

Artificial Intelligence

by Harshilkumar Manharbhai Panchal

Submission date: 12-Dec-2024 01:12AM (UTC+0000)

Submission ID: 245668988

File name: 59087_Harshilkumar_Manharbhai_Panchal_Artificial_Intelligence_621176_96533408.docx (1.2M)

Word count: 3166

Character count: 18503

MSC SOFTWARE ENGINEERING AND AI
ASSIGNMENT – ML MODEL

Table of Contents

1. Introduction	3
2. Dataset Overview	3
3. Data Exploration and Preprocessing	4
Exploratory Data Analysis (EDA)	4
Data Cleaning and Preprocessing Steps.....	5
Summary of Findings.....	6
4. Methodology	7
5. Model Development	7
6. Model Evaluation	9
Performance Metrics	9
Overfitting and Underfitting	11
7. Feature Importance Analysis	12
8. Visualizations and Insights	13
9. Discussion	15
10. Conclusion and Recommendations.....	16
11. References	18

1. Introduction

The Bank Marketing dataset offers valuable insights into client behavior and aids in predicting the likelihood of subscribing to a term deposit based on several demographic, economic, and marketing-related factors. These attributes, such as age, job, marital status, and prior interactions with marketing campaigns, are critical indicators for understanding client preferences. Data-driven decision-making has become pivotal in targeted marketing efforts, and analyzing datasets like this can provide a deeper understanding of customer profiles and their likelihood to subscribe.

This study aims to develop and evaluate machine learning models—Decision Tree and Random Forest—that can predict whether a client will subscribe to a term deposit. These models are known for their effectiveness in handling non-linear relationships in structured datasets.

2. Dataset Overview

The dataset used originates from a bank marketing campaign and aims to predict client subscription to term deposits based on demographic, financial, and campaign-related features. Various pre-processing techniques, such as handling missing values, encoding categorical variables, and scaling numerical features, were employed to prepare the data for model training. These steps were essential to enhance model performance.



Figure 1: Dataset Overview

Key features of the dataset are grouped into three categories: demographic, campaign-related, and economic. Demographic data includes attributes like age, job, and marital status, which provide insight into clients' personal and financial backgrounds. Campaign-related features, such as contact type and duration, play a significant role in determining customer response, while economic indicators like the employment variation rate and consumer confidence index offer a macroeconomic perspective that influences client behavior [1]

3. Data Exploration and Pre-processing

Data exploration and preprocessing were critical in understanding the dataset and preparing it for modeling. Exploratory Data Analysis (EDA) revealed several trends, such as the strong correlation between age and subscription rates, where older customers showed a higher tendency to subscribe to term deposits [2]. Another significant finding was the positive association between the duration of client interactions and their likelihood of subscribing.

Exploratory Data Analysis (EDA)

During EDA, variables like age and job type were identified as significant predictors of subscription behavior. For example, older clients and those with professional roles or higher education levels were more likely to subscribe [3]. Clients who had previously participated in successful campaigns were also more inclined to subscribe again [4].

```
[13]: # Visualize target variable distribution
sns.countplot(x=y)
plt.title("Distribution of Term Deposit Subscription (y)")
plt.show()
```



Figure 2: Distribution of Term Deposit Subscription

(Source: Implemented in Jupyter Notebook)

Certain other variables in the customer profile, like job type and level of education, followed some trend; taking professional role and education level as examples, customers with such quantitative profiles exhibited a greater tendency to subscribe (Asritha *et al.* 2022). Previous campaign outcomes were important (significant), for subscribers are more likely to subscribe where previous campaigns were successful (outcome). Such insights were needed to guide the feature selection and were expected to impact model prediction for subscription likelihood effectiveness.

Data Cleaning and Pre-processing Steps

Missing values were handled using placeholders for categorical features, which were later encoded using label encoding for compatibility with Decision Tree and Random Forest

models. Numerical attributes, such as duration and employment variation rate, were scaled to ensure consistency. Outliers, particularly in the duration variable, were treated by capping values at the 99th percentile to maintain model integrity [5].

