Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 6, 2025

PREDICTIVE REGRESSION MODEL FOR SUBSCRIPTION FORECASTING THROUGH TELEMARKETING

[1]Inzamam Shahzad, [2]Muhammad Abdur Raphay Zia, [3]Muhammad Tanveer Meeran , [4]Salahuddin ,[5]Zainab Tariq

[1]*School of Computer Science & School of Cyberspace Science, Xiangtan University, Xiangtan, Hunan, China*
[2]*i2c Inc Lahore, Pakistan.*
[3]*Faculty of computer science and mathematics, Universiti Malaysia Terengganu, Malaysia*
[4]*Department of Computer Science, NFC Institute of Engineering and technology, Multan, Pakistan.*
[5]*Department of Computer Science & IT, Govt.Graduate College, Burewala, Pakistan.*
**Corresponding Author: Salahuddin. Email: msalahuddin8612@gmail.com*

DOI: https://doi.org/

Keywords
(Logistic Regression, Predictor, Bayes, Telemarketing Success, Customer Data Analysis, Socio-economic Factors, Deep Learning).

Article History
Received on 18 May 2025
Accepted on 10 June 2025
Published on 19 June 2025

Corresponding Author: * **Salahuddin**

Abstract

In this study, a data mining technique was employed to predict the success of telemarketing calls aimed at promoting long-term bank deposits, a key challenge for financial institutions aiming to optimize their marketing strategies. The dataset used for this analysis was sourced from a Portuguese retail bank, containing 45,211 records and 17 different attributes, including demographic and behavioral information about the clients. The primary objective was to develop a predictive model that could accurately determine whether a client would subscribe to a term deposit based on the features available in the dataset. To build the predictive model, logistic regression was applied, a statistical method well-suited for binary classification tasks like this one. The model was trained to estimate the probability that a client would respond positively to the marketing campaign. A crucial step in improving model performance was the feature selection process, where 22 distinct sets of features were evaluated. This helped identify the most relevant attributes that contributed to the prediction, ensuring that the final model was both efficient and interpretable. The final model achieved a precision of 0.74 and a recall of 0.74, indicating that it performed well in both identifying positive responses (precision) and capturing as many positive responses as possible (recall). Specifically, the model made 11,294 correct predictions, including 6,124 true positives and 5,170 true negatives, demonstrating its ability to accurately classify both successful and unsuccessful subscription cases. However, it also made 4,047 incorrect predictions, which consisted of 2,505 false positives and 1,542 false negatives. These errors reflect the inherent challenges in predictive modeling, particularly in distinguishing between clients who are likely to subscribe and those who are not.

INTRODUCTION

When an organization plans for growth, telemarketing often becomes a key strategy for customer outreach. Telemarketing, where sales agents directly contact potential customers via phone, remains one of the most widely used direct marketing techniques. Numerous marketing campaigns have leveraged it effectively. The performance of such campaigns, especially in the banking sector, can be evaluated using modern analytical approaches. In this study, a data mining approach is proposed to predict the success of telemarketing calls made to promote long-term deposit products. The aim of this research is to assess the accuracy and performance of various classification models. Logistic regression, a commonly used predictive model, has been employed in this analysis using 16 independent variables. Feature selection techniques were applied to identify the most significant subsets of variables. Subsequently, different classification algorithms were evaluated to compare their accuracy and performance. The regression model derived from feature selection was then assessed in terms of classification performance. Logistic regression is particularly suitable for this type of binary classification task.

To support future sales forecasting, time-series data from a company can be analyzed using models such as linear and logistic regression. In this research, time-series sales data were examined to predict the company's future sales, both overall and for specific products. The models were trained, tested, and validated using appropriate datasets. Python, with the Spider IDE, was used for coding and analysis due to its powerful data science libraries and ease of use.

The dataset for this research was obtained from the University of California, Irvine (UCI) Machine Learning Repository. It reflects real-world data collected from a Portuguese retail bank's direct marketing campaigns conducted between May 2008 and November 2010. The dataset contains 45,211 records and 17 attributes, including a response variable indicating whether the client subscribed to a term deposit ("yes" or "no"). Often, multiple contacts with the same client were required to reach a final decision. After preprocessing the dataset, a logistic regression model was implemented in Python, and the outcomes demonstrated better accuracy compared to previously reported results.

1.1 Objectives

This study highlights various data mining techniques as powerful tools in the decision-making process. Real-world data from the UCI repository was used to evaluate model performance. The research was conducted in two main phases:

- **Feature Selection:** Using statistical methods including best-subset selection, logistic regression, and random forest to identify the most impactful features.
- **Model Evaluation:** The selected feature subsets were then used with different classification algorithms such as Support Vector Machines (SVM), Decision Trees, Random Forest, and Artificial Neural

Networks (ANN) to analyze their predictive performance.

The goal was to determine whether reduced models with fewer variables could achieve accuracy comparable to full models. The findings suggest that a reduced model using only 10 key variables can perform nearly as well as the full model.

Key Objectives:

- Analyze existing methods for predicting the outcome of bank telemarketing campaigns.
- Assess the performance of the logistic regression model.
- Compare logistic regression with other classification techniques.

1.2 Research Questions

This research applies data mining techniques to forecast the outcome of telemarketing calls aimed at promoting long-term bank deposits—a topic of growing importance in the field of machine learning. While earlier studies primarily relied on decision trees, this research focuses on logistic regression for prediction.

In the banking sector, an effective marketing campaign is essential, and term deposits represent a critical product line. Telemarketing has proven to be a powerful method for reaching potential clients. Data mining has gained significant traction for handling and interpreting large volumes of data, especially in the age of big data analytics. The primary objective of this study is to develop a robust prediction model to evaluate the success of bank telemarketing efforts.

Research Questions:

- What predictive models currently exist for analyzing telemarketing data?

What types of analysis do these models perform?

Which predictive model is most effective for telemarketing data?

How can the performance of a prediction model be accurately measured?

Literature Review

Forecasting plays a vital role in the banking sector, especially in areas like credit card fraud detection, where accurate predictions can significantly reduce financial losses. Another crucial aspect is understanding and anticipating customer behavior, which enables banks to build long-term relationships. Predictive modeling in such contexts often takes the form of classification problems, where accuracy is essential. Inaccurate predictions can lead to either missed opportunities or unnecessary costs, significantly affecting marketing outcomes [1]. Intelligent systems and analytics applications are now fundamental in transforming raw business data into actionable strategies. These applications enhance both performance and interpretability of decision-making models. Many modern data-driven models offer flexibility and better performance than traditional statistical models. However, a notable drawback is their "black-box" nature, where interpretability is limited or absent [2].

