Enhancing Bank Direct Marketing Through Machine Learning

Report submitted to SASTRA Deemed to be University As per the requirement for the course

CSE425 : MACHINE LEARNING ESSENTIALS

Submitted by

Pothireddy Chandrahas Reddy
(Reg No: 125018015, B. Tech Computer Science and Business Systems)

OCTOBER- 2024

SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act,1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

SCHOOL OF COMPUTING

THANJAVUR, TAMIL NADU, INDIA – 613 401

Table of Contents

Abstract	1
Introduction	1
Related Work	2
Methodology	3
Learning Outcome	8
Result	9
Code	10

ABSTRACT

This project focuses on enhancing bank telemarketing performance by predicting the success of calls aimed at selling long-term deposits. Using data mining techniques and a dataset from a Portuguese bank, machine learning models such as Logistic Regression (LR), KNN and Random Forest are implemented to classify customers based on their likelihood of subscription. The results provide valuable insights for banks to improve their telemarketing campaigns, reducing costs while improving effectiveness.

INTRODUCTION

Importance of Dataset:

The dataset is crucial for understanding patterns in client subscription behavior for term deposits.

Project Objective:

The objective of this project is to develop a predictive model that can estimate the likelihood of success in a telemarketing campaign targeting long-term bank deposits. With a focus on optimizing marketing efforts, the predictive model aims to help the bank prioritize customers who are more likely to subscribe to the offered product, thereby reducing wasted efforts and costs in unsuccessful calls. The study also explores how different machine learning models perform in predicting customer behavior based on historical data collected from the bank's telemarketing efforts. The key goal is to identify the most efficient machine learning algorithm and the most relevant features that drive successful telemarketing outcomes.

Problem Formulation:

Bank telemarketing campaigns face the challenge of reaching out to thousands of customers, many of whom may not be interested in the offered products. Randomly contacting customers without predictive insights leads to inefficient use of resources, increased customer dissatisfaction, and high operational costs. Thus, a data-driven approach is necessary to accurately predict which customers are more likely to respond positively to a telemarketing call, specifically for long-term deposit products.
The primary problem is a binary classification task: determining whether a telemarketing call will result in a "success" (customer subscribes to the deposit) or "failure" (customer declines or is unreachable). The study aims to address this issue by analyzing a dataset from a Portuguese bank and applying machine learning models to predict outcomes before calls are made, ensuring more effective targeting.

Related Work

References:

https://archive.ics.uci.edu/dataset/222/bank+marketing

https://www.semanticscholar.org/paper/cab86052882d126d43f72108c6cb41b295cc8a9e

Methodology

Dataset:

The dataset used for this project is the Bank Marketing dataset from the UCI Machine Learning Repository. It contains 52,944 records collected from a Portuguese retail bank between 2008 and 2013. Each record corresponds to a telemarketing call and includes features related to both the customer and the telemarketing process.

The dataset includes various types of features, such as:

Client information: Age, job, marital status, education level, etc.

Call details: Duration of the call, whether the call was inbound or outbound, and previous contact history.

Economic context: Indicators such as the Euribor 3-month rate, consumer confidence index, and employment variation rate.

The dataset is highly imbalanced, with only 12.38% of the records corresponding to successful outcomes (i.e., where the customer subscribed to the deposit). To address this, appropriate data processing techniques were applied to ensure robust model training and evaluation.

Machine Learning Models:

Four different machine learning models were implemented in this study:

Logistic Regression (LR): Logistic regression is a simple and interpretable model used for binary or multi-class classification problems. It predicts the probability that a given input belongs to a specific class by modeling the relationship between the independent variables and the probability of the target class using the logistic function. This algorithm is effective for problems where the classes are linearly separable, making it a popular choice for tasks like medical diagnosis or customer churn prediction. However, it struggles with complex, non-linear relationships in data, as it assumes a linear boundary between the classes.

K-Nearest Neighbor (KNN): K-Nearest Neighbors (KNN) is an intuitive and straightforward algorithm used for both classification and regression tasks. It classifies a new data point based on the majority label of its K nearest neighbors, where distance is typically measured by Euclidean distance. Since KNN does not require training, it is easy to implement, but it can be computationally expensive during prediction, especially for large datasets. KNN works best for small datasets with well-defined clusters but struggles with high-dimensional data and is sensitive to the choice of K and distance metrics.

Random Forest (RF): Random Forest is a powerful ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each tree in the forest is built on a random subset of the data and features, and the final prediction is based on the majority vote (for classification) or average (for regression) of the individual trees. This approach allows Random Forest to capture complex, non-linear patterns in the data and handle large datasets effectively. While it is robust and highly accurate, Random Forest can be computationally intensive and is less interpretable compared to simpler models like logistic regression.

