STATS5099 Data Mining and Machine Learning

Comparative Analysis of Classification Models for Predicting Term Deposit Subscriptions

Group 14

Guanjie Wang, Kumud Sharma, Tejaswi, Zilu Wang

Introduction

The study aims to predict the factors that impact the subscription of term deposit (yes/no). The data set contains 10,000 records from a Portuguese banking institution's directed marketing campaigns conducted via phone calls. These campaigns targeted clients to promote bank term deposits. The data set comprises information from 17 campaigns spanning May 2008 to November 2010.

Data Preprocessing

The process below describes the detailed data wrangling techniques used. On examination of the data set, it is observed that there are empty cells present. These cells are substituted by NA. The categorical columns are converted to factor type. From the summary statistics, missing values are observed to be present in categorical columns. These missing values are imputed by Mode. The column *default* has 2151 missing values. Since imputation in this column is done by using Mode, there are 9999 'no' values and 1 'yes' values. This imputation does not provide any relevant information. Hence, *'default'* will be omitted.

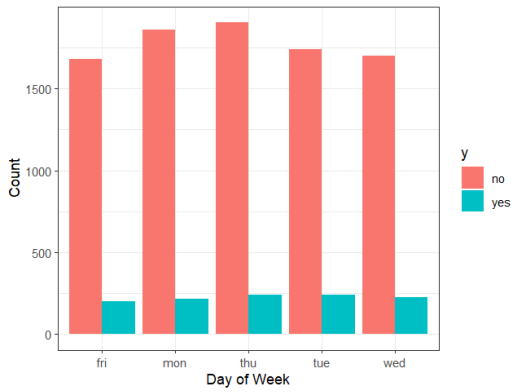

Figure 1 Histogram of Day of Week

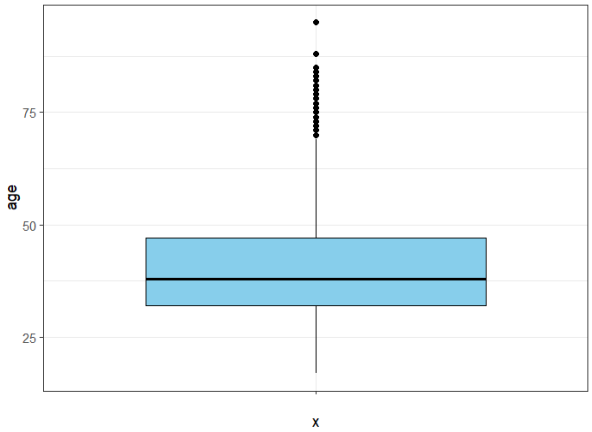

Figure 2 Boxplot of Age

From the Figure 1, the days of the week do not affect the response variable distribution. Due to this, *day_of_week* is omitted. This can further be verified by using chi-squared test of independence. From the summary statistics and the boxplot fig 2, *age* shows a large variation and many outliers. Logarithmic transformation is applied to facilitate a robust analysis. For column *pdays* it can be observed that is highly skewed, with most values concentrated in a specific range around '999'. This can reduce the effectiveness of the variable in splitting the data into homogeneous groups, which is essential for some classification models. Considering this, this column is not included henceforth. Binary variable is allotted to the response variable *y*.

Exploratory Data Analysis

The data frame is divided into a training set, validation set and test set using 50%-25%-25% division. The Exploratory data analysis is performed on the training data set.

Summary Statistics

Table 1 Summary of Categorical Data

skim_type	skim_variable	n	factor.n_unique	factor.top_counts
factor	y	5000	2	no: 4449, yes: 551
factor	job	5000	11	adm: 1326, blu: 1126, tec: 810, ser: 470
factor	marital	5000	3	mar: 3047, sin: 1377, div: 576
factor	education	5000	7	uni: 1696, hig: 1157, bas: 713, pro: 625
factor	housing	5000	2	yes: 2724, no: 2276
factor	loan	5000	2	no: 4263, yes: 737
factor	contact	5000	2	cel: 3139, tel: 1861
factor	month	5000	10	may: 1723, jul: 856, aug: 715, jun: 637
factor	poutcome	5000	3	non: 4330, fai: 515, suc: 155