Modern marketing campaigns often rely on established direct marketing techniques, such as telephone-based outreach. Many organizations use centralized customer management systems to interact with clients efficiently. These systems help businesses better understand customer needs and provide timely responses. A common approach includes the use of integrated contact centers that manage inbound and outbound calls. Among various strategies, telephone marketing remains one of the most widely used methods. These centers serve as the operational hub for

client engagement and are often referred to as merchandising centers due to their external focus [3].

Decision Support Systems (DSS) utilize information technology to assist in complex decision-making processes. These systems are broadly categorized into several types, including individual and intelligent DSS. While smaller organizations may rely on personal DSS for singular decision-making, larger enterprises often use intelligent systems powered by artificial intelligence to support more complex, multi-stakeholder decisions [4]. Business Intelligence (BI) is another umbrella term that includes a range of technologies such as databases, data warehouses, and data mining, all aimed at improving strategic business decisions [5].

Data mining plays a central role in both BI and DSS by enabling semi-automated extraction of meaningful insights from large datasets. One of its primary applications is classification, which involves developing data-driven models that categorize data points based on specific input variables [6][7]. Popular classification techniques include decision trees, logistic regression (LR), neural networks (NN), and support vector machines (SVM). While LR and decision trees offer interpretability and ease of understanding, more complex models like NN and SVM are less transparent. These models often require sensitivity analysis to estimate how input features influence outcomes—this is especially important for evaluating "black-box" models [8].

Support Vector Machines (SVMs) have gained prominence in recent years for their effectiveness in predictive modeling, particularly in banking applications involving long-term deposits. Studies have shown that data collected from direct marketing campaigns between 2011 and 2014 revealed SVMs often outperformed traditional models. For comparison, researchers applied SVM alongside Decision Trees and Naïve Bayes algorithms. Results indicated that while SVM achieved high accuracy, it did not drastically outperform all other techniques. Long calls—meaning extended conversations with clients—were found to correlate with higher success rates. Another study by Moro et al. used data from 2008 to 2013 involving 150 attributes. The models tested included Decision Trees, Neural Networks, SVM, and Logistic Regression. Among these, Neural Networks showed the highest prediction accuracy, with Decision Trees being most effective in identifying customer satisfaction. SVMs showed particularly strong performance when integrated with statistical learning techniques [9–12].

This research specifically focuses on telemarketing efforts designed to promote long-term deposits. Campaigns often involve both outbound calls to prospective clients and inbound calls from interested customers. The success of these campaigns is measured by whether or not a term deposit is subscribed to. As discussed, this study is grounded in real-world data from a Portuguese bank. In recent years, companies have intensified efforts to attract customers using innovative messaging and promotional techniques. Recent marketing research has explored the psychological impact of media exposure, brand loyalty, and advertising engagement on customer decisions [13][14].

However, many of these studies face limitations due to their survey-based approaches and may not fully capture real-time digital interactions. Therefore, the current study emphasizes analyzing online behavioral data to improve forecasting accuracy. One method involves collecting user engagement data from online platforms and applying machine learning models such as Neural Networks, Decision Trees, and Logistic Regression to predict customer responses.

These models are trained using features like call duration, engagement frequency, and message content. Some research has also explored how audio and video media influence customer purchasing behavior, indicating that multimedia marketing can significantly impact decision-making [15–17].

Fundamental Study

3.1 Regression Model

Regression analysis is a powerful statistical technique used to examine the relationships between one or more independent variables and a dependent variable. There are various types of regression analysis, but at their core, they all focus on understanding how independent variables influence the dependent variable. Regression analysis provides valuable insights that can be applied to improve products and services.

It is a reliable method for identifying which factors affect the variable of interest. The process of performing regression allows one to determine which factors are most important, which can be disregarded, and how these factors interact with each other. To fully comprehend regression analysis, it is essential to understand the following key terms:

- Dependent Variable: The primary variable you are trying to understand or predict.
- Independent Variables: The factors that are believed to have an impact on the dependent variable.

3.1.1 Best-Subset Logistic Regression

Best-subset logistic regression is a method that operates similarly to backward stepwise regression. Stepwise regression is a technique for fitting regression models where variables are added or removed at each step based on certain criteria. There are different types of stepwise regression, such as forward selection, backward elimination, and a hybrid approach.

Forward Selection: The process starts with an empty model (no predictors) and progressively adds variables that contribute most to reducing the residual sum of squares. This process continues until no further variables significantly improve the model.

Backward Elimination: This method starts with a full model and removes the variables with the highest p-values (those least statistically significant). The process continues until only significant variables remain in the model.

Hybrid Approach: This combines both forward and backward approaches. At each step, the model evaluates whether to add or remove variables based on their p-values, optimizing the model's performance.

Best-subset selection is a crucial step in regression modeling. In scenarios where the dependent variable is categorical (such as success or failure), and the independent variables can be a mix of continuous and categorical factors, best-subset logistic regression is employed. When the dependent variable is binary, logistic regression is the appropriate method to use. In such cases, the best-subset approach helps identify the most significant variables by applying logistic regression.

Signal (1978) introduced a methodology for effectively screening nonlinear regression models, which set the foundation for the current best-subset methods used in modern software. These techniques are widely applied in advanced statistical packages. For example,

consider a situation where the response variable (Y) has two possible outcomes (coded as 0 or 1), and the vector of covariates x0=(x0,x1,x2,...xp)x_0 = (x_0, x_1, x_2, ... x_p)x0 =(x0 ,x1 ,x2 ,...xp) needs to be estimated without error. Logistic regression models are used to model this relationship, and the best-subset selection process aims to determine which predictors are the most impactful in explaining the binary outcome.

$$Pr(Y = 1 \mid x) = \pi(x) \quad (1)$$

where

$$\pi(x) = e^{g(x)} / (1 + e^{g(x)}) \quad (2)$$

and

$$g((x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p \quad (3)$$

Then the function of likelihood data will be

$$L(\beta) = [\sum_{i=1}^{n} yi \ln(\pi i) + (1 - yi) \ln(1 - \pi i)] \quad (4)$$

Where $\pi_i = pi(x_i)$

The majority of the measurable bundles utilize the reweighed slightest square strategy to get the most extreme probability estimator of β. it tends to be communicated as

$$B = (X'VX)\text{-}1\ X'\ V\ z \quad (5)$$

3.1.2 Random Forest for Variable Selection

Random Forest (RF) is a non-parametric statistical method based on decision trees, commonly used for both regression and classification tasks. It operates with minimal assumptions about the model and relies heavily on the available data. The key idea behind RF is to build multiple decision trees and aggregate their results. Each tree in the forest is built using random subsets of the data (through a process called bootstrapping) and a random selection of predictor variables at each node.

