A Machine Learning Approach to Identify Customers for Subscribing Term Deposit

Sunny Guha[1], Moloy Dhar[2], Bidyutmala Saha[3], Nirupam Saha[4]

[1] Department of Computer Science and Engineering, Guru Nanak Institute of Technology
Kolkata, India

[2,3,4] Department of Computer Science and Engineering, Guru Nanak Institute of Technology
Kolkata, India

Abstract. The motivation for this study stems from the increasing importance of targeted marketing campaigns in banking and the potential of machine learning (ML) to revolutionize this domain. The need for banks to efficiently identify potential customers for term deposits, thereby optimizing their marketing efforts and resources, is emphasized. Our paper discusses the evolving role of ML in the banking sector, particularly in marketing strategies. The problem statement identifies the challenge of accurately predicting customer behaviour in the context of bank term deposits. The statement highlights how traditional marketing approaches may fall short and posits that a ML approach could offer more accurate predictions and insights. It also touches on the need for more comprehensive studies integrating various ML algorithms for this specific application in the banking sector. The objectives include developing a ML model to predict the likelihood of customers subscribing to term deposits, comparing the effectiveness of algorithms like Support Vector Machines (SVM), Random Forest, Decision Tree, Regression, and KNN, and evaluating the model's performance using the datasets from the Portuguese banking institution. Another objective is to contribute to the body of knowledge in applying ML to banking marketing strategies. Here, we delve into the specific ML algorithms pertinent to customer identification in banking and discuss various algorithms' strengths, weaknesses, and applicability in banking.

Keywords: ML, banking , SVM, fraud detection, risk assessment, supervised learning.

1 Introduction

The introductory section sets the stage for the paper by presenting the context and significance of the research. It aims to capture the reader's interest and provide a clear roadmap for the study. This section outlines the broader context of the research. It starts by discussing the evolving role of ML in the banking sector, particularly in marketing strategies. The motivation for this paper stems from the increasing importance of targeted marketing campaigns in banking and the potential of ML to revolutionise this domain. The need for banks to efficiently identify potential customers for term deposits, thereby optimising their marketing efforts and resources, is emphasised. The background sets the stage for why this research is timely and significant. The problem statement concisely describes the specific issue addressed by the thesis. It identifies the challenge of accurately predicting customer behaviour in the context of bank term deposits. The statement highlights how traditional marketing approaches may fall short and posits that ML approach could offer more accurate predictions and insights. It also touches on the need for more comprehensive studies integrating various ML algorithms for this specific application in the banking sector. This part outlines the primary goals of the research. The objectives include developing a ML model to predict the likelihood of customers subscribing to term deposits, comparing the effectiveness of algorithms like SVM, Random Forest, Decision Tree, Regression, and KNN, and evaluating the model's performance using the datasets from the Portuguese banking institution. Another objective is to contribute to the body of knowledge in applying ML to banking marketing strategies. The scope defines the boundaries of the research. It clarifies that the study focuses on using specific ML algorithms to analyse data from a Portuguese bank's direct marketing campaigns. The limitations section addresses potential constraints, such as the applicability of the results to other banking institutions or regions and the reliance on the datasets accuracy and representativeness. It may also discuss the computational limitations in implementing and testing the algorithms. This final section of the introduction provides an overview of the paper. It briefly describes each chapter: the literature review covers existing research in the field; the methodology explains the data collection, pre-processing, and selection of ML algorithms; the implementation chapter discusses the application of these algorithms; the results and discussion chapter interprets the findings; and the conclusion sums up the study with future research directions. The thesis ends with references and appendices, including code listings and additional data tables.

2 Literature Review

This section provides a comprehensive overview of how ML technologies are applied in the banking sector. It explores various use cases, such as fraud detection, risk assessment, customer segmentation, and personalised banking services. The discussion highlights how ML has transformed traditional banking practices, emphasising efficiency, accuracy, and improvements in customer experience. Include references to seminal works and recent studies to illustrate the evolution and current state of ML in banking. This subsection reviews literature specific to customer identification processes in banking using ML. This includes studies on customer due diligence,

