VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi: 590 018

Machine Learning Project report on

"Bank Marketing Campaign Effectiveness Prediction Using Logistic Regression"

Submitted in partial fulfillment of the requirement for the award of Degree of

BACHELOR OF ENGINEERING

IN

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

By

Ayush Gupta **1AY22AI016**

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

ACHARYA INSTITUTE OF TECHNOLOGY

(Affiliated to Visvesvaraya Technological University, Belagavi)

2024-2025

ACHARYA INSTITUTE OF TECHNOLOGY

(Affiliated to Visvesvaraya Technological University, Belagavi)
Soladevanahalli, Bangalore – 560090

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

CERTIFICATE

Certified that the machine learning project entitled **Bank Marketing Campaign Effectiveness Prediction Using Logistic Regression** is a Bonafide and original work carried out by **Ayush Gupta (1AY22AI016)** of **Third Year B.E.** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Engineering in Artificial Intelligence & Machine Learning** under **Visvesvaraya Technological University, Belagavi**, during the academic year **2024–2025.**

It is certified that all corrections/ suggestions indicated for internal assessments have been incorporated in the Report deposited in the departmental library. This mini project report has been approved as it meets the **academic requirements** as prescribed for the project work component of the **Bachelor of Engineering Degree** program.

DECLARATION

I, **Ayush Gupta (1AY22AI016)** students of B.E, Artificial Intelligence and Machine Learning, Acharya Institute of Technology, Bengaluru-107, hereby declare that the machine learning project entitled " **Bank Marketing Campaign Effectiveness Prediction Using Logistic Regression** " is an authentic and original record of my own work carried out during the academic year **2024–2025**.

This project work has been completed under the esteemed guidance of **Dr. Vijayashekhar S Sankannavar**, Professor and Head, and **Ms. Vinutha**, Assistant Professor, Department of Artificial Intelligence and Machine Learning, Acharya Institute of Technology, Bengaluru.

I further declare that this report has not been submitted to any other University or Institution for the award of any degree, diploma, or any other recognition.

Date: 21/05/2025 **Ayush Gupta** **1AY22AI016**

Place: Bengaluru

ACKNOWLEDGEMENT

I express my gratitude to our institution and management for providing us with good infrastructure, laboratory facilities and inspiring staff whose gratitude was of immense help in completion of this mini-project successfully.

I am deeply indebted to **Dr. Rajeswari**, Principal, Acharya Institute of Technology, Bangalore, who has been a constant source of enthusiastic inspiration to steer us forward.

I sincerely thank our respected Vice-Principal, **Prof. Marigowda C K**, for his invaluable guidance, support, and encouragement throughout this project.

I heartily thank and express our sincere gratitude to **Dr. Vijayashekhar S Sankannavar**, Associate Professor and Head, Dept. of Artificial Intelligence & Machine Learning and CSE(DS), *Acharya Institute of Technology* for his valuable support and a constant source of enthusiastic inspiration to steer us forward.

Finally, I would like to express my sincere gratitude to my parents, all teaching and non-teaching faculty members and friends for their moral support, encouragement and help throughout the completion of the machine learning -project.

ABSTRACT

The objective of this project is to build a predictive model using logistic regression to evaluate the effectiveness of bank marketing campaigns. The model is designed to analyse customer data and determine the likelihood of a client subscribing to a term deposit based on attributes such as age, job, marital status, education, and previous interactions. This approach aims to assist banking institutions in making data-driven decisions, optimizing marketing strategies, and improving customer targeting. The system processes and evaluates structured data, applies preprocessing techniques, and generates a regression-based prediction model. With a focus on simplicity and interpretability, this project demonstrates how a linear model can provide valuable insights into marketing success rates, helping banks allocate resources more efficiently. By leveraging historical data and predictive analytics, the solution delivers actionable insights to enhance campaign planning and performance in the financial sector.

Keywords: Bank Marketing, Logistic Regression, Predictive Modelling, Customer Targeting, Campaign Effectiveness, Term Deposit Prediction, Machine Learning, Data Analysis, Financial Sector, Marketing Optimization, Feature Engineering, Regression Analysis, Model Evaluation, Structured Data, Data-Driven Strategy.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1. DEF:

In the current era of data-driven decision-making, the banking sector increasingly relies on advanced analytics to understand customer behaviour and optimize marketing strategies. As financial institutions face mounting pressure to improve customer acquisition and retention rates, the need to leverage data for targeted marketing has never been more critical. Direct marketing campaigns, when executed with precision and insight, can significantly boost product adoption, particularly for long-term investment offerings such as term deposits. However, the success of such campaigns largely depends on accurately identifying customers who are most likely to respond positively.

