Optimizing Bank Marketing Strategies through Predictive Analytics

Venkata Ram Pranith Garlapati (vg28691n@pace.edu)
CS 668 - Analytics Capstone Project - Fall 2024
Professor: *Dr. Christelle Scharff*
Pace University, Seidenberg School of Computer Science & Information Systems

GitHub

Abstract:

The project deals with optimizing bank marketing strategies from predictive analytics and customer segmentation. Employing machine learning models, this study identifies crucial factors affecting customers' willingness to subscribe to term deposits and increases targeting efficiency in telemarketing campaigns. Ensemble models, including **XGBoost** and **Random Forest**, were applied to predict customer behavior with good performance. Model outputs were interpreted by means of a **SHAP** analysis, with insightful recommendations to marketing managers. Results show that call duration, contact type, and previous campaign outcomes have substantial impacts on conversion rates. The study illustrates how advanced predictive models can improve marketing efficiencies, shave costs, and help in increasing customer conversions.

Research Questions:

1) What are the most influential factors in predicting customer conversion during bank telemarketing campaigns?
2) Which machine learning models are most effective in forecasting customer subscription to term deposits?
3) How can customer segmentation be optimized to increase conversion rates and reduce marketing costs?
4) What role does model interpretability play in improving marketing decision-making and strategy formulation?
5) Can predictive modeling help identify and retain high-value customers by understanding their behavior patterns?

Related Works:

Moro and others (2014) analyzed customer lifetime value via a neural network modeling process to predict bank deposit subscription, firmly establishing the utility of machine-learning models for campaign optimization.
The authors, namely Rogić et al. (2022), handled class imbalance with the help of balanced support vector machines (B-SVM) thereby gaining advantageous effects in the campaigns with low response rates.
Singh et al. (2024) used machine learning models to predict customer churn, establishing that ensemble methods like XGBoost outperform baseline models on financial datasets.
Peng et al. (2023) underscored the interpretability of the model by SHAP analysis, meaning insights on feature importance can greatly facilitate strategic decision-making.

Dataset:

Data for the project were derived from the **Bank Marketing Dataset** available from the UCI Machine Learning Repository. These data originate from telemarketing campaigns to promote term deposits. The dataset contains **41,188** records and **17** features which are basically demographic, economic, and campaign-related. The input features included demographic variables such as **age**, **job type**, **marital status**, and **education level**, as well as **economic attributes** such as **average balance**, **default statu**s, housing loan, and **personal loan**. Many campaign features such as the last contact's duration, contact communication type (telephone or cellular), and days since carrying out the last contact (**pdays**), as well as previous campaign outcome existence. The target variable, **y**, indicates whether or not a customer subscribes to a term deposit ("Yes" or "No").The most significant challenge facing the dataset was imbalance; the bulk of customers did not subscribe to the term deposit, thus affecting model performance.

Methodology:

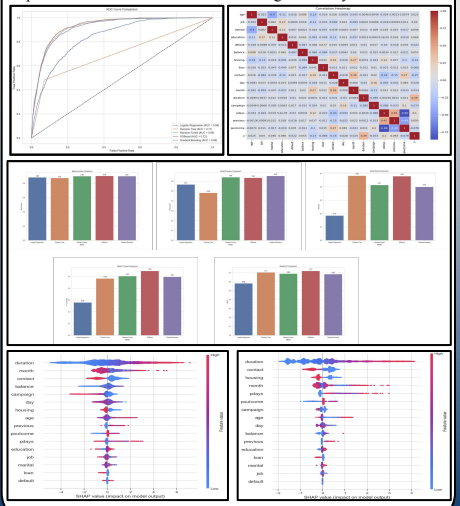
The study went a long way towards establishing the procedures to preprocess data, build models, and evaluate them. The missing values were sort of attended to, with categorical features such as job and marital status encoded, and some numerical variables like age and balance normalized for a balance of scales. The Synthetic Minority Over Sampling Technique **(SMOTE)** was applied to combat the imbalance, ensuring positive and negative classes had some balance going on. During exploratory data analysis **(EDA)**, major relationships and trends were visualized, and features deemed important were identified, namely call duration, type of contact, and outcome of the previous campaign. Features, especially interaction features created by combining job type and education, were driven to strengthen the accuracy of the model.

A baseline model using such standards as **logistic regression** and Decision Tree are used as benchmarks with **Random Forest**, **Gradient Boosting**, and **XGBoost** as more advanced ensemble models. Classifying the model in terms of **performance**, **precision**, **recall**, **F1-score**, **accuracy**, and **AUC** (Area Under the Curve) map onto an assumed cause of common variance; and k-Fold Cross-Validation is applied to ascertain this process of assurance and generalization. Analysis based on **SHAP** was introduced for model interpretability wherein individual feature contributions to its predictions could be gleaned and impart fruitful interpretations.

Results:

This study shed light on crucial aspects of predicting customer conversion and optimizing bank marketing strategies. Key features influencing customer subscription identified from the study are call duration, contact type, previous campaign outcome, and pdays. Among the various models tried, **XGBoost** emerged on top with an **AUC** of **0.72** and an **F1-Score** of **0.55**, second best to **Random Forest** and **Gradient Boosting** at **AUC = 0.69**. The baseline models, such as **Logistic Regression**, did quite poorly at an **AUC** of **0.58**.

SHAP analysis shows that the call duration impact on subscription predictions is the highest: longer calls mean positive conversion. **Confusion Matrix:** The **XGBoost** model is reducing false negatives, meaning the subscribers were correctly predicted using this algorithm. The following insights come from these results: targeting customers who had favorable past interactions can enhance marketing efficiency.



Conclusion:

It optimizes bank marketing strategies with machine learning models, finding the best performance with **XGBoost**. The AUC for **XGBoost** is **0.72**, while its **F1-Score** is **0.55**. The most important predictors in determining customer subscription identified during feature engineering include call duration, contact type, and previous campaign outcome. **SHAP** analysis thus puts forth that targeting customers who have longer interactions and better histories of campaigns would definitely bring higher conversion rates and reduce the costs involved in marketing.

Future Work:

In future works, more features will be integrated into the model, such as customer transaction data and satisfaction scores, to enhance the accuracy of prediction. Class imbalance can also be further addressed with techniques like SMOTE, while better performance can be achieved with hyperparameter optimization on robust computation platforms. Finally, the integration of the predictive models into CRM systems will allow real-time analytics for dynamic campaign adjustments and enable data-driven decision-making.

References:

1. Moro, S., Cortez, P., & Rita, P. (2014). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing and Applications*, 26(1), 131–139. https://doi.org/10.1007/s00521-014-1703-0
2. Moro, S., Rita, P., & Cortez, P. (2014). Bank Marketing [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5K306.
3. Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. *Data Science and Management*, 7(1), 7–16. https://doi.org/10.1016/j.dsm.2023.09.002
4. Rogić, S., Kašćelan, L., & Pejić Bach, M. (2022). Customer response model in direct marketing: Solving the problem of unbalanced dataset with a balanced support vector machine. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(3), 1003-1018. https://doi.org/10.3390/jtaer17030051
5. Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. *PLOS ONE*, 18(12). https://doi.org/10.1371/journal.pone.0289724