KYC (Know Your Customer) procedures, and anti-money laundering efforts. Our paper focus on improving customer identification processes and any limitations or challenges they have encountered. Here, we delve into the specific ML algorithms pertinent to customer identification in banking and discuss various algorithms' strengths, weaknesses, and applicability in banking. These might include supervised learning algorithms like SVM or Random Forests, unsupervised learning techniques like K-means clustering, or advanced deep learning models. The aim is to provide a thorough understanding of how these algorithms work and why they are suitable (or not) for customer identification tasks in banking. The final subsection should synthesise the earlier information to highlight significant contributions made by existing literature in the ML field in banking, particularly in customer identification. Equally important is to identify and discuss gaps or areas that are under-explored. This could include limitations in current methodologies, challenges in data privacy, or the need for more robust algorithms. Identifying these gaps not only sets the stage for our research objectives but also underscores the relevance and necessity of our study. Emphasise [1,4] using predictive analytics to forecast bank term deposit subscriptions, showcasing how these insights drive marketing strategies, where [2,3] offer comparative analyses of ML techniques to identify potential long-term deposit customers , enhance targeting accuracy and enhancing Model Effectiveness and Interpretability. [16] Explore improvements in bank telemarketing success through predictive and interpretability analyses to enhance model transparency. [6] Focus on explainable ML models to improve banking processes for predicting potential term deposit customers. [7] Discuss the adaptation of ML in different regional contexts, such as Portuguese banks and Tunisia, highlighting cultural and economic impacts on model performance. [8,12] Present case studies on forecasting and predicting the success of banking telemarketing campaigns, respectively. [5,15] Analyse the application of specific algorithms, such as decision trees and ensemble learning, in bank telemarketing. [14] Explore advanced techniques such as bagging and LightGBM combined with SMOTE, enhancing the prediction of potential customers. [18,19] Provide insights from practical implementations of ML in bank marketing, addressing operational challenges and solutions. [11,15] Examine transaction-based predictive models and assess ML performance on banking term deposits.

3 Methodology

The data collection phase is crucial in ML research, as the quality and nature of the data significantly influence the study's outcomes. For this paper, the data was gathered from https://archive.ics.uci.edu/dataset/222/bank+marketing . This section has described the criteria for data selection, ensuring that the data is relevant, comprehensive, and representative of the study's context. It also details the dataset's size, format, and variety, along with any ethical considerations or permissions required. Data pre-processing is critical in preparing the dataset for practical ML analysis. This section outlines the various pre-processing methods employed, such as cleaning (removing noise and correcting inconsistencies), normalisation (scaling features to a range), and transformation (converting data into a suitable format). This part describes rationale behind choosing specific pre-processing techniques and how they are expected to improve the dataset's quality for the subsequent analysis.

3.1 Model Training

3.1.1 Data Cleaning and Preparation

This section details the practical steps to clean and prepare the data for analysis. It describes the initial state of the dataset, including any inconsistencies, missing values, or irrelevant features. Then, outline the data cleaning process, which may involve techniques like imputation for missing data, outlier removal, and feature selection or extraction. Also, detail the data preparation steps, such as normalisation or encoding categorical variables, ensuring the data is in a suitable format for the ML algorithms. This section should be detailed enough to allow others to replicate the process. This section explains the steps taken to optimise and tune the models. This involves:

- Parameter Tuning: Detail how we adjusted the parameters of each algorithm to improve performance. This might include grid search, random search, or other optimisation techniques.
- Feature Selection and Engineering: Discuss any techniques for selecting the most relevant features or engineering new features that could improve model performance.
- Cross-Validation: it explains the cross-validation method used to assess the models' performance and ensure they are not over fitting.
- Optimisation Metrics: State the metrics used to guide the optimisation process, such as accuracy, precision, recall, F1-score, etc., and explain why they were chosen.

3.1.2 **System Design via ER Diagram**

The system design is illustrated through an ER diagram that encapsulates the database schema, highlighting the relationships between different data entities crucial to understanding the data flow and storage mechanism used in our paper. This visual representation (Fig. 1 & Fig. 2) aids in grasping the complex interactions between various data points and their roles in the predictive models.