In essence, a random forest consists of numerous decision trees, each with leaf nodes representing the predicted output, and internal nodes representing the decision criteria. The RF algorithm improves predictive accuracy by combining the outputs from many individual trees, reducing overfitting, and improving generalization. Additionally, by using random sampling for each tree, it provides robustness to variance in the dataset.

3.2 Telemarketing

Telemarketing is a direct marketing strategy in which sales representatives contact potential customers to promote products or services, typically over the phone, or in some cases through online meetings or in-person appointments scheduled during the call. It can also include pre-recorded sales pitches, which are automatically dialed and played to the recipient.

Telemarketing involves reaching out to potential clients, assessing their suitability, and persuading them to purchase or learn more about products or services. It is distinct from

other forms of marketing, such as direct mail or email marketing, which do not involve real-time communication.

3.2.1 Abnormal Behavior Therapy for Supply

This section refers to a psychological intervention method that uses systematic behavior modification to improve outcomes in specific settings. For instance, therapeutic techniques like cognitive-behavioral therapy are applied to assess and modify the thought patterns and behaviors of individuals, often aimed at reducing churn or customer dissatisfaction. The therapy involves understanding emotional and cognitive processes that influence behavior, with a focus on developing effective strategies to promote desired changes.

The effect of abnormal behavior therapies on churn rates or customer retention can be seen through statistical evaluation. Results show that structured interventions, when applied correctly, can reduce churn rates by about 25%, indicating an improved client retention process. The strategy involves closely monitoring customer satisfaction and refining the approach to offer better outcomes over time.

3.2.2 Availability Bias and One-sided Thinking

Tversky and Kahneman (1973) discussed availability bias, a cognitive heuristic where people judge the likelihood of events based on how easily examples come to mind. This bias leads individuals to overestimate the frequency of events they have recently or frequently encountered. For instance, if a person often hears about airplane crashes in the media, they might overestimate the risk of flying, even though statistically, air travel is much safer than other modes of transportation.

This type of thinking significantly impacts decision-making, as individuals may weigh recent or highly memorable events more heavily than statistical reality. Media plays a substantial role in amplifying these biases, as repeated exposure to certain types of news or events can skew an individual's perception of their frequency or importance.

3.2.3 Customer Churn

Customer churn refers to the loss of clients over a specific period, typically calculated as a percentage of the customer base. It is influenced by various factors, both internal (such as dissatisfaction with the product or service) and external (such as changes in the market or competitive actions). Several studies have identified key drivers of customer churn, including poor service quality, unmet expectations, and competitive pricing.

Research has shown that customer churn is closely related to levels of customer engagement and satisfaction. Dissatisfaction is often linked to negative experiences, such as unresolved service issues or inadequate product offerings. By improving customer relationships, businesses can reduce churn and foster customer loyalty, which is vital for long-term success.

3.2.4 Decision Support Systems (DSS) in Marketing

Decision Support Systems (DSS) are essential tools in modern marketing that aid managers in making informed decisions. These systems leverage various analytical techniques, including simulations, knowledge-based

systems, and optimization models, to provide insights into marketing strategies. The primary purpose of DSS is to help marketers evaluate potential outcomes of different strategies and make decisions before committing resources.

For example, marketing managers can use DSS to conduct "what-if" analyses, which examine the potential impact of different strategies on sales or customer satisfaction. These systems help evaluate factors such as product pricing, customer segmentation, and promotional strategies, allowing marketers to optimize their decisions and minimize risk.

In the past few decades, the evolution of DSS tools has been significant, especially with the rise of machine learning (ML) and artificial intelligence (AI) technologies. These technologies have enabled more accurate predictions and deeper insights into consumer behavior, providing marketers with the tools to refine their strategies and improve customer targeting. The use of data mining, combined with statistical analysis, has become a cornerstone of modern marketing decision-making.

Proposed Work:

This section outlines the structure of the proposed methodology. The model framework is depicted in Figure 4.1, which highlights the two primary components of the process: data preprocessing and modeling. Initially, the raw data are presented and described in detail.

The objective of this study is to analyze the relationship between the success of bank telemarketing campaigns and various factors such as customer demographics, social and economic context, and other relevant characteristics. Additionally, the study aims to predict the success rate of bank telemarketing efforts.

The data for this analysis was obtained from a study conducted by Moro and is publicly available for download. The dataset, which was collected from a Portuguese retail bank, focuses on the success of telemarketing campaigns. In this dataset, multiple contacts were made to the same customer to determine whether they would sign up for a term deposit. The dataset consists of 4,119 instances and 21 attributes. Excluding the target variable (whether the customer subscribed to the term deposit), there are 20 features, including age, occupation, education, and previous credit history.

