DSCI 445 Final Report

Machine Learning Models on A Bank Marketing Dataset

Sandra Afrifa, Tigist Kefelew , Mussa Hassen, Waddah Alqahtani

Motivation

Bank telemarketing campaigns are a major channel through which financial institutions promote term deposit products. These marketing campaigns come with a cost and are often inefficient due to low client response rates. Prior work by Moro et al. (2014) discover that client characteristics and campaign related factors strongly influence subscription outcomes. Motivating a data-driven approach to improve targeting and decision-making in bank marketing. In their approach a Portuguese retail bank was studied using campaign data collected between 2008 and 2013.

Our dataset was a publicly available dataset, thus had less variables for privacy concerns of the bank customers. 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. The dataset consists of sixteen variables describing demographic, financial and campaign related variables that enlist individual clients contacted during direct marketing efforts. A list of the 16 variable found below...

- age: Age in years
- job: Occupation (Categorical)
- marital: Marital status (Categorical)
- education: Highest level of education (Categorical)
- default: has credit in default? (binary)
- balance: average yearly balance
- housing: has housing loan? (binary)
- loan: has personal loan? (binary)
- Contract: contact communication type (Categorical)
- day: last contact day (of the month)
- month: last contact month of year (month; 1=January)
- duration: last contact duration, in seconds
- campaign: number of contacts performed during this campaign and for this client
- pdays: number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted)
- previous: number of contacts performed before this campaign and for this client
- poutcome: outcome of the previous marketing campaign (Categorical)
- y: Indicator of whether the client subscribed a term deposit (Binary target)

The response variable indicates whether a client subscribed to a term deposit which defines a binary classification task.

Exploratory analysis shows that the numeric predictors (balance, duration, pdays, previously contacted) are heavily skewed. Also, analysis reveals substantial class imbalance, in the response variable, which motivates

the use of appropriate modeling and evaluation for our classification problem. Visualizations of subscription rates by job education marital status contact type and previous outcome reveal variations in client responsiveness across different factors.

Exploratory Data Analysis (EDA)

Data Characteristics and Class Imbalance

Our exploratory analysis examined 45,211 client records from the Portuguese banking institution. Analysis of the target variable revealed severe class imbalance, with only 11.7% of clients subscribing to term deposits compared to 88.3% who declined (Figure 1). This imbalance represented a key challenge for predictive modeling.

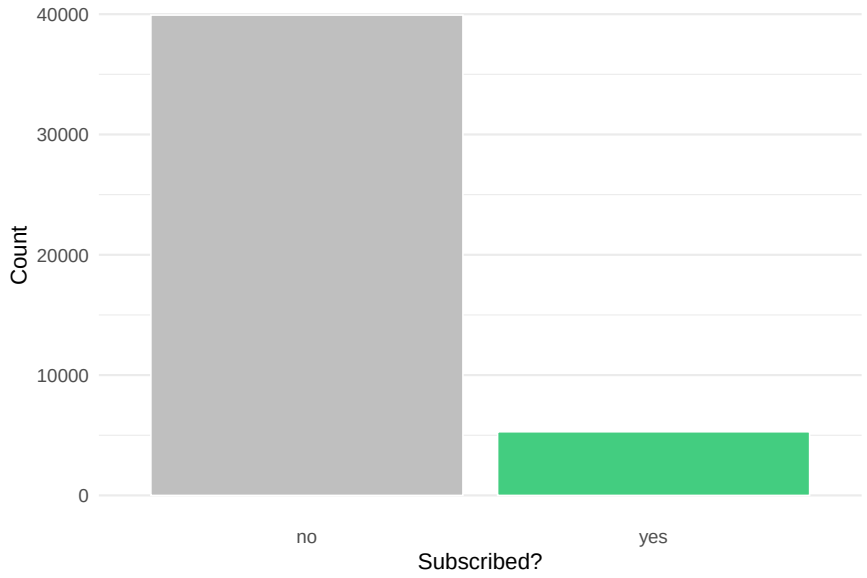


Figure 1: Class imbalance in term deposit subscriptions. Only 11.7% of clients subscribed, presenting a significant modeling challenge.

Demographic Patterns

Demographic patterns emerged from the data: retired clients and students were most likely to subscribe, while blue-collar workers were least likely. Higher education levels correlated with increased subscription rates.