Table 2 Summary of Numeric Data

Variables	Mean	Median	Std. Dev	Min	Max	IQR	Sample Size
age	3.66	3.64	0.25	2.83	4.48	0.38	5,000.00
campaign	2.56	2.00	2.70	1.00	39.00	2.00	5,000.00
previous	0.17	0.00	0.51	0.00	7.00	0.00	5,000.00
emp.var.rate	0.07	1.10	1.56	−3.40	1.40	3.20	5,000.00
cons.price.idx	93.58	93.80	0.58	92.20	94.77	0.92	5,000.00
cons.conf.idx	−40.58	−41.80	4.62	−50.80	−26.90	6.30	5,000.00
euribor3m	3.60	4.86	1.74	0.63	5.00	3.62	5,000.00
nr.employed	5,165.89	5,191.00	72.66	4,963.60	5,228.10	129.00	5,000.00

Table 1 shows the summary statistics of the categorical variables and Table 2 displays the summary statistics of the numerical variables. It is evident that the continuous variables exhibit significant variability, with notable differences observed in dispersion of their minimum and maximum values.

Data Visualisation

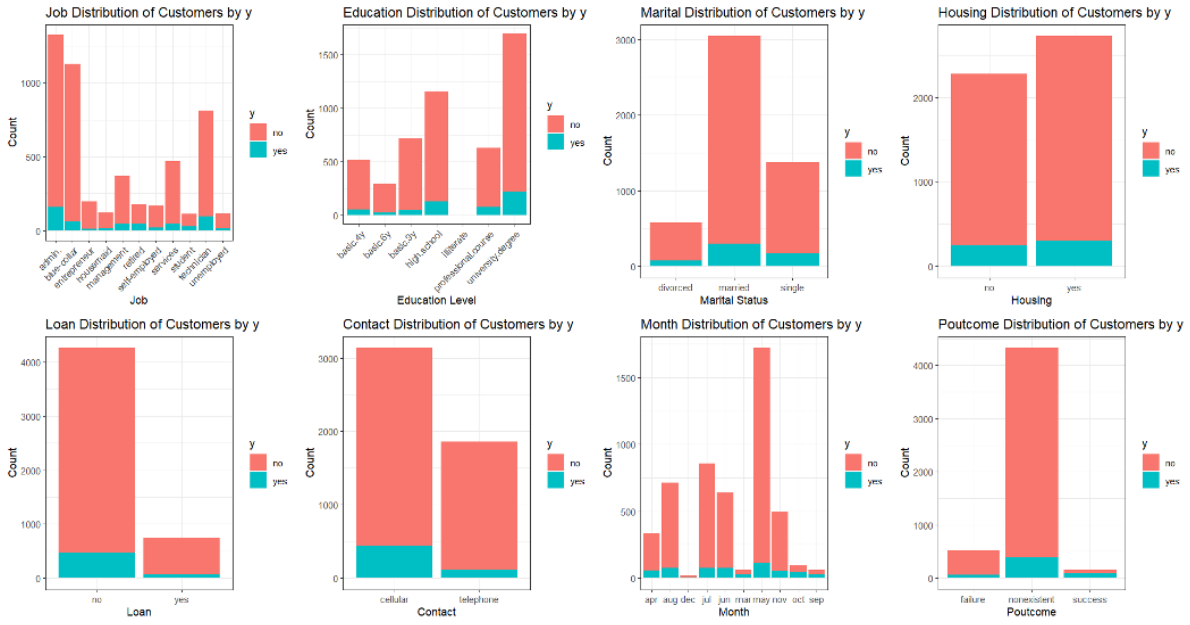

Figure 3. Ratio of yes and no in categorical variables

The visualisation to observe the data structures for all the variables are constructed in Figure 3. It can be observed:

- ***job***: More number of customers have jobs in admin followed by blue-collar
- ***education***: The maximum number of customers have a university degree. There are almost negligible illiterate customers (only 2)
- ***marital***: The marital status of maximum number of customers is married
- ***housing***: More customers have a housing loan, but the difference compared to those without is not significant
- ***loan***: Most customers do not have a personal loan but few customers do.
- ***contact***: More customers were contacted by the type 'cellular'
- ***month***: May was month of the year in which the most customers were contacted last
- ***poutcome***: The outcome of the previous marketing campaign is 'nonexistent' for the most customers
- **y**: It can be observed that the response variable 'y' is quite imbalanced