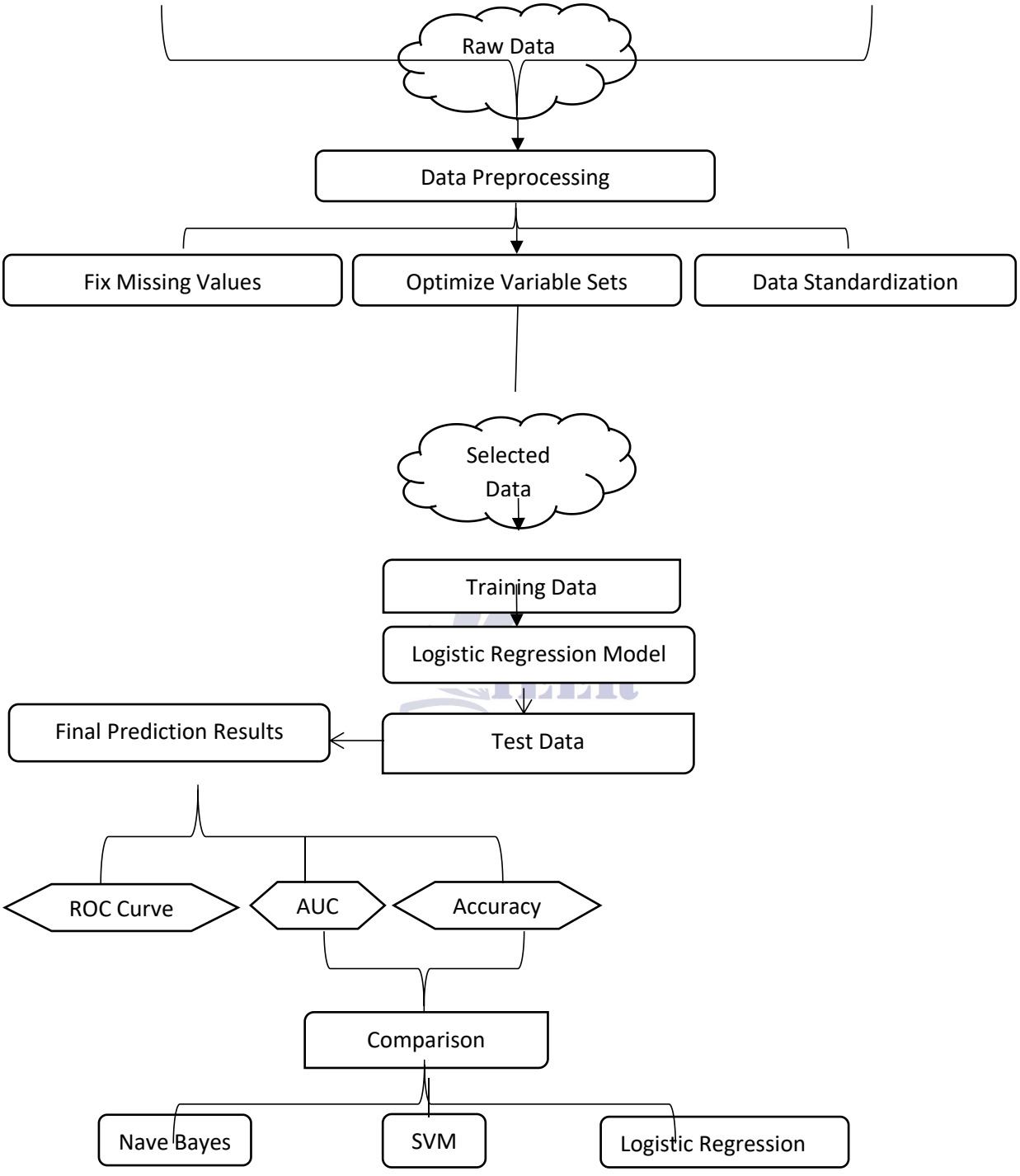

Figure 4.1 Model Framework

4.1 Data Collection Preprocessing

Initially, raw data are collected in various formats, comprising 45,211 records and 17 attributes. This dataset is publicly available through the UCI Machine Learning Repository [42]. To enhance the predictive power of our model, the raw data, which are often redundant, inconsistent, or incomplete, need to be cleaned and optimized. Therefore, data preprocessing is essential before any modeling can begin. The preprocessing steps are carried out to ensure the data is ready for the model.

4.2 Handling Missing Values

The dataset contains several missing values, particularly for attributes such as "job," "marital status," and other key customer information. Out of 4,119 records, 1,012 contain missing values. These records are labeled as "unknown," which could potentially introduce biases or noise into the analysis. Proper preprocessing is crucial, as raw data are rarely perfect. Issues such as invalid ages, missing values, or outliers must be addressed to avoid skewing the results. One common technique to manage missing data is imputation, where missing values are substituted with averages or calculated based on existing data.

4.3 Normalizing the Data

Normalization refers to transforming the data into a standard format, ensuring consistency and minimizing redundancy. Since the attributes in the dataset vary in magnitude and units, normalization helps align them. This step ensures that each feature has a comparable scale, making the dataset suitable for modeling and machine learning techniques. By transforming the data, we make it easier to process and interpret, improving the accuracy and efficiency of the model.

4.4 Removing Irrelevant Features

Next, we perform a feature selection process to discard irrelevant or weakly correlated variables. In this case, any variables with a correlation coefficient lower than 0.05 are removed, resulting in the dataset being reduced to five core attributes: "income," "job," "age," "marital status," and "previous outcome." Feature selection aims to improve the model's performance by removing noise and irrelevant data, ensuring that only the most important predictors are used for training the model.

4.5 Data Selection

When dealing with large datasets, missing values can skew results. One option is to remove records with missing values entirely, although this can lead to a loss of potentially useful data. A more effective approach is to use imputation methods, where missing values are filled using statistical techniques such as the mean, median, or most frequent value. For categorical variables, missing data might be replaced with the most common category. Selecting the appropriate imputation method depends on the nature of the data and the specific task at hand.

4.6 Data Testing and Training

After collecting and preprocessing the data, we use logistic regression (LR) for training the model. In contrast to previous studies that used a 70% training and 30% testing split, I perform an 85% training and 15% testing split. This adjustment yields a 93% accuracy rate, which is 1% higher than the results from earlier research. For this task, I reclassified data from 9 levels to 5 levels, including categories

such as "unfit," "rich," "middle-aged," etc. Logistic regression is a probabilistic classification technique that provides insights into the relationships between independent variables and the target variable.

4.7 Model Comparison

In this study, I compare different machine learning models to predict the success of telemarketing campaigns. I use Python to implement logistic regression, and compare it with other common models, such as Support Vector Machines (SVM) and Naive Bayes. A 20% testing split is used, and the model's performance is evaluated based on metrics such as accuracy, area under the ROC curve (AUC), and the precision-recall trade-off. Logistic regression shows superior performance in comparison to the other models, as evidenced by the higher AUC value, indicating better predictive capabilities.

5: Results & Discussion

In the context of financial marketing, effective campaigns play a crucial role, and term deposits are a significant aspect of a bank's business strategy. Telemarketing has proven to be one of the most effective methods for reaching potential clients. This research leverages data mining techniques to predict the success of bank telemarketing campaigns. The dataset consists of customer information from a Portuguese bank collected between 2008 and 2013. The model is trained using 85% of the data, achieving an accuracy rate of 93%.

5.1 Logistic Regression Results

5.2 Discussion of Results

We analyzed a large dataset of customer information from a Portuguese bank, which included 45,211 records and 17 attributes. Through feature selection, we reduced the dataset to 22 key features. The logistic regression model was applied to calculate the key metrics, including the Receiver Operating Characteristics (ROC) area.

The analysis was conducted on a Windows 10 server using an Intel Xeon 5500 processor, and the model was run 20 times to ensure the reliability of the results. Table 5.1 provides a detailed summary of the input factors used in the model.