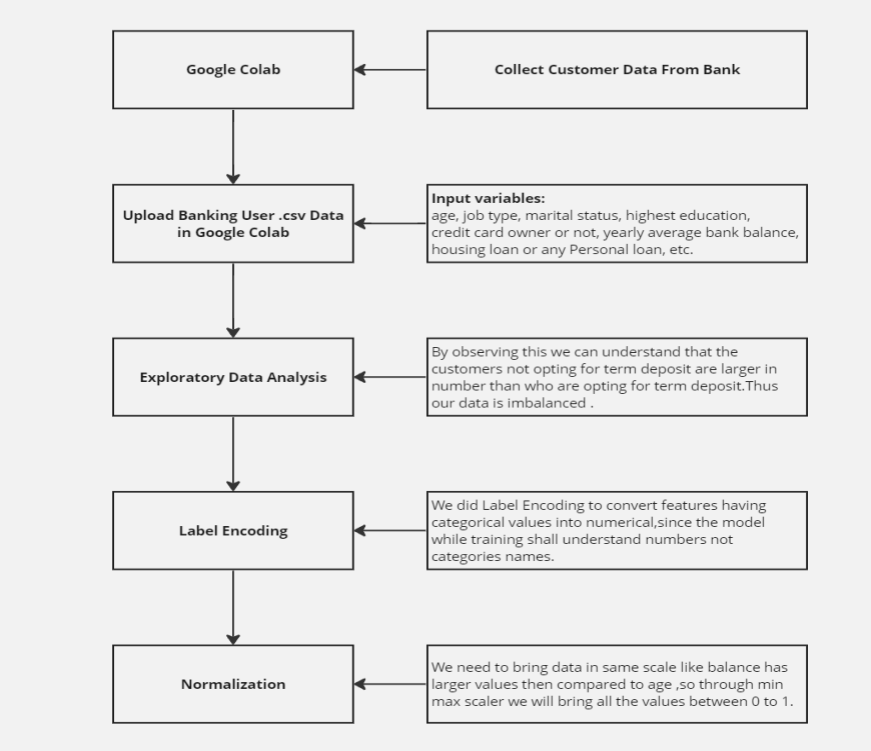

Fig. 1

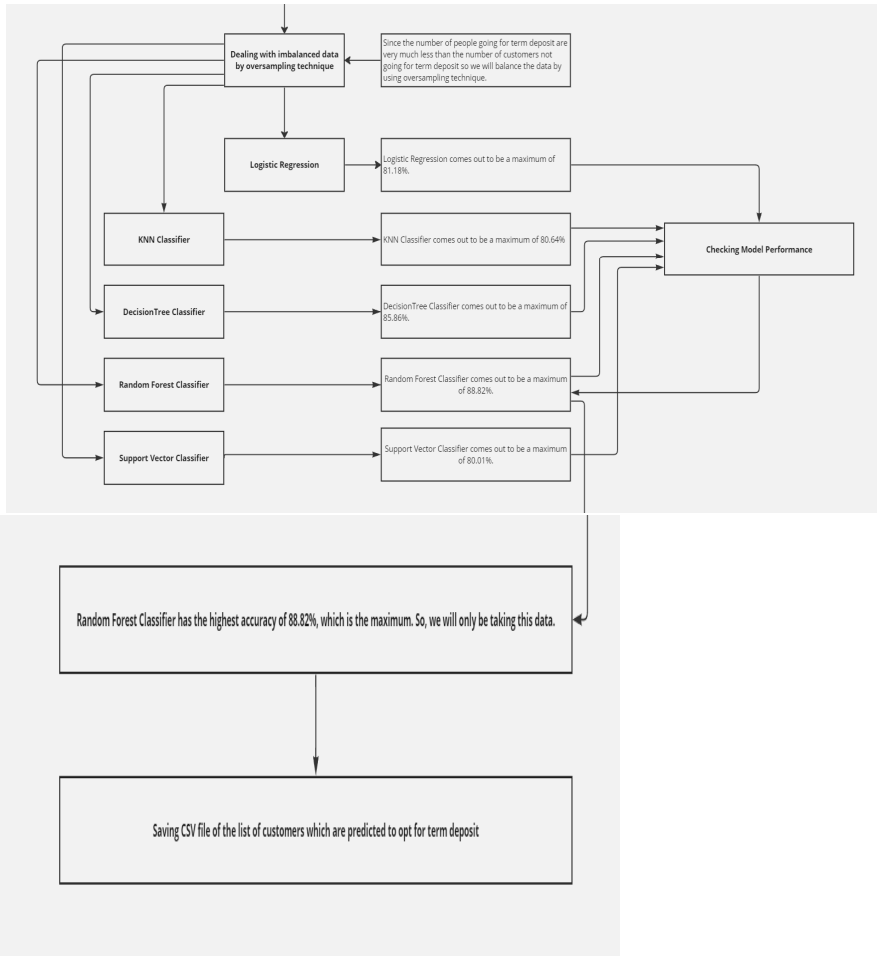

Fig. 2

3.1.3 Model Training

This section details training the ML models used in this study. It covers the data preparation steps, including cleaning and pre-processing, and describes algorithm training, optimisation, and tuning methodologies to ensure optimal model performance.

4 Results

Evaluating ML models is essential to determine their effectiveness and accuracy. This section defines the metrics and criteria used to assess the performance of the ML algorithms. Standard metrics include accuracy, precision, recall, and F1 score. We must understand why these metrics are suitable for our research and how they will be used to assess and compare the performance of different models. Also, we will discuss any statistical methods or validation techniques (like cross-validation) employed to ensure the reliability and generalizability of the results. This methodology section outlines the systematic approach which is consider in our paper, ensuring the study is structured, transparent, and replicable. This section should present a detailed analysis of the performance of each ML model used in the study. Include a variety of performance metrics, such as accuracy, precision, recall, F1-score, and any others relevant to your research. For each model, provide visual representations like confusion matrices or other pertinent plots that help interpret the results. We will discuss the performance of each model in the context of the specific problem we are addressing, highlighting any unique strengths or weaknesses observedinthemodels.
We conduct a comparative analysis of the algorithms (Fig. 3- Fig. 9) based on the performance metrics and visualisations presented earlier. This comparison should focus on which algorithm performed best and also on why certain algorithms performed better or worse in this specific context. Consider factors like algorithm complexity, the nature of the data, and each algorithm's suitability for the problem at hand. This comparison helps us understand each algorithm's nuances and practical applicability in real-world scenarios.