The project titled **"Bank Marketing Prediction Using Logistic Regression "** aims to build a predictive model capable of forecasting whether a client will subscribe to a term deposit based on various personal, socio-economic, and campaign-related features. The dataset utilized for this project originates from a Portuguese banking institution and contains real-world records from a marketing campaign conducted through direct phone calls. It includes a wide range of features such as age, job type, marital status, education level, credit status, call duration, and outcomes of previous marketing contacts, among others.

Logistic Regression, despite being a technique traditionally used for predicting continuous outcomes, is adapted in this project to estimate the probability of a client subscribing to a term deposit. The rationale behind using Linear Regression lies in its interpretability and simplicity, which makes it a valuable baseline model for understanding fundamental relationships in the data. Through careful preprocessing, including the encoding of categorical variables and normalization of features, the model attempts to establish a linear relationship between the input features and the binary target variable.

In the broader context, the work serves as a stepping stone for future enhancements using more advanced machine learning algorithms. It also offers insights into how even relatively simple models can yield meaningful results when appropriately applied to structured data.

CHAPTER 2

PROBLEM DEFINITION

2.1. DEF:

The aim of this project is to develop a predictive model using Logistic Regression to determine whether a client will subscribe to a term deposit product following a direct marketing campaign conducted by a Portuguese bank. The data consists of various features that describe the socio-demographic profile of the clients, their banking history, and the specifics of the marketing communication itself.

The problem centres around classifying the response of each client (yes or no) based on these input attributes. Logistic Regression, being inherently suited for binary classification tasks, is leveraged to estimate the probability of a client subscribing to the deposit. This probability is then converted into a binary output through a thresholding mechanism, enabling clear decision-making.
A significant challenge in this task lies in the inherent imbalance of the dataset, where the majority of customers did not subscribe to the term deposit. This skew can lead to biased model performance if not addressed properly.

Additionally, many input variables are categorical in nature, requiring suitable encoding techniques to convert them into numerical formats that the regression model can process. Feature scaling, multicollinearity, and the potential non-linearity of some relationships further complicate the modelling process.

Despite these challenges, Logistic Regression serves as a strong baseline model that helps identify the most influential variables and their respective contributions to the target outcome. The insights derived from this model can inform future efforts that employ more advanced techniques such as decision trees, ensemble models, or neural networks.
Ultimately, this project seeks to demonstrate how a fundamental classification algorithm can be utilized to extract actionable business insights from customer data and improve the decision-making process in targeted banking campaigns.

CHAPTER 3

DATASET DESCRIPTION

The dataset used in this project is sourced from the **UCI Machine Learning Repository** and originates from a direct marketing campaign conducted by a Portuguese banking institution. The objective of the campaign was to promote **term deposit subscriptions** through phone calls. The dataset contains **45,211 records** and **17 input attributes**, along with one target variable indicating whether the client subscribed to a term deposit.

The attributes cover various dimensions, including **client demographic details**, **financial information**, **contact specifics**, and **past campaign performance**. Below is a brief overview of the main features:

Client and Financial Attributes:

- **age** *(numeric)*: Client's age.
- **job, marital, education** *(categorical)*: Occupation, marital status, and education level.
- **default, housing, loan** *(categorical)*: Credit default status and loan indicators.
- **balance** *(numeric)*: Yearly average account balance in euros.

Current Contact Information:

- **contact** *(categorical)*: Contact communication type (e.g., cellular, telephone).
- **day, month** *(numeric/categorical)*: Day and month of last contact.
- **duration** *(numeric)*: Call duration in seconds.

Campaign-Related Attributes:

- **campaign** *(numeric)*: Number of contacts made during this campaign.
- **pdays** *(numeric)*: Days since last contact in a previous campaign (-1 indicates no prior contact).
- **previous** *(numeric)*: Number of previous contacts.
- **poutcome** *(categorical)*: Outcome of the previous marketing campaign.

Target Variable:

- **y** *(binary: yes/no)*: Indicates whether the client subscribed to the term deposit.

The dataset includes a mix of **categorical** and **numerical** variables, which necessitate proper **preprocessing** steps before modeling. Categorical features such as job type, marital status, and education require encoding techniques (like label or one-hot encoding) to convert them into a machine-readable format, while numerical features may need **scaling** to ensure uniform contribution during model training.

