Prediction Modeling for Bank Deposit Subscription

Wilbert Rodriguez
Computer Science
California State University, Fullerton
Fullerton, USA
wilbertrodriguez@csu.fullerton.edu

Tommy Le
Computer Science
California State University, Fullerton
Fullerton, USA
tommyle@csu.fullerton.edu

Kanika Sood
Computer Science
California State University, Fullerton
Fullerton, USA
kasood@fullerton.edu

***Abstract*—Machine Learning, a sub-specialty of AI, has a wide range of applications that aim to address various real-world scenarios effectively. These applications span many fields such as medicine, education, business, and more. For instance, in the banking and financial sector, machine learning techniques can prove useful. In this case, we focus on analyzing data from direct bank marketing campaigns conducted via phone by a Portuguese banking institution between 2008 and 2013. The primary objective is to predict whether clients would subscribe to a term deposit based on the available information. To achieve accurate predictions, we have chosen to employ a combination of powerful machine learning algorithms, including Gradient Boosting, Logistic Regression, Naive Bayes, and Random Forest. By leveraging the strengths of these techniques, we can uncover valuable insights and improve decision-making processes in the banking domain.**

***Index Terms*—Bank deposits, Machine Learning, Data, Marketing**

I. Introduction and Background

The idea of computers doing the tasks of humans was once a dream, but as time went on, that dream became a reality, and that led to the creation of Artificial Intelligence. Artificial Intelligence (AI) is a part of the computer science field that works with the idea of computers performing tasks that people would otherwise do themselves [3]. A way we can improve its performance would be through machine learning which has the ability to be able to "learn" to find patterns within data and do tasks that do not need complex programming. Machine learning can be applied to many domains, such as hospitals, businesses, and fast-food places. Another domain that it can be extended to successfully is the financial area. There, we have to find out what the goal was, and it was to find out how many would subscribe to a term deposit. We use the Bank Marketing data from a Portuguese retail bank from 2008 to 2013 from UCI Machine Learning Repository. In this work, we propose a data mining approach to predict it.

The marketing campaigns are based on phone calls. Since the goal was to predict if the client will subscribe to a term deposit, it would be classification. There were four datasets: bank-additional-full.csv with all examples and 20 inputs, bank-additional.csv with 10 percent of the examples, bank-full.csv with all examples, and 17 outputs, and bank.csv with 10 percent and 17 inputs. For the features, there are those like age, duration, and poutcome. In this paper, we will analyze the data and have the machine be able to predict if a client would subscribe to a term deposit or not based on the data we got from the client. We hope through our analysis, we can predict depending on their background that they will either subscribe to the term deposit of a bank or not. By understanding this, this can open up new opportunities for innovation.

Predicting whether a client will sign up for a term deposit based on the information available is really important in the banking and financial industry. It helps banks and financial institutions in many ways. Firstly, by accurately figuring out who is likely to subscribe, they can make their marketing campaigns more effective. This means they can reach out to the right people and use their resources wisely. It also helps them improve their strategies for acquiring new customers and keeping existing ones.

Understanding the factors that influence a client's decision to sign up for a term deposit allows banks to customize their offerings and communication. They can provide better options and information that match the customers' needs and preferences. This personalized approach makes customers happier and strengthens their relationship with the bank. It can even help the bank make more money by attracting more customers and increasing profits. Using machine learning to solve this problem also has practical benefits. It helps reduce the risks associated with unsuccessful marketing campaigns. By accurately predicting who is likely to subscribe, banks can avoid wasting time and resources on people who are not interested. This saves money and allows them to focus on potential customers who are more likely to be interested in the term deposit.

II. Bank Marketing Dataset

A. Attributes Information

The dataset has have a lot of input variables. There was only one output variable which is y which is about if the client has subscribed a term deposit which the answers being either 'yes' or 'no'. The input variables are age, job, marital, education, default, housing, loan, contact, month, day of the week, duration, campaign, pdays, previous, poutcome. Age, job, marital, education, month, and day of the week should be self-explanatory. Default is if they have credit in default, housing is having a housing loan or not and loan is having a personal loan or not. Contact is communication type, month and the day of the week is the time last contacted, duration is last contact duration in seconds, campaign is number of

contacts performed during this campaign and for this client, pday is number of days passed by after the client was last contacted, previous is number of contacts performed before this campaign and for this client. Poutcome is the outcome of the previous marketing campaign. Below are a few datapoints from the dataset.