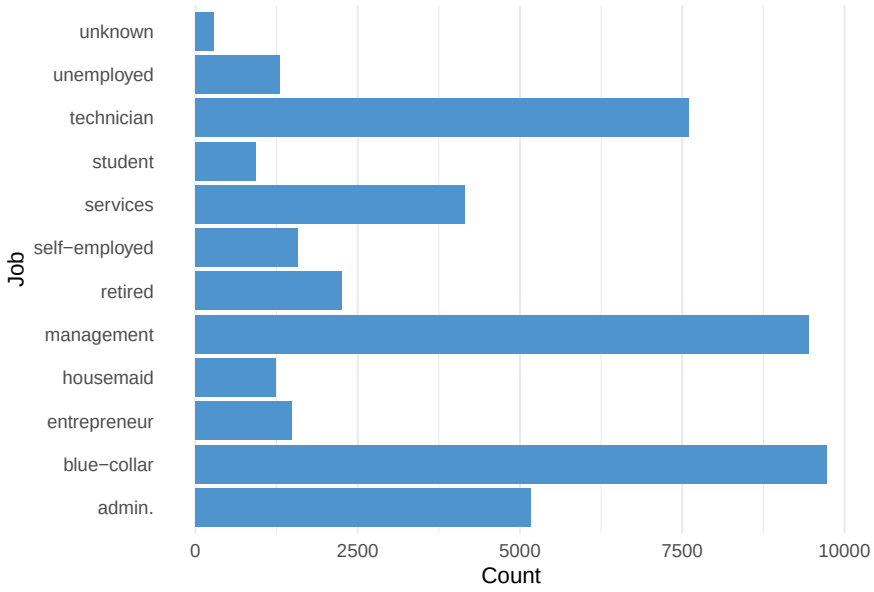


Figure 2: Distribution of job categories in the dataset. Administrative, blue-collar, and technician occupations are most common.

Historical Patterns

Analysis of previous campaign outcomes revealed powerful patterns: clients with previous successful subscriptions demonstrated approximately 65% likelihood of subscribing again, compared to only 15% among those with previous unsuccessful outcomes.

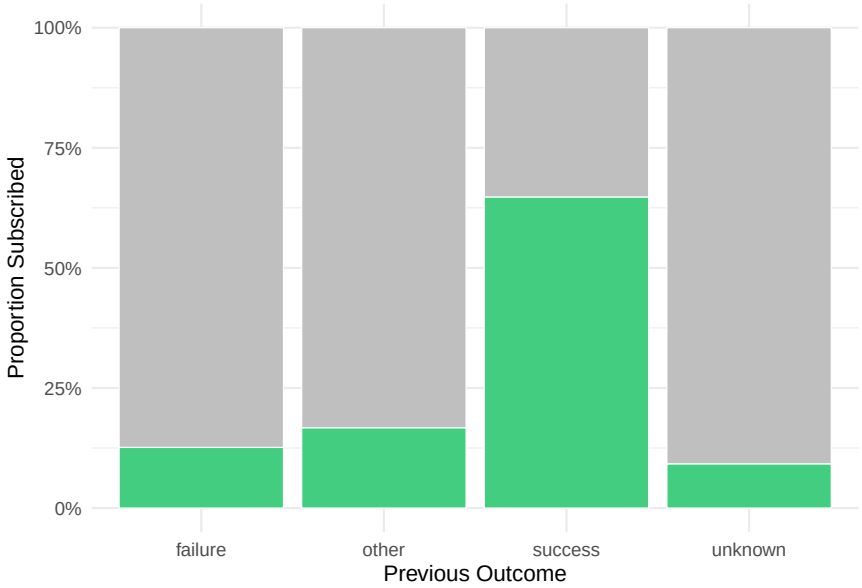


Figure 3: Previous campaign outcome influence. Clients with previous success show 65% subscription rates compared to 15% for previous failures.

Campaign Dynamics and Diminishing Returns

Campaign persistence analysis showed a clear downward trend in success rates as contact frequency increased, with substantial decline after five attempts and minimal success beyond thirty contacts.