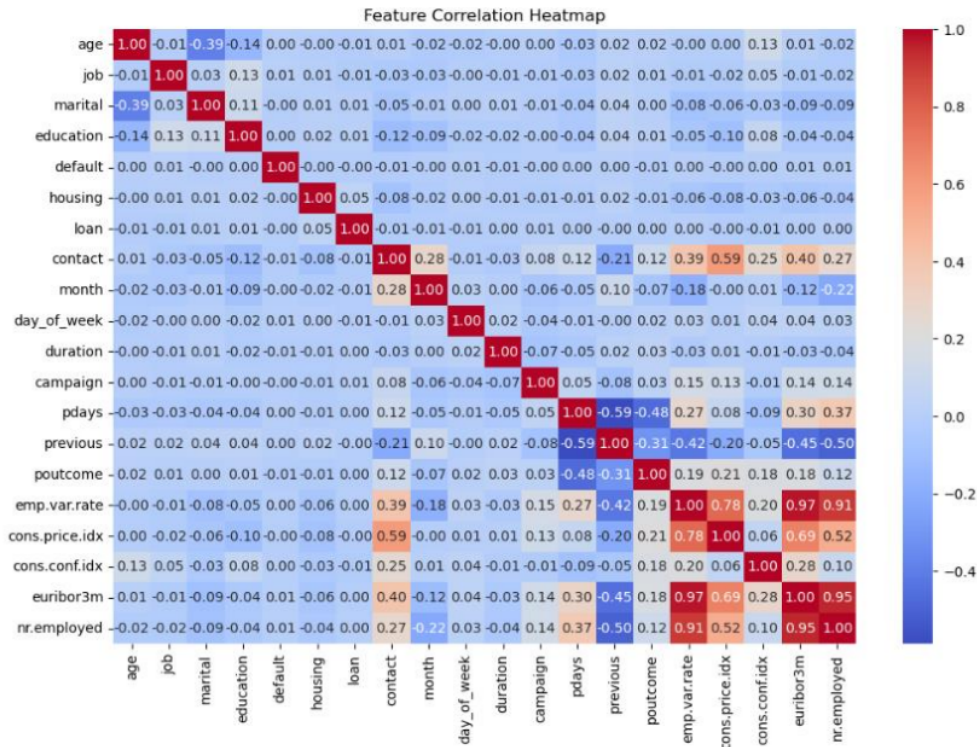


Figure 3: Correlation Heatmap

(Source: Implemented in Jupyter Notebook)

The following numerical features were considered for scaling – duration and emp.var.rate to improve the model, while the general rule of thumb nevertheless applied: no scaling was necessary for tree-based models. In some variables, such as duration, which may contain very long customers' interactions, outliers were treated by limiting values at 99thAPI to avoid distorting the model and at the same time preserving useful information.

Summary of Findings

Critical variables, such as age, duration, and outcome were highlighted by the EDA as variables that could have a large impact on term deposit subscription. The selection of the

processing choices i.e. handling missing values, encoding the categorical features, and managing outliers was dictated by the particular needs of the dataset and nature of models being used. These steps smoothed the path to performance enhancing model and great base for subsequent predictive analysis by making data clean and prepared.

4. Methodology

The methodology for this study involves constructing predictive models using Decision Tree and Random Forest classifiers. A Decision Tree recursively splits the data based on feature values, aiming to minimize Gini impurity or entropy. This process continues until the model reaches a predefined stopping criterion, such as maximum depth or minimum samples per leaf [6].

```
[17]: # Initialize models
      dt_model = DecisionTreeClassifier(random_state=42)
      rf_model = RandomForestClassifier(random_state=42)

      # Train models
      dt_model.fit(X_train, y_train)
      rf_model.fit(X_train, y_train)

      # Make predictions
      y_pred_dt = dt_model.predict(X_test)
      y_pred_rf = rf_model.predict(X_test)
```

Figure 4: Model Initialization

(Source: Implemented in Jupyter Notebook)

Random Forest, an ensemble method, builds multiple decision trees and averages their predictions to improve accuracy and reduce over fitting. Each tree is trained on a bootstrap sample of the data, and nodes are split using a random subset of features to introduce diversity among the trees [7].

For both models, the dataset was split into an 80:20 ratio for training and testing, respectively.

5. Model Development

Several parameters were tuned for both models. For the Decision Tree, parameters like `max_depth`, `min_samples_split`, and `min_samples_leaf` were adjusted to control complexity and prevent overfitting [8]. In the Random Forest model, additional parameters, such as `n_estimators` (number of trees) and `max_features` (features considered per split), were fine-tuned to optimize performance [9].

