UNIVERSITÀ
di VERONA

Machine Learning Project Report

- **Project Title:** Predicting Bank Term Deposit Subscription
- **Master Degree:** Artificial Intelligence
- **Course:** Machine Learning and Deep Learning
- **Academic Year:** 2024/2025
- **Matricula Number:** VR528190
- **Author:** Shah Rukh Aleem

Table of Contents

1. Motivation and Rationale

Predicting customer behavior has become a critical component of financial marketing plans. The project will define the prediction of whether a customer will choose a bank's term deposit. Such predictions give power to the banks to design specific marketing campaigns, allocate resources appropriately, and increase conversion rates. The purpose for this project is to use data-driven insights to improve decision-making and reduce costs associated with ineffective outreach activities. Identifying clients who are likely to subscribe will increase campaign efficacy while adhering to ethical and responsible marketing norms.

2. State of the Art

Machine learning has emerged as a critical tool in financial services predictive analytics, allowing organizations to extract actionable insights from complicated datasets. Ensemble techniques, particularly Random Forest and Gradient Boosting, are frequently used owing to their resilience, scalability, and better performance in dealing with non-linear patterns and unbalanced data.

Traditional models, such as Logistic Regression, although simple to understand, often struggle with unbalanced datasets and complicated interactions owing to linear assumptions (Pedregosa et al., 2011). These models tend to favor the dominant class, resulting in low recall for the minority class, which is often the most important class in campaign targeting.

Considering the popularity of ensemble models, many difficulties remain underappreciated in the literature, such as:

- Class imbalance: Majority class dominance distorts model learning, resulting in high accuracy but low minority class recall.
- Threshold optimization: Standard decision thresholds (e.g., 0.5) may not be optimum in skewed datasets, masking the model's real capacity.
- Interpretability trade-offs: While tree-based techniques provide some transparency, deeper models present explainability issues for corporate stakeholders.

This research overcomes these constraints by combining Random Forest with tailored preprocessing (such as Tomek Links for under-sampling) and decision threshold tweaking. The purpose is to assess performance using a wide range of variables and present a realistic, explainable, and successful model for real-world financial marketing applications.

3. Objectives

3.1. General Objective: Develop a reliable machine learning model to predict bank term deposit subscriptions.

3.2. Specific Objectives:

- Comprehend and preprocess the dataset to guarantee quality inputs for the modeling process.
- Effectively handle missing data and encode categorical variables.
- Use advanced methods to balance the dataset and address class imbalance.
- Train and compare different classifiers (Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest) using optimized hyperparameters to get the best model.
- Evaluate each model's performance using multiple metrics, examine feature importance for Random Forest, and compare evaluation findings to help with model selection.

4. Methodology

4.1. Dataset Description

The dataset is sourced from the Bank Marketing Dataset available in the UCI Machine Learning Repository. It comprises 45,211 entries and 17 variables, with the target variable denoted as y, which indicates whether a customer has subscribed to a term deposit ('yes' or 'no'). The features encompass demographic data (such as age, occupation, and marital status), financial information (including account balance), and campaign-specific variables (like contact duration and the number of previous contacts).

4.2. Data Preprocessing

4.2.1. Handling Missing Values:

- Columns containing "unknown" values, such as 'job' and 'education', were filtered to exclude rows with these entries.
- Non-essential columns, including 'contact' and 'poutcome', were removed because they had a large proportion of unknown values (28.79% and 81.74% respectively), which could introduce noise into the model.

4.2.2. Encoding Categorical Data:

- Categorical variables (`job`, `marital`) were label encoded using sklearn's LabelEncoder.

4.2.3. Feature Scaling:

- Numerical features were standardized using StandardScaler to enhance model convergence.

4.2.4. Handling Class Imbalance:

Tomek Links was used to increase recall for the minority class ('Yes') by reducing overlapping majority class occurrences. Unlike SMOTE, which produces synthetic samples, Tomek Links retains data authenticity while potentially lowering dataset size a trade-off justified by its efficiency in minimizing false negatives (Tomek 1976).

