Factor characterization for term deposit through marketing campaign

ISM 6136 Data Mining Project Report

APRIL 21, 2024

Vladomir Jungkman, Gang Wei, Ann Nyawira, Tasnim Shairy

Factor characterization for term deposit through marketing campaign

Main Objective: Predict customers' response to bank's telemarketing campaign and identify which customers' type is more likely to make term deposits.

Introduction

This project predicts customer responses to bank telemarketing campaigns using machine learning models. The analysis identifies key factors influencing term deposit subscriptions, enabling banks to optimize their marketing strategies and improve campaign effectiveness.
Countries like Portugal, where bank telemarketing is a popular method for attracting new customers and staying competitive in the market. The banks focus on targeting a potential customer base to grow their banks, particularly for products like term deposits, to maximize profits.
Banks reach out directly to potential customers through phone calls with an aim of analyzing customer behavior, building connections, and earning long-term trust to encourage current and future investments. Effective telemarketing strategies are crucial for achieving several key objectives:

1. Customer Acquisition: Bringing in new customers is vital for the growth and success of any financial institution.
2. Personal Engagement: Building personal connections with customers fosters trust and loyalty, leading to stronger relationships.
3. Cost-Effectiveness: Using resources efficiently ensures that the bank's efforts are yielding the best possible results.
4. Feedback and Marketing Insights: Understanding customer feedback and market trends allows for continuous improvement and better decision-making.
5. Relationship Building: Nurturing relationships with customers are essential for long-term success and loyalty.

This traditional method of bank telemarketing, that the Bank of Portugal is currently implementing, has its challenges. The most common issue is reaching potential customers who may not be available or willing to engage with the telemarketing campaign. This can result in a low response rate and wasted resources. Additionally, telemarketers often struggle to understand the factors that contribute to the success of their campaigns.
To address these challenges, we propose a solution that leverages advanced data analytics and machine learning models. By analyzing customer preferences, behaviors, and likelihood to respond to campaigns, banks can better target their outreach efforts. This data-driven approach allows for more effective telemarketing strategies that focus on individuals who are most likely to be interested in joining the bank or investing in term deposits.

Data Dictionary

Data Source: https://archive.ics.uci.edu/dataset/222/bank+marketing from UC Irvine Machine Learning Repository

Data Attributes: Table of 45,211 instances and 17 features

Updated Input variables:

Input variables:

1. Age(int): How old person is.
2. Job(categorical): Type of job, categories are either admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown.
3. Martial (Categorical): Person's marriage status categories. Either divorced, married, single, or unknown.
4. Education (Categorical): Person's education categories. Either basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, or unknown.
5. Default (Binary): Whether a person has gone 90 days without paying scheduled payment for credit. 0-no 1-yes.
6. Balance (Int): Average yearly balance of a person in euros.
7. Housing (Binary): Whether a person has a housing loan or not. 0-no 1-yes.
8. Loan (Binary): Whether a person has a personal loan or not. 0-no 1-yes.
9. Contact (Categorical): Persons contact method categories. Either cellular or telephone.
10. Day_of_week(int): last contact day of the person through marketing campaign.
11. Month(string): last contact month of the person through marketing campaign.
12. Duration (int): Duration of the last contact with the person through a marketing campaign.
13. Campaign (int): Number of marketing campaigns that this person has been included in.
14. Pdays (int): Number of days between the last contact day and the day they were contacted.
15. Previous (int): Number of contacts before this campaign and for this client.
16. Poutcome (categorical): Previous outcome categories. Either failure, nonexistent, or success.

Output variables:

1. Results(string): Whether the person will subscribe to making a term deposit with this bank.

Data Updates

Based on the data definition from the proposal, transformations of the data had to be made. To ensure that one hot encoding would not skew the data to be unpredictable at all, there had to be grouping of certain attributes within the raw data. An example of this would be the season in which the client was given the call from the bank campaign. Originally in the raw data file, season was not an attribute, but instead it was the month. We grouped 12 months into 4 seasons, there was skewness in the data that was reduced. This was applied to other columns as well. Most of the character related attributes were grouped in this manner to ensure that one hot encoder's variance would not be an issue for the models. The other columns grouped and transformed in this manner were the marital status, job, and education as well as season as stated before. With this stated, when using the getting dummies function to our categorical variables, it expanded the number of columns from 17 to 31 which will impact the classification methods used for this project.