TABLE I
PART OF THE BANK MARKETING DATA

Age	Job	Marital	Education	Default	Balance	Y
58	Management	Married	Tertiary	No	2143	No
44	Technician	Single	Secondary	No	29	No
33	Entrepreneur	Married	Secondary	No	2	No
47	Blue-collar	Married	Unknown	No	1506	No
33	Unknown	Single	Unknown	No	1	No
35	Management	Married	Tertiary	No	231	No
28	Management	Single	Tertiary	No	447	No
42	Entrepreneur	Divorced	Tertiary	Yes	2	No
58	Retired	Married	Primary	No	121	No
43	Technician	Single	Secondary	No	593	No

III. PREPARATION AND ANALYSIS

Once we have obtained the dataset, we embark on a crucial phase, which is preprocessing the data to ensure its readiness for testing. The process of analyzing and preparing the data holds immense significance. In the case of a dataset centered around the Semantic Segmentation of 3D Mobile LiDAR Point Clouds, data preparation entails selecting a meaningful cluster of points from the outdoor 3D LiDAR point clouds to serve as input for the neural networks [9]. Acquiring high-quality datapoints is of utmost importance to ensure accurate predictions. The accuracy of the entire process hinges on this stage, as any errors or inaccuracies during data preparation can propagate throughout the subsequent analysis, leading to flawed outcomes. Therefore, meticulous attention must be given to this critical step to lay a solid foundation for the subsequent stages of the analysis.

A. Preprocessing

Preprocessing in the dataset would be visualizing the data to gain insights and looking for correlations. There is also cleaning up the data, which means handling categorical data and outliers. Data preprocessing is similar to SmartData which is another , and refers to the challenge of transforming raw data into quality data, as one of the most important stages in the data mining process since it will clean up the data so it is more efficient and more accurate[2]. The goal of preprocessing is to transform the data into a format that is suitable for machine learning algorithms and also improve the quality of the data. The steps are usually data cleaning which means removing incorrect values or missing values, data transformation which means encoding it or converting tdata into a format that the algorithm can use, feature selection which is choosing the most relevant features, and feature engineering which involves creating new features from existing data. We actually had to transform part of the data because they were categorical. We used one-hot encoding to work with the categorical features in our dataset. It is a popular encoding method that is used on categorical variables that contain binary vectors that turn variables of categorical features into numerical values such as either 0 or 1 [1]. Also, just in case, we removed any ages less than 1. There was no missing values when we checked on Juypter Notebook.

B. Models

After preprocessing and ensuring the dataset is as efficient as possible, we build models based on four machine-learning techniques that we believe are the best for this dataset. That being logistic regression, gradient boosting, random forest, and Naive Bayes and why we believe it would be good is explained below:

- Logistic regression is a commonly used method in machine learning that build a model to distinguish two or more categories of the samples [4]. It is also simple as it provides results that is easy to understand how it makes predictions, efficient, and robust. It is ideal for large datasets with many features and knowing the banking industry, it is very suited to handle this. It can handle noise very well. It can also handle binary classification and since we are testing the output of either ”yes” or ”no”, this will also be another reason why this is one of the better techniques. Adding on to that, it can also handle continuous data at times. We made it based on our data, and it looks like this. However, we have been working to improve the plot, too, so there is another plot that has been through feature selection. We also use feature selection methods, the number of features can be reduced by removing redundant and irrelevant attributes from datasets [6].

