Predictive Analysis for Term Deposit : Enhancing Long-Term Investments

Kent State University

Ambassador Crawford College of Business and Entrepreneurship

Course - Fundamentals of Machine Learning

Batch – Spring'24

Instructor - Chaojiang (CJ) Wu

Name – Vaishnavi Mangesh Donde

Email Id – vdonde@kent.edu

KSU ID - 811300129

Project Objective :

Confronted with a decrease in revenue attributed to inadequate customer engagement in long-term deposit investments, the Portuguese Bank is gearing up to elevate its strategic marketing endeavors. This endeavor aims to pinpoint customers with a higher propensity to engage in long-term deposit investments, enabling the bank to optimize its marketing strategies more precisely. To accomplish this objective, we will construct and refine a predictive model utilizing sophisticated data analysis techniques. Our emphasis lies in employing and contrasting various predictive methodologies, such as K-Nearest Neighbors (K-NN) and Naive Bayes, to identify the most suitable model tailored to our dataset's unique attributes. The overarching objective is to empower the bank with data-centric insights that drive an uptick in long-term deposit subscriptions and, consequently, revenue.

Here's the list of central objectives for the study:

Identifying Factors Influencing Long-Term Deposit Subscription: The objective is to determine the factors that influence customers' decisions to subscribe to long-term deposits offered by the Portuguese bank. This involves analyzing various customer attributes, interactions, and campaign-related factors to understand their impact on subscription behavior.

Segmentation of Customers: Another objective is to segment the bank's customer base based on their likelihood to subscribe to long-term deposits. This segmentation will help prioritize marketing efforts by targeting customers who are more predisposed to subscribing, thereby maximizing the efficiency of marketing campaigns.

Recommendation of Targeted Marketing Strategies: Based on the analysis, the study aims to provide actionable insights and recommendations for the bank to implement targeted marketing strategies. This includes identifying specific customer segments or characteristics that are associated with higher subscription rates and suggesting tailored approaches to engage these segments effectively.

Improving Revenue and Long-Term Deposit Performance: Ultimately, the overarching objective is to assist the Portuguese bank in reversing the revenue decline attributed to insufficient long-term deposit investments. By identifying and targeting customers with a higher propensity to subscribe to long-term deposits, the study aims to contribute to improving the bank's revenue and long-term deposit performance.

Model Selection: The main goal of the project is to find the best model for the dataset. Two algorithms, Naive Bayes and KNN, were tested. Although Naive Bayes showed high sensitivity, the KNN model performed better overall, with higher accuracy and better specificity. Therefore, the KNN model was chosen as the best option for predicting potential term deposit subscribers among existing bank customers.

Overview of the Dataset :

The dataset pertains to direct marketing campaigns conducted by a Portuguese banking institution. These campaigns primarily relied on phone calls, often necessitating multiple contacts with the same client to ascertain whether they would subscribe ('yes') or not ('no') to the product, specifically the bank term deposit.Contained within our dataset, labeled “train.csv,” are a total of 32,950 entries. Each entry encompasses 16 features, which encompass the target feature as well.

Feature	Feature Type	Description
age	numeric	age of a person
job	categorical	type of job ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
marital	categorical	marital status ('divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
education	categorical	('basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
default	categorical	has credit in default? ('no','yes','unknown')
housing	categorical	has housing loan? ('no','yes','unknown')
loan	categorical	has personal loan? ('no','yes','unknown')
contact	categorical	contact communication type ('cellular','telephone')
month	categorical	last contact month of year ('jan', 'feb', 'mar', …, 'nov', 'dec')
day_of_week	categorical	last contact day of the week ('mon','tue','wed','thu','fri')
duration	numeric	last contact duration, in seconds . Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no')
campaign	numeric	number of contacts performed during this campaign and for this client (includes last contact)
pdays	numeric	number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)
previous	numeric	number of contacts performed before this campaign and for this client
poutcome	categorical	outcome of the previous marketing campaign ('failure','nonexistent','success')

Target variable (required output):

Feature	Feature Type	Description
y	binary	has the client subscribed a term deposit? ('yes','no')

The Approach (Methodology) :

Data Exploration:

We explored the dataset to understand its structure and contents.

Data Preparation:

We cleaned the data by handling missing values, and removing outliers to ensure the reliability of our analysis.

Feature Selection:

We transformed categorical variables, employed linear regression to identify significant predictors and selected relevant features for predictive modeling.