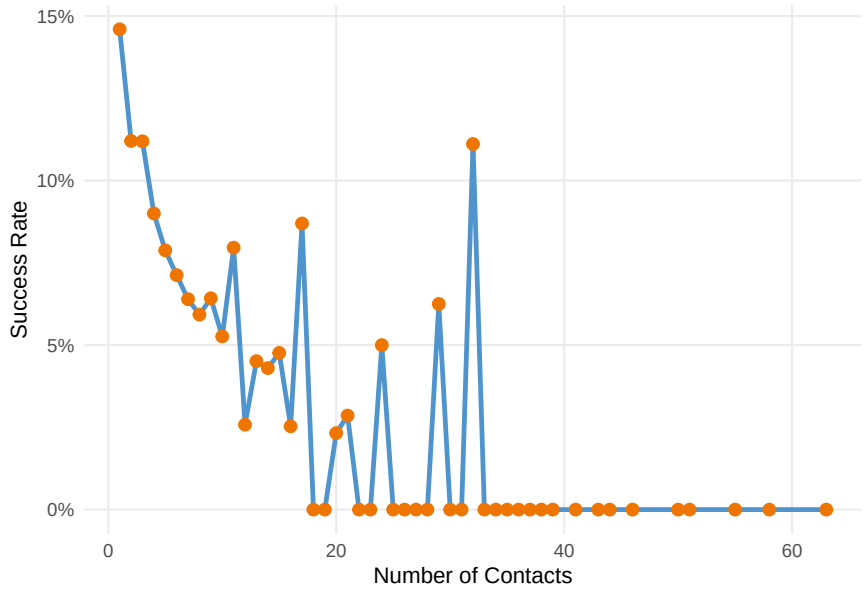


Figure 4: Diminishing returns from repeated contacts. Success rates drop significantly after 5 contacts.

Data preprocessing involved converting categorical variables to factors and adding derived variables such as pdays to account for clients not previously contacted. Converting the categorical variables (initilay character type) to factors creates an indicator for each class in the variable (i.e. if job has 20 factors, 20 indicator variables will be considered in the model fit). As for the previously contaced variable (days since previously contacted), customers who haven't been prreviously contacted will recieve a value of -1. This can cause improper modeling of customers, since the variable now takes on the value of negative numbers. Instead we set -1 to zeros and created a new variable to indicate customers that have not been previously coontacted before. These steps ensure that the dataset is suitable for statistical learning models and enable accurate assessment of predictive performance.

Problem Statement

This study is motivated by the availability of bank marketing datasets that enable comparison of machine learning methods for predicting term deposit subscription behavior.

The specification of the problem begins with the recognition that we are addressing a binary classification task where the goal is to predict whether a client will subscribe to a term deposit based on demographic financial and campaign features. According to the framework presented in Introduction to Statistical Learning (ISLR), the fundamental goal is to estimate the unknown function, f(x) that relates the predictors, X to the response, Y. Some models are highly flexible and can capture complex non-linear relationships between predictors and the response. In most Cases these models achieve higher predictive accuracy. However, these flexible models often suffer from lower interpretability, making it difficult to understand the influence of individual features. Simpler models such as logistic regression are less flexible but provide coefficients that can be directly interpreted in terms of odds ratios, offering insight into feature importance.

Chapter 5 of Introduction to Statistical Learning emphasizes the importance of obtaining an unbiased estimate of a model's test error to assess its predictive performance on unseen data. Test error quantifies how well a model generalizes beyond the data used for training, which is critical for selecting among competing models. The chapter presents several methods for estimating test error, including the simple training/test split, K-fold cross validation, and the leave-one-out cross validation (LOOCV) approach. Each method balances bias and variance differently, with cross validation generally providing a lower-variance estimate compared to a single split. These techniques are essential for model selection, tuning hyperparameters, and comparing the expected performance of alternative models. For the Bank Marketing dataset, estimating test error accurately is particularly important due to the class imbalance and the presence of both categorical and numeric predictors, which can influence model generalization.

In the following chapter of Introduction to Statistical Learning focuses on model selection and regularization techniques to improve predictive performance and reduce overfitting. It introduces methods such as subset selection, shrinkage (Ridge and LASSO regression), and dimension reduction to assess and enhance model accuracy while controlling variance. These approaches allow practitioners to identify parsimonious models that generalize well to new data, providing complementary strategies to traditional test error estimation. We apply these concepts to the models we fit on the bank data, to reduce overfitting and ultimately improve the generalization of our models to new clients.

