and duration, in predicting client subscription likelihood.

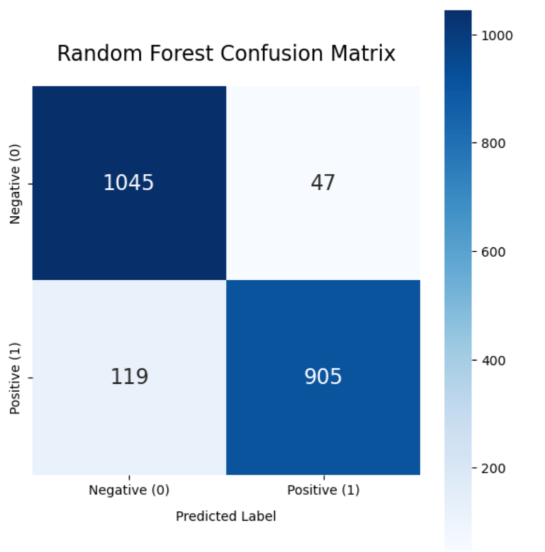

Fig. 10. Random Forest Confusion Matrix.

VII. Conclusion

The Random Forest model demonstrated superior performance in predicting client subscription to term deposits. The application of explainability techniques like SHAP and LIME contributed to understanding the decision-making process of the models, ensuring transparency and accountability.

Potential directions for future research include:

1) **Feature Engineering:** Developing new features to better capture client behavior and preferences.
2) **Ensemble Methods:** Exploring advanced ensemble techniques like XGBoost or Gradient Boosting Machines to improve accuracy further.
3) **Deep Learning:** Applying neural network architectures to capture complex patterns in the data.

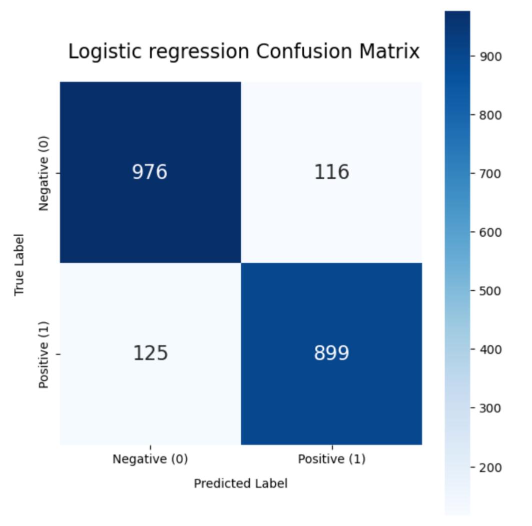

Fig. 11. Logistic Regression Confusion Matrix.

4) **Cost-Sensitive Learning:** Addressing business-specific objectives by optimizing for false positives and negatives.
5) **Real-Time Prediction Systems:** Implementing dynamic models to enable responsive marketing strategies.

VIII. Dataset and Python Source Code

1) **Python code:** https://github.com/karanzijm/MLExploratoryDataAnalaysis
2) **Slides:** https://github.com/karanzijm/MLExploratoryDataAnalaysis
3) **Paper:** https://www.overleaf.com/project/675b2e8819a6887e34374cc9

References

[1] M. Lichman, "Uci machine learning repository: Bank marketing dataset," 2013, [Online]. Available: https://archive.ics.uci.edu/dataset/222/bank+marketing.

[2] H. Liu *et al.*, "Research on methods for balancing component sample categories under rule constraints," in *2024 International Conference on Intelligent Computing and Data Mining (ICDM)*, 2024, pp. 81–85.

[3] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, [Online]. Available: https://proceedings.neurips.cc.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[5] A. B. Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[6] A. Adadi and M. Berrada, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities, and challenges toward responsible ai," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[7] Y.-S. Jeon and D.-J. Lim, "Psu: Particle stacking undersampling method for highly imbalanced big data," *IEEE Access*, vol. 8, pp. 131 920–131 927, 2020.

[8] W. Tang and H. Zhu, "Ensemble learning for bank marketing campaign success prediction," *PLOS ONE*, vol. 19, no. 5, pp. 1–15, May 2024, [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0294759.

[9] G. Vilone and L. Longo, "Explainable artificial intelligence: A systematic review," arXiv preprint arXiv:2006.00093, 2020, [Online]. Available: https://arxiv.org/abs/2006.00093.

[10] S. Ghosh, A. Sinha, and P. Bhattacharya, "A comparative study on bank market prediction using machine learning," in *Lecture Notes in Computer Science*. Springer, Cham, 2018, pp. 369–384.