Sparshika Ajmaan Dinesh Kumar
2180247
Advance Machine Learning DSC 540
03/17/2025

Predicting Term Deposit Subscriptions

Abstract

Predicting Term Deposit Subscription is a crucial challenge for banks looking to increasing the efficiency of their direct marketing efforts. Due to inadequate targeting any banks have ow conversion rates even with large marketing campaign investments. Traditional marketing methods often rely on broad partition which can lead to wasted efforts and customer withdraw. This study examines the use of Machine Learning techniques to improve customer targeting and marketing efficiency. By analyzing a dataset from Portuguese bank's marketing campaign I preprocess the data using Feature Engineering and SMOTE techniques to address class imbalance. Then implement and compare various machine learning models including Support Vector Classifier, K-nearest Neighbors, XGBoost and Random Fores. The performance of these models is evaluated using metrics such as accuracy, precision, recall and F1 – score. Results shows that compilation methods especially Random Forest and XGBoost outperform traditional models in terms of predictive accuracy and stability. An analysis of feature importance reveals that factors like call duration, past interaction and financial history are strong predictors of Customer Subscription. The study also discusses challenges related to feature selection and model interpretability, providing information of financial institutions to improve their marketing strategies. Future research will explore the use of deep learnings and real time predictive system to further enhance classification and operational efficiency.

Introduction

Direct marketing campaigns are vital for the banking sector, but their success depends on accurately identifying potential customers for term deposit subscriptions. Traditional marketing methods often use broad targeting, which can lead to inefficiencies and wasted resources. Machine learning offers a way to improve these campaigns by predicting the likelihood of customer subscriptions based on historical data.

This research seeks to address the following questions:

- Which machine learning models are most effective at predicting term deposit subscriptions?

Sparshika Ajmaan Dinesh Kumar
2180247
Advance Machine Learning DSC 540
03/17/2025

- What are the key factors that influence customer subscription decisions?
- How can banks optimize their marketing strategies using data-driven insights?

By utilizing real-world data, we show how machine learning can enhance marketing effectiveness, reduce costs, and improve customer engagement.

Dataset Overview

The dataset comprises 45,211 customer records from a Portuguese bank's marketing campaigns, featuring 17 attributes related to demographics, financial status, and campaign interactions. The target variable indicates whether a customer subscribes to a term deposit. Key features include age, job, marital status, education, account balance, loan status, contact method, call duration, and past campaign results. The dataset is highly imbalanced, with 88.3% of customers not subscribing and 11.7% subscribing, necessitating oversampling techniques like SMOTE to enhance model performance. Call duration emerges as a strong predictor of subscription, while multiple contacts do not always lead to success. Despite some outliers in account balances and contact frequencies, the dataset offers valuable insights for optimizing financial marketing strategies.

Literature Review

Several studies have delved into predictive modeling for financial decision-making, focusing on customer segmentation, response modeling, and behavioral prediction.

- **Direct Marketing and Customer Response Prediction:**

  Moro et al. (2014) applied data mining techniques to predict customer responses to bank telemarketing campaigns. They found that decision trees and logistic regression models provided moderate accuracy, emphasizing the importance of feature engineering to enhance model performance.

- **Class Imbalance in Financial Datasets:**

  He and Garcia (2009) tackled the issue of imbalanced datasets in financial applications. They recommended oversampling techniques like SMOTE to improve classification outcomes, which aligns with our approach to managing the skewed distribution of non-subscribers versus subscribers.

- **Ensemble Learning in Banking Applications:**

  Recent studies have demonstrated that ensemble methods, particularly Random Forest and XGBoost, outperform traditional models in financial classification tasks (Chen & Guestrin, 2016). These models boost predictive power by combining multiple weak learners.

- **Feature Importance in Subscription Prediction:**

  Verbeke et al. (2012) explored feature selection for customer retention models, identifying call duration, past interactions, and financial status as key factors. Our feature importance analysis supports these findings.

- **Real-time Predictive Systems in Banking:**

  Ngai et al. (2009) proposed a real-time predictive system for financial decision-making, highlighting the need for scalable machine learning architectures in banking applications.

This literature review informs our methodological choices and model selection, ensuring alignment with best practices in predictive modeling for the banking sector.