Table 5.1: Input Factors

Factors	Description

Logit Regression Results

Dep. Variable:	y	No. Observations:	4521
Model:	Logit	Df Residuals:	4509
Method:	MLE	Df Model:	11
Date:	Mon, 24 Dec 2018	Pseudo R-squ.:	-0.2677
Time:	16:24:27	Log-Likelihood:	-2048.0
converged:	True	LL-Null:	-1615.5
		LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
previous	-0.1868	0.030	-6.168	0.000	-0.246	-0.127
job_blue-collar	-2.2574	0.128	-17.687	0.000	-2.508	-2.007
job_retired	-0.6482	0.174	-3.734	0.000	-0.988	-0.308
job_services	-1.9535	0.176	-11.123	0.000	-2.298	-1.609
job_student	-0.8832	0.280	-3.159	0.002	-1.431	-0.335
month_aug	-1.8360	0.124	-14.797	0.000	-2.079	-1.593
month_dec	-0.3869	0.513	-0.754	0.451	-1.393	0.619
month_jul	-1.9696	0.138	-14.318	0.000	-2.239	-1.700
month_nov	-1.9183	0.174	-11.033	0.000	-2.259	-1.578
month_oct	0.1183	0.255	0.464	0.642	-0.381	0.618
month_sep	-0.6314	0.330	-1.912	0.056	-1.279	0.016
poutcome_success	2.3357	0.239	9.771	0.000	1.867	2.804

Accuracy of logistic regression classifier on test set: 0.93
15-fold cross validation average accuracy of navebayse is : 0.856
15-fold cross validation average accuracy svm is: 0.892

Factors	Description
Interest	Whether the bank offers monthly interest to customers
Gender	Customer's gender (male/female)
Bank Profiling Indicators	Customer's loyalty and interaction with bank services

Table 5.2: Correlation Between Attributes and Descriptions

Attribute	Description
Age	Customer's age
Job	Customer's occupation (admin, blue-collar, etc.)
Marital Status	Customer's marital status (single, married, etc.)
Education	Customer's education level (primary, secondary, tertiary, etc.)
Housing Loan	Whether the customer has a housing loan (yes/no)

Conclusion:

In conclusion, this research demonstrates the utility of machine learning models, particularly logistic regression, in predicting the success of telemarketing campaigns. The results show that with appropriate preprocessing and feature selection, models can achieve high accuracy in predicting customer behavior. Data mining techniques, such as logistic regression, continue to be valuable tools in financial marketing, providing insights that help organizations optimize their marketing strategies and improve customer engagement.

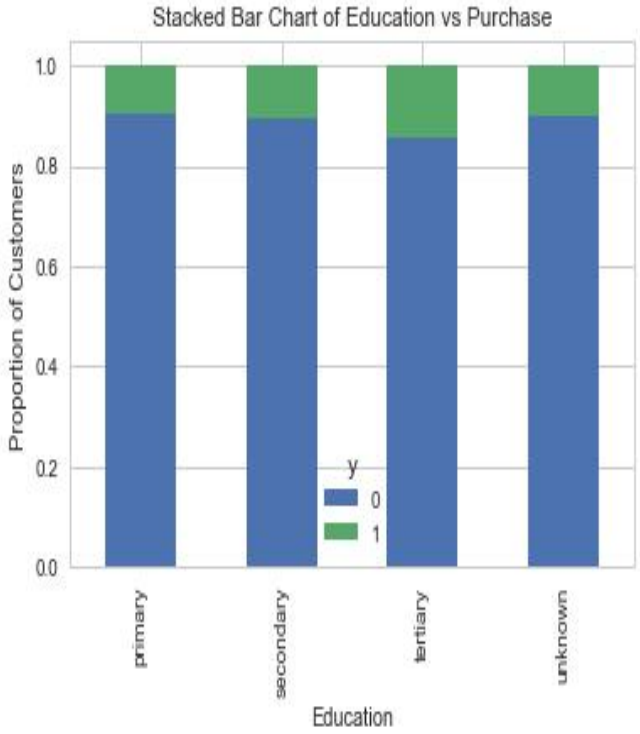

Figure 5.1 Stacked Bar chart of Education vs purchase

Figure 5.1 illustrates a method of regression, where the basic summary of the outcome is presented before additional factors are revealed to the audience. This leads to a model of prediction that can be easily applied in real-time scenarios, such as banks with limited data availability. Logistic regression, as a predictive technique, has gained increasing attention in recent years, particularly among Human Resource managers in Europe, who are integrating advanced information technologies into their processes.

Some may prefer to adopt more straightforward and traditional methods, but logistic regression, with its proven ability to handle complex datasets, can still be adapted for practical use. By leveraging predictive

factors, logistic regression can generate a version of the predicted outcome, known as the "logistic regression score." This score can be calculated using the following formula:

Predicted mortality = e (ß0+ßiXi)1+e(ß0+ßiXi) .....(6)

This formula, which predicts outcomes like mortality rates, is particularly useful for medical professionals. Heart doctors, for instance, can rely on these predictions without needing to perform complicated calculations themselves. The model is easy to implement, validate, and use at the point of care, making it highly practical in real-world settings.

Due to its increasing popularity and the simplicity of its application, the logistic regression model is becoming a valuable tool for individualized risk assessments. Its results can be used to refine decision-making in high-risk situations, contributing to more accurate and effective patient care or business strategies.

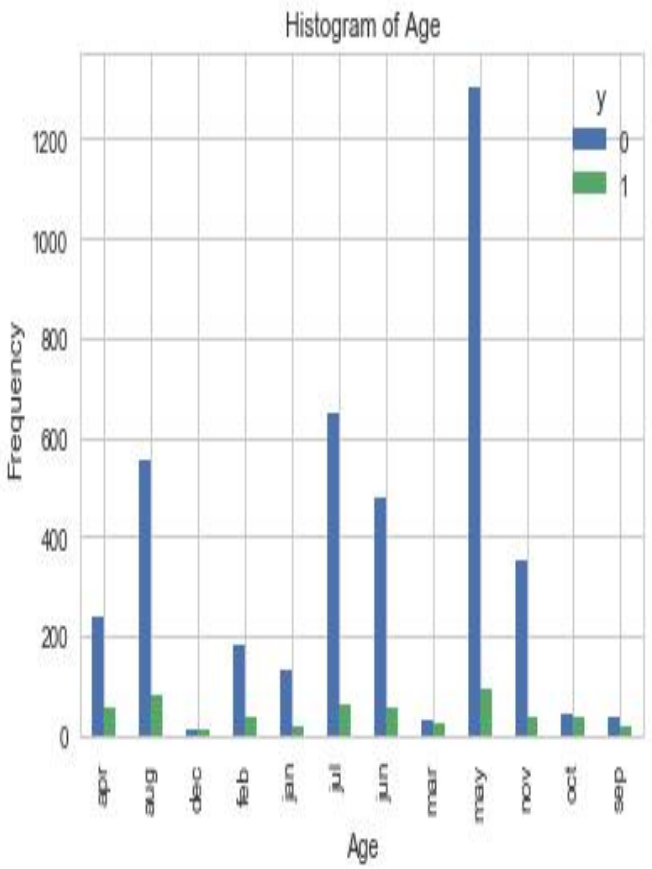

Figure 5.2 Histogram of Age

Figure 5.2 shows that the codebook used here is similar to others presented in this textbook. In fact, there are no significant changes to how the continuous variables (such as Age, Income, and Cigarettes) are represented. However, when considering logistic regression processing, it is important to note that the attributes of the categorical variables differ in the following ways:

The values for each categorical variable are encoded in a way that allows for changes to be more clearly visualized.