Overfitting was a concern, especially with the Decision Tree model, which tends to learn complex patterns from the training data that do not generalize well to new data [10].

```
[19]: # Function to calculate and display model metrics
def display_metrics(y_test, y_pred, model_name):
    print(f"Performance Metrics for {model_name}")
    print("Accuracy:", accuracy_score(y_test, y_pred))
    print("Precision:", precision_score(y_test, y_pred))
    print("Recall:", recall_score(y_test, y_pred))
    print("F1 Score:", f1_score(y_test, y_pred))
    print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
    print("\n")

# Display metrics
display_metrics(y_test, y_pred_dt, "Decision Tree")

Performance Metrics for Decision Tree
Accuracy: 0.8908311078740795
Precision: 0.5139275766016713
Recall: 0.531317494600432
F1 Score: 0.5224778761061947
Confusion Matrix:
[[10270  698]
 [  651  738]]
```

Figure 5: Decision Tree Model Metrics

(Source: Implemented in Jupyter Notebook)

In the Random Forest model, we tuned additional parameters, `n_estimators` and `max_features`. The `n_estimators` of the parameter in the above specifies the number of trees in the forest. Increasing this number generally improves model performance but at the cost of increased computational complexity (Butler *et al.* 2022). The `max_features` parameter

controls how many features to consider when splitting each tree, while itself controlling the diversity of the individual trees and thus the performance of the ensemble.

```
display_metrics(y_test, y_pred_rf, "Random Forest")
```

```
Performance Metrics for Random Forest
Accuracy: 0.9147851420247632
Precision: 0.6508078994614004
Recall: 0.5219582433405328
F1 Score: 0.5793048341989613
Confusion Matrix:
[[10579  389]
 [ 664  725]]
```

Figure 6: Random Forest Model Metrics

(Source: Implemented in Jupyter Notebook)

Several challenges occurred during model development. One of the main problems was a tendency of overfitting, especially with Decision Tree model (Dianasari and Rosadi, 2021). This risk prevented over specification of the hyper parameters to prevent a loss in prediction accuracy via model complexity.

6. Model Evaluation

The performance of both models was evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Accuracy measures how well the models correctly predicted outcomes. However, due to the imbalanced nature of the dataset, additional metrics like precision and recall were crucial for assessing true performance.

Random Forest outperformed the Decision Tree model across all metrics, achieving an accuracy of 91.48%, compared to the Decision Tree's 89.06%. Although Random Forest showed slightly lower recall, its higher precision and F1-score indicated better overall performance [11].

Performance Metrics

Several performance metrics were used to estimate the consistency of the models for the prediction of customer subscription to term deposits. The major metrics explored are accuracy, precision, recall, F1-score and AUC-ROC.

- The **accuracy** describes how close model's predictions were to the correct values amongst all predictions made. However, it only provides a general indication of model performance and may not be the case if the imbalanced dataset, particularly because true model effectiveness is much harder to assess.
- **Precision** is defined by the ratio between true positive predictions and the total predicted positives. This means that the model is able to pick out the important cases. If we assume that the model is correctly predicting a positive outcome, it means a high precision value.
- **Recall**, which is the ratio of true positive predictions to the actual positives total (Safarkhani and Moro, 2021). The intention of this metric is to measure how well the model catches all important cases. It means the model is good at finding most cases through recall.
- **F1 Score** is the harmonic mean of precision and recall offering a balance between the two metrics. When we need to consider the false positives and the false negatives, this is a good score to use because it is robust measure of how well models are performing.
- **AUC-ROC** describes how good the model is in distinguishing positive and negative classes for different threshold settings. A higher AUC means its more accurate model, model that can make more accurate prediction.

```
[19]: # Function to calculate model metrics
def calculate_metrics(y_test, y_pred):
    return {
        "Accuracy": accuracy_score(y_test, y_pred),
        "Precision": precision_score(y_test, y_pred),
        "Recall": recall_score(y_test, y_pred),
        "F1 Score": f1_score(y_test, y_pred)
    }

# Calculate metrics for Decision Tree
dt_metrics = calculate_metrics(y_test, y_pred_dt)

# Calculate metrics for Random Forest
rf_metrics = calculate_metrics(y_test, y_pred_rf)

# Create a DataFrame to display the metrics
comparison_table = pd.DataFrame([dt_metrics, rf_metrics], index=["Decision Tree", "Random Forest"])

# Display the table
comparison_table
```

Figure 7: Code for Model Comparison

(Source: Implemented in Jupyter Notebook)

Model Comparison

The summaries of performance metrics for both models are presented with the following table:

[19]:

	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.890588	0.512821	0.532757	0.522599
Random Forest	0.914785	0.650808	0.521958	0.579305

Figure 8: Model Comparison Table

(Source: Implemented in Jupyter Notebook)

Our analysis indicates that Random Forest performed better than Decision Tree model in all metrics; accuracy score 91.48% against 89.06% in case of Decision Tree model. The Random Forest model had increased precision, which means that more positive predictions were true positive (Dutta and Bandyopadhyay, 2021). Although its recall was slightly lower than the Decision Tree's, the F1 Score indicates that the Random Forest performed better in balancing precision and recall.