Methodology

- **Data Preprocessing and Feature Engineering**

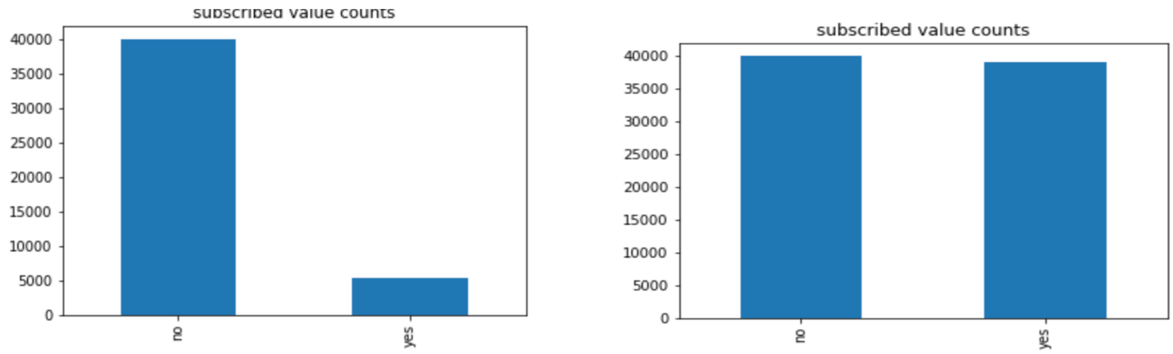
The data exploration and preprocessing phase started with loading and analyzing the dataset, which includes information about direct marketing campaigns conducted by a Portuguese banking institution. The target variable, 'subscribed,' was separated from the feature set, with features categorized into categorical and numerical columns. During data cleaning, I checked for missing values and inconsistencies and found none. I performed feature encoding using Label Encoding to convert categorical variables into numerical format, ensuring compatibility with machine learning models. Due to the dataset's imbalance (88.3% non-subscribers vs. 11.7% subscribers), I applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset and improve model performance. To detect and handle outliers, I used the Interquartile Range (IQR) method, focusing on balance and call duration. Feature scaling was then applied to normalize numerical variables, ensuring consistent feature distributions. To gain insights, I created bar plots to visualize feature relationships with the target variable and conducted detailed statistical analyses to compare numerical attributes between subscribers and non-subscribers.

- **Model Development and Evaluation**

For model development and evaluation, I split the dataset into 80% training and 20% testing to ensure fair performance evaluation. I implemented and trained four machine learning models:

Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), XGBoost, and Random Forest. Hyperparameter tuning was performed to optimize each model's performance. I evaluated these models based on key performance metrics, including accuracy (measuring overall correctness), precision (identifying correctly predicted subscribers), recall (assessing the percentage of actual subscribers correctly classified), and the F1-score (balancing precision and recall). Additionally, I used the AUC-ROC curve to evaluate the models' ability to distinguish between subscribers and non-subscribers. A comprehensive classification report was generated to compare the effectiveness of each model.