Tabular Analysis

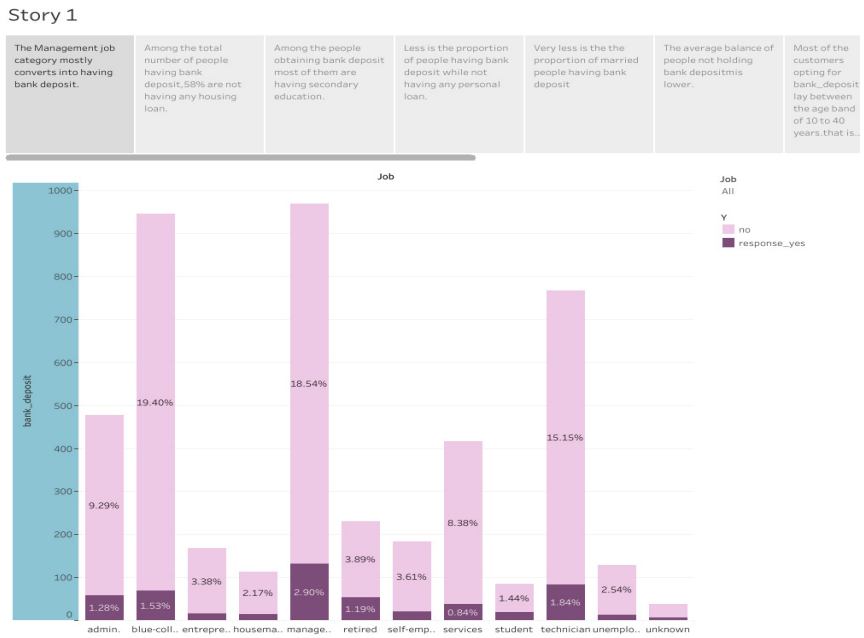
Story 1
The Management job category mostly converts into having bank deposit.
Among the total number of people having bank deposit,58% are not having any housing loan.
Among the people obtaining bank deposit most of them are having secondary education.
Less is the proportion of people having bank deposit while not having any personal loan.
Very less is the the proportion of married people having bank deposit
The average balance of people not holding bank depositmis lower.
Most of the customers opting for bank_deposit lay between the age band of 10 to 40 years.that is..
Job
bank_deposit
1000
900
800
700
600
500
400
300
200
100
0
9.29%
19.40%
3.38%
2.17%
18.54%
3.89%
3.61%
8.38%
1.44%
15.15%
2.54%
1.28%
1.53%
2.90%
1.19%
0.84%
1.84%
admin.
blue-coll..
entrepre..
housema..
manage..
retired
self-emp..
services
student
technician
unemplo..
unknown
Job
All
Y
no
response_yes

Fig. 3

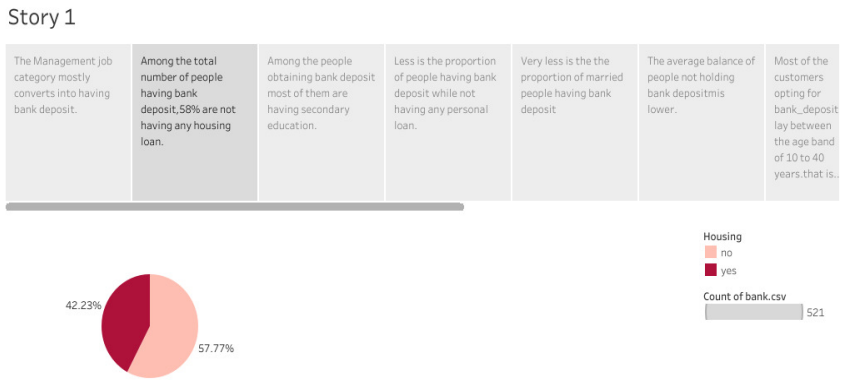
Story 1
The Management job category mostly converts into having bank deposit.
Among the total number of people having bank deposit,58% are not having any housing loan.
Among the people obtaining bank deposit most of them are having secondary education.
Less is the proportion of people having bank deposit while not having any personal loan.
Very less is the the proportion of married people having bank deposit
The average balance of people not holding bank depositmis lower.
Most of the customers opting for bank_deposit lay between the age band of 10 to 40 years.that is..
42.23%
57.77%
Housing
no
yes
Count of bank.csv
521