MATH BEHIND THE MODELS:

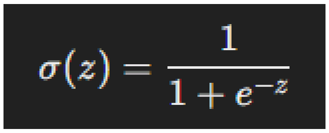
1. Logistic Regression:

Logistic Regression is a linear model used for classification, especially binary classification. Despite its name, it’s a classification algorithm, not a regression one.

Key Concepts:

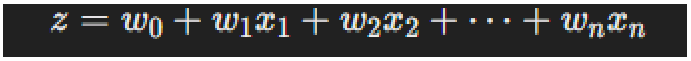
- **Logistic Function (Sigmoid Function):**

  The output of logistic regression is modeled using a sigmoid function, which maps any real-valued number to the range (0, 1).



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

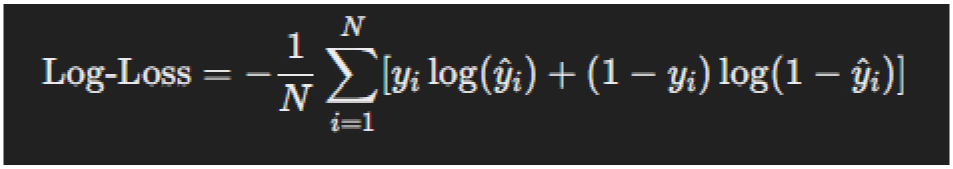
  where z is the linear combination of input features:



$$z = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

- **Loss Function (Log-Loss or Cross-Entropy Loss):**

  Logistic regression is typically optimized using the **log-loss** or **binary cross-entropy loss** function:



$$\text{Log-Loss} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

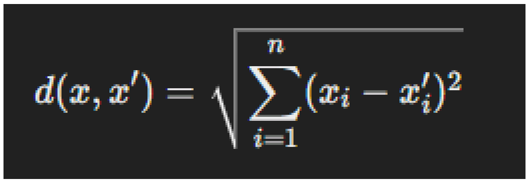
2. K-Nearest Neighbors (KNN):

KNN is a **non-parametric, instance-based learning** algorithm used for classification and regression. It works based on proximity (distance) between data points.

Key Concepts:

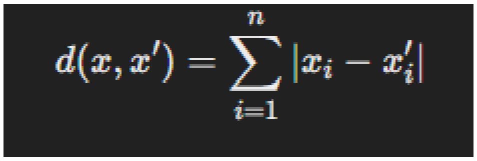
- **Distance Metrics:** KNN classifies a data point based on the majority class of its KKK nearest neighbors in the feature space. The most commonly used distance metrics are:

- **Euclidean Distance** (for continuous data):



$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

- **Manhattan Distance**:



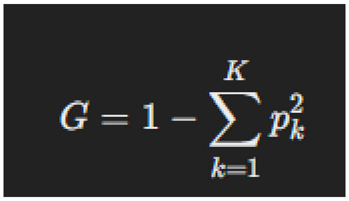
$$d(x, x') = \sum_{i=1}^{n} |x_i - x'_i|$$

3. Random Forest:

Random Forest is an **ensemble** learning method that uses multiple decision trees to improve classification or regression results.

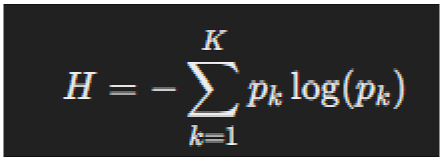
Key Concepts:

- **Decision Tree**: A decision tree splits the data into subsets using feature thresholds, aiming to maximize homogeneity within the subsets (e.g., using **Gini Impurity** or **Entropy**).
  - **Gini Impurity**:



$$G = 1 - \sum_{k=1}^{K} p_k^2$$

  - **Entropy**:

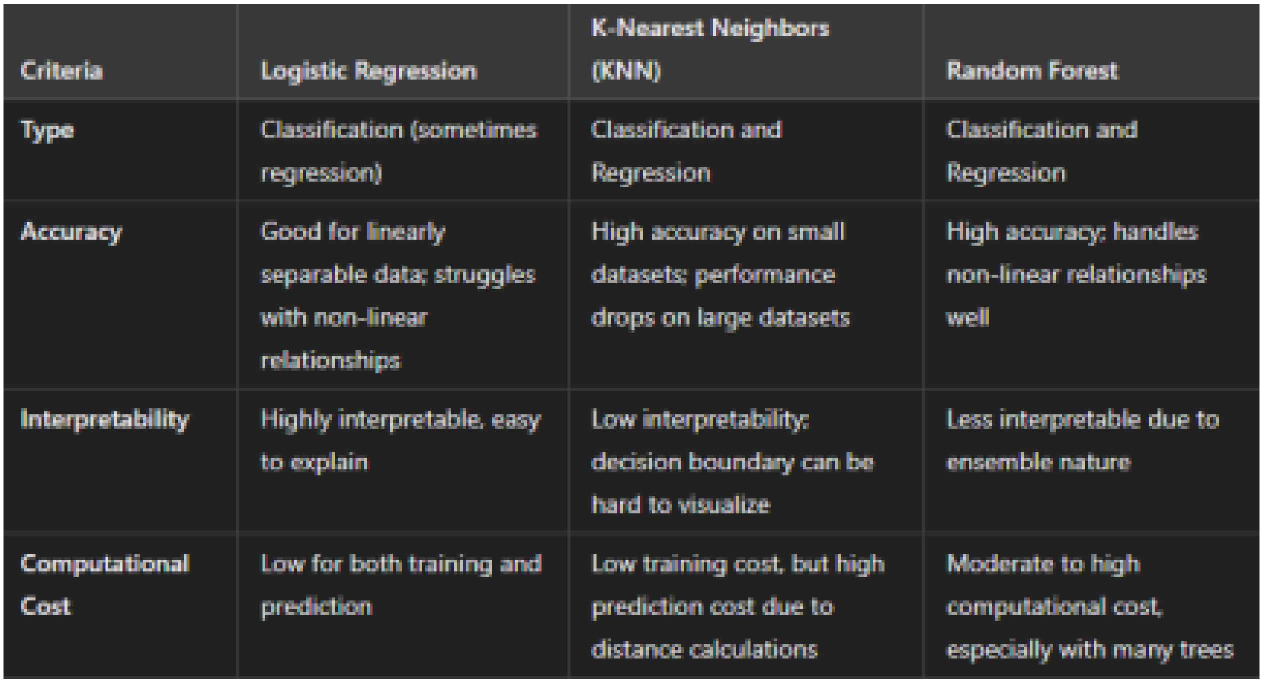


$$H = -\sum_{k=1}^{K} p_k \log(p_k)$$

Data Cleaning:

The dataset does not contain any missing or incomplete values. Any entries left as inaccessible or incomplete by clients are labeled as "unknown." Rather than removing these missing values, our goal is to identify correlations in the data that can help impute or fill these unknown entries. It's important to mention that unknown values make up less than 5% of all features. Given this small proportion, we can rely on correlations to fill in the gaps. As for outliers, since the data involves personal information and there are no unrealistic values, these outliers will be treated as specific cases. For example, clients older than 80 years should not be removed from the dataset.

Model Comparison:



Criteria	Logistic Regression	K-Nearest Neighbors (KNN)	Random Forest
Type	Classification (sometimes regression)	Classification and Regression	Classification and Regression
Accuracy	Good for linearly separable data; struggles with non-linear relationships	High accuracy on small datasets; performance drops on large datasets	High accuracy; handles non-linear relationships well
Interpretability	Highly interpretable, easy to explain	Low interpretability; decision boundary can be hard to visualize	Less interpretable due to ensemble nature
Computational Cost	Low for both training and prediction	Low training cost, but high prediction cost due to distance calculations	Moderate to high computational cost, especially with many trees

Learning Outcomes

1. **Skills Used:**

- Data analysis and preprocessing.
- Data visualization.
- Machine learning modeling and evaluation.
- Handling missing data and performing exploratory data analysis (EDA).

2. **Tools Used:**

- Python libraries: Pandas, NumPy, Seaborn, Matplotlib.
- Machine learning tools: Scikit-learn.
- Dimensionality reduction with t-SNE.
- Google Colab for coding and computation.