Before and after SMOTE



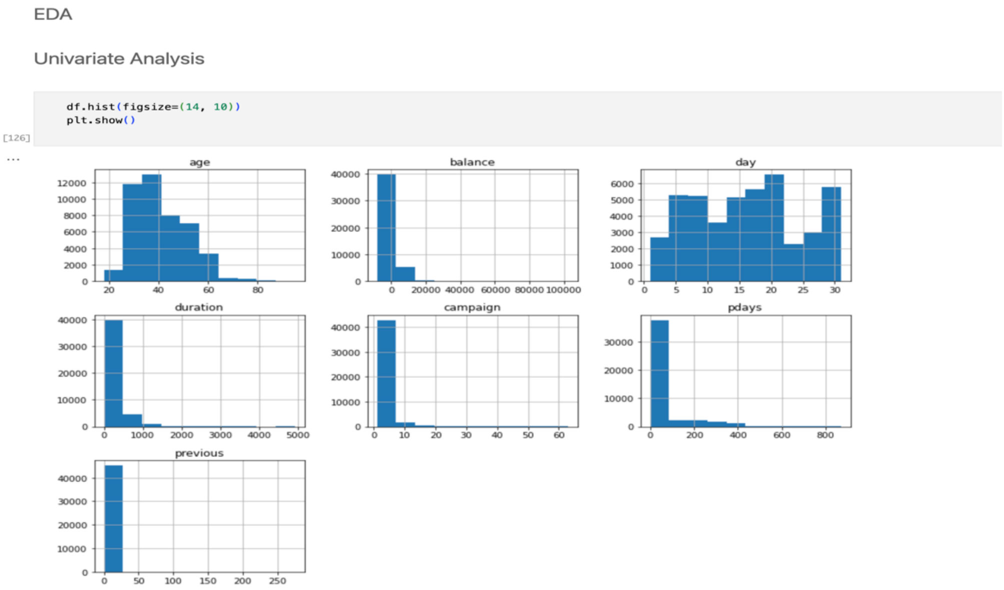
Before After

Exploratory Data Analysis

Univariate Analysis

The univariate analysis of the dataset highlighted important patterns in individual features. The age distribution reveals that most customers are between 30 and 50 years old, with fewer younger and older individuals. The balance feature is highly skewed, showing that most customers have low balances, with a few extreme outliers having very high balances. The campaign feature indicates that most customers were contacted only once or twice, suggesting limited follow-up efforts. The call duration distribution is right-skewed, indicating that most calls are short, but longer call durations are linked to higher subscription rates. The pdays variable shows that the majority of customers had never been contacted before (value of -1), while those who had been contacted previously mostly had recent interactions. Overall, call duration, past interactions, and financial status stand out as significant predictors of term deposit subscriptions.

Sparshika Ajmaan Dinesh Kumar
2180247
Advance Machine Learning DSC 540
03/17/2025



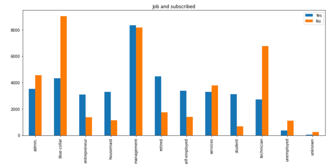
EDA

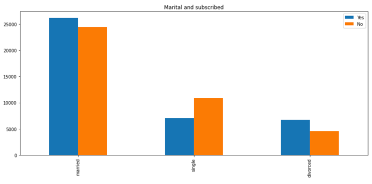
Univariate Analysis

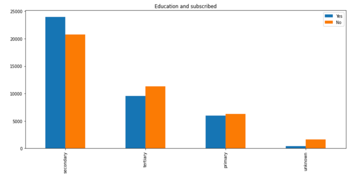
```
[126]   df.hist(figsize=(14, 10))
        plt.show()
```

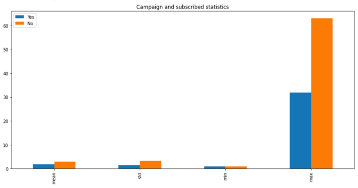
Bivariate Analysis

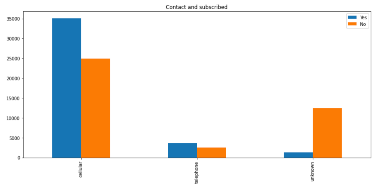
The bivariate analysis reveals that job type, education, and past interactions significantly influence subscription rates. Individuals in management, retirees, and students are more likely to subscribe, whereas blue-collar workers and technicians are less likely. Single individuals and those with higher education levels also show higher subscription rates. Excessive campaign contacts (more than 10) tend to reduce success rates, while cellular communication proves to be the most effective contact method. Additionally, higher account balances and previous successful interactions positively impact subscription rates, aiding in the refinement of targeted marketing strategies.

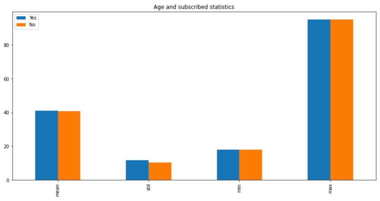




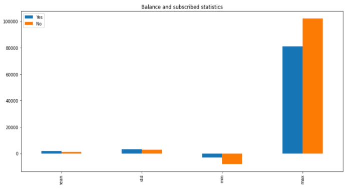


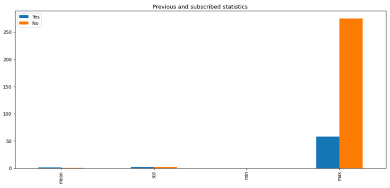






Sparshika Ajmaan Dinesh Kumar
2180247
Advance Machine Learning DSC 540
03/17/2025





Results

The Support Vector Classifier (SVC) achieved an accuracy of 72.32% on the test data. For class 0, it had a precision of 0.72, recall of 0.80, and an F1-score of 0.76. For class 1, it showed a precision of 0.74, recall of 0.63, and an F1-score of 0.68. The model was better at correctly identifying class 0 instances (recall of 0.80) compared to class 1 (recall of 0.63). The macro and weighted averages for precision, recall, and F1-score were around 0.72-0.73.