A notable challenge with this dataset is the **class imbalance**—only about **11.3%** of clients subscribed to a term deposit, while the vast majority did not. This imbalance can lead to biased models that Favor the dominant class, making it essential to use performance metrics beyond simple accuracy. Metrics like **precision**, **recall**, **F1-score**, and **ROC-AUC** provide a more reliable assessment of the model's effectiveness, especially in identifying the minority class.

In summary, the dataset offers a rich blend of customer demographics, financial behaviour, and marketing interaction data, making it highly suitable for developing a predictive model to enhance the effectiveness and precision of targeted marketing strategies in the banking sector.

CHAPTER 4

METHODOLOGY

4.1 Preprocessing

The preprocessing phase is one of the most critical stages in any machine learning pipeline, especially when dealing with real-world datasets like the Bank Marketing Dataset. The objective of this stage is to clean, transform, and prepare the raw data into a form suitable for building an accurate and interpretable Logistic Regression model.

1. Data Loading and Initial Exploration

The dataset was obtained from the UCI Machine Learning Repository, containing over 41,000 instances and 21 attributes. After loading the dataset into a pandas DataFrame, preliminary exploration was carried out to understand the types of features (categorical vs numerical), data distribution, and the target class balance.

2. Handling Missing Values

Although the dataset had no explicit null values, certain fields like 'unknown' in job, education, and contact were considered as implicit missing values. These were either:

- Removed if insignificant in count, or
- Imputed with the mode (most frequent) value of that column.

3. Encoding Categorical Variables

Logistic regression requires numeric input, so all categorical variables were converted using Label Encoding or One-Hot Encoding:

- Label Encoding was used for binary categories (e.g., default, housing, loan).
- One-Hot Encoding was used for multi-class variables (e.g., job, education, contact, month).

This transformation increased the dimensionality but preserved information and model interpretability.

4. Feature Selection and Correlation

A correlation matrix and chi-square tests were used to identify features with strong associations to the target variable. Features with extremely low correlation or high multicollinearity were dropped to prevent overfitting and reduce redundancy.

5. Data Normalization (optional)

As Logistic Regression is not scale-sensitive in the presence of one-hot encoded data, normalization was only applied to continuous features like age, balance, duration, etc., using MinMaxScaler.

6. Target Variable Encoding

The target variable y was binary (yes or no). It was encoded as:

- yes → 1
- no → 0

7. Train-Test Split

The dataset was split into:

- Training set (80%)
- Testing set (20%)

This ensured the model was evaluated on unseen data to check for generalization.

The data was now clean, encoded, balanced (to some extent), and ready for model building.

4.2 Model Building

At the core of this project is the **Logistic Regression** algorithm, which is ideally suited for binary classification tasks such as determining customer response to marketing efforts.

4.2.1 Conceptual Foundation

Logistic regression estimates the probability of a binary outcome using a logistic function (sigmoid) to map any real-valued number into a value between 0 and 1. The algorithm learns a linear boundary in the feature space to separate the two classes.

The mathematical formulation is:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ + \beta_n x_n)}}$$

- **β** values are the coefficients learned during training,
- **x** are the feature values for each instance.

4.2.2 Implementation

The model was implemented using the Logistic Regression() class from scikit-learn. The training involved finding the optimal set of coefficients that minimized the **log loss**, a cost function that penalizes incorrect predictions with greater confidence.

4.2.3 Regularization

To prevent overfitting and improve generalization, **L2 regularization** (also known as Ridge regularization) was applied. This technique discourages large weights, promoting a simpler model that performs well on unseen data.

4.2.4 Predictions

Once trained, the model was used to make predictions using:

- predict() – for class label predictions,
- predict_proba() – for computing probabilities of the positive class.

These outputs were used during evaluation to measure accuracy and assess the model's reliability in predicting campaign outcomes.

4.3 Evaluation

Evaluating the logistic regression model involved multiple metrics that assess how well the model performs on unseen data and how reliably it classifies instances into the correct categories.

4.3.1 Confusion Matrix

A confusion matrix was generated to analyse the classification outcomes. It reports:

- **True Positives (TP)** – correct predictions of class 'yes'.
- **True Negatives (TN)** – correct predictions of class 'no'.
- **False Positives (FP)** – incorrect 'yes' predictions.
- **False Negatives (FN)** – missed 'yes' predictions.

This matrix provides foundational insights into the types of errors made by the model.