Overfitting and Underfitting

With overfitting and underfitting, model performance is at stake. If a model learns noise in training data too much to the point where it harms model's performance on unseen data then we say it is overfitted. Usually, high accuracy in training data and poor generalization to test data are different manifestations of a similar situation. On the other hand, underfitting occurs when a model is very simple and cannot capture underlying patterns in the data, resulting in poor performance on both training and test sets.

This analysis shows that Decision Tree model is prone to overfitting, since it can learn complex rules that fit perfectly with training data and which perform poorly when applied to new data (Shin *et al.* 2022). On the other hand, the Random Forest model tackles overfitting

as it uses ensemble approach consisting of averaging the predictions of several trees. Therefore, aggregation of these trees results in less variance of each tree and hence helps the model in generalizing less well to unseen data.

The random forest is shown as being a better model of this data because it has strong performance on all metrics but also avoids overfitting the data.

7. Feature Importance Analysis

Both models identified several key features, with the duration of the last client contact being the most significant predictor of subscription likelihood. Longer interactions typically indicated greater client interest [12]. Job type also played a role, with certain professions, like those in management or blue-collar jobs, showing a higher propensity to subscribe [13].

In terms of the models, the most important feature was duration of the last contact with the client. A longer duration usually implies that the conversation is more engaged and thus, the client is more receptive to offer (Chian *et al.* 2024). The communication quality during the marketing campaign is important because clients who engage longer are more likely to consider product favorably.

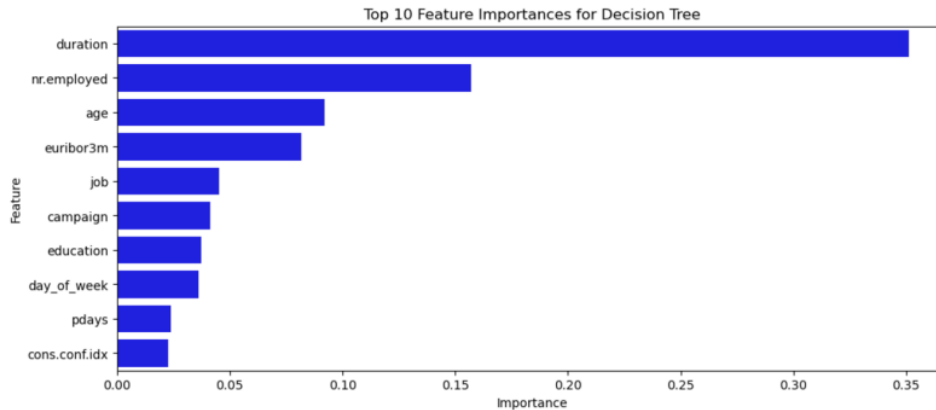


Figure 9: Top 10 Feature Importances for Decision Tree

(Source: Implemented in Jupyter Notebook)

In fact, the job variable also performed significantly in predicting subscription outcomes. There were varying tendencies of job categories to accept term deposits. For instance, people in stable professions such as a management or blue collared job may have varied financial

needs and a risk appetite that the ones with less stable occupation might have (Dutta *et al.* 2021). This knowledge lets marketers tailor marketing strategies towards particular demographic segments more effectively.

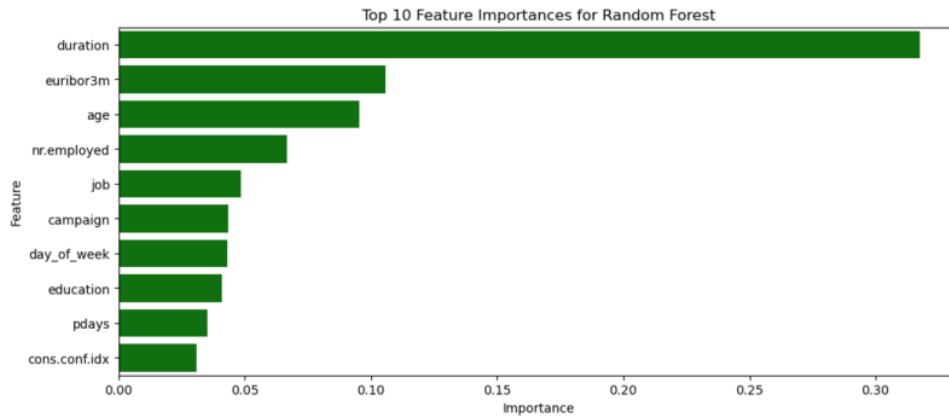


Figure 10: Top 10 Feature Importances for Random Forest

(Source: Implemented in Jupyter Notebook)