Fig. 4

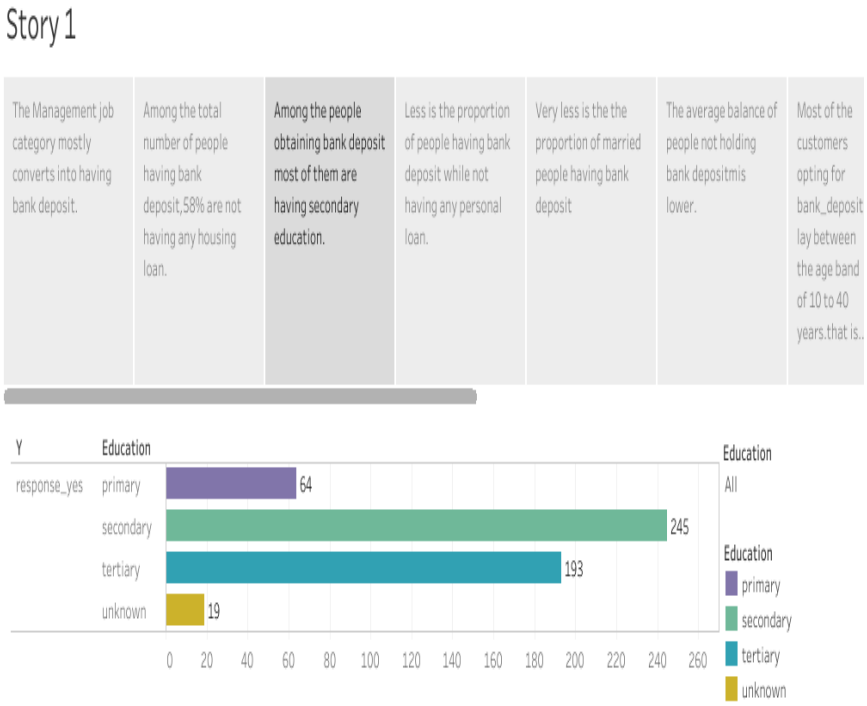
Story 1
The Management job category mostly converts into having bank deposit.
Among the total number of people having bank deposit,58% are not having any housing loan.
Among the people obtaining bank deposit most of them are having secondary education.
Less is the proportion of people having bank deposit while not having any personal loan.
Very less is the the proportion of married people having bank deposit
The average balance of people not holding bank depositmis lower.
Most of the customers opting for bank_deposit lay between the age band of 10 to 40 years.that is..
Y
Education
response_yes
primary
secondary
tertiary
unknown
64
245
193
19
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Education
All
Education
primary
secondary
tertiary
unknown

Fig. 5

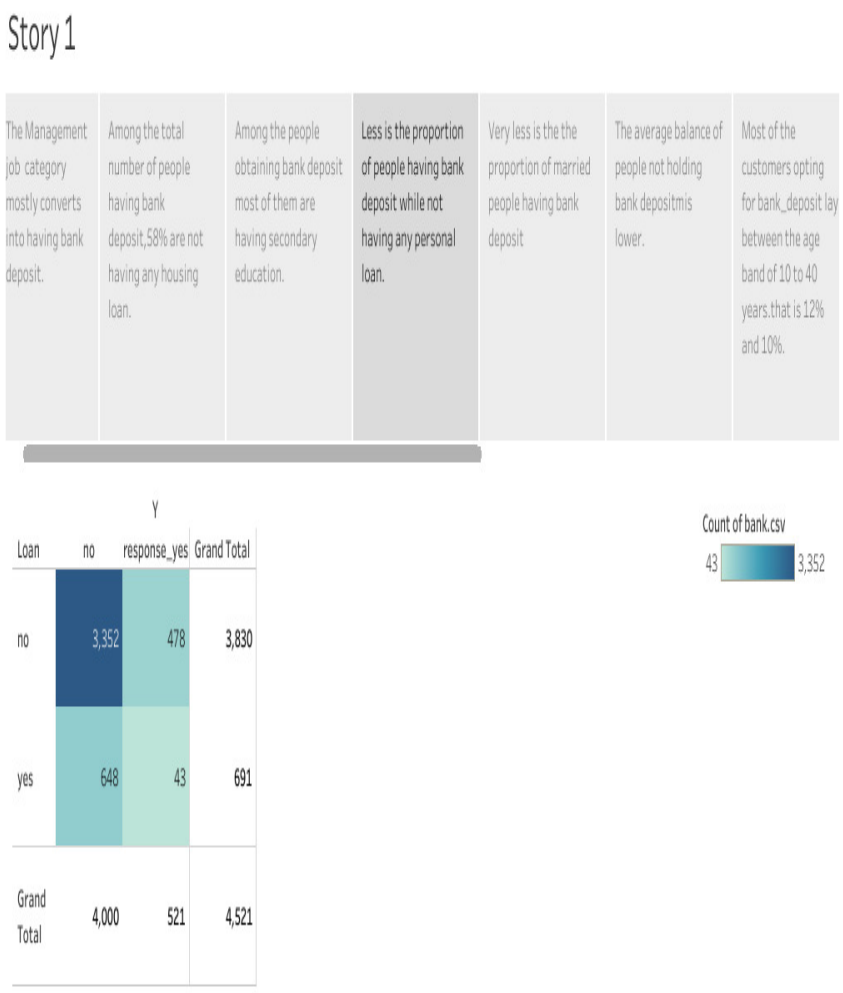
Story 1
The Management job category mostly converts into having bank deposit.
Among the total number of people having bank deposit,58% are not having any housing loan.
Among the people obtaining bank deposit most of them are having secondary education.
Less is the proportion of people having bank deposit while not having any personal loan.
Very less is the the proportion of married people having bank deposit
The average balance of people not holding bank depositmis lower.
Most of the customers opting for bank_deposit lay between the age band of 10 to 40 years.that is 12% and 10%.
Y
Loan
no
response_yes
Grand Total
no
3,352
478
3,830
yes
648
43
691
Grand Total
4,000
521
4,521
Count of bank.csv
43
3,352