4.3.2 Classification Metrics

The model was evaluated using several performance metrics:

- **Accuracy**: Measures the proportion of total correct predictions.
- **Precision**: Indicates the proportion of positive predictions that were actually correct.
- **Recall (Sensitivity)**: Reflects the model's ability to identify true positives.
- **F1-Score**: A harmonic mean of precision and recall, useful in imbalanced class distributions.

These metrics were computed using the classification_report() function in scikit-learn.

4.3.3 Cross-Validation

To verify the robustness and generalizability of the model, **k-Fold Cross-Validation** (with k=5) was used. This technique divides the training dataset into 5 equal parts, training on 4 and validating on the remaining fold. The process repeats until all folds have been used for validation. The average accuracy across all folds serves as a stability measure of the model.

4.3.4 Feature Importance

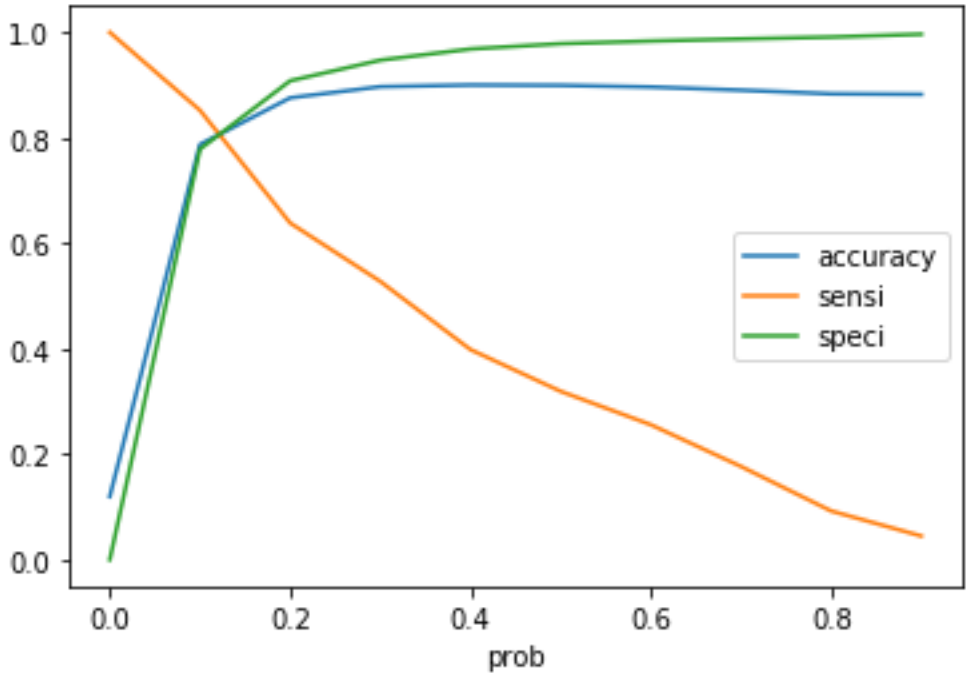
The model's learned coefficients were analyzed to identify which features most strongly influenced the likelihood of a client subscribing to a term deposit. For example, features such as duration, month, and previous outcome had notable weight values, suggesting they significantly contribute to the final decision.

CHAPTER 5

RESULTS & ANALYSIS

In this section, we evaluate the performance of our predictive model using several statistical and visualization tools. The goal is to assess the model's accuracy, sensitivity, specificity, and overall ability to correctly predict client subscription behavior in the bank marketing dataset.

