DSCI 445 Final Report

Machine Learning Models on A Bank Marketing Dataset

Sandra Afrifa, Tigist Kefelew , Mussa Hassen, Waddah Alaahtani

Motivation # TT

Bank Marketing dataset The Bank Marketing dataset was collected by a Portuguese banking institution during a series of direct telemarketing campaigns. These campaigns were done to encourage clients to subscribe to a long-term deposit product. Telemarketing is an expensive and time-consuming process; therefore, the bank would want to know in advance which customers are most likely to subscribe before any more investment takes place. Predictive modeling bears great value in this scenario, as it helps the bank focus resources on only those customers who depict a higher possibility of subscribing. In the subscription rate of this dataset, the problem is even more difficult as it is very low. Most individuals say “no,” and only a small percentage decides to open a term deposit. If one class is rare, human judgment or simple rules fail to identify those clients who could actually be interested. A well-constructed model may present potential improvements in targeting accuracy, reduce the number of unsuccessful calls, and raise the return on investment. That is why it is necessary to research this dataset and look at which modeling approaches work best.

Description of dataset #TT Each entry of the dataset represents one client contacted in one of the bank’s telemarketing campaigns. This dataset contains 16 predictors carrying significant information about the client’s background, financial situation, and interactions with previous campaigns. These variables fall into several key categories: Demographics include such factors as age, occupation, marital status, and education. These help describe who the clients are and often relate to financial behavior or economic stability. Financial indicators: Loan status, credit history, and average account balance. These variables describe the client’s existing financial commitments and capacity, which may influence their willingness to invest in a long-term product like a term deposit. Marketing campaign features include the type of contact: cellular or telephone, month of contact, and day of contact, and duration of the call. These are operational decisions made by the marketing team and may influence a client’s likelihood of subscribing. For example, longer call duration sometimes is indicative of higher client engagement. Campaign outcomes: Number of previous contacts and if the past interactions have resulted in a successful outcome. These variables highlight whether the client has been interested in the past or has been contacted many times without success. The data includes both numerical and categorical variables, and this can serve as a good example to compare the many different classification methods with respect to how each treats different data types: for instance, linear models require the conversion of categorical variables, whereas tree-based or distance-based methods would naturally handle those in different ways. This allows us to explore how each approach might respond to the same set of features. The response variable, y, tells whether the client has subscribed to the term deposit or not. Only a small fraction of the contacted clients actually subscribe, therefore this is a strongly imbalanced dataset. The class imbalance is because, in general, marketing datasets are imbalanced since the majority of customers do not respond positively when outreach is performed. This skewness biases models toward the “no” prediction since this prediction would naturally be correct for most of the cases. Therefore, careful handling of this bias will be critical to generate models capable of identifying the small minority of clients who are likely to subscribe-the group that the bank cares most about.

Problem Statement/Assumptions #TT It is now required to develop statistical models based on the available information in the dataset that will predict with accuracy whether a client will subscribe to a term deposit. This problem is of a binary classification nature, and our work focuses on comparing several models that differ in assumptions, flexibility, and interpretability. Understanding how such models perform is useful to a marketing team for making informed decisions about which clients to contact in future campaigns. We focused our attention on well-performing models that give insightful knowledge of which factors drive the choice of a client. In the banking industry, understanding why a model has yielded some prediction is crucial. This supports ethical marketing practices, helps managers explain decisions, and can be important for regulatory compliance. For instance, highly flexible models may perform very well but may not indicate which client characteristics are actually relevant. More interpretable models do not always have the strongest accuracy but can guide marketing strategies more directly. We make the following assumptions to keep the analysis focused and manageable: Independence of observations: We treat the observations as independent, even though multiple contacts with the same client may appear across different campaigns. This assumption makes the training and evaluation of the models easier, even if some repeated-measure structure in the data is ignored. No time-series modeling: variables such as campaign month or day are available within the dataset, which can be considered to hold temporal patterns. However, we do not model this data as a time series. This keeps the focus on cross-sectional relationships between client characteristics and subscription behavior. Use of class-relevant models: We restrict our approach to models presented in class because this enables us to contrast multiple concepts, including linear decision boundaries, distance-based approaches, and flexible tree-based structures. This renders the analysis suitable for a course project while still answering substantial questions regarding classification performance. Generalization from sample to population: We assume that this dataset is a good representation of the client population to which the bank will offer its services in forthcoming campaigns. If we couldn't assume this, our predictions wouldn't generalize properly to real-world decision-making. By selecting these assumptions, we are assured of a standard framework that will allow the comparison of different classification techniques on equal footing. The aim is not only to identify which model performs the best but also to understand how different modeling strategies behave when applied to imbalanced marketing data; this may provide practical insights into how a bank would structure its future campaigns and what features of clients are most relevant for predicting subscription behaviors.