Fig. 6

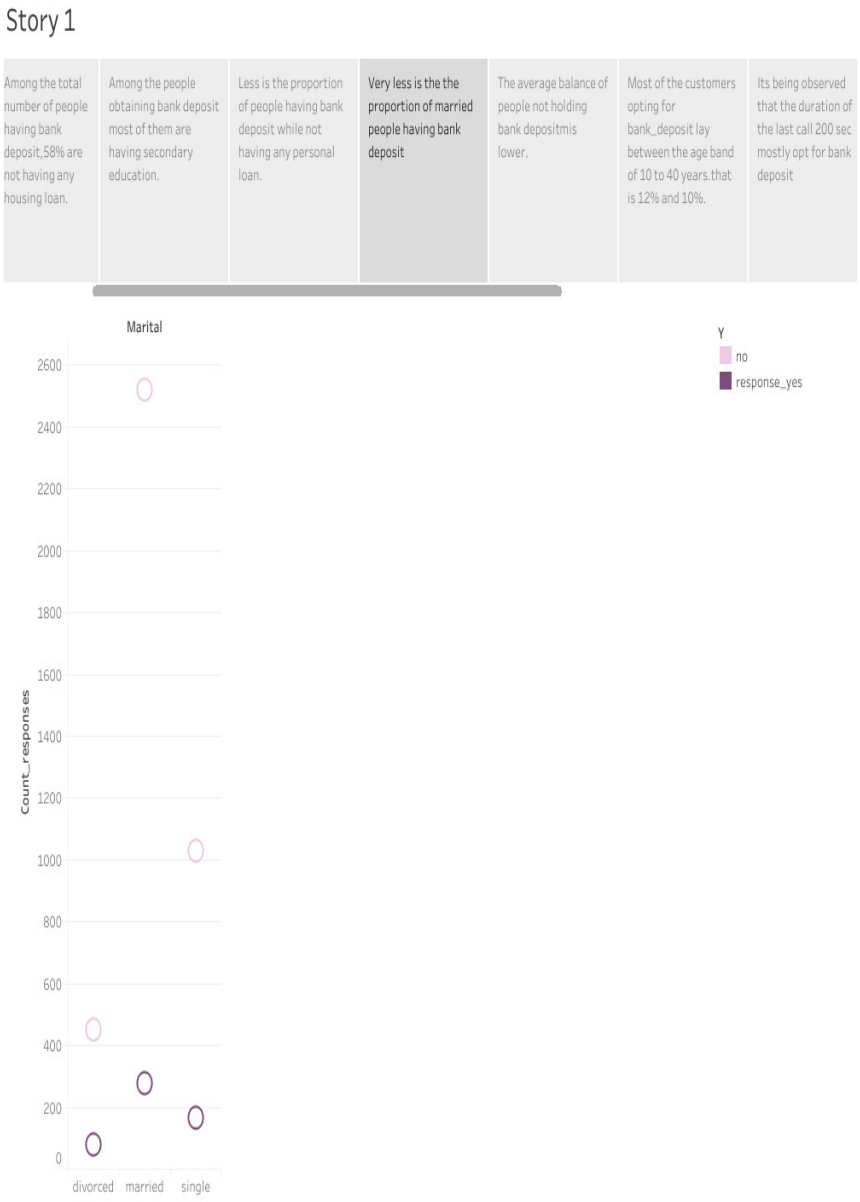
Story 1
Among the total number of people having bank deposit,58% are not having any housing loan.
Among the people obtaining bank deposit most of them are having secondary education.
Less is the proportion of people having bank deposit while not having any personal loan.
Very less is the the proportion of married people having bank deposit
The average balance of people not holding bank depositmis lower.
Most of the customers opting for bank_deposit lay between the age band of 10 to 40 years.that is 12% and 10%.
Its being observed that the duration of the last call 200 sec mostly opt for bank deposit
Marital
Y
no
response_yes
Count_responses
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
divorced
married
single