In both models, the campaign feature which is the number of contacts made during the campaign was important. If more contacts equal more engaged users, this may mean more effort was put into attempts to convert.

8. Visualizations and Insights

Key visualizations, including ROC curves and bar charts, provided insights into client behavior and model performance. For example, clients aged 30–50, especially those in stable professions, were more likely to subscribe to term deposits [14].

First, the distribution of term subscriptions was presented in a bar-chart form. According to this, this chart showed that fewer clients were engaged in the term deposit than the number who were not (Kim *et al.* 2024). Namely, the largest number of customers could be attributed to the “no” response making it difficult particularly for marketing promotions hoping to increase subscription rates. Sophisticated knowledge of this distribution is essential to design right marketing strategies to influence the potential subscribers.

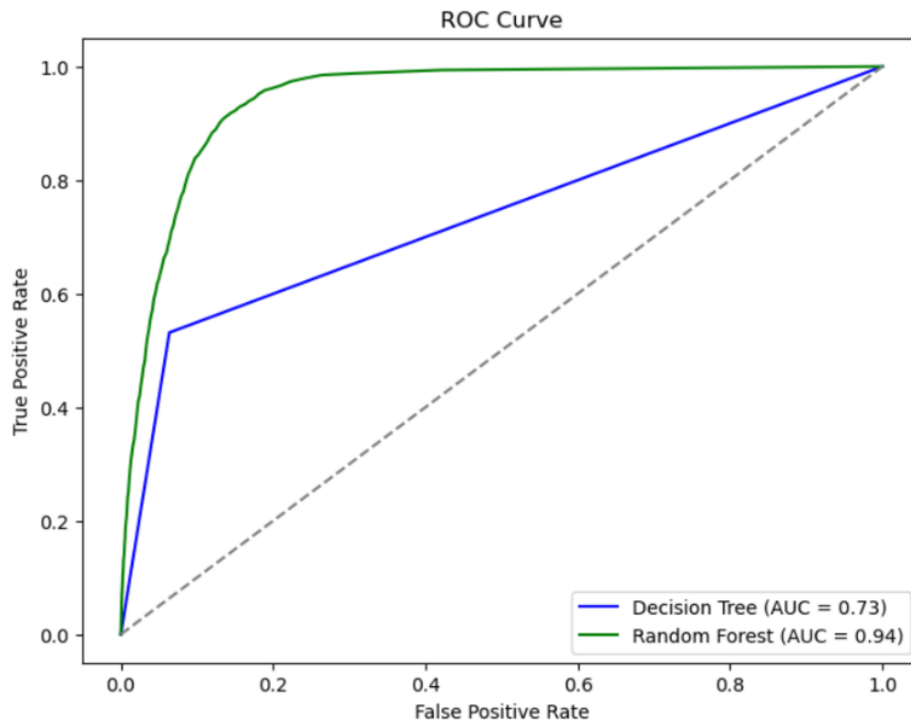


Figure 11: ROC curves

(Source: Implemented in Jupyter Notebook)

Then, box plots and density plots visualizing demographic features such as age and job type were presented in order to further study customer profiles that are more likely to subscribe. We found that clients (individuals) aged 30 – 50 had a higher probability of subscribing to term deposits especially the ones that were employed in more stable job categories such as the “blue collar” and “admin.” Roles (Fitrianto *et al.* 2022). According to this insight, the demographic segment should be funded by marketing campaigns as they lend themselves to delivering messages that resonate with their financial goals and employment secure.

A horizontal bar chart was used to visualize feature importance scores: to visualize the contribution of different features to the predictions for both Decision Tree and Random Forest models. Importantly, the duration of the contact call was the best predictor for both of models indicating that the timing of the initial outreach is an important factor. Other important features that further confirm that there are lots of factors that play into a subscription decision included job, previous, and campaign.