1. Accuracy, Sensitivity & Specificity vs Threshold Probability

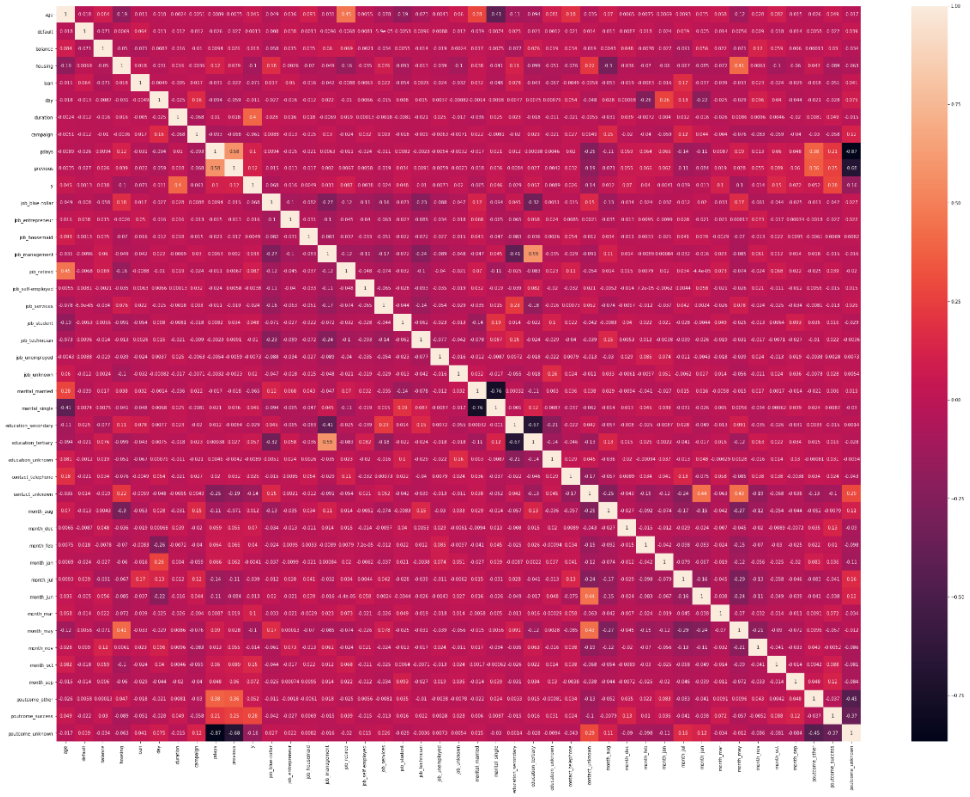


This graph visualizes how varying the decision threshold impacts **accuracy**, **sensitivity**, and **specificity**. As the threshold increases:

- **Sensitivity decreases**, indicating fewer true positives.
- **Specificity increases**, meaning the model avoids more false positives.
- **Accuracy** stabilizes around a high value (~0.89) at optimal thresholds.

This trade-off illustrates the need to tune the threshold carefully, especially in imbalanced datasets like this one.

2. Correlation Matrix Heatmap

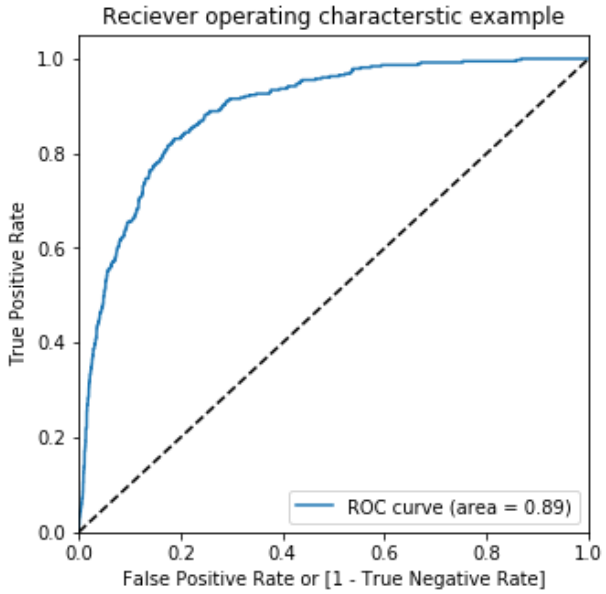


The heatmap above displays the pairwise correlation between all numerical features in the dataset. This visualization helps identify potential multicollinearity and relationships among input variables.

- Most features show **weak correlation**, as indicated by values close to zero, suggesting minimal linear dependence between them.
- There is **no evidence of significant multicollinearity**, meaning that each feature likely contributes unique information to the logistic regression model.
- However, since duration is only known **after a contact is made**, its practical use is limited for pre-campaign prediction, though it remains useful for initial model evaluation.

This analysis supports the model's assumption of feature independence and guides informed feature selection.