Fig. 7

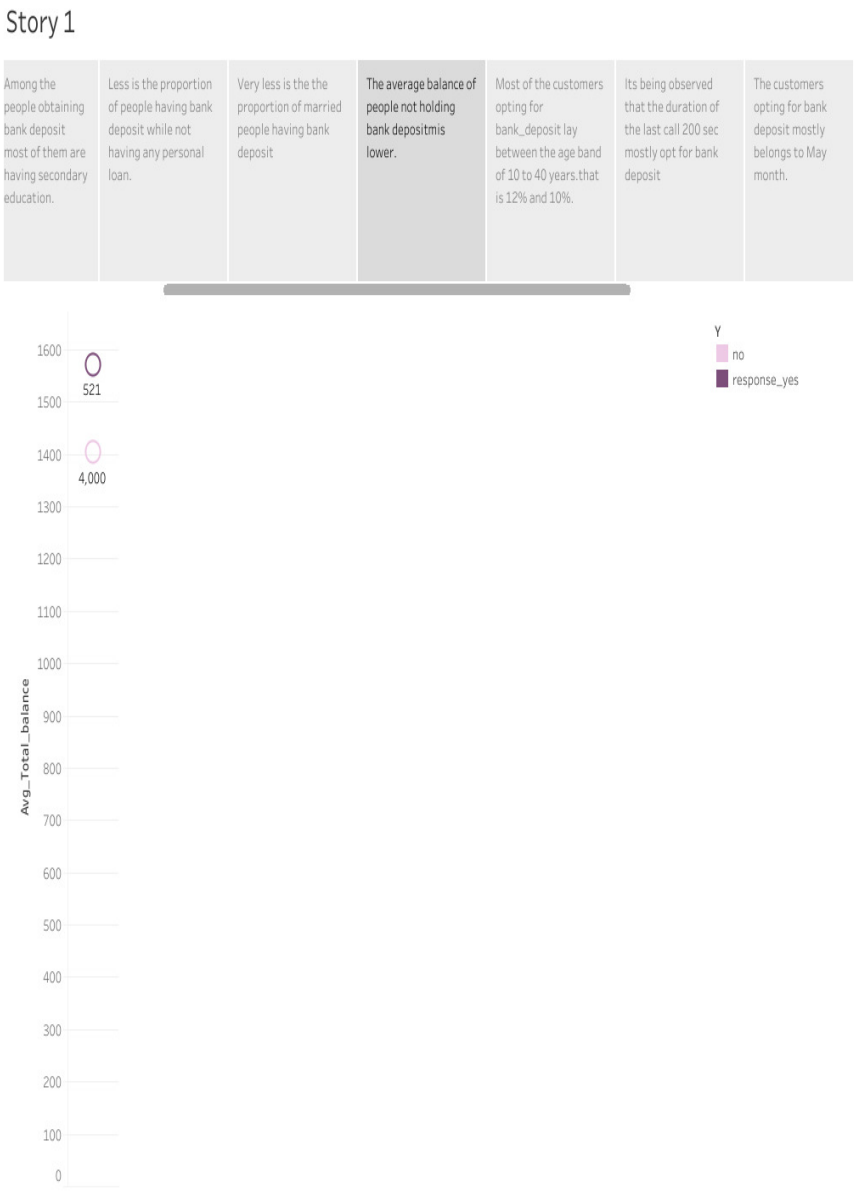

Fig. 8

Tabular Analysis: Final Dashboard

Fig. 9

5 Conclusion

Our paper adeptly illustrated the efficacy of ML algorithms in forecasting clients' propensity to subscribe to term deposits, utilising a dataset obtained from a Portuguese banking institution. The process entailed rigorous data collection, pre-processing, and the deployment of several algorithms, such as SVM, Logistic Regression, Decision Tree, Random Forest Classifier, and KNN. This comprehensive approach determined the most efficacious models and enriched the practical understanding of banking marketing strategies. Furthermore, this thesis generated a .csv output delineating data equal to 88% of the data designed for direct marketing applications by the bank. We can also use this data to show online remarketing ads via Facebook and Google Ads to these particular users. This strategic output optimises the bank's resource allocation and operational efficiency, enhancing engagement rates.