We start with a logistic regression (described in methodology section), then build up adjustments and regularization to the logestic regression to make our baseline. We then move to another interpretable model: KNN. This model is considered transparent in a unique, due to it's neighboring factor. Finally we fit a Boosting model. This state-of-the-art model is not considered interpretable. Although, there are methods used to extract the contributing variables in such models. We included this model is used to gage how much accuracy we are sacrificing for interpretability.

The authors in (Moro et al.) employed time ordered data splitting rolling window evaluation. For feature enrichment, they implemented a semi-automatic approach including external intuitive knowledge from domain experts (bank manager) to optimize features. They compared Logistic regression with Decision trees, SVMs, and Neural networks; all complex 'black-box' models. While (Moro et al.) prioritized maximizing classification performance, our approach also examines model simplicity, automatic feature selection, and diagnostic checks for appropriate classifications (from confusion matrices).

Methodology

For each of our models we implement a k-fold CV approach, with k=5 training/test split to maintain the distribution of the response variable across the sets. Our models do not apply time-series modeling, since our data does not include sequential variables. Although variables such as campaign month or day are available within the dataset, which can be considered to hold temporal patterns, we do not model this data as a time series. This keeps the focus on cross-sectional relationships between client characteristics and subscription behavior. An 80/20 random train–test split was used, and all modeling decisions, including feature selection and hyper-parameter tuning, were made using only the training data. This prevented information leakage and ensured that the test set provided an honest evaluation of generalization performance.

1. **Logistic regression**

Logistic regression is used to model the probability of a binary outcome as a function of predictor variables. The model estimates the probability that a client subscribes to a term deposit, $P(Y = 1 \mid X)$. This model fits with assumptions of independent observations and normality.

Formula:

$$f(x) = P(Y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

Coefficients are estimated using the maximum likelihood method, which finds the parameter values that maximize the likelihood of observing the training data (ISLR).

Interpretability

To understand the meaning of the coefficients, we can transform the formula into the log-odds (logit) form:

$$\log\left(\frac{P(Y=1 \mid \mathbf{x})}{1-P(Y=1 \mid \mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Consider a single coefficient, say $\beta_{balance}$, which relates to a client's average bank account balance. If the estimated coefficient for balance is $\beta_{balance} = 0.001$, this means that a one-unit increase in the client's balance is associated with a change of 0.001 in the log-odds of subscribing to a term deposit, assuming all other variables remain constant. This change in log-odds translates to a multiplicative change of $e^{0.001} \approx 1.001$ in the odds. Therefore, for every unit increase in balance, the odds of subscription increase by a factor of 1.001. This shows a clear relationship between the coefficient and the target variable.

Without loss of generality, we can generalize this interpretation and estimation process to logistic regression models with multiple variables and their variations, such as incorporating Lasso regularization (for variable selection and complexity control) or applying polynomial transformations to the predictor variables (to capture non-linear relationships).

Model fit

Feature selection is performed using backward stepwise selection to identify a parsimonious subset of predictors that minimizes estimated test error as measured by AIC. Variables such as duration and previous marketing outcomes remain in the model due to their strong predictive power, whereas less informative variables are excluded to reduce variance. Categorical predictors are converted to factors, and a new binary variable previously_contacted is created to handle clients with no prior campaign contact. This preprocessing ensures that the data is suitable for statistical modeling and facilitates accurate interpretation of model coefficients. We resulted 12 features:

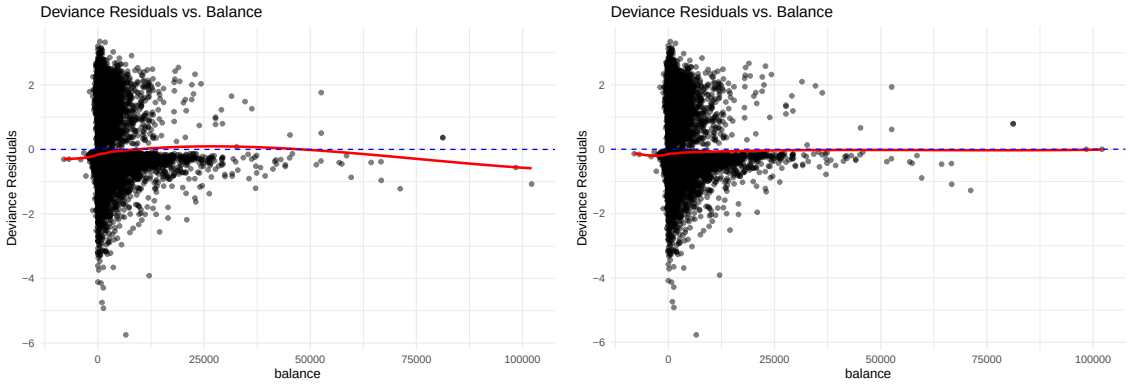
```
## y ~ job + marital + education + balance + housing + loan + contact +
##     day + month + duration + campaign + poutcome
```