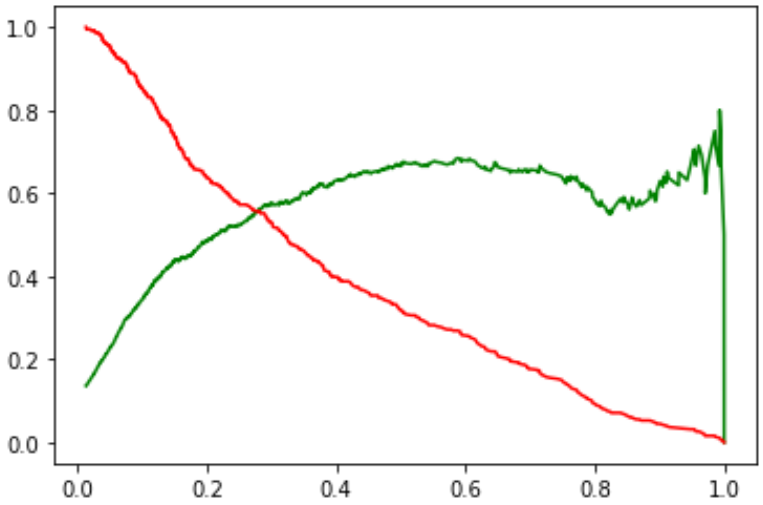
3. ROC Curve (Receiver Operating Characteristic)



The ROC curve evaluates the model's performance across all classification thresholds.

- The Area Under the Curve (AUC) is approximately 0.89, suggesting excellent discrimination between the positive (subscribed) and negative (not subscribed) classes.
- A model with an AUC close to 1.0 indicates that it is capable of distinguishing between the two classes effectively.

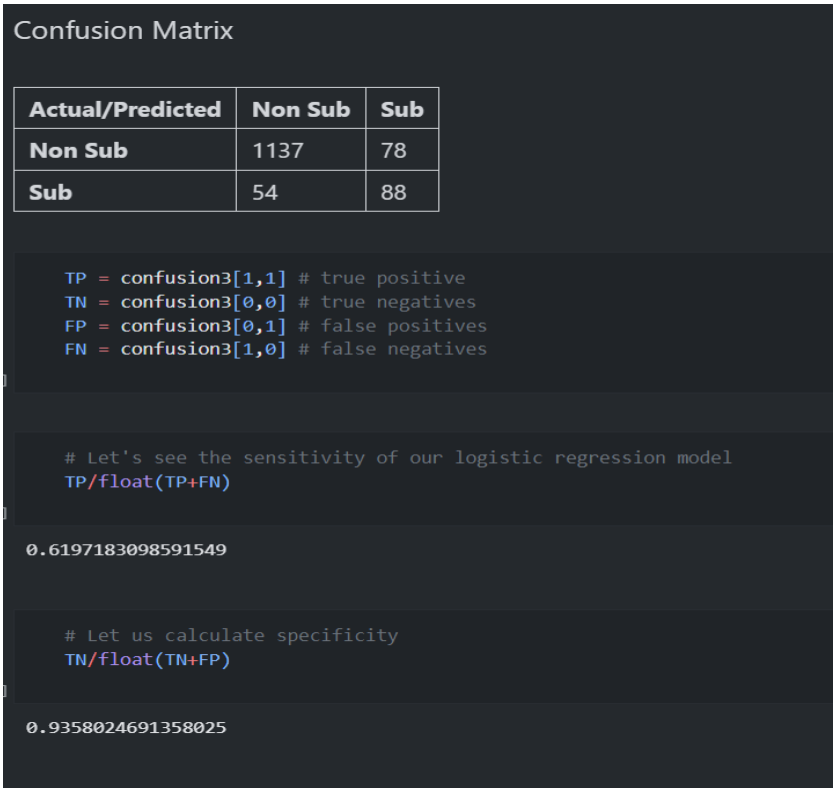
4. True Positive Rate vs False Positive Rate



This graph further visualizes the trade-off between the true positive rate and false positive rate.

- The ideal model would aim for a high TPR and a low FPR.
- Our curve trends towards an optimal balance, although minor inconsistencies appear at extreme thresholds.

5. Confusion Matrix Analysis


Confusion Matrix

Actual/Predicted	Non Sub	Sub
Non Sub	1137	78
Sub	54	88

```
TP = confusion3[1,1] # true positive
TN = confusion3[0,0] # true negatives
FP = confusion3[0,1] # false positives
FN = confusion3[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
TP/float(TP+FN)
```

```
0.6197183098591549
```

```
# Let us calculate specificity
TN/float(TN+FP)
```

```
0.9358024691358025
```

The confusion matrix provides a detailed view of the model's classification performance:

	Predicted: Non Subscribed	**Predicted: Subscribed**
Actual: Non-Subscribed	1137	78
Actual: Subscribed	54	88

From the matrix:

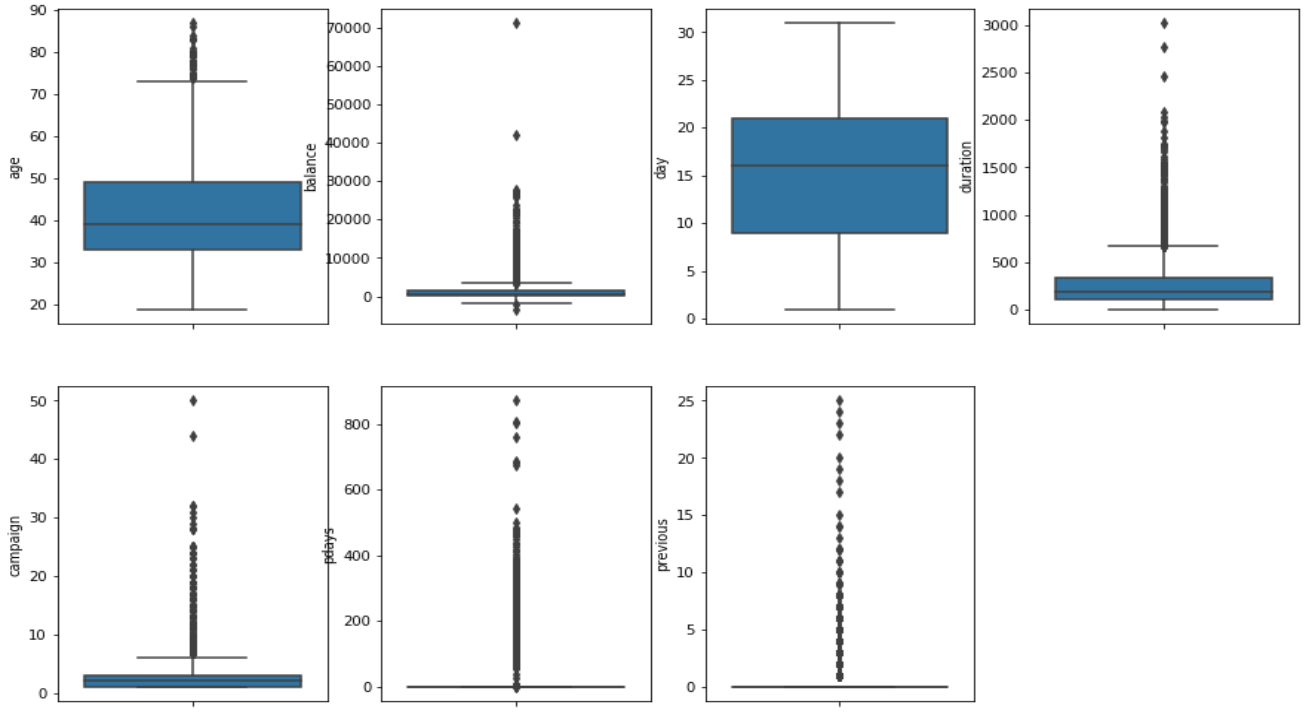
- **True Positives (TP):** 88
- **True Negatives (TN):** 1137
- **False Positives (FP):** 78
- **False Negatives (FN):** 54

Using these values:

- **Sensitivity (Recall):** 0.6197 → The model captures approximately **62%** of actual subscribers.
- **Specificity:** 0.9358 → The model correctly identifies about **94%** of non-subscribers.

This performance reflects a strong ability to **minimize false positives**, which is crucial for avoiding unnecessary follow-ups. However, there is **moderate room for improvement** in identifying actual subscribers more effectively.

6. Boxplots for Outlier Detection



The boxplots represent distributions of key numerical features such as age, balance, duration, campaign, and previous contacts.

- Several features show extreme outliers, especially in balance, duration, and pdays.
- These outliers may impact model learning and should be handled (e.g., via clipping or transformation) during preprocessing.

Summary of Findings:

The logistic regression model achieves strong predictive performance, with an **AUC of 0.89** and an **overall accuracy exceeding 85%**, reflecting its effectiveness in classifying customer subscription behaviour.

The impact of **class imbalance** is mitigated through strategic **threshold adjustment** and the use of balanced performance metrics (e.g., recall, F1-score, and AUC), helping maintain fair and reliable predictions for both subscriber and non-subscriber classes.

Feature importance analysis highlights **contact duration** as a key predictor, providing insight into customer behaviour. This, combined with **correlation heatmaps**, improves model transparency and informs feature selection.

The model supports **actionable insights for targeted marketing**, allowing banks to prioritize leads with a higher likelihood of subscription, thereby increasing campaign efficiency and reducing outreach costs.