References

1. Zaki, A. M., Khodadadi, N., Lim, W. H., & Towfek, S. K. (2024): Predictive Analytics and ML in Direct Marketing for Anticipating Bank Term Deposit Subscriptions. American Journal of Business and Operations Research, 11(1), 79-88.
2. Rony, M. A. T., Hassan, M. M., Ahmed, E., Karim, A., Azam, S., & Reza, D. A. (2021, December): Identifying long-term deposit customers: a ML approach. In 2021 2nd International Informatics and Software EngineeringConference (IISEC) (pp.1-6).IEEE.

3. Singh, M., Dhanda, N., Farooqui, U. K., Gupta, K. K., & Verma, R. (2023, July). Prediction of Client Term Deposit Subscription Using ML Check for updates: In Proceedings of the 4th International Conference on Communication, Devices and Computing: ICCDC 2023 (Vol. 1046, p. 83). Springer Nature.

4. Hou, S., Cai, Z., Wu, J., Du, H., & Xie, P. (2022): Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription. International Journal of Business Analytics (IJBAN), 9(1), 1-14.

5. Borugadda, P., Nandru, P., & Madhavaiah, C. (2021): Predicting the success of bank telemarketing for selling long-term deposits: An application of Machine Learning algorithms. St. Theresa Journal of Humanities and Social Sciences,7(1),91-108.

6. Khan, M. Z., Munquad, S., & Rao, T. S. M. (2022, April): A Study on Improving Banking Process for Predicting Prospective Customers of Term Deposits using Explainable Machine Learning Models. In Proceeding of International Conference on Computational Science and Applications: ICCSA 2021 (pp. 93-103). Singapore: Springer Nature Singapore.

7. Alexandra, J., & Sinaga, K. P. (2021, October): Machine Learning approaches for marketing campaign in portuguese banks. In 2021 3rd International Conference On Cybernetics And Intelligent System (ICORIS) (pp. 1-6). IEEE.

8. Olajide Olajide, B., & Ishaku Wreford, A. (2023): Bank Term Deposit Service Patronage Forecasting using ML. Vallis Aurea (International Journal Vallis Aurea), 9(2), 53-64.

9. Tuan, N. M. (2022). ML Performance on Predicting Banking Term Deposit. In ICEIS (1) (pp. 267-272).

10.Vongchalerm, L. (2022): Analysis of predicting the success of the banking telemarketing campaigns by using Machine Learning techniques (Doctoral dissertation, Dublin, National College of Ireland).

11. Sonkar, N., Chaurasiya, R. K., Choubey, M., Prasad, R., & Baghel, R. K. (2024, February): Prediction of Potential Customers for Term Deposit: Analysis Using Bagging with Bank Marketing Data. In 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-6). IEEE.

12. Hayder, I. M., Al Ali, G. A. N., & Younis, H. A. (2023): Predicting reaction based on customer's transaction using Machine Learning approaches. International Journal of Electrical and Computer Engineering, 13(1), 1086.

13. Patwary, M. J., Akter, S., Alam, M. B., & Karim, A. R. (2021): Bank deposit prediction using ensemble learning.ArtificialIntelligenceEvolution,42-51.

14. Xie, C., Zhang, J. L., Zhu, Y., Xiong, B., & Wang, G. J. (2023): How to improve the success of bank telemarketing, Prediction and interpretability analysis based on Machine Learning. Computers & Industrial Engineering, 175,108874.

15. Wang, D. (2020, October): Research on bank marketing behavior based on Machine Learning. In Proceedings of the 2nd international conference on artificial intelligence and advanced manufacture (pp. 150-154).

16. Li, Z., Xu, Z., & Zhou, Y. (2024): Application of Machine Learning on Client Prediction in Bank Marketing. In Economic Management and Big Data Application: Proceedings of the 3rd International Conference (pp. 1064-1076).

17. Gafrej, O. (2024): Predicting customer deposits with Machine Learning algorithms: evidence from Tunisia. Managerial Finance,50(3),578-589.

18. Saeed, S. E., Hammad, M., & Alqaddoumi, A. (2022, March): Predicting Customer's Subscription Response to Bank Telemarketing Campaign Based on Machine Learning Algorithms. In 2022 International Conference on Decision Aid Sciences and Applications (DASA) (pp.1474-1478). IEEE.

19. Dong, Y. (2024): Potential Customer Prediction of Telecom Marketing based on Machine Learning. Highlights in Science, Engineering and Technology, 92, 138-145.

20. Gupta, A., Raghav, A., & Srivastava, S. (2021, February): Comparative study of Machine Learning algorithms for Portuguese bank data. In 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 401-406). IEEE.