Dataset	Class 0 (No)	Class 1 (Yes)	Imbalance Ratio
Before Tomek	38,172	5,021	7.6:1
After Tomek	25,878	3,515	7.4:1

Table 1: Impact of Tomek Links on Class Distribution and Imbalance Ratio

Tomek Links reduced the majority class by 32.2% while retaining 70% of the minority samples, somewhat lowering the imbalance ratio from 7.6:1 to 7.4:1. This tradeoff favored recall over dataset size.

5. Model Selection and Training

This research examined four machine learning algorithms: logistic regression, support vector machine (SVM), K-nearest neighbors (KNN), and random forest. To guarantee fair comparison and optimum performance, each model had its hyperparameters optimized using GridSearchCV.

5.1. Hyperparameter Tuning:

- **Logistic Regression:**
  `{'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}`
  (L2 regularization with a small C value helps prevent overfitting.)

- **SVM:**
  `{'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}`
  (RBF kernel with higher C allows fitting more complex decision boundaries.)

- **KNN**:
  `{'n_neighbors': 7, 'weights': 'distance', 'metric': 'manhattan'}`
  (Uses distance-weighted voting and Manhattan metric for improved robustness.)

- **Random Forest**:
  `{'max_depth': None, 'min_samples_split': 2}` *(default settings)*
  (Offers flexibility and avoids overfitting while maintaining performance.)

Furthermore, a threshold change of 0.5 to 0.25 was made to the Random Forest model to boost recall on the minority class ('Yes'), enabling it to discover more affirmative instances at the expense of accuracy.

5.2. Model Selection:

Random Forest outperformed the other models on various parameters (AUC, recall, and accuracy). Its robustness, scalability, and capacity to handle non-linear interactions make it the best fit for this classification assignment. Furthermore, it gives useful insights about feature relevance, which aids in interpretation and future enhancements.

6. Experiments and Results

6.1. Performance Comparision of Models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.891	0.576	0.252	0.351	0.870
SVM	0.897	0.596	0.356	0.446	0.864
KNN	0.896	0.598	0.331	0.426	0.833
Random Forest	**0.901**	**0.617**	**0.385**	**0.474**	**0.919**
Random Forest Adjusted Threshold	0.876	0.480	0.782	0.595	0.919

Table 2: Performance Comparison of Machine Learning Models

Random Forest topped the other three models in terms of accuracy, recall, F1 score, and ROC AUC. To enhance minority class detection, the Random Forest model's decision threshold was adjusted from 0.5 to 0.25. This significantly increased recall from 0.385 to 0.783, a 103% improvement, while precision dropped from 0.617 to 0.481. Despite a slight drop in accuracy (from 0.901 to 0.877), the F1 score improved to 0.596, making this version of Random Forest

the most effective at identifying customers likely to subscribe. The ROC AUC remained stable (0.919), indicating no degradation in model discrimination. These measures demonstrate its high capacity to accurately identify positive instances, which is particularly crucial in an unbalanced classification problem.

While SVM and KNN had somewhat greater accuracy (0.596 and 0.598, respectively), they fell short on recall and F1 score, indicating that they are more cautious in predicting the positive class. Logistic Regression had the lowest recall (0.252) and F1 score (0.351), although having a reasonable ROC AUC (0.870), showing that it differentiates across classes well but has difficulty properly identifying the minority class. Overall, Random Forest provided the best balance between precision and recall, making it the most suitable model for this problem.