Descriptive Statistics:

We employ various techniques such as mean, median, mode, standard deviation, and range to provide a succinct overview of the data's characteristics. Furthermore, we utilize both statistical methods and visual exploration approaches to delve deeper into the dataset's distribution, interrelationships among variables, and overarching patterns. Through the generation of descriptive statistics and the creation of visualizations like histograms or box plots, we not only formulate hypotheses but also make informed decisions regarding data preprocessing and model selection. This holistic approach ensures that we attain a comprehensive understanding of the dataset, guiding our subsequent analytical endeavors with precision and insight.

Predictive Analytics:

We normalized the data and applied two classification algorithms, Naive Bayes and K-Nearest Neighbors (KNN), to predict term deposit subscriptions. We evaluated the models' performance using confusion matrices and analyzed their accuracy, sensitivity, and specificity.

Naïve Bayes :

Naive Bayes emerges as a probabilistic machine learning technique rooted in Bayes' theorem, operating on the assumption of feature independence within a dataset. Despite its simplistic nature, Naive Bayes proves remarkably effective in classification tasks, especially in scenarios featuring a plethora of features or limited training data.

The significance of Naive Bayes lies in its simplicity, efficiency, and robustness. It exhibits computational efficacy and excels in environments with high-dimensional data, making it suitable for real-time applications and large-scale datasets. Overall, Naive Bayes offers a practical and reliable approach to tackle classification challenges, establishing itself as a preferred choice in the realm of machine learning applications.

After selecting the essential variables, we applied the Naive Bayes algorithm for classification on our prediction dataset. The results of the classification, including the confusion matrix and error metrics such as accuracy and sensitivity, are succinctly summarized below.

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1 5290  421
         2  488  405

               Accuracy : 0.8624
                 95% CI : (0.8538, 0.8706)
    No Information Rate : 0.8749
    P-Value [Acc > NIR] : 0.99892

                  Kappa : 0.3922

 Mcnemar's Test P-Value : 0.02859

            Sensitivity : 0.9155
            Specificity : 0.4903
         Pos Pred Value : 0.9263
         Neg Pred Value : 0.4535
             Prevalence : 0.8749
         Detection Rate : 0.8010
   Detection Prevalence : 0.8648
      Balanced Accuracy : 0.7029