Correlation

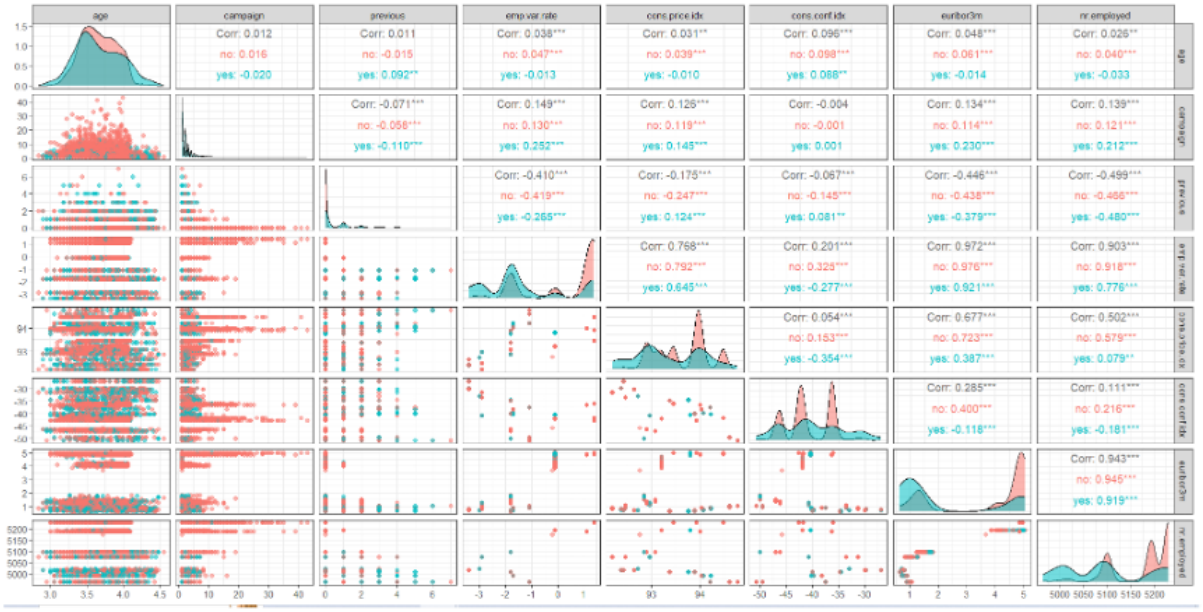

Figure 4. Correlation Matrix of Numeric Variables

The above displays the correlation between continuous variables. It can be observed there are mostly weak correlations. However, strong correlation can be noticed between *euribor3m* and *emp.var.rate*, *nr.employed* & *emp.var.rate* , *euribor3m* & *nr.employed*. Moderate correlation can be observed between *cons.price.idx* & *emp.var.rate*, *euribor3m* & *cons.price.idx*.

Test for Independence

To test if there is a significant association between the two categorical variables, chi-square test and fisher tests are conducted. For both the tests, the null hypothesis assumes that the variables

are independent, while the alternative hypothesis suggests otherwise. All the variables have $p < 0.05$ except *loan* and *housing*. This suggests that there is no significant association between the two categorical variables.

Outliers

The data is further explored to check if there are any outliers in the continuous data. Few outliers are observed in *age*. Logarithmic transformation has already reduced the skewness to certain extent. In *cons.price.idx* columns, 53 observations were outliers but for the values -26.9. Since the outliers are few, it is decided to keep them in the dataset. As housing and loan are found to be insignificant, they are omitted from the data sets. To conduct a statistical analysis for a data set containing both categorical and continuous variables, one-hot encoding is applied to categorical variables. In this method, the categorical variables are assigned a binary value of 1 or 0 so as to achieve a new column for each category of each categorical variable.