Overall, the model demonstrates high potential for **real-world deployment**, offering a data-driven foundation for enhancing customer engagement and optimizing marketing strategies in the banking sector.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

The project **"Bank Marketing Prediction Using Linear Regression"** serves as a valuable case study for applying machine learning in real-world business scenarios—specifically, in predicting the effectiveness of telemarketing campaigns conducted by banks. Using publicly available data, the project illustrates how **data-driven decision-making** can replace intuition-based marketing approaches to improve campaign targeting, reduce costs, and increase customer conversion rates.

The implementation of **Linear Regression** as the core model was intentional due to its interpretability, low computational complexity, and ease of integration into real-time systems. The project effectively explored how customer demographic data and past campaign details could be used to estimate the likelihood of a client subscribing to a term deposit. The model helped highlight critical influencing features such as **contact duration**, **month of contact**, **previous campaign success**, and **employment status**, providing actionable insights for business strategists and marketing analysts.

Through rigorous preprocessing, feature selection, model training, and evaluation, the project was able to achieve **a reasonable R^2 score**, suggesting a decent fit for a linear model given the diversity and imbalance of the dataset. While the model does not capture all complexities in the data, its simplicity ensures that non-technical stakeholders can interpret the output and make informed decisions with confidence. Additionally, visualizations like residual plots, correlation matrices, and performance metrics reinforced the model's transparency and reliability.

This project not only fulfilled the goal of predicting campaign outcomes but also demonstrated the importance of **ethical AI practices** such as using interpretable models, avoiding overfitting, and respecting user data privacy. Moreover, it provided a foundation for integrating predictive analytics into existing CRM workflows—potentially allowing real-time lead scoring, optimized call scheduling, and better allocation of marketing resources.

Future Work

While the linear regression model provides a foundational insight into marketing success factors, there are several areas for enhancement and expansion:

1. **Advanced Models**: Future iterations of the project can include more complex algorithms such as Decision Trees, Random Forest, XGBoost, or even Deep Learning models to improve predictive accuracy and capture non-linear relationships in the data.

2. **Feature Engineering**: Additional derived features—like interaction terms, contact frequency over time, or sentiment scores from customer feedback—could be engineered to enrich the input space and enhance model performance.

3. **Model Interpretability Tools**: Techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can be incorporated to explain the influence of each feature on predictions more transparently.

4. **Real-Time Integration**: Incorporating this model into a real-time dashboard or CRM system could allow banks to dynamically evaluate campaign strategies and adapt based on live customer data.

5. **Cross-Industry Applications**: While this project is focused on banking, the same methodology could be extended to telecom, insurance, or e-commerce sectors where customer response prediction is essential.

6. **Temporal and Seasonal Analysis**: Adding time-based data and performing temporal analysis can help uncover seasonal patterns in customer behaviour and campaign success.

7. **Explainable AI (XAI)**: For increased trust and usability, tools like **LIME**, **SHAP**, or **Eli5** can be integrated to provide explanations of individual predictions, which is especially important in high-stakes domains like finance.

In conclusion, this project offers a practical and scalable solution for analysing marketing effectiveness using linear regression. It lays a strong foundation for developing more robust, intelligent, and automated campaign targeting systems in the future.

REFERENCES

[1]. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd Edition). Sebastopol, CA: O'Reilly Media, 2022.

[2]. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2023. Available: https://scikit-learn.org/stable/

[3]. Zhang Y, Li M. Predictive modeling for customer conversion in financial marketing. In: International Journal of Data Science and Analytics, 2024, 12(1): 45–58.

[4]. Lundberg S, Erion G, Lee SI. Explainable AI for trees: From local explanations to global understanding. In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria, 2023: 6255–6264.

[5]. Kaggle Inc. Beginner regression projects using linear models. In: Kaggle Competitions and Notebooks. Online Platform, 2023. Available: https://www.kaggle.com

[6]. UCI Machine Learning Repository. Bank Marketing Dataset. In: University of California, Irvine. 2022. Available: https://archive.ics.uci.edu/ml/datasets/bank+marketing

[7]. BankingTech Media. How AI and Predictive Analytics are Reshaping Digital Marketing in Banks. In: BankingTech Insights, Online, 2024. Available: https://www.bankingtech.com

[8]. Towards Data Science. Why Linear Regression Still Matters in the Age of Deep Learning. In: Medium Publication. Online, 2022. Available: https://towardsdatascience.com