       'Positive' Class : 1
```

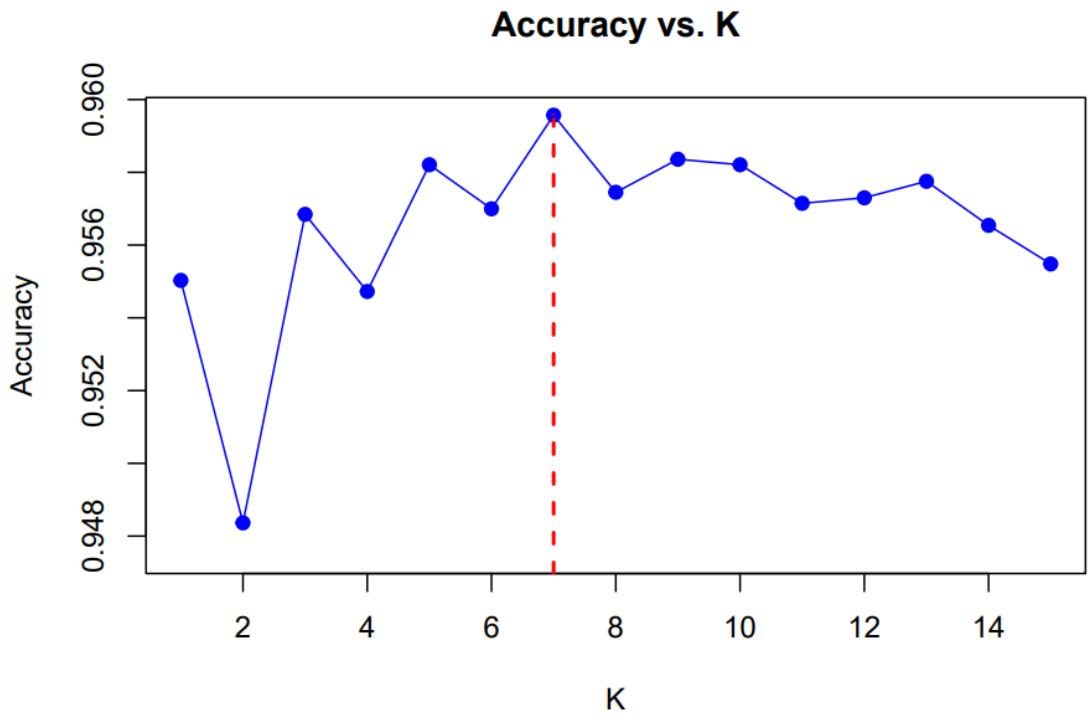
Summary of Naïve Bayes model:

The model's accuracy is about 86.24%, meaning it gets most predictions right. The model exhibits high sensitivity (True Positive Rate) of 91.55% which indicates that it correctly identifies a high proportion of actual positive instances.The specificity (True Negative Rate) is relatively low at 49.03% which says that the model has difficulty in correctly identiyfying negative instances. True positives and true negatives show how well the model sorts each class correctly, while false positives and false negatives reveal where it messes up. To improve accuracy, the model might need some adjustments to reduce these mistakes and improve the overall accuracy.

KNN Method :

The KNN (K-Nearest Neighbors) technique is a versatile machine learning algorithm employed for both classification and regression tasks. It functions based on the concept of similarity, where the prediction for a new data point is determined by the majority vote or average of its closest neighbors in the feature space. One of the primary advantages of KNN is its simplicity and straightforward implementation, making it accessible to individuals at all proficiency levels in the field. It proves particularly valuable in scenarios where the decision boundary is irregular or challenging to define using conventional methods, as it does not rely on any assumptions regarding the underlying data distribution.

Identifying the optimal value of k in KNN is crucial, as it significantly impacts the model's performance, directly influencing its ability to accurately classify data points based on their nearest neighbors. As a result, we have initially calculated the optimal k value, and subsequently, we will utilize the KNN model to evaluate its accuracy.



Summary of KNN model:

As shown in the graph, we find that the best k value is 7. The model boasts an impressive accuracy of 95.96%, meaning it accurately predicts the class for most cases. It shows a high sensitivity (True Positive Rate) of 99.17% and a moderate specificity (True Negative Rate) of 73.49%. This indicates that it's good at spotting true positives and handling false positives fairly well.

Conclusion :

The project aimed to tackle a decline in revenue within a Portuguese bank by formulating strategies to boost long-term deposit subscriptions among customers. Initial exploration and data cleansing revealed a dataset spanning from May 2008 to November 2010, comprising 16 variables of both numerical and categorical types. Following feature selection and normalization, predictive analytics techniques were applied, including Naive Bayes and k-Nearest Neighbors (KNN) classifiers, to analyze the relationship between predictors and term deposit subscriptions.

While the Naive Bayes classifier achieved an accuracy rate of 86.24%, the KNN classifier surpassed it with a 95.96% accuracy rate. Despite the notable sensitivity of the Naive Bayes model, the KNN model exhibited superior performance, demonstrating higher accuracy and enhanced specificity. These findings suggest that the KNN model emerges as the preferred option for predicting potential term deposit subscribers among existing bank customers, offering actionable insights to address the revenue decline and promote long-term growth.

Source of Data :

1. Banking Dataset Classification - Predicting if the client will subscribe to a term deposit : https://www.kaggle.com/datasets/rashmiranu/banking-dataset-classification?resource=download

How this Data satisfy the condition for being real-world :

The data from the Portuguese banking institution's marketing campaigns satisfies the condition for being real-world due to the following reasons:

Origin and Purpose: The data originates from direct marketing campaigns conducted by a real banking institution in Portugal. These campaigns were aimed at promoting bank term deposits, which reflects real-world business objectives and strategies.

Multiple Contacts: The campaigns involved multiple contacts with the same client to determine if they would subscribe to the bank term deposit or not. This multi-contact approach mirrors real-world marketing practices where follow-up communications are often necessary to persuade potential customers.

Temporal Aspect: The data spans from May 2008 to November 2010, covering a substantial timeframe. This temporal aspect adds to the real-world relevance as it captures variations and trends in customer behavior over time, allowing for more accurate analysis and decision-making.

Volume and Variety: The dataset contains a significant number of examples (32950) and 16 inputs, including various features related to the customers and their interactions with the bank. This volume and variety reflect the complexity and richness of real-world data, which often consists of multiple variables and a large number of instances.

Target Feature: The target feature indicates whether the customer subscribed ('yes') or did not subscribe ('no') to the bank term deposit. This reflects the real-world outcome of interest for the banking institution, as it directly relates to its goal of increasing long-term deposits.

Overall, the data from the Portuguese banking institution's marketing campaigns meets the criteria for being real-world, as it closely resembles the dynamics and challenges faced by businesses in a practical setting. This authenticity enhances the relevance and reliability of any analysis or recommendations derived from the data.