Methodology

Data Wrangling and Cleaning: We began by loading the dataset, converting categorical variables into factors, and identifying "unknown" levels used in place of missing data. We evaluated several approaches to handling these values—including keeping "unknown" as its own category and treating them as missing, but ultimately selected the method that preserved the most information without distorting class balance.

A train/test split was created, and all models were trained using the same set of cleaned predictors for fair comparison.

Models Considered:

We evaluated five main models from the course:

- Logistic Regression — A linear classifier that estimates the log-odds of subscription. It is highly interpretable and provides coefficients that can be directly translated into marketing insights.
- K-Nearest Neighbors (KNN) — A nonparametric method that makes predictions by finding the k most similar clients. It captures nonlinear relationships but can be sensitive to scaling and high-dimensional data. #TT

K-Nearest Neighbors is a nonparametric classification method that predicts an outcome for a new observation by looking at the k most similar clients in the training dataset. Rather than estimating a set of coefficients or assuming any functional form between predictors and the response, KNN makes predictions based purely on proximity in predictor space. This makes the method particularly well-suited when the underlying relationship between predictors and subscription behavior is nonlinear or too complicated for a parametric model to capture. Because KNN relies on distance calculations, feature scaling is an important preprocessing step. Otherwise, predictors that are measured on a larger scale-e.g., account balance or call duration-would dominate the distance metric and distort which observations are considered “nearest.” In our implementation, we standardized all numeric variables so that each feature was given equal weight in determining the similarity between any two customers. This preprocessing step is especially important for marketing datasets which often involve heterogeneous variables. A major modeling decision in KNN, which involves a choice, is the value of k. For small values of k, the classifier is highly flexible and closely follows the training data with a potential risk of overfitting noise. Larger values produce smoother, more stable decision boundaries; however, this has a potential risk of missing the important patterns. We used 5-fold cross-validation to identify the value of k that minimized the estimated test error, balancing the bias-variance trade-off. While KNN presents an intuitive method of finding clients with similar profiles, it has several limitations in this context. The performance may suffer for high-dimensional data, with points becoming further apart and distances becoming less informative-the “curse of dimensionality”. Second, KNN is less interpretable than models like logistic regression, where the effect of every predictor can be expressed quantitatively. It does, however, provide a useful baseline when assessing the performance of alternative nonlinear classification methods for predicting term deposit subscriptions.

- Boosting (e.g., AdaBoost or Gradient Boosting) — An ensemble method that sequentially builds weak learners to minimize classification error. Boosting is often superior in predictive accuracy but is less interpretable.

Model Training, Formulation, and Validation

All models were trained using k-fold cross-validation to estimate out-of-sample performance. This approach reduces variance and ensures that results do not depend on a single train/test split.

We also carried out: - Feature selection using stepwise regression and subset selection for logistic regression - Regularization techniques (ridge/lasso) when appropriate - Dimensionality reduction (e.g., PCA) if needed to improve KNN performance - Hyperparameter tuning for KNN (choice of k)

This section of the paper will include the formal model equations and the details of the validation procedure.

Results

Once our models are trained, we will compare their performance using test error rate, confusion matrices, and metrics such as accuracy, recall, and AUC. Because the dataset is imbalanced, we expect recall and AUC to be especially important for evaluating how well each model identifies clients who subscribe.

We will also create ROC curves to compare models visually and use information criteria (AIC/BIC) for parametric models where appropriate. After summarizing these results, we will determine which model performs best overall and which provides the most useful insights for guiding future marketing decisions

Discussion

Since we are still exploring the dataset and fitting our models, we have not yet determined which classification method will perform best. However, based on the structure of the Bank Marketing data and findings from similar studies, we expect that boosting is likely to give the best predictive accuracy, while logistic regression will likely be the most interpretable and actionable for managers.

In the final version of this section, we will compare the models based on their test performance, interpretability, and robustness to class imbalance. We will also discuss any limitations of our approach and suggest future improvements, such as trying resampling methods or evaluating additional machine learning models.

References

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Bank Marketing Dataset. https://archive.ics.uci.edu/dataset/222/bank+marketing