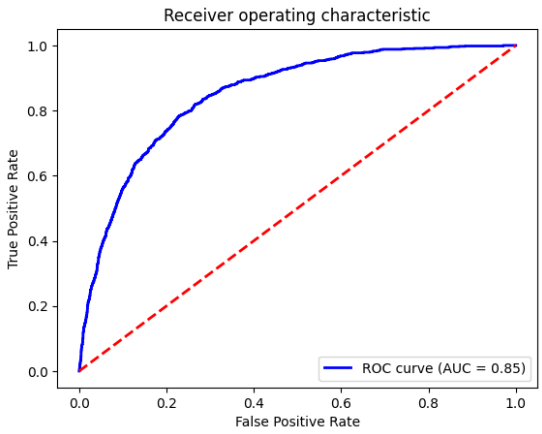

Fig. 1. Logistic Regression Model

- Gradient boosting. It is a supervised machine learning technique and an emerging machine learning method for time series forecasting in recent years [8]. There is a plot where we did feature selection as well. It is well-suited to predict from many different areas which includes bank marketing. It can handle complex relationships between

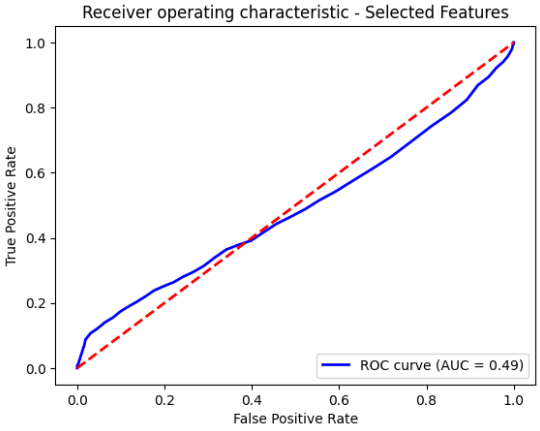

Fig. 2. Logistic Regression Model with Feature Selection

the features and the outcome. In banking, there are many factors that can influence the choices of the customers including age, employment status, if they have a loan, and income. Gradient Boosting can learn these relationships and make accurate predictions. It can also handle both categorical and continuous variables which is another plus. It can also handle missing data and although our dataset didn't have any missing values, it can be used if the data becomes inconsistent. Since our dataset is huge, it is good that gradient boosting can handle large datasets. It can also handle class imbalance. This is why we decided to use this technique as well.

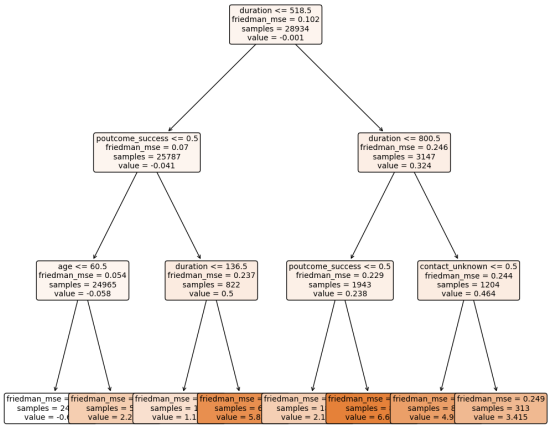

Fig. 3. Gradient Boosting Model

- Random Forest models are basically collections of prediction trees that when together, they sort of perform a "forest" [11]. They grow many prediction trees that will convert into optimal splits. We have done it as well, but it seems like a lot still. It can also handle complex relationships like the previous one as it can capture the relationships between the features and the outcomes. The

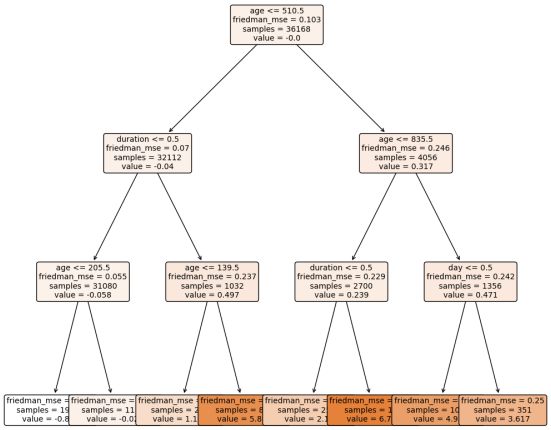

Fig. 4. Gradient Boosting Model with Feature Selection

features we have like age, martial status, default, housing loan, personal loan, and balance. It can deal with complex relationships and make accurate predictions. Overfitting can also be a very huge issue so this is another reason this technique is good. It reduces overfitting as it creates many decision trees with different random subsets of features and data samples. It can provide a more robust and accurate prediction than a single decision by averaging the predictions of these trees. It can also handle both categorical and continuous variables, like the other techniques. It can provide a measure of feature importance. This can be useful in bank marketing to understand which features are the most important in predicting customer behavior. It can also handle missing data like the others, and it can handle large datasets.