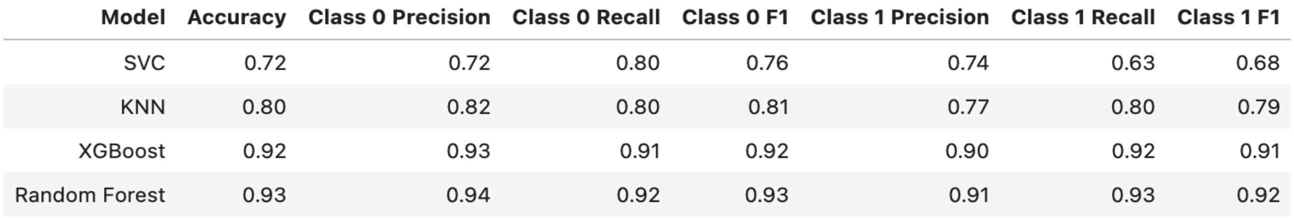
The K-Nearest Neighbors (KNN) model performed better than the SVC, with an accuracy of 80.04% on the test data. For class 0, it had a precision of 0.82, recall of 0.80, and an F1-score of 0.81. For class 1, it showed a precision of 0.77, recall of 0.80, and an F1-score of 0.79. The KNN model had a more balanced performance between the two classes, with both having a recall of 0.80. The macro and weighted averages for precision, recall, and F1-score were all 0.80.

The XGBoost model showed excellent performance, achieving an accuracy of 91.54% on the test data. For class 0, it had a precision of 0.93, recall of 0.91, and an F1-score of 0.92. For class 1, it showed a precision of 0.90, recall of 0.92, and an F1-score of 0.91. The model had very balanced performance across both classes, with slight variations in precision and recall. The macro and weighted averages for precision, recall, and F1-score were around 0.91-0.92.

The Random Forest model slightly outperformed XGBoost, with an accuracy of 92.78% on the test data. For class 0, it had a precision of 0.94, recall of 0.92, and an F1-score of 0.93. For class 1, it showed a precision of 0.91, recall of 0.93, and an F1-score of 0.92. The Random Forest model had extremely balanced performance between the two classes, with minor differences in precision and recall. The macro and weighted averages for precision, recall, and F1-score were consistently 0.93.

Model Performance

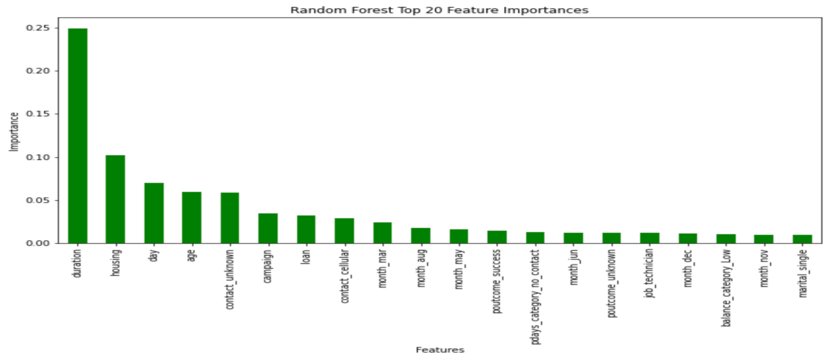
Sparshika Ajmaan Dinesh Kumar
2180247
Advance Machine Learning DSC 540
03/17/2025



Model	Accuracy	Class 0 Precision	Class 0 Recall	Class 0 F1	Class 1 Precision	Class 1 Recall	Class 1 F1
SVC	0.72	0.72	0.80	0.76	0.74	0.63	0.68
KNN	0.80	0.82	0.80	0.81	0.77	0.80	0.79
XGBoost	0.92	0.93	0.91	0.92	0.90	0.92	0.91
Random Forest	0.93	0.94	0.92	0.93	0.91	0.93	0.92