6.2. Confusion Matrix

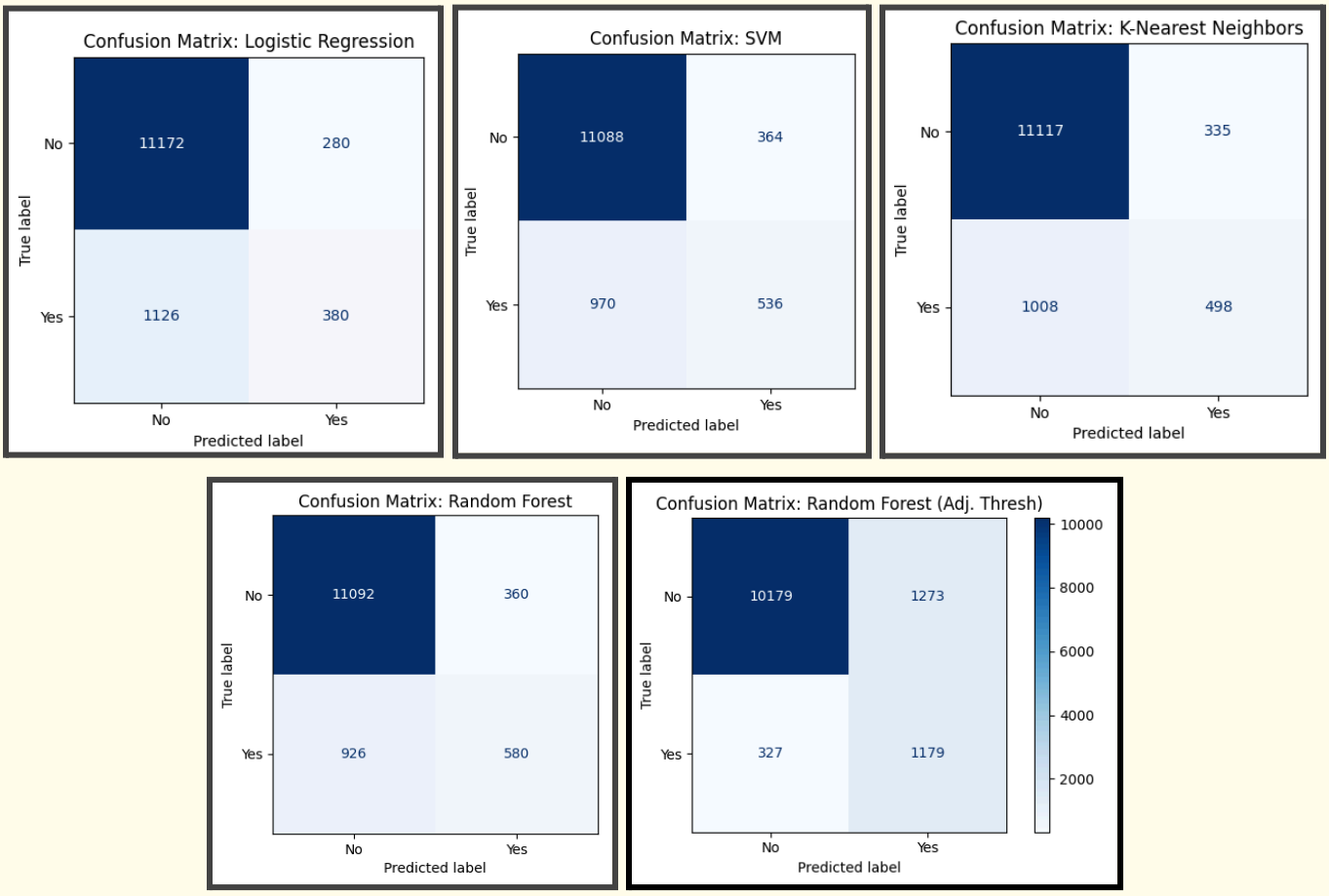

Figure 1. Confusion Matrices of All Models

The confusion matrices help to understand each model's classification performance. Random Forest obtained the greatest balance, with 580 true positives and the fewest false negatives (926). In contrast, Logistic Regression produced the most false negatives (1126), while SVM and KNN performed moderately. Whereas, threshold adjusted Random Forest significantly improved recall, achieving 1179 true positives and reducing false negatives to 327 the lowest among all models. This validates Random Forest's better ability to accurately identify affirmative situations.

6.3. Evaluation Metrics

The performance of the model was measured using metrics like accuracy, precision, recall, F1-score, and ROC-AUC. These measures give a notion of the model's quality and the balance involved in its classification.