The numbering of the categories for each variable starts from 0, rather than 1.

For the categorical variables such as Sex and Class, the first category (0) is defined as the

baseline group, which will be further explained in the results section.

- For the target variable (status), the higher category (1, which refers to "unsuccessful outcome") is treated as the reference group for this logistic regression model.

In general, when working with logistic regression, it is often assumed that the model will mainly consist of categorical variables. If there are no more than one or two continuous variables present in the model, issues like multicollinearity are not a concern, as there would be little to no correlation among different predictor variables.

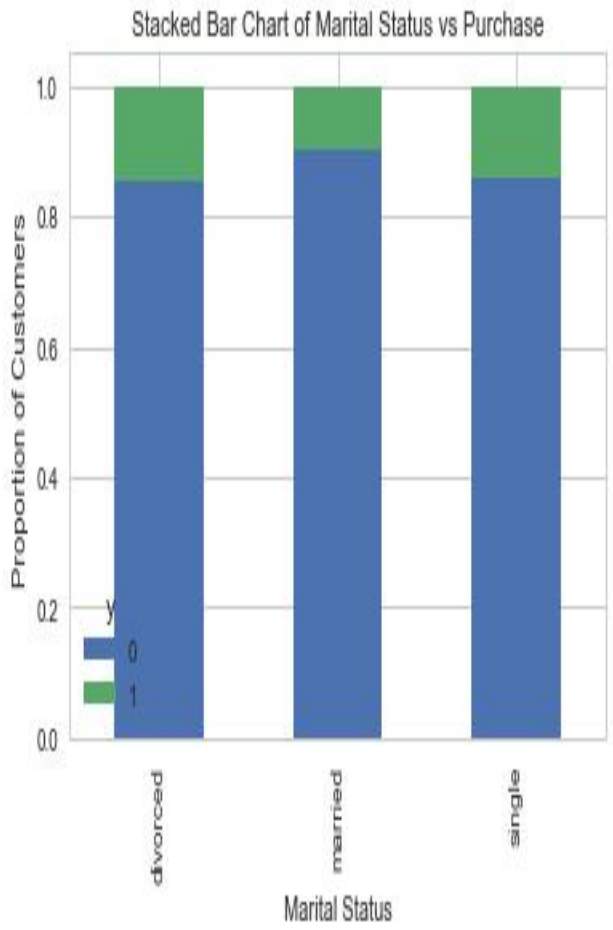

Figure 5.3 Stacked bar chart of Marital Status vs. Purchases

Figure 5.3 demonstrates the application of Binary Logistic Regression. In this method, both binary independent variables and a dependent variable are utilized. The dependent variable is set to a binary outcome (yes-no or 0-1) to answer the question, "Was the outcome successful?" The independent variables include age, gender, marital status, education, income, debt-to-income ratio, and length of time with the bank. The data collected from the bank was processed using SPSS 23.0 software, and the Logistic Regression (LR) Model was applied to analyze the results. The logistic regression function used to describe the model is as follows:

Logistic regression can handle binary, continuous, and categorical data. This flexibility makes it a robust and effective tool for regression analysis. In logistic regression, both the independent and dependent variables are analyzed to understand the cause-effect relationships, particularly with binary outcomes (yes/no, positive/negative, etc.). LR Analysis is essential when the dependent variable consists of two or more categories, and it helps to assess the relationship between the dependent and independent variables.

The logistic regression model is structured such that the predicted probability of the event occurring, denoted as f(x), must lie between 0 and 1. A key concept in logistic regression is the logit transformation, which introduces "odds." The odds are calculated as the ratio of the probability of an event occurring (f(x)) to the probability of it not occurring (1 - f(x)). This transformation is critical for interpreting the model in terms of probability.

$$g(y) = \log p(y)(1-p(y)) = \mu + ß1x1 + ß2x2 + BJxJ = ß \times 1\mu + x'b \quad (7)$$

When x increases one unit, the odds ratio increases exp (ß) object.

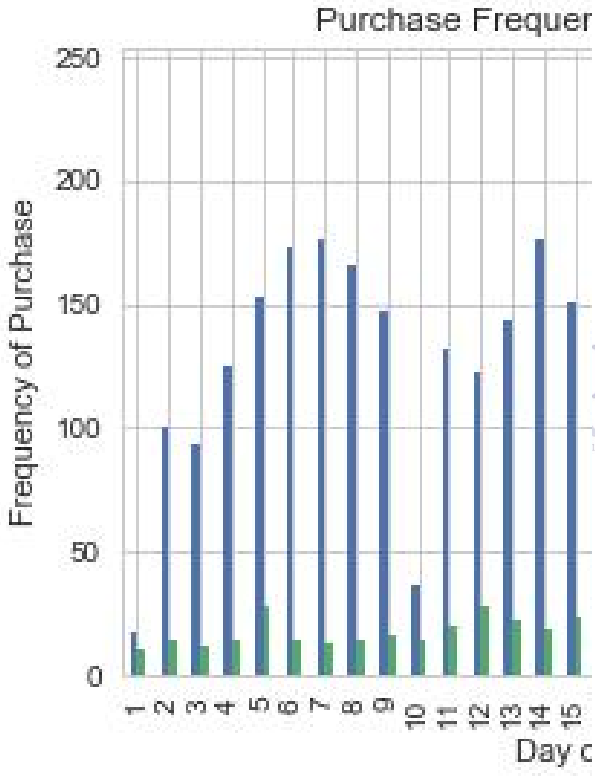

Figure 5.4 Purchase Frequency for Day of week

Figure 5.4 illustrates the results of applying **Logistic Regression** to the dataset, highlighting significant variables. Upon evaluating the model, we found that variables such as **age**, **interest (old)**, **campaign_cat**, and **cons_price_index_cat** were not statistically significant at the alpha = 0.05 level, as shown in Table 25 of the odds ratio statistics.