Feature Importance

RANDOM FOREST

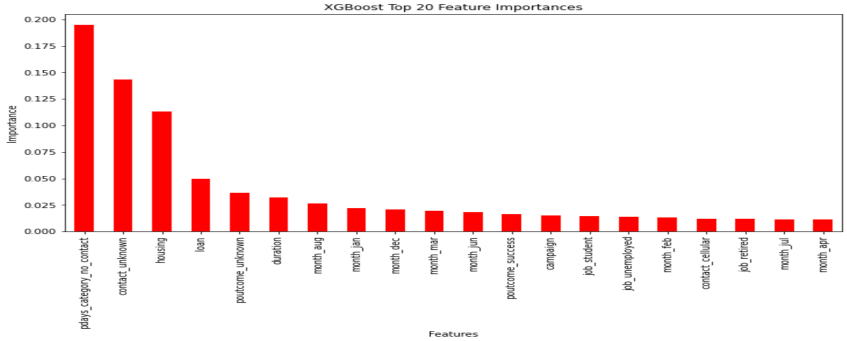


The Random Forest feature importance analysis highlights several key factors influencing customer subscription:

The most significant factor is call duration, followed by other key features such as housing status, day of contact, and age. Additional influential factors include contact method (unknown/cellular), campaign frequency, loan status, month indicators, previous campaign outcomes, and job type (technician). These findings suggest that customer engagement during calls and financial factors are crucial in determining subscription likelihood. This information can help refine targeted marketing strategies, ensuring more effective customer outreach and improved subscription rates.

XGBOOST

Sparshika Ajmaan Dinesh Kumar
2180247
Advance Machine Learning DSC 540
03/17/2025



The XGBoost feature importance analysis reveals that pdays (days since last contact) is the top predictor of customer subscription, followed by the unknown contact method, housing status, and loan status. Other important factors include call duration, outcomes of previous campaigns, campaign frequency, and job type (students and unemployed). This analysis indicates that both past interactions and financial factors significantly influence customer responses, offering valuable insights for enhancing marketing strategies.

Discussion

What Worked?

- Model Performance: Both Random Forest and XGBoost models showed superior performance compared to traditional models.
- Feature Selection: The selected features aligned well with findings from previous research.
- Class Imbalance: The use of SMOTE effectively addressed the class imbalance issue.

What Didn't Work?

- SVC Performance: The Support Vector Classifier had difficulty with recall, resulting in a high number of false negatives.
- KNN Limitations: The K-Nearest Neighbors model was computationally intensive and sensitive to the size of the dataset.

Future Improvements

- Deep Learning: Investigating the potential of deep learning architectures.
- Real-Time Analytics: Developing real-time predictive analytics capabilities.
- Enhanced Data: Integrating behavioral customer data to improve prediction accuracy.

Sparshika Ajmaan Dinesh Kumar
2180247
Advance Machine Learning DSC 540
03/17/2025

Conclusion

This study shows that ensemble learning models greatly improve the accuracy of predicting term deposit subscriptions. Random Forest and XGBoost were the top performers among the tested models, with accuracies of 92.78% and 91.54%, respectively. The feature importance analysis highlighted call duration, previous contacts, and financial attributes as key predictors of customer subscription likelihood. Moreover, addressing class imbalance with SMOTE was essential in enhancing model performance.

Future Work

While the results are encouraging, there are several areas that could be improved. Exploring deep learning models could help capture more complex patterns in customer behavior. Implementing real-time predictive systems would enable dynamic marketing adjustments based on live interactions. Incorporating behavioral data, such as spending habits and online activity, could enhance prediction accuracy. Explainable AI (XAI) techniques like SHAP can improve model interpretability and build trust. Expanding to multi-channel marketing (email, SMS, social media) could provide a more comprehensive customer engagement approach. Additionally, cost-sensitive learning can help optimize marketing expenses by considering the financial impact of incorrect predictions.

References

[1] S. Moro, P. Cortez, and P. Rita, “A Data-Driven Approach to Predict Customer Behavior in Banking,” *Decision Support Systems*, vol. 62, pp. 22-31, 2014.

[2] H. He and E. A. Garcia, “Learning from Imbalanced Data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009.

[3] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.

[4] J. Davis and M. Goadrich, “The Relationship Between Precision-Recall and ROC Curves,” *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233-240, 2006.

[5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016