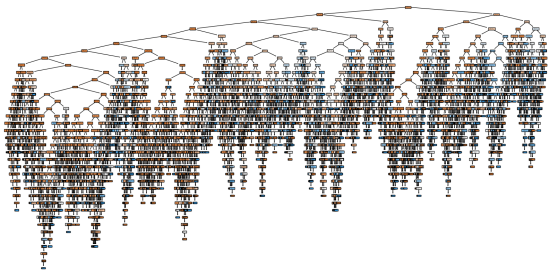

Fig. 5. Random Forest Model

- Naive Bayes is a fast and mature algorithm that is used in fault diagnosis [9]. It is compatible with very large datasets to build and for further analysis [5]. It can handle categorical and continuous variables and it is also fast and efficient. It can be trained quickly on large datasets, which is good for bank marketing as analyzing data quickly is good for an area like that. It can also handle irrelevant features as it assumes all features are independent of each other. That means its can handle

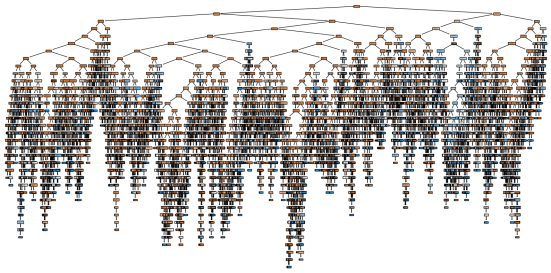

Fig. 6. Random Forest Model with Feature Selection

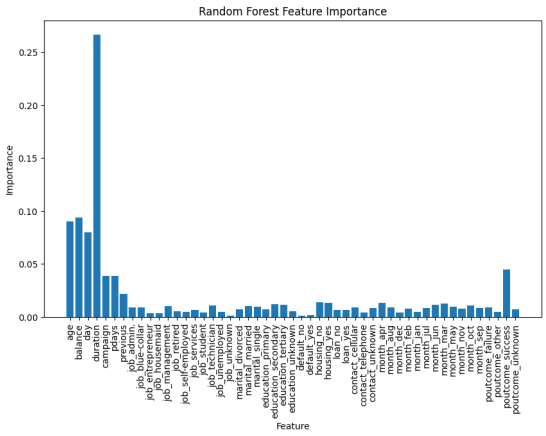

Fig. 7. Feature Importance

noise and it can also work well with small datasets. It also provides interpretable results that can help explain the predictions that were made. The plots are shown down below:

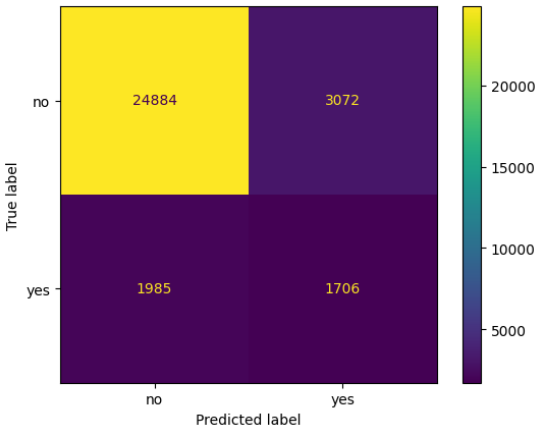

Fig. 8. Naive Bayes Model

Conclusion and Evaluation

With all that data collected, we created Precision-Recall Curve, ROC Curve, and Confusion Matrix for each technique. ROC, also known as the Relative operating characteristic, measures the relative discrimination between the True Positive