Classification Methods

k-Nearest Neighbours (KNN)

K-Nearest Neighbours is a popular classification method in machine learning. It learns from the training data by memorising the features of each point, calculating the distance between points, choosing optimal k and finally makes predictions on validation data by predicting the class as the class label of majority of the k nearest points (neighbours).

An interesting point worth noticing is that KNN does not assume a parametric model for the underlying data. This characteristic facilitates easy understanding of predictions. It is utilised in this study to categorise the client subscription for term deposits. The method's capacity for pattern recognition is another reason for its application. Given its training on various factors, its principle that similar data points belong to the same class is useful for forecasting whether or not a client would subscribe to a term deposit. It should also be noted that KNN assumptions of feature scaling, stable choice of distance and careful consideration of k optimal have been addressed in the analysis.

It was previously observed that the features in the data have different ranges and standard deviations. Since features with wide ranges and standard deviations might dominate the

computation, adopting Euclidean Distance raises certain challenges. To deal with this, all features have been standardised so that they have zero mean and a variance of one.

Moving forward the most important decision to be made when predicting classes using KNN is the choice of k. Given that the response variable is binary, an odd values for K should be chosen to avoid ties. To begin with, the correct classification rate on the validation set for different values of K is evaluated.

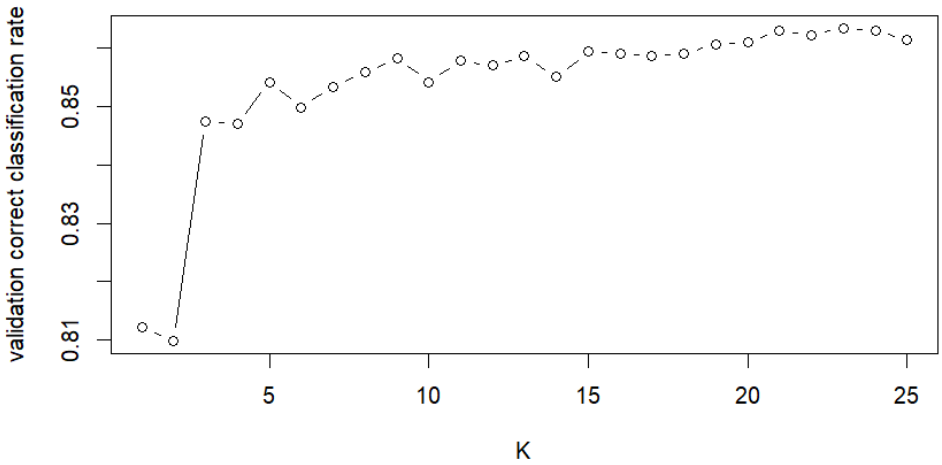

Figure 5. Correct Classification rates on the validation set for different values

Figure 5 plots the validation accuracy against different values of k to visually analyse the significance of k on model performance. The optimal k value would be the one which gives the highest correct classification rate, in our case, $k = 23$ is the optimal choice. This step also makes predictions on the validation set which is utilised in assessing the performance of KNN as a classification method.

Table 3 Confusion Matrix for KNN

	Actual.Yes	Actual.No
Predicted.Yes	46	22
Predicted.No	251	2181

Table 4 Performance *Metrics for KNN*

Metric	Value
Accuracy	0.89
Precision	0.68
Recall	0.15
F1	0.25

Table 4 shows high overall accuracy of 0.89, indicating it correctly classified 89.08% of the cases in the validation set. This suggests the model is, in general, good at distinguishing between “yes” and “no” classes. A precision of 67.64% further indicates a moderate proportion of positive predictions were actually positive. However, a low recall of 15.49% indicates

missing substantial proportion of actually positive cases. Overall, the low F1 score indicates even though the method avoids many false positive, it misses many true positives. This indicates a weak overall performance of KNN on the bank marketing data.

Linear Discriminant Analysis/ Quadratic Discriminant Analysis

Linear Discriminant Analysis is a classification method used to define linear relationships between features. It assumes that the features are normally distributed and begins by modelling the features, x, in classes G=g. The most important step in LDA is checking the validity of assumptions before actually applying LDA.