After excluding these non-significant variables and re-running the regression analysis, we observed that all remaining variables became significant at the 0.05 alpha level. This resulted in our final regression model, which includes 14 significant predictors. The **Pearson residuals** of the model were mostly unremarkable, with 62 instances showing residuals greater than 3 standard deviations from 0, as shown in Figure 26. These large residuals occurred when predicted probabilities were near 0 or 1, which was expected due to the nature of the logistic model.

The model revealed that the **predicted probabilities** exhibited an increasing trend when the **probability prediction** approached 0 and then became more pronounced as the predicted probabilities neared 1. This pattern of behavior is not surprising, considering the distribution of residuals. Since the occurrence of such large residuals was minimal (1.5% of the total observations), it was not considered a significant issue. The residuals indicated that the model fit was acceptable overall.

Additionally, the **DF beta plots** clearly showed that the impact of each variable was within

reasonable bounds, with no significant **DF beta** values exceeding 0.6, which would indicate a problematic predictor. This suggests that no single predictor disproportionately influenced the model's coefficients, reinforcing the stability of the regression model.

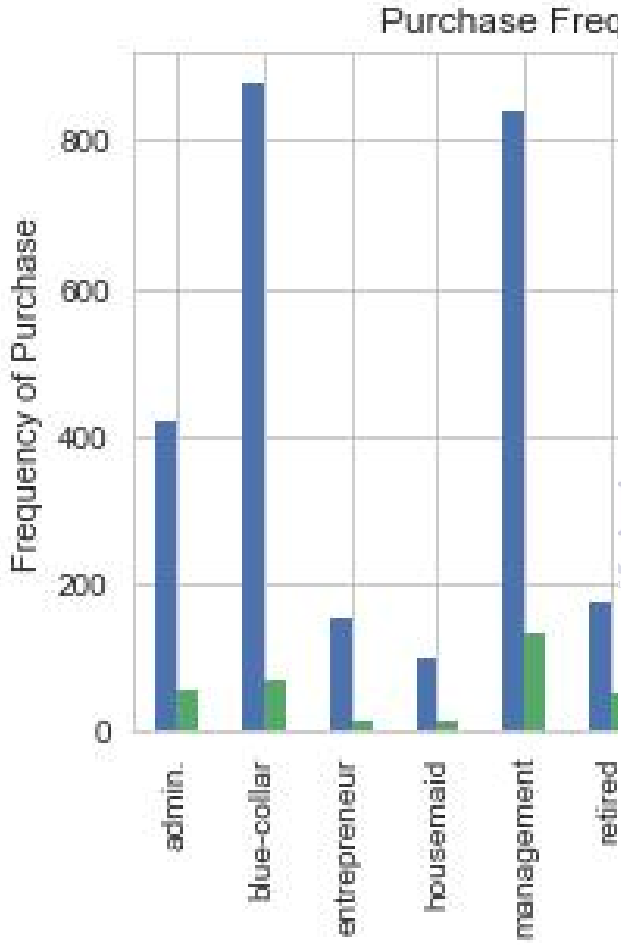

Figure 5.5 Purchase Frequencies for Job Title

Figure 5.5 demonstrates that changes in any of the predictor variables affect the log-odds ratio of the outcome. The log-odds ratio is used to calculate the change in the odds (P[yes]/P[no]). The odds ratios for all predictor variables, where the 95% confidence interval does not contain 1, are listed in Table 29. Variables that include 1 within their 95% confidence interval are recorded in Table 30, indicating that these variables do not significantly impact the outcome. Variables with confidence intervals containing 1 are not considered meaningful predictors.

Among the predictor variables, the most significant factors influencing the odds ratios are **job**, **break**, and **word**. For the **job** variable, the reference category is "unemployed." Most job categories show an odds ratio between 2.5 and 5, meaning that individuals in these job categories are 2.5 to 5 times more likely to experience a certain outcome than those who are unemployed. However, self-employed individuals and small business owners were found to be nearly the same as unemployed individuals in terms of the likelihood of experiencing the event.

For the **client data** variable, someone reporting "no message" (lack of information) is 3 times more likely to respond positively than someone whose status is unknown. Additionally, **client system** categories show that people who do not have a fixed contract or commitment are 3 times more likely to engage with the service.

Among the time-related variables, **month** is a critical factor, with September, October, and December showing the lowest likelihood of engagement. November stood out, with clients being 5% more likely to sign up in November than in September. **Lens** was another key factor, with customers contacted through **Group Sound** being 2.4 times more likely to purchase compared to those contacted via other methods like telecom.

For **client behavior**, those who had previously made a purchase were 5.7 times more likely to make another purchase. Similarly, customers who had been contacted before had a 3.9 times higher chance of engaging with the service.

The **social and economic variables** also played a significant role. For example, individuals in the highest category of **euribor3m_cat** were 24 to 28 times more likely to make a deposit than those in the lowest categories. Additionally, **receiver and plot** variables were important, with the smallest levels of **employed_cat** and **emp_var_rate_cat** showing the highest odds ratios. At the highest levels, professionals were 6 to 17 times more likely to make a decision compared to those at lower levels.

Finally, the **cons_price_index_cat** variable at its highest value showed a significant increase in the likelihood of a positive response, being twice as likely to result in a purchase than at its lowest value.

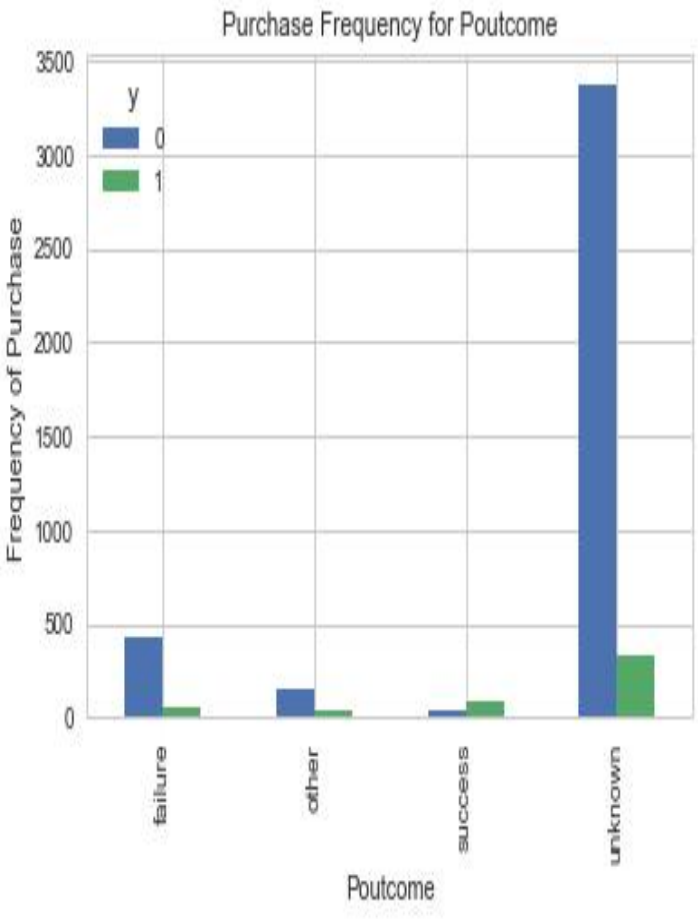

Figure 5.6 Purchase Frequency for pout come

Figure 5.6 shows the analysis of the Odds Ratio for each explanatory variable at the 0.05 significance level for the grouped dataset. The results are displayed in **Table 24**. Each variable is analyzed based on its odds ratio estimate. The null hypothesis for each explanatory variable is tested, where the outcome of each variable's levels is compared against the base category.