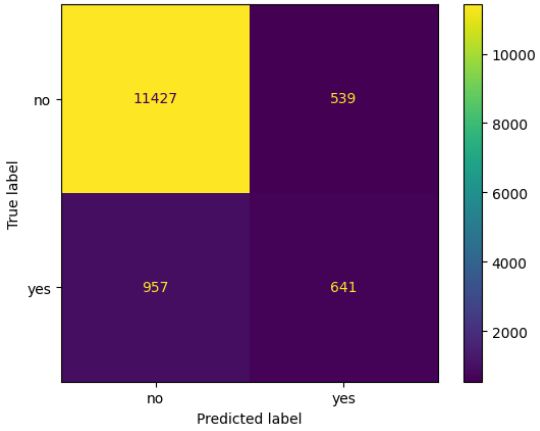

Fig. 9. Naive Bayes Model with Feature Selection

Rate (TPR) and the False Positive Rate (FPR) [11]. It is a graphical representation of the performance of a binary classification algorithm. TPR or True Positive Rate is the proportion of actual positive instances that are actually correct while FPR or False Positive Rate is the proportion of negative instances that are actually positive by the classifier. TPR and FPR are calculated at various points along the curve. ROC is useful as it provides a comprehensive view of the binary classification algorithm performance across all possible threshold values. A good classifier would have the curve try to be as close to the top left corner of the plot with TPR of 1 and FPR of 0 at all thresholds. The AUC or area under the ROC curve is used to evaluate the performance of the binary classification algorithms as it is the measure of the overall performance of the algorithm across all possible threshold values. An AUC of 0.8 or higher is considered a good performance.

A precision-recall curve plots the precision against recall [12]. Precision is the proportion of true positive instances among all instances that are positive, while recall is the proportion of true positive instances that are corrected and identified by the classifier among all actual positive instances. This is useful because it provides a different perspective on the performance of a binary classification algorithm than a ROC curve. Precision-recall curve provides more informative insights when the positive class is rare or went he cost of false positive and false negative is not equal. It can also be used to compare the performance of different algorithms or to select the optimal classification threshold based on the tradeoff between precision and recall. The area under the precision-recall curve is used to evaluate the performance of binary classification algorithms and ranging between 0 and 1 and if it was higher, it would be stating it is better performance.

The Confusion Matrix is a table summarizes a binary classification model and shows off the predicted values to the actual values. It counts the number of true positive, true negative, false positive, and false negative predictions made by the algorithm. TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives,

and TN is the number of true negatives. It is important because it provides a detailed summary of the performance of a binary classification algorithm and it allows us to calculate many different performance metrics such as accuracy, precision, recall, F1-score, and others. They provide insights into the strengths and weaknesses of the algorithm and help make informed decisions about how to improve the performance.

We have analyzed the accuracy, precision, recall, F1 score, and AUC, as well as the ROC curve, Precision-Recall Curve, and Confusion Matrix. To summarize the others, accuracy measures the proportion of correct predictions made by the model compared to the total number of predictions made, while precision is the proportion of true positive predictions among the total positive predictions. Accuracy is calculated as the ratio of the number of correct predictions to the total number of predictions and precision is calculated as the ratio of the number of true positive predictions to the sum of true positive and false positive predictions. F1 is the mean of precision and recall. It provides a balance between precision and recall and it is used when the distribution of positive and negative instances in the dataset is skewed. Each metric provides a different aspect of the performance of the model and together they provide a comprehensive evaluation.

For Gradient Boosting, the accuracy is 0.895, the Precision is 0.620, the Recall is 0.341,and F1 score is 0.440. The Confusion Matrix is 7724, 228, 719, and 372. The Precision-Recall Curve seems to fall fast to around 0.5 on Precision when Recall is 0.0. It does increase a bit but eventually it curves down at 1.0. The ROC Curve seems aright as it doesn't dip back but the area where False Positive Rate is between 0.2 and 0.4 isn't as close the 1.0 of the True Positive Rate which is what we strive for.

For Logistic Regression, the accuracy is 0.887, the Precision is 0.583, the Recall is 0.214, and F1 is 0.313. The Confusion Matrix is 7.8e+03, 1.7e+02, 8.6e+02, and 2.3e+02. The ROC Curve is a smaller curve where around 0.2 to 0.4 for False Positive Rate, it is around 0.7 for True Positive Rate. The Precision-Recall curve has its drops to around 0.6 for Precision when Recall is 0.0.