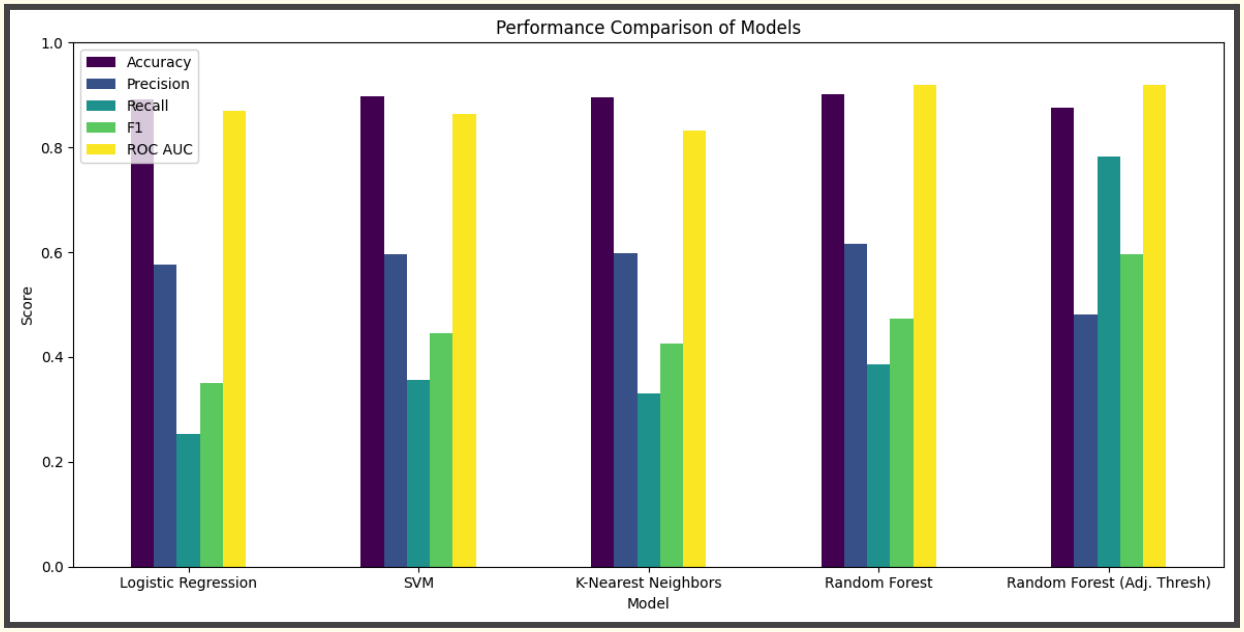

Figure 2: Bar Graph Comparing Model Performance

6.4. Feature Importance:

The most important feature, 'duration,' is a strong predictor but is only available after a customer interaction. A variant model excluding this feature could help build a real-time prediction (before calling the customer) decision tool, other features like balance, age, and previous campaign performance must be relied upon.