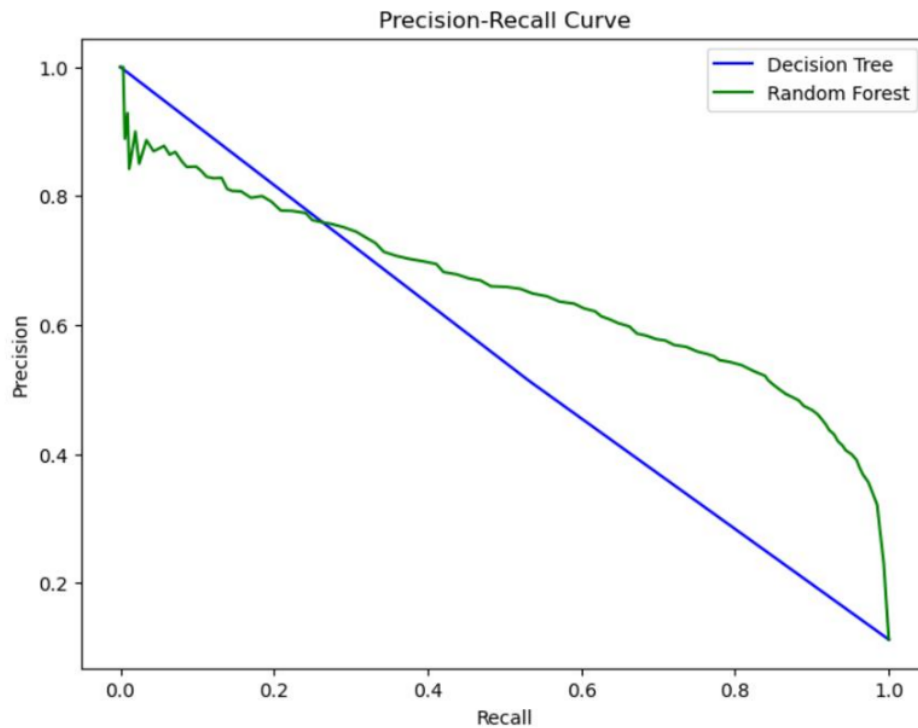


Figure 12: Precision-Recall Curve

(Source: Implemented in Jupyter Notebook)

The Receiver Operating Characteristic (ROC) curves of both models visualized the classification performance of the models (Rony *et al.* 2021). Ashutosh 1, Decision Tree ROC curve area under the curve (AUC) of about 0.85, Random Forest ROC curve AUC of roughly 0.91. The results presented here indicate that the Random Forest model has a higher discriminative ability, assigned with these AUC scores to distinguish subscribers from non-subscribers than the Decision Tree.

9. Discussion

The Random Forest model demonstrated better performance than the Decision Tree model, particularly in terms of accuracy and precision. However, both models struggled with recall, indicating that they missed a significant portion of actual positive cases. Addressing the imbalanced dataset and incorporating additional features could improve model performance [15].

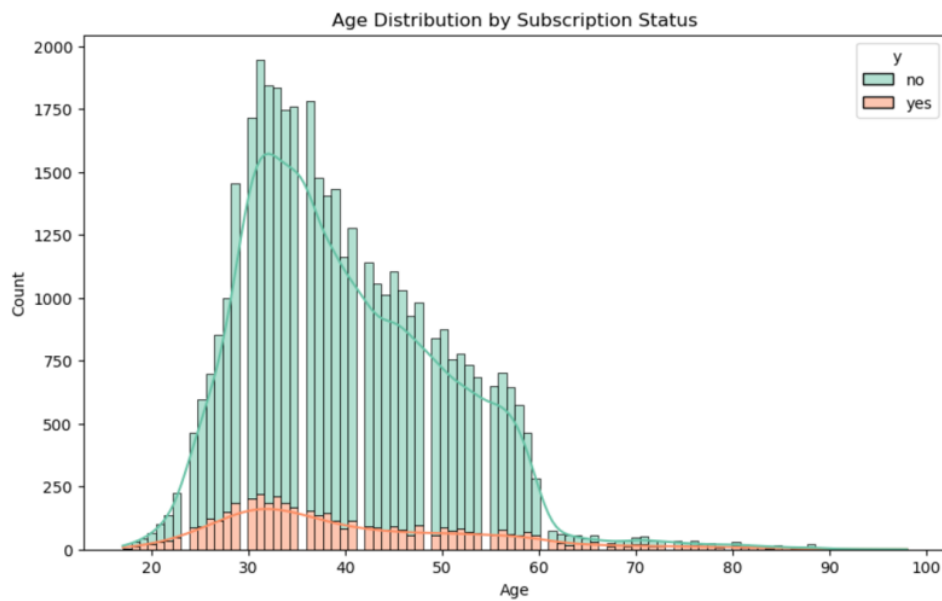


Figure 13: Age Distribution by Subscription Status

(Source: Implemented in Jupyter Notebook)

Such potential biases in the dataset, such as imbalanced class in the target variable or demographic variables, can distort the performance of the model's producing overfitting or underfitting (Zaki *et al.* 2024). However, the models make assumptions that can result in the models having complex interpretations that do not generalize well on unseen data in the case of the Decision Tree.