In the context of the bank marketing data, it is observed in EDA that very few variables like *age*, *cons.price.idx* and *cons.conf.idx* fairly follow a Gaussian distribution while others do not. To further analyse the validity of LDA the class-covariance matrix is examined. In this case, it is observed that there is large difference in variances indicating QDA might be a better option.

Quadratic Discriminant Analysis (QDA) is an extension of Linear Discriminant Analysis which models categorical variables, especially those having non-linear decision boundary and unequal variance in grouped features. The violations thus deem LDA as an inappropriate choice for the bank marketing dataset. However, it would be interesting to observe LDA, followed by QDA on three of the features which satisfy the Gaussian distribution. These features are age, *cons.price.idx* and *cons.conf.idx*.

After building the two cases on training data and predicting the results on validation data, it is observed that approximately 12.56% of the predictions made by the QDA model on the validation dataset were incorrect, which is higher as compared to LDA with 11.88%. On further analysis of confusion matrix, it is observed that out of 2203 clients who did not subscribe for the term deposit, QDA misclassifies 82 (3.72%), which implies slightly worse results as compared to LDA which misclassified zero. However, out of 297 clients who said yes to subscription, QDA incorrectly predicts 232 (78.11%) of them as no, this indicates better results than LDA which misclassifies all.

To understand the results visually, the ROC curves for both the methods are visualised below.

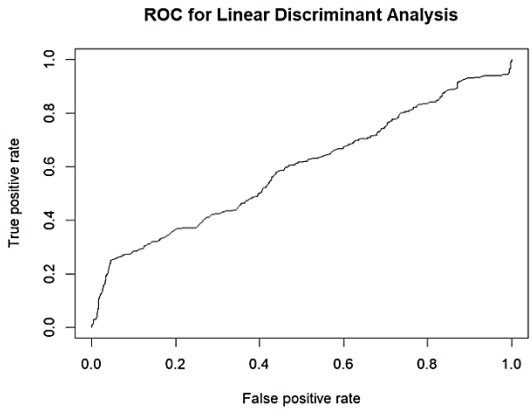

Figure 6. ROC for LDA

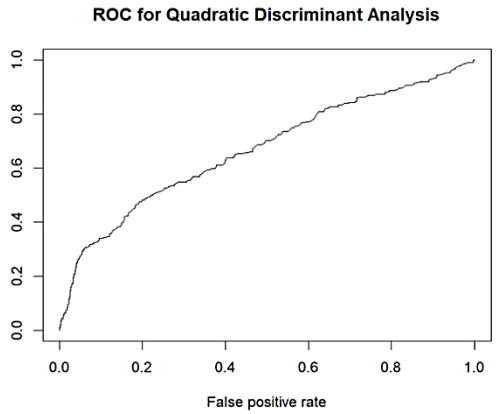

Figure 7 ROC for QDA

The ROCs indicate that the methods are far from a good classification technique, it indicates slightly better results for QDA as compared to LDA. The AUC for the data subset being considered is closer to random guess for QDA, 0.67, but still better than that of LDA 0.59. Hence providing evidence of QDA being a better classifier relative to LDA, however not so good a classifier in general.

In conclusion, it is extremely important to consider that LDA and QDA analysis was conducted only on a subset of features which fairly followed a multivariate Gaussian distribution. Hence, they cannot be considered as an overall classifier for the subscription responses. Even for the few variables of interest, the results don't seem very impressive. Although QDA does a slightly better job than LDA, both of them are not very good classifiers for the whole data in general as the assumption of normality is violated in the data.

Support Vector Machines (SVM)

The purpose of a Support vector machines (SVMs) is to fit a hyperplane that divides the points of the two classes into distinct zones. By kernel trick, it can be mapped to a new space to solve the classification problem of the original space: the resulting function is linear in new space, but non-linear in the primitive. This method is intended for the binary classification, which is suitable for this data set and it is compatible to deal with the categorical variables, so there is no requirement to convert them to numerical variables.

The choices of kernel functions are linear kernel (model 1), polynomial kernel (model 2), radial basis function (RBF) kernel (model 3) and sigmoid kernel, the first three are chosen and the

classification for different values of parameters is fitted, the one with the smallest validation prediction error is finally chosen for each function.