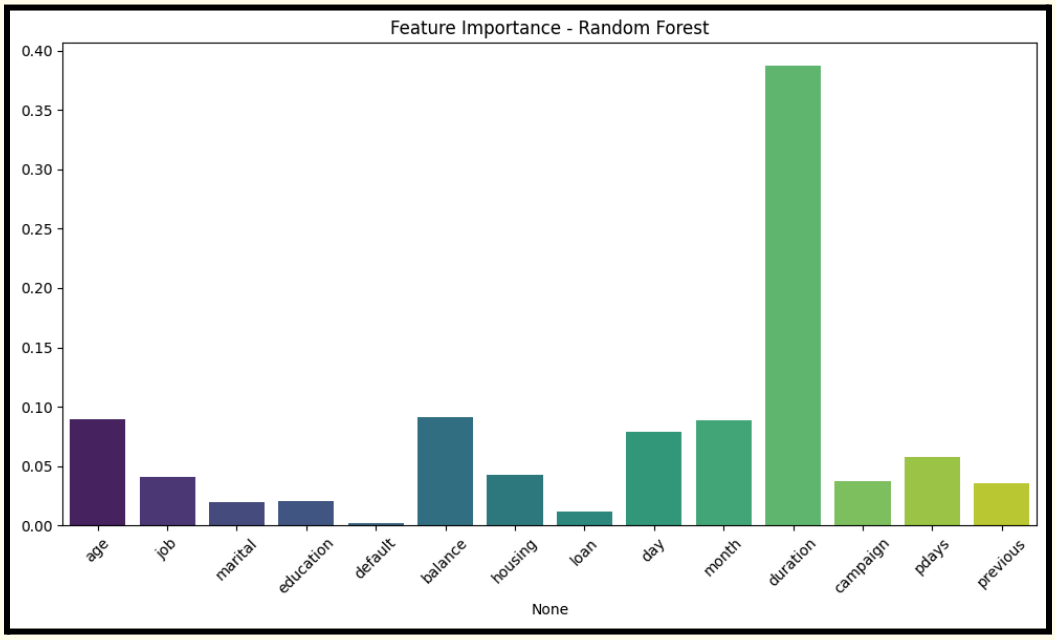

Figure 3: Feature Importance from Random Forest Model

6.5. Visualizations:

ROC Curve: The ROC curve represents each model's ability to discriminate between the two classes. As indicated, Random Forest has the greatest AUC (0.92) and hence the best overall performance, followed by Logistic Regression (0.87), SVM (0.86), and KNN (0.83). Random Forest's curve regularly outperforms the others, demonstrating its competence in dealing with class imbalances and producing correct forecasts.

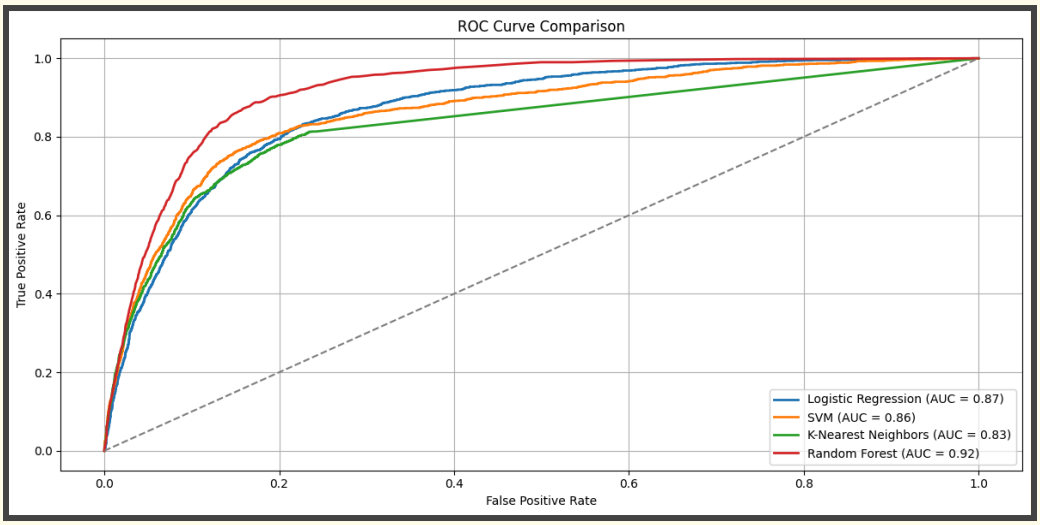

Figure 4: ROC Curve Comparison of Classifiers

Precision-Recall Curve: The Precision-Recall curve demonstrates performance under class imbalance. Random Forest beats other models, with an AUC of 0.56, followed by SVM (0.51), KNN (0.49), and Logistic Regression (0.46). This shows that Random Forest retains superior accuracy across different recall levels, making it more appropriate for unbalanced classification in this scenario.

