Predicting Customer Deposits in Banking: An Analysis Using the SEMMA Data Mining Approach

Koluguri Ruchitha Reddy

September 2023

Abstract

Handling vast volumes of data for insightful analysis requires a structured approach, and the SEMMA (Sample, Explore, Modify, Model, Assess) methodology provides just that. In our study, "Predicting Customer Deposits in Banking: An Analysis Using the SEMMA Data Mining Approach," we delve into a banking dataset to forecast whether customers will make deposits. This research systematically applies SEMMA to unravel patterns and predictors related to customer deposits. The outcomes hold the potential to empower the banking sector with strategic insights for targeted marketing campaigns and enhanced customer retention.

1 Introduction

The modern banking landscape is inundated with data, and extracting valuable insights from this vast reservoir of information demands a structured approach. The SEMMA (Sample, Explore, Modify, Model, Assess) methodology provides a systematic framework for effective data mining. In this research paper, "Predicting Customer Deposits in Banking: An Analysis Using the SEMMA Data Mining Approach," we delve into a dataset from a banking institution with the goal of predicting the likelihood of a customer making a deposit. Through a systematic application of the SEMMA methodology, this study aims to uncover patterns and predictors associated with customer deposits, offering actionable insights that can drive targeted marketing campaigns and customer retention strategies in the banking sector.

2 SEMMA Methodology

2.1 Sample

The first step in the SEMMA methodology is "Sample." Here, we select and obtain data for analysis. This phase involves data collection, extraction, and ensuring that the data is representative of the problem at hand. It lays the foundation for the subsequent analytical process.

2.2 Explore

In the "Explore" step, data is thoroughly examined and visualized. Exploratory data analysis (EDA) is performed to understand the dataset's characteristics, identify patterns, and detect any outliers or missing values. This phase helps in gaining insights into the data's structure and content.

2.3 Modify

The "Modify" phase focuses on data preparation and transformation. Here, data is cleaned, preprocessed, and transformed to make it suitable for modeling. Tasks in this phase may include handling missing values, encoding categorical variables, scaling, or creating new features to enhance the dataset's quality.

2.4 Model

The "Model" step involves the development and training of predictive or analytical models. This is where machine learning algorithms are applied to the prepared dataset to build models that can make predictions or uncover patterns. Model selection and training are key tasks in this phase.

3 Implementation

3.1 Sample

In the initial phase of SEMMA, we load the dataset and gain an understanding of its structure and content. The dataset provides information on bank customers, including attributes such as age, job, marital status, education, balance, and others. Our goal is to predict whether a customer will make a deposit based on their details.

3.2 Explore

Exploration of the dataset reveals crucial insights. The dataset comprises 11,162 records and exhibits no missing values. Key variables such as age, balance, duration, campaign, pdays, and previous are numerical, with statistics including mean, median, and standard deviation obtained. Categorical variables, including the target variable "deposit," are identified.

3.3 Modify

In the modification phase, we prepare the dataset for modeling. Categorical variables like job, marital status, education, default, housing, loan, contact, month, and poutcome may need encoding. Additionally, numerical variables can benefit from normalization or scaling. Feature engineering may also be applied to enhance the dataset's quality.

3.4 Model

The modeling step involves the creation of predictive models. Data is split into training and testing sets. In this research, we initiated a predictive model using a Decision Tree classifier. The model was trained using the training data.

3.5 Assess

To evaluate the model's performance, additional metrics beyond accuracy are considered. Precision, recall, and F1-score are used to assess the model's ability to predict customer deposits, especially considering potential class imbalances. This detailed evaluation provides insights into the model's performance for each class, "Deposit" and "No Deposit."

4 Conclusion

In the realm of banking and customer relations, the application of the SEMMA methodology has unveiled significant insights. Through the systematic steps of Sampling, Exploration, Modification, Modeling, and Assessment, we embarked on a journey to predict customer deposits. The Decision Tree classifier showcased an accuracy of 76.40

References

1. https://www.kaggle.com/datasets/vikrishnan/iris-dataset
2. PyCaret. (2020). An open-source, low-code ML library in Python.
3. https://github.com/ruchithareddy269/Dm-Assignment-3