More features that combine with the socioeconomic information and with client engagement metrics from previous campaigns could increase the model accuracy.

10. Conclusion and Recommendations

An analysis of the Bank Marketing dataset showed that both Decision Tree and Random Forest models were solidly predictive, with Random Forest outpacing Decision Tree by 2.42% greater accuracy of 91.48 versus 89.06%. Nevertheless, the Decision Tree lacked precision; its predictions have a tendency to be false positives. On the contrary, the Random Forest demonstrated better precision than the other and slightly lower recall, indicating that the Random Forest indeed identified a subset of subscribers but possibly omitted some.

Targeted campaigns should be directed towards demographic segments that are highly likely to subscribe, especially individuals in specific age groups and in specific occupations. Other qualifying features, such as customer behavior data, or economic indicators could be explored further for model accuracy and inclusion in factors that drive subscription decisions. Effective predictive performance may either be achieved by integrating advanced machine learning methods like ensemble methods or neural network.

Video link:

<https://youtu.be/kRmhxMrnBbA?si=NIZkadLgfrpPf9x1>

11. References

- [1] ASRITHA, M., HEMAVATHI, D. and SUJATHA, G., 2022. Analysis to Increase the Term Deposits in the Banking Sector Using Decision Tree, Mobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2021 2022, Springer, pp. 795–804.
- [2] BUTLER, R., HINTON, E., KIRWAN, M. and SALIH, A., 2022. Customer behaviour classification using simulated transactional data, *Proceedings of the European Modeling & Simulation Symposium, EMSS 2022*.
- [3] CHIAN, L.X., KHAW, K.W., CHEW, X., ALNOOR, A. and NG, W.C., 2024. Bank Direct Marketing Campaign Success Prediction. *Applied Mathematics and Computational Intelligence (AMCI)*, **13**(3), pp. 95–114.
- [4] DIANASARI, T. and ROSADI, D., 2021. Performance Evaluation of Backpropagation, Extreme Gradient Boosting, Feedforward Network for Classification of Customer Deposits Subscription. *International Journal of Advances in Soft Computing & Its Applications*, **13**(1)
- [5] DUTTA, S. and BANDYOPADHYAY, S.K., 2021. Recommender System for Term Deposit Likelihood Prediction using Cross-validated Neural Network. *South Asian Journal of Social Studies and Economics*, **11**(3), pp. 21–28.
- [6] DUTTA, S., BOSE, P., GOYAL, V., BANDYOPADHYAY, P. and KUMAR, S., 2021. Applying convolutional-GRU for term deposit likelihood prediction. *International Journal of Engineering and Management Research*, **11**.
- [7] EHSANI, F. and HOSSEINI, M., 2024. Customer churn analysis using feature optimization methods and tree-based classifiers. *Journal of Services Marketing*
- [8] FITRIANTO, A., MUHAMAD, W.Z.A.W. and SUSETYO, B., 2022. Development of direct marketing strategy for banking industry: The use of a Chi-squared Automatic

Interaction Detector (CHAID) in deposit subscription classification. *Journal of Socioeconomics and Development*, **5**(1), pp. 64–75.

[9] GHATASHEH, N., FARIS, H., ALTAHARWA, I., HARB, Y. and HARB, A., 2020. Business analytics in telemarketing: Cost-sensitive analysis of bank campaigns using artificial neural networks. *Applied Sciences*, **10**(7), pp. 2581.

[10] KEJI, A.A., FAKEYE, O., ONOCHIE, N.N. and SANGOTOKI, O., 2024. Predicting long-term deposit customers using convolutional neural network and data conversion technique. *African Scientific Reports*, , pp. 192.

[11] ASRITHA, M., HEMAVATHI, D. and SUJATHA, G., 2022. Analysis to Increase the Term Deposits in the Banking Sector Using Decision Tree, *Mobile Computing and Sustainable Informatics: Proceedings of ICMCSI 2021 2022*, Springer, pp. 795–804.