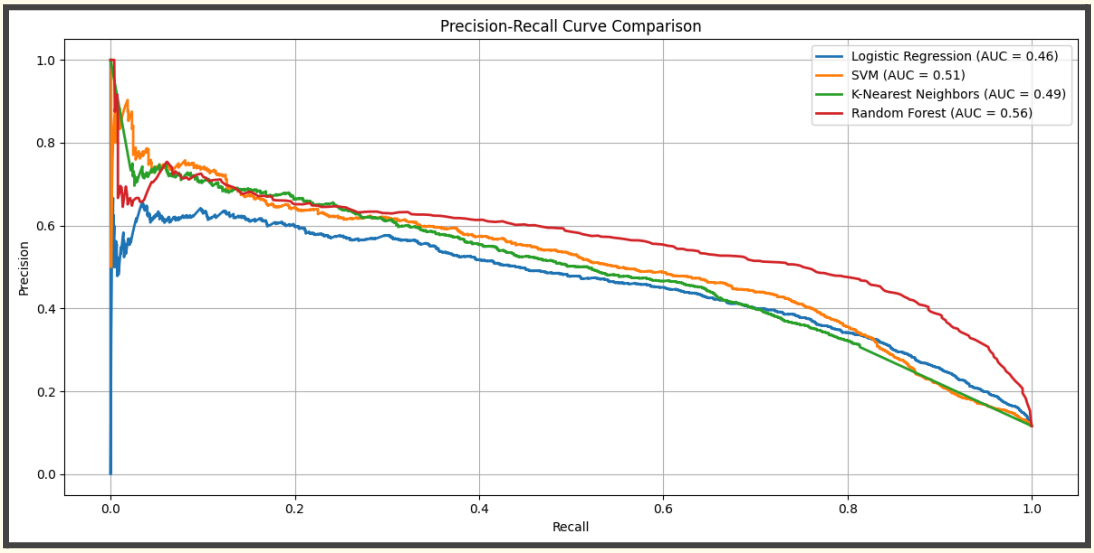

Figure 5: Precision-Recall Curve Comparison of Classifiers

6.6. Discussion

The comparative analysis of multiple classification models reveals the strengths and weaknesses of each technique in estimating bank term deposit subscriptions, especially in the setting of an unbalanced dataset.

The Random Forest classifier outperformed the conventional models in terms of accuracy (90.1%) and ROC AUC (0.919). It achieved a balanced trade off between accuracy (0.617) and recall (0.385), surpassing Logistic Regression, SVM, and KNN across all measures. Despite this high overall performance, recall was still quite low, suggesting that many real subscribers were being overlooked.

To solve this, we designed a threshold adjusted Random Forest model, which reduced the decision threshold from 0.5 to 0.25. This considerably increased recall to 0.783, which nearly doubled the capacity to accurately identify positive situations when compared to the previous model. Although this resulted in lower precision (0.481) and a minor decrease in accuracy (87.7%), the F1 score increased significantly to 0.596, suggesting a better balance of precision and recall. Importantly, the ROC AUC remained constant (0.919), indicating that the model's core classification ability was unaltered.

This threshold-adjusted RF model also had the lowest false negative rate among all models, as seen by its confusion matrix, which revealed a significant increase in real positive detections (1179 vs. 580 in conventional RF). This is a critical improvement in practical terms, since accurately identifying prospective subscribers is usually more beneficial than eliminating false positives in marketing situations.

Overall, our results highlight the significance of post training threshold tweaking, particularly in situations where recall is more important than accuracy. The adjusted threshold Random Forest is a more effective and practical method for subscription prediction for term deposits, with the objective of maximizing customer acquisition.

7. Conclusions

Multiple machine learning models were used and tested in this research to predict whether a consumer will sign up for a bank term deposit. Four classifiers were trained and evaluated using meticulous preprocessing, class imbalance management, and hyperparameter tuning: Logistic Regression, SVM, K-Nearest Neighbors, and Random Forest.

Random Forest consistently demonstrated superior performance comparetively to other models in AUC and overall balance among precision and recall. Furthermore, the adjusted-threshold variation had the greatest recall (0.783) and F1 score (0.596), making it particularly successful at detecting future subscribers in an unbalanced context. This shows how threshold adjustment may improve real world applicability without retraining the

model. This is critical in real world circumstances because recognizing prospective subscribers is more important than merely increasing accuracy. The research illustrates how, with adequate data preparation, model tweaking, and assessment, machine learning can give considerable insights and decision assistance for banks marketing efforts.

8. Bibliography

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
2. Lichman, M. (2013). UCI Machine Learning Repository: Bank Marketing Dataset. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
4. Tomek, I. (1976). Two Modifications of CNN. IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(11), 769–772.
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
6. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Springer.