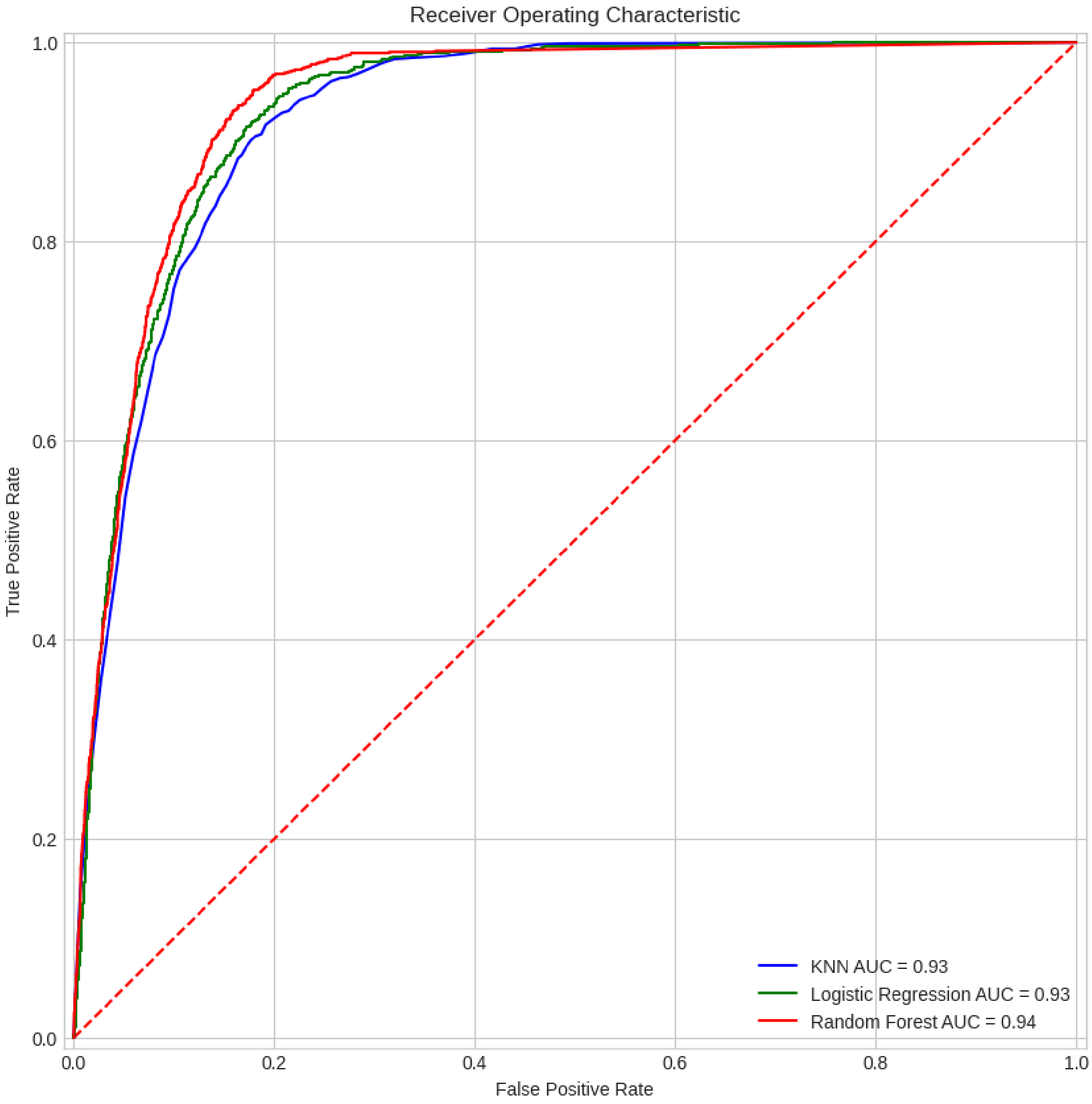
3. **Dataset Used:**

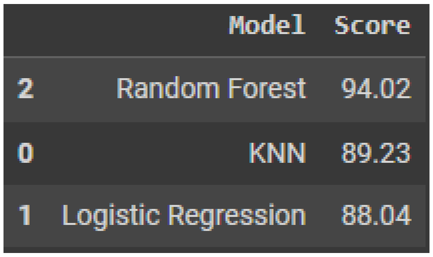
The "Bank Marketing" dataset (as seen from bank-additional.csv) includes features related to customer information and marketing campaign responses which are used for classification tasks such as predicting customer responses to a marketing campaign.

4. **Topic Learnt:**

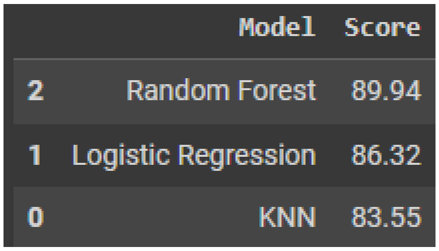
- Data preprocessing and visualization techniques.
- Application of machine learning algorithms (likely classification) on marketing datasets.
- Techniques for evaluating model performance (e.g., accuracy, precision, recall).

Result:





	Model	Score
2	Random Forest	94.02
0	KNN	89.23
1	Logistic Regression	88.04



	Model	Score
2	Random Forest	89.94
1	Logistic Regression	86.32
0	KNN	83.55

Code: MLproject.ipynb