In the logistic regression model, the variables **age**, **campaign_cat**, **cons_price_index_cat**, and **large_cheating_plan** were not significant at the alpha = 0.05 level. For further details on the odds ratio statistics, see **Table 25**. When these variables were removed from the analysis, the remaining variables became significant, as shown in **Table 3**, reflecting the key factors influencing the outcome.

The **Pearson residuals** of the logistic regression model are fairly distributed, showing no significant skew in the residuals. However, some values with residuals greater than 3 were observed. These high residuals are associated with the **predicted event** probabilities close to 0 or 1, indicating potential discrepancies in the data. This is not unexpected in a real-world scenario, as such residuals often occur when the predicted probabilities approach extremes (0 or 1). Since the occurrence of these high residuals is relatively rare (approximately 1.5% of the observations), they are not deemed a major concern.

The **DF Beta** plots also reveal that the influence of the explanatory variables is minimal. No DF Beta values were found to be greater than 0.6 for any variable, suggesting that none of the predictor variables are disproportionately affecting the model's outcome.

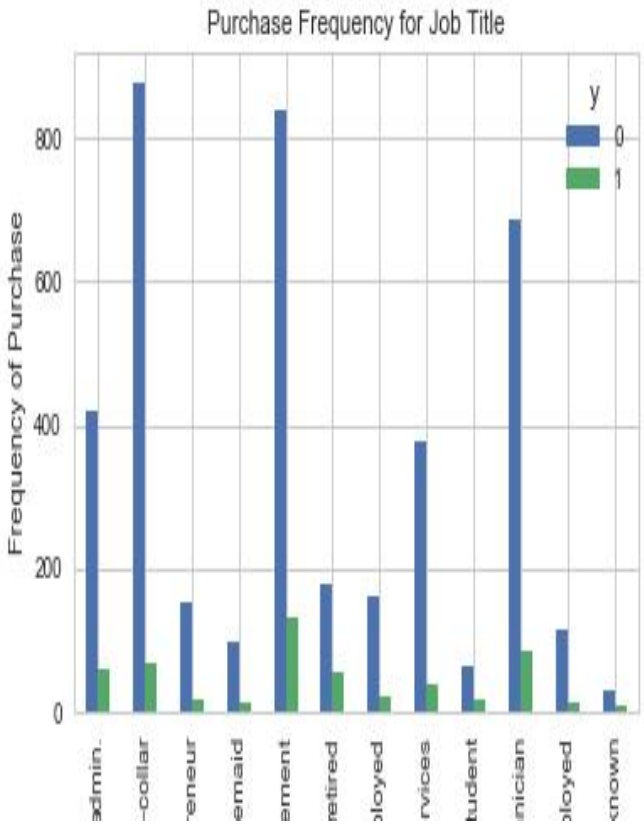

Figure 5. 7 Purchase frequencies for job title

Figure 5.7 illustrates the distribution of customers by job title. As shown in the pie chart, the largest portion of customers is from managerial positions, accounting for 26% of the total. This is followed by employees in various other sectors, including technicians and workers.

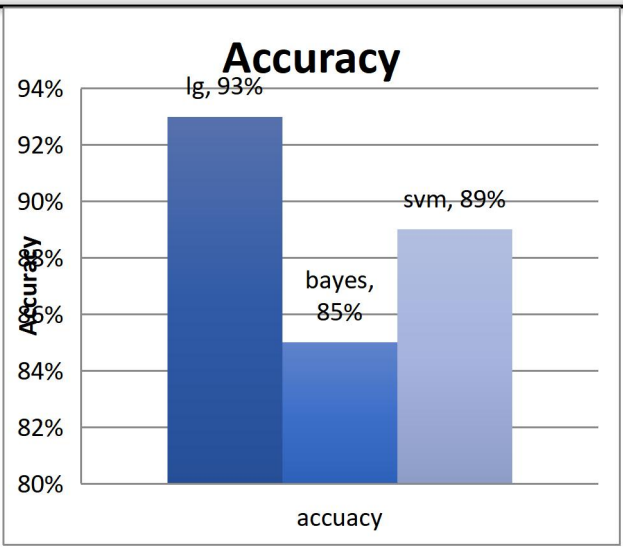

Figure 5.8 Accuracy

Figure 5.8 compares the performance of different models, including Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes (NB). The accuracies for each model are as follows:

- **Logistic Regression**: 93%
- **Support Vector Machine**: 89%
- **Naive Bayes**: 85%

This result shows that Logistic Regression outperforms both SVM and Naive Bayes. A comparison with previous research (Yiyan Jiang, 2018) demonstrates that LR has improved by 1% over the previously reported accuracy. Specifically:

Author	LR Accuracy	NB Accuracy	SVM Accuracy	Findings
Yiyan Jiang, 2018	92.03%	86.11%	90.70%	LR p well wit
Proposed Work	93%	85.6%	89.2%	accuracy LR accuracy improved by 1%
Improvement	+1%	-1%	-1%	LR provides the best result

In this study, 85% of the data was used for training, while 15% was used for testing. This approach resulted in a 93% accuracy, which is 1% higher than the previous work.

Conclusion:

This research develops a Logistic Regression (LR) model to analyze the relationship between customer data, socio-economic factors, and the success of bank telemarketing efforts. The model is used to predict whether a customer will respond positively to a marketing campaign, and LR performs better than other models in terms of prediction accuracy.

Despite the promising results, the study is limited by the data, which was sourced from a single Portuguese retail bank. Therefore, the findings may not be universally applicable. The relatively small dataset and the limited scope of customer information constrain the generalizability of the results.

Overall, LR is shown to be a reliable tool for predicting telemarketing success and can be easily applied in real