- The model fulfils the assumptions
- 'type = C-classification' is used since the aim is to classify
- All functions specify the cost parameter
- Polynomial kernel adds a parameter Degree, when degree > 1, we are fitting a hyperplane in a higher-dimensional space involving polynomials of degree
- In RBF kernel, gamma is positive constant which defines the influence of a single sample when mapping.

Table 5 Performance of models based on different kernels

Metric	Model_1	Model_2	Model_3
Sensitivity	0.645	0.741	0.645
Specificity	0.907	0.897	0.907
Accuracy	0.899	0.895	0.899
F1 Score	0.283	0.141	0.283

According to Table 5, if we care most about Sensitivity, then it is better to choose the polynomial kernel. But if F1 score is set to be used to compare with other models, then linear kernel or radial kernel may be chosen.

The linear kernel and radial kernel perform so similarly on classification decision boundaries may because that if the number of features is large, one may not need to map data to a higher dimensional space. That is, the nonlinear mapping does not improve the performance. Using the linear kernel is good enough, and one only searches for the parameter C.

Decision Trees

Decision trees are versatile supervised learning algorithms, constructing hierarchical structures that partition data based on feature values to predict outcomes. In this section, a decision tree model was constructed and tuned using various stopping criteria. Additionally, tree pruning techniques, including minimum error reduction and smallest tree size, are applied to prevent overfitting and simplify the model. Evaluating these models using metrics like sensitivity, specificity, and AUC helps in understanding their predictive power and generalizability, guiding the selection of the most effective model configuration.

A base decision tree was constructed using the rpart package in R. Then, parameter tuning is conducted to enhance its performance. The tuning process involved exploring a grid of

parameters, also referred as stopping criteria, including cp, minsplit, minbucket, and maxdepth. Two tree pruning strategies were utilized to reduce overfitting and simplify the model, minimum error and smallest tree.

Table 6 Performance Metrics for Tree Models

Model	Accuracy	Sensitivity	Specificity	F1_Score	AUC
Base Tree	0.892	0.185	0.988	0.290	0.713
Base Tree Tuned	0.892	0.185	0.988	0.290	0.716
Pruned Tree Min Error	0.892	0.185	0.988	0.290	0.713
Pruned Tree Smallest	0.892	0.172	0.989	0.274	0.702

The performance metrics show that tuning and pruning minimally affect the decision tree models' accuracy and specificity, which remain high across all versions. However, there is a slight drop in sensitivity and F1 score from the base to the smallest pruned tree, indicating a diminished ability to identify positive cases. The AUC, which measures overall class distinction, varies little, with a small dip for the smallest pruned tree, suggesting that pruning simplifies the model but marginally lowers predictive performance.

Bagging and Random Forest

In advancing the ensemble methods, bagging was initially employed to reduce variance and enhance model robustness. Building on this, the Random Forest method was applied, incorporating feature selection randomness at each split. This added layer of randomness aimed to reduce over-fitting further and improve model generalization. Parameters within the Random Forest, including the number of trees and the maximum number of features considered at each split, were systematically tuned based on model performance metrics.

Bagging

A bagging technique was utilized to reduce variance and enhance robustness by generating multiple trees from different data set samples. The model was then fine-tuned to improve the F1-score, thereby optimizing the balance between precision and recall.

Random Forest

The Random Forest method builds on bagging by introducing randomness in feature selection at each tree split, enhancing diversity, and reducing over-fitting. This process improves model generalization by preventing reliance on any single feature, allowing for the capture of complex relationships.

Table 7 Performance Metrics for Bagging and Random Forest

Model	Accuracy	Sensitivity	Specificity	F1_Score	AUC
Tree Bagging	0.886	0.256	0.971	0.349	0.749
Tree Bagging Tuned	0.888	0.249	0.975	0.347	0.760
Random Forest	0.890	0.226	0.980	0.328	0.769
Random Forest Tuned	0.892	0.242	0.980	0.348	0.777