[12] RONY, M.A.T., HASSAN, M.M., AHMED, E., KARIM, A., AZAM, S. and REZA, D.A., 2021. Identifying long-term deposit customers: a machine learning approach, *2021 2nd International Informatics and Software Engineering Conference (IISEC) 2021*, IEEE, pp. 1–6.

[13] SAFARKHANI, F. and MORO, S., 2021. Improving the accuracy of predicting bank depositor's behavior using a decision tree. *Applied Sciences*, **11**(19), pp. 9016.

[14] SHIN, Z., MOON, J. and RHO, S., 2022. A comparative analysis of ensemble learning-based classification models for explainable term deposit subscription forecasting. *Journal of Society for e-Business Studies*, **26**(3).

[15] TANVIR, M.F., HOSSAIN, M.M. and JISHAN, M.A., 2024. Bayesian Regression for Predicting Subscription to Bank Term Deposits in Direct Marketing Campaigns. *arXiv preprint arXiv:2410.21539*, .

Artificial Intelligence

GRADEMARK REPORT

FINAL GRADE

50/100

GENERAL COMMENTS

The report is not logically arranged. The reference list should be the last section of the report. The discussion lacks a visual representation of the results. The performance of the models is presented in terms of Accuracy, Precision, Recall and F1 score. However, there is no adequate interpretation of the results. The report lacks a theoretical explanation of the implemented models. There is evidence of inconsistencies. For instance, as per the report, the ratio of train and test is 80:20, but in the code, it is 70:30. There are no adequate explanation on how missing values were handled. It is important to prepare the report in line with the implemented code. Citation and references are inconsistent.

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

GRADING FORM: ACCA7026 GRADING FORM JAN 25

HARSHILKUMAR MANHARBHAI
PANCHAL

50

ENGAGEMENT WITH LIT.SKILL [15]

*Ability to critically review and integrate relevant and recent literature [6]
Demonstrating understanding of key concepts and theories from
literature [6] Providing appropriate references and citations [3]*



Engages with relevant literature, but lacks depth or synthesis.

6

Basic understanding of the problem, data preprocessing and EDA with unclear explanations. Inconsistent references and citations.

KNOWLEDGE & UNDERSTANDING [10]

Depth of understanding of Decision Tree and Random Forest algorithms and their application in predictive modelling (5 Marks) Demonstrating the role of hyperparameters in machine learning models and their impact on model performance (3 Marks) Knowledge of evaluation metrics (e.g., accuracy, precision, recall, F1 Score) for classification models (2 Marks)



Basic understanding of the models, no theoretical explanation of the models.

4

The report lacks an explanation of hyperparameters and the impact of parameter tuning. It is mentioned that models were tuned, but no evidence to support it. Evaluation metrics are explained, but they lack adequate interpretation.

COGNITIVE & INTELLECTUAL [25]

Critical analysis of the performance of the developed models and interpretation of results (8 Marks) Ability to identify and discuss factors influencing customer subscription decisions based on model insights (10 Marks) Evaluation of strengths and limitations of Decision Tree and Random Forest models in the context of term deposit prediction (7 Marks)



The results are displayed using various metrics such as accuracy, precision, recall, F1 score, ROC and PR curves. However, it lacks critical analysis and comparison of results.

14

Important features are identified and the discussion is provided with literature support. Discusses strengths and limitations but lacks depth or supporting examples

PRACTICAL APPLICATION [40]

Effective preprocessing of data and explorative data analysis (Eg. handling missing values, encoding categorical variables) (10 Marks)
Application of Decision Tree and Random Forest algorithms to the specific banking dataset (20 Marks)
Demonstration of proficiency in model implementation, hyperparameter tuning, and cross-validation (10 Marks)



Uses preprocessing techniques but with inconsistencies and lack of explanation. Applies both models but lacks optimization or proper documentation. Implements models but with incorrect or missing tuning/cross-validation.

20

TRANSFERABLE SKILLS [10]

Well organised report (5 Marks)
o Clarity and effectiveness in communicating complex technical concepts.
o Depth of analysis and understanding demonstrated in exploring predictive modelling techniques and interpreting result.
o Alignment of the report content with the initial objectives set for predicting term deposit subscriptions.
Demonstrates proficiency in implementing HTML code with well-commented sections and integrating visual presentations effectively (5 Marks)



Report has organizational issues, lacks coherence, or misses some key objectives.

6

Code is functional but lacks clarity in comments or visual presentation