Cross validation is used to select optimal tuning parameters for these regularized models and to estimate the expected test error. Confusion matrices, sensitivity, specificity, and ROC/AUC metrics are computed to assess predictive performance, especially given the class imbalance where non-subscribers are the majority. This model performed will with mean 5-fold CV test error metrics:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  7786  710
##        yes  198  347
##
##                Accuracy : 0.8996
##                  95% CI : (0.8932, 0.9057)
##     No Information Rate : 0.8831
##     P-Value [Acc > NIR] : 3.496e-07
##
```

```
##                   Kappa : 0.3842
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.32829
##             Specificity : 0.97520
##          Pos Pred Value : 0.63670
##          Neg Pred Value : 0.91643
##              Prevalence : 0.11691
##          Detection Rate : 0.03838
##    Detection Prevalence : 0.06028
##       Balanced Accuracy : 0.65174
##
##        'Positive' Class : yes
##
```

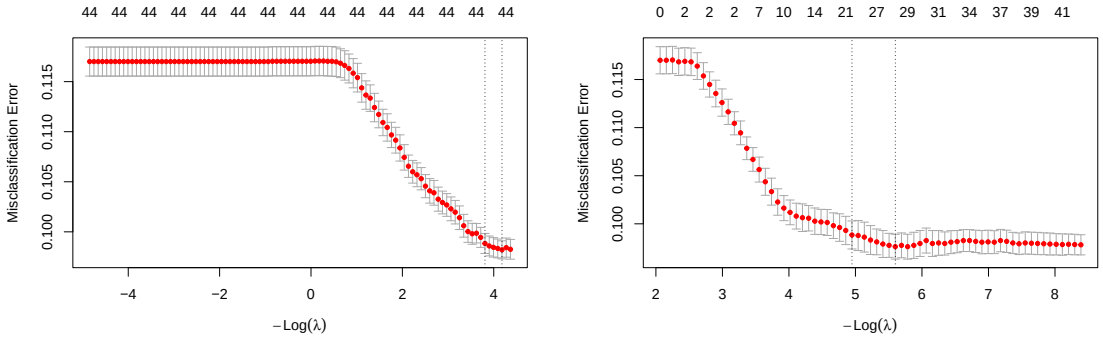
Next, we check for potential non-linearity in continuous predictors from our predictors remaining after back-ward stepwise selection. We plotted deviance residuals against numeric features. Based on residual diagnostics, we incorporated a polynomial term (using the poly function: *poly(balance, 4)*) for balance to account for non-linear effects. The Balance squared and power 4 term produced significant coefficients (Last 4 on the list of coefficients).




```
##                       Estimate    Pr(>|z|)
## poly(balance, 4)1   5.5981797 0.198257351
## poly(balance, 4)2 -23.0245708 0.002841268
## poly(balance, 4)3  -0.9137324 0.908272540
## poly(balance, 4)4 -19.8046012 0.001559676
```

Although this polynomial transformation didn't provide desired metrics in expense of the nonlinear effect. From the deviance residual plots we can observe there weren't significant deviance from a normal residual line. We will not include this model in our comparison.

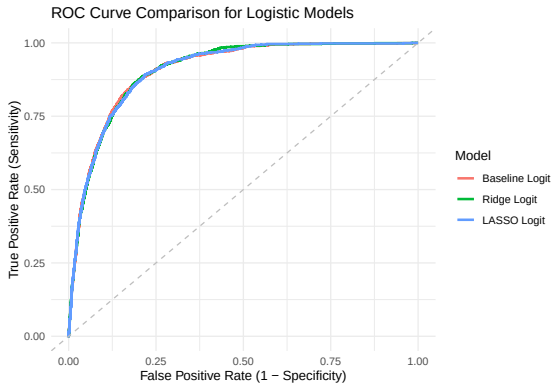
Regularization techniques such as Ridge (ℓ_2 penalty) and LASSO (ℓ_1 penalty) are employed to reduce variance and perform feature selection, particularly in the presence of correlated predictors. LASSO (Least Absolute Shrinkage and Selection Operator) regression uses an ℓ_1 penalty ($\alpha = 1$). A key property of the LASSO is that it forces the coefficients of some variables to be exactly zero, thereby performing automatic feature selection (ISLR Chapter 6). The LASSO plot similarly uses cross-validation to select the optimal $\lambda_{\min}$, 0.0037. Unlike Ridge, the LASSO is expected to yield a more parsimonious model by excluding irrelevant predictors entirely.



Here is a comparison of the models fitted thus far:

```
##            Model  Accuracy Sensitivity Specificity
## 1 Baseline Logit 0.8996792   0.3301798   0.9750752
## 2    Ridge Logit 0.8973565   0.2686850   0.9805862
## 3    LASSO Logit 0.8981307   0.3008515   0.9772044
```

Ridge and LASSO models yield similar results, indicating that additional shrinkage does not meaningfully improve classification of subscribers. We assume this is due to our modeling being majority indicator variables. This binary aspect of the indicator variable causes it's coefficient to be less affected by regularization techniques.



All three models exhibit a high overall Accuracy (around 0.89), which is expected given the significant class imbalance (Prevalence ≈ 0.11). However, the critical metric, Sensitivity (the True Positive Rate for the minority class, 'yes'), is very low (around 0.17). This result, consistent across all three models, indicates that they struggle to correctly identify subscribers, leading to many False Negatives (FN). Conversely, the high Specificity (True Negative Rate ≈ 0.98) means the models are excellent at identifying non-subscribers. The regularization methods (Ridge and LASSO) failed to provide any significant improvement in these metrics, suggesting that the primary limitation is one of high bias (due to insufficient flexibility) rather than high variance.

Models Considered:

We evaluated five main models from the course:

2. **K-Nearest Neighbors (KNN)**

- 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.

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.

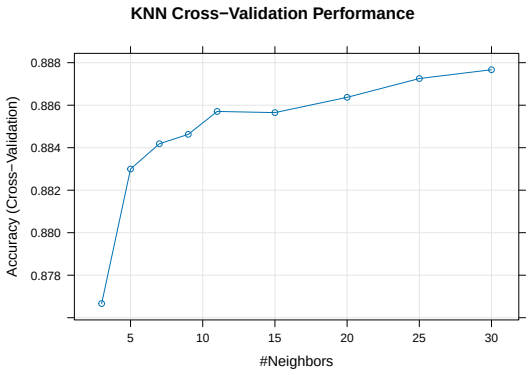
While KNN presents an intuitive method of finding clients with similar profiles, it has several limitations in this context. The performance may suffer for high-dimensional data, with points becoming further apart and distances becoming less informative-the "curse of dimensionality". Second, KNN is less 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.

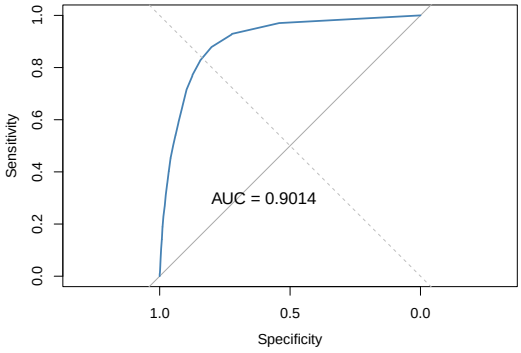
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  7850  790
##        yes  134  267
##
##                Accuracy : 0.8978
##                  95% CI : (0.8914, 0.904)
##     No Information Rate : 0.8831
##     P-Value [Acc > NIR] : 5.031e-06
##
##                   Kappa : 0.3227
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.25260
##             Specificity : 0.98322
##          Pos Pred Value : 0.66584
##          Neg Pred Value : 0.90856
##              Prevalence : 0.11691
##          Detection Rate : 0.02953
##    Detection Prevalence : 0.04435
##       Balanced Accuracy : 0.61791
##
##        'Positive' Class : yes
```

```
##
```

```
##
## Optimal number of neighbors (k): 30
```

```
## Cross-validation accuracy: 0.8877
```





3. **Boosting**

Boosting was used as the most flexible modeling approach in this study to address the strong class imbalance and complex predictor relationships present in the Bank Marketing dataset. Unlike logistic regression and KNN, boosting does not rely on a fixed functional form and can naturally capture nonlinear effects and interactions among demographic, financial, and campaign-related variables.