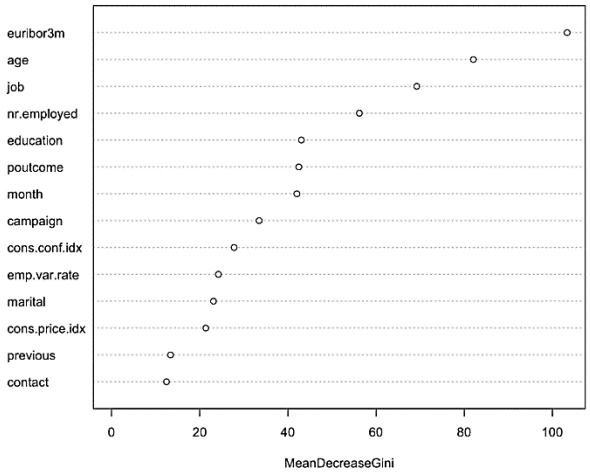

Figure 8. Feature Importance for Random Forest

Top Features: The variables *nr.employed, euribor3m, age, job*, and *education* appear to be the most significant in influencing the model's predictions. These features are likely driving the majority of the predictive power of the model, reflecting economic conditions (*nr.employed* and *euribor3m*), as well as demographic and job-related factors (*age, job, education*).

Economic Indicators: The prominence of *nr.employed* and euribor3m highlights the impact of economic factors on the decision to subscribe to term deposits. This aligns with expectations as economic conditions often play a significant role in financial decision-making.

Demographic and Job-related Factors: The significance of *age, job*, and *education* underscores the im-portance of demographic and employment characteristics in influencing a client's likelihood to subscribe, suggesting targeted marketing strategies could be effectively.

Results

In the evaluation of machine learning models, the Random Forest model outperformed KNN, SVM, and Decision Tree variants in our banking dataset analysis. It showcased the highest F1 score (0.348) and a well-balanced sensitivity (0.242) and specificity (0.978).

Table 8: Performance Metrics for Chosen Model (Random Forests)

Metric	Value
Accuracy	0.902
Sensitivity	0.278
Specificity	0.980
F1 Score	0.387
AUC	0.783

This led to its selection as the optimal model. When applying this model to the test set, the Random Forest model achieved an accuracy of 0.902, sensitivity of 0.278, specificity of 0.980, an F1 score of 0.387, and AUC of 0.783, confirming its robustness and effectiveness for the task.

Conclusion

In our comparative study, KNN, LDA/QDA, SVM and Tree-based models are employed to predict whether a client is subscribed to a term deposit. The Random Forest model emerged as the most effective with the highest F1 score, which is particularly significant given the imbalanced nature of our dataset, showcasing a good balance between predictive metrics. The feature importance analysis in our Random Forest model highlights the impact of economic indicators, such as *'nr.employed'* and *'euribor3m'*, and demographic factors like *'age'*, *'job'*, and *'education'* on bank deposit subscriptions. These findings suggest that both broader economic conditions and personal attributes significantly influence customer decisions, guiding targeted strategies for marketing bank products. The Random Forest model excels due to its ensemble approach, which combines multiple decision trees to reduce variance and prevent overfitting, leading to more reliable predictions. Its effectiveness is further enhanced by its ability to handle diverse variable scales without pre-processing, crucial for our dataset with its wide-ranging economic and demographic factors.

References

1. Harrison, O. (2018) *Machine Learning Basics with the K-Nearest Neighbors Algorithm.* [Online Article]. Available from: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.
2. IBM. (2023) *K-Nearest Neighbors (KNN).* [Online Resource]. Available from: https://www.ibm.com/docs/en/db2oc?topic=procedures-k-nearest-neighbors-knn.
3. Skand, K. (2017) *kNN(k-Nearest Neighbour) Algorithm in R.* [Online Article]. Available from: https://rstudio-pubs-static.s3.amazonaws.com/316172_a857ca788d1441f8be1bcd1e31f0e875.html.
4. Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. (2016) *A Practical Guide to Support Vector Classification.* [Department of Computer Science, National Taiwan University]. Taipei 106, Taiwan, May 19. Available from: http://www.csie.ntu.edu.tw/~cjlin.
5. Xiaochen Yang. (2024) *Data Mining and Machine Learning 2023-24.* [Online Resource]. Available from: https://moodle.gla.ac.uk/course/