For Naive Bayes, the accuracy is 0.0567, the Precision is 1.0, the Recall is 0.0567, and F1 is 0.107. The Confusion Matrix is 11427, 539, 957, and 641. The ROC curve at 0.0 to 0.2 for False Positive Rate is near 0.84 for True Positive Rate. The Precision-Recall Curve after the drop from 1.0 to 0.5 for Precision, it's highest point before the drop again is at 0.09 for Recall, it has a 0.7 for Precision.

For Random Forest, the accuracy is 1.0, the precision is 1.0, the recall is 1.0, and the F1 score is 1.0. The Confusion Matrix is 7.8e+03, 1.8e+02, 6.9e+02, and 4e+02. The Precision-Recall Curve drops until 0.9 for Precision at 0.01 for Recall and ends at 0.2 at 1.0 for (Precision, Recall). The ROC Curve has its closest point to the left corner at around 0.9 for True Positive Rate at False Positive Rate at 0.1.

After carefully examining and comparing the performance of different techniques using evaluation metrics such as the ROC curve, Precision-Recall Curve, and Confusion Matrix, we have arrived at the conclusion that Random Forest outperforms the other methods. Notably, Random Forest demonstrates a more favorable precision-recall curve, maintaining higher values and showing a slower decline compared to the alternative techniques. This indicates its superiority in accurately predicting positive instances. Moreover, when comparing it with logistic regression, Random Forest displays a higher True Positive rate, further emphasizing its effectiveness.

In addition, the ROC plot of Random Forest reveals its performance, with the curve closely hugging the top-left corner of the plot closer than the others. This suggests a higher true positive rate while keeping the false positive rate at a minimum. Overall, based on the evidence derived from the data, we confidently assert that Random Forest is the most promising technique thus far for predicting term deposit subscriptions. It is worth noting that the data used in our analysis exhibited a high level of accuracy. As we continue to refine and develop this approach, there is a strong potential to harness it for the creation of a robust system capable of accurately predicting the likelihood of clients subscribing to a term deposit. By leveraging the power of Random Forest and additional data and features, we can further increase the system's accuracy and utility.

Furthermore, a successful predictive model for term deposit subscriptions are fantasic. Financial institutions can utilize this information to optimize their marketing strategies, targeting individuals who are more likely to subscribe. This targeted approach not only maximizes the efficiency of promotional efforts but also minimizes costs associated with reaching out to uninterested clients. Additionally, such predictive models enable banks to personalize their offerings and communication, thereby improving customer satisfaction and strengthening customer relationships.

In summary, Random Forest has proven to be the most effective technique for predicting term deposit subscriptions based on the available data. As we continue to refine and expand upon this approach, we anticipate the significant potential for creating a practical and accurate system that can benefit the banking and financial sector by optimizing marketing efforts, enhancing customer relationships, and improving overall operational efficiency.