We implemented gradient boosting using a Bernoulli loss function for binary classification. The model was trained on the same 80% training set as the other methods. To encourage gradual learning and prevent overfitting, we used a small learning rate (shrinkage = 0.01) and shallow trees (interaction depth = 3). The model was initially trained with 1,500 trees to allow sufficient flexibility.

The optimal number of trees was selected using out-of-bag (OOB) error estimation, which serves a similar role to cross-validation in ensemble methods. The OOB procedure identified the point at which additional trees no longer meaningfully improved predictive performance. Final predictions were computed using the optimal number of trees selected by this procedure.

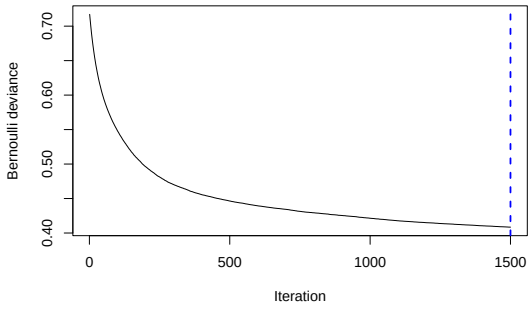
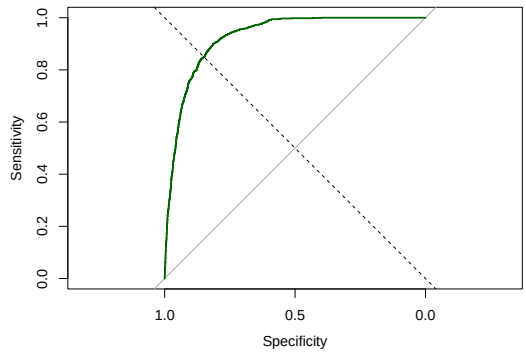


Figure 5: Out-of-bag error used to select the optimal number of boosting trees.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  7744  622
##        yes  240  435
##
##                Accuracy : 0.9047
##                  95% CI : (0.8984, 0.9106)
##     No Information Rate : 0.8831
##     P-Value [Acc > NIR] : 2.946e-11
##
##                   Kappa : 0.4524
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.41154
##             Specificity : 0.96994
##          Pos Pred Value : 0.64444
##          Neg Pred Value : 0.92565
##              Prevalence : 0.11691
##          Detection Rate : 0.04811
##    Detection Prevalence : 0.07466
##       Balanced Accuracy : 0.69074
##
##        'Positive' Class : yes
##
```



When evaluated on the test set, boosting achieved the strongest overall performance among all models considered. It produced the highest accuracy (0.905), sensitivity (0.412), and AUC (0.927). In particular, boosting correctly identified over 40% of subscribing clients while maintaining high specificity. This represents a substantial improvement over logistic regression, KNN, and GAM, all of which struggled to identify the minority "yes" class. These results highlight the advantage of flexible ensemble methods in reducing bias and capturing complex nonlinear patterns in imbalanced marketing data.

Model Training, Formulation, and Validation

K-Nearest Neighbors required additional preprocessing due to its reliance on distance-based calculations. All numeric predictors were standardized to ensure equal contribution to the distance metric. The number of neighbors, k, was selected using 5-fold cross-validation by identifying the value that minimized classification error on the training folds.

Boosting was trained using a large number of trees to allow sufficient model flexibility, with the optimal number of trees selected using out-of-bag error estimation. This tuning procedure served a similar role to cross-validation by identifying the point at which additional complexity no longer improved performance. Across all models, final performance was assessed on the held-out test set using accuracy, sensitivity, and AUC. Because the dataset is highly imbalanced, particular emphasis was placed on sensitivity and AUC as measures of a model's ability to identify the minority "yes" class.

Results

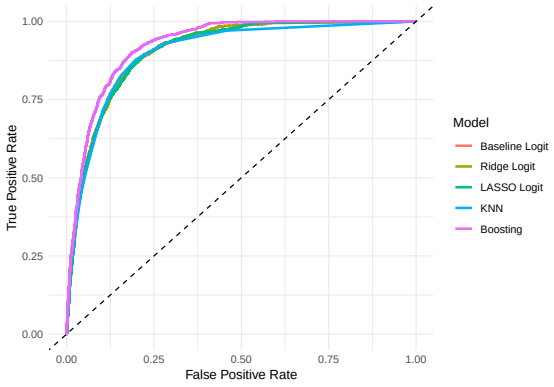
Model performance was evaluated on a held-out test set using accuracy, sensitivity, and AUC. Because the dataset is highly imbalanced, particular emphasis was placed on sensitivity and AUC as measures of a model's ability to identify the minority "yes" class.