Results and Model Interpretation

Based on the preprocessed dataset we have, we implemented seven different supervised machine learning algorithms to find the best model for the bank's phone promotion as below:

1. Decision Tree: a non-parametric algorithm for the classification task of the bank. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes, and leaf nodes.
2. KNN (k nearest numbers): a non-parametric algorithm to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.
3. Gradient Boosting: a machine learning ensemble technique that combines the predictions of multiple weak learners, typically decision trees, sequentially.
4. Random Forest: another commonly used machine learning algorithm that combines the output of multiple decision trees to reach a single result.
5. Logistic Regression: a classic machine learning algorithm widely used for binary classification tasks, such as identifying whether the bank's customer would accept or decline the promotion.
6. Support Vector Machines (SVM): are supervised max-margin models with associated learning algorithms that analyze data for classification and regression analysis.
7. Naive Bayes: is a generative learning algorithm which could be used to model the distribution of inputs of a given class or category.

Model Name	True Positives	Recall Score
Naïve Bayes	912	57.76%
Gradient Boost	619	39.20%
Random Forest	566	35.85%
Decision Tree	564	35.72%
Logistic Regression	287	18.18%
Knn	232	14.69%
SVM	11	.70%

As the table shown above, the models are ordered descending by True Positive numbers and the corresponding recall scores. The True Positive number means the accurately predicted customers who really take the promotion, and the recall score is the ratio of True Positive divided by the total actual customers willing to take the promotion.

In the phone promotion campaign, the bigger True positive, the higher the recall score, the better, it means more deposits and less missed real customers to the bank.

There are 912 customers predicted accepting the offer would accept it in fact with Naive Bayes model in our testing database. The recall score is 57.76%, which means out of the total customers willing to take the promotion, 57.76% of them are predicted correctly. Compared to the industry average, Naive Bayes is outstanding, brings the bank high recall score, large true positive number, and gains powerful competitive advantages in the market.

Key Findings

1. Training Time: The time it takes to train a model can be critical, especially if the bank needs to iterate quickly or deploy the model frequently. Naive Bayes typically trains very quickly compared to more complex models like Gradient Boosting or Support Vector Machines.
2. Resources: Naive Bayes is generally less resource-intensive compared to models like Gradient Boosting and SVMs that require more resources for training and inference, especially with large datasets.
3. Data Ambiguity: Naive Bayes assumes independence between features, which might not hold true in all cases. Models like Decision Trees are prone to overfitting if the data contains noise or irrelevant features.

Implications

Choosing Naive Bayes as the best model for predicting customers' responses to the bank's telemarketing campaign has several implications for the telemarketing strategy:

1. Targeted Outreach: With this model, the bank can easily identify which customers are more likely to respond positively to the telemarketing campaign hence allowing for more targeted outreach efforts, focusing resources on individuals with a higher probability of conversion.

2. Personalized Messaging: By understanding the factors influencing customers' responses, the bank can easily tailor telemarketing messages that better resonate with each customer segment hence increasing engagement and improving the likelihood of a positive response.

3. Resource Allocation: With insights from the model, the bank can optimize resource allocation by prioritizing telemarketing efforts towards customers who are more likely to convert leading to cost savings and improved efficiency in the marketing budget.

4. Campaign Optimization: Continuous monitoring and analysis of the model's predictions will allow for ongoing optimization of the telemarketing campaign. The bank can refine their approach over time to maximize effectiveness by identifying patterns and trends in customer responses.

Conclusion

Based on the evaluation of supervised machine learning algorithms to find the best model for the bank's telemarketing project, it is evident that the Naive Bayes model stands out as the top-performing option with a recall score of 57.76%. The model demonstrates the highest capability in identifying potential customers likely to respond positively to the campaign. Its efficient training time and lower computational requirements further enhance its suitability for deployment, ensuring optimal resource allocation. Moreover, Naive Bayes offers interpretability, which is crucial for regulatory compliance and stakeholder understanding. Therefore, deploying the Naive Bayes model in the telemarketing campaign presents a compelling choice, promising to maximize the campaign's effectiveness in targeting prospective customers and ultimately driving positive outcomes for the bank in a timely manner.

References

1. UCI Machine Learning Repository. (n.d.). Archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/222/bank+marketing