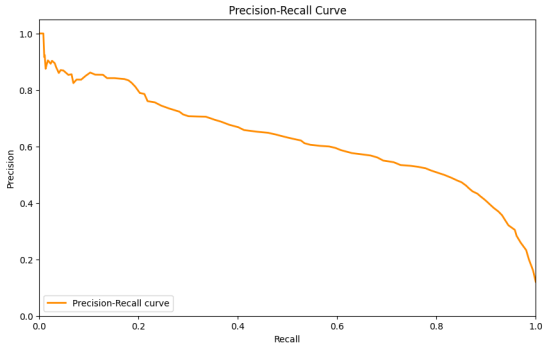

Fig. 10. Precision-Recall Curve

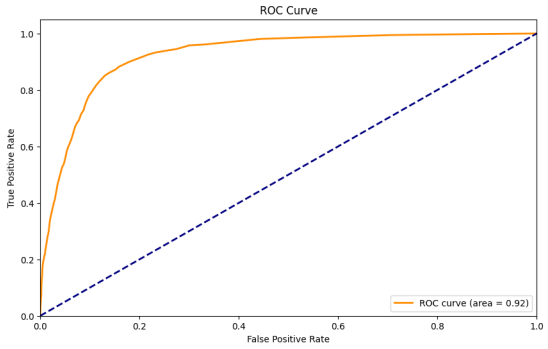

Fig. 11. ROC Curve

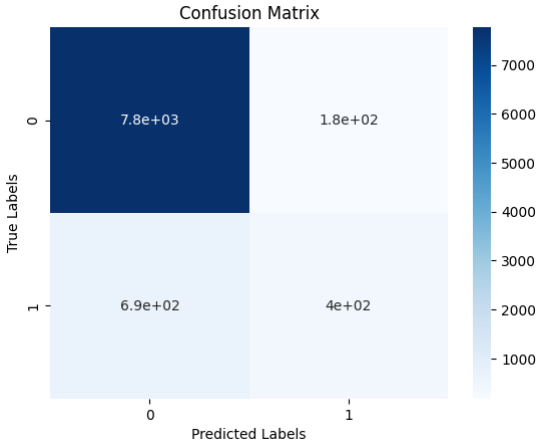

Fig. 12. Confusion Matrix

REFERENCES

[1] Al-Shehari, Taher, and Rakan A. Alsowail. "An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques." Entropy: an International and Interdisciplinary Journal of Entropy and Information Studies., vol. 23, no. 10, 2021, https://doi.org/10.3390/e23101258.

[2] Cordón, Ignacio, et al. "Smartdata: Data Preprocessing to Achieve Smart Data in R." Neurocomputing, vol. 360, 2019, pp. 1–13, https://doi.org/10.1016/j.neucom.2019.06.006.

[3] Dang, Amit, et al. "Extent of Use of Artificial Intelligence Machine Learning Protocols in Cancer Diagnosis: A Scoping Review." Indian Journal of Medical Research, vol. 157, no. 1, 2023, https://doi.org/10.4103/ijmr.IJMR 555 20.

[4] Huang, Yongfen, et al. "Prediction Model of Bone Marrow Infiltration in Patients with Malignant Lymphoma Based on Logistic Regression and XGBoost Algorithm." Computational Mathematical Methods in Medicine, June 2022, pp. 1–7. EBSCOhost, https://doi-org.lib-proxy.fullerton.edu/10.1155/2022/9620780.

[5] Jackins, V., et al. "AI-Based Smart Prediction of Clinical Disease Using Random Forest Classifier and Naive Bayes." Journal of Supercomputing, vol. 77, no. 5, May 2021, pp. 5198–219. EBSCOhost, https://doi-org.lib-proxy.fullerton.edu/10.1007/s11227-020-03481-x.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Kumar, Prashant, et al. "Feature Selection Using PRACO Method for IDS in Cloud Environment." Journal of Intelligent Fuzzy Systems, vol. 43, no. 5, Nov. 2022, pp. 5487–500. EBSCOhost, https://doi-org.lib-proxy.fullerton.edu/10.3233/JIFS-212196..

[8] Mahmoudi Kouhi, Reza, et al. "Data Preparation Impact on Semantic Segmentation of 3D Mobile LiDAR Point Clouds Using Deep Neural Networks." Remote Sensing, vol. 15, no. 4, 2023, https://doi.org/10.3390/rs15040982.

[9] Noorunnahar, Mst, et al. "A Tree Based EXtreme Gradient Boosting (XGBoost) Machine Learning Model to Forecast the Annual Rice Production in Bangladesh." PLoS ONE, vol. 17, no. 3, Mar. 2023, pp. 1–15. EBSCOhost, https://doi-org.lib-proxy.fullerton.edu/10.1371/journal.pone.0283452.

[10] Rachakonda, Aditya Ramana, and Ayush Bhatnagar. "A: Extending Area Under the ROC Curve for Probabilistic Labels." Pattern Recognition Letters, vol. 150, 2021, pp. 265–71, https://doi.org/10.1016/j.patrec.2021.06.023..

[11] Wang, Yi, et al. "Generator Fault Classification Method Based on Multi-Source Information Fusion Naive Bayes Classification Algorithm." Energies (19961073), vol. 15, no. 24, Dec. 2022, p. 9635. EBSCOhost, https://doi-org.lib-proxy.fullerton.edu/10.3390/en15249635.

[12] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.