```
##            Model Accuracy Sensitivity    AUC
## 1 Baseline Logit   0.8997      0.3302 0.9078
## 2    Ridge Logit   0.8974      0.2687 0.9078
## 3    LASSO Logit   0.8981      0.3009 0.9072
## 4            KNN   0.8978      0.2526 0.9014
## 5       Boosting   0.9047      0.4115 0.9266
```

Boosting outperformed all other models across all key performance metrics. It achieved the highest accuracy (0.9047), sensitivity (0.4115), and AUC (0.9266), indicating superior ability to identify clients likely to subscribe to a term deposit. This improvement in sensitivity is particularly important in the bank marketing context, where failing to identify potential subscribers represents a missed business opportunity.

The logistic regression models (baseline, Ridge, and LASSO) achieved similar accuracy levels (approximately 0.898–0.900) and strong AUC values (around 0.907), indicating good overall discrimination. However, their sensitivity remained relatively low, suggesting that these models frequently misclassified subscribing clients as non-subscribers. Regularization did not substantially improve performance, indicating that model bias rather than variance was the primary limitation.

KNN performed the weakest among all models, with the lowest AUC (0.617), reflecting poor discriminatory power in this high-dimensional setting. GAM demonstrated improved AUC (0.854) but still failed to substantially improve sensitivity, indicating limited gains from smooth nonlinear terms alone.



The ROC curves reinforce the numerical results. Boosting dominates across most thresholds, confirming its superior ranking ability. Logistic regression models cluster closely together, while KNN performs only slightly better than random classification.

Discussion

The results of this study highlight an important trade-off between predictive performance and interpretability when modeling imbalanced bank marketing data. Among all models evaluated, boosting achieved the strongest overall predictive performance, particularly in terms of sensitivity and AUC. This indicates that boosting is the most effective method for identifying clients who are likely to subscribe to a term deposit, which is the primary objective of the bank. By capturing complex nonlinear relationships and interactions among predictors, boosting substantially reduced the number of missed potential subscribers compared to simpler models.

However, while boosting offers superior predictive accuracy, it lacks transparency. In contrast, regularized logistic regression provides more interpretable results that can be directly translated into business insights. Coefficient estimates from the logistic models reveal how specific client characteristics influence subscription probability. For example, higher account balances were associated with lower marginal increases in subscription probability at extreme values, housing loan status was negatively associated with subscription, and student job status showed a strong positive association. These interpretable effects allow marketing teams to better understand customer behavior and design targeted strategies, even if the model itself is not the most accurate.

Interpretability is particularly important in the banking sector, where predictive decisions may be subject to regulatory oversight and ethical scrutiny. Models that can clearly justify why a client was targeted or excluded are easier to audit and defend. From this perspective, logistic regression remains a valuable tool despite its lower sensitivity. A practical strategy for the bank may involve using boosting as a primary screening model to identify high-potential clients, followed by logistic regression to provide explanations and support decision-making. Several limitations should be noted. The analysis did not incorporate time-aware modeling, even though campaign timing may influence client responses. Additionally, no resampling techniques were used to directly address class imbalance. Future work could explore methods such as SMOTE or cost-sensitive learning to further improve minority-class detection. Other extensions may include experimenting

with alternative ensemble methods or adjusting decision thresholds to better align predictions with business costs.

Overall, this study demonstrates that flexible ensemble models offer substantial gains in predictive performance for imbalanced marketing problems, while interpretable models remain essential for transparency and actionable insight. Balancing these two objectives is critical for effective and responsible deployment of predictive models in real-world banking applications.

References

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

James, Gareth, et al. An Introduction to Statistical Learning : With Applications in R. Springer, 2013.

Moro, Sérgio, et al. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, vol. 62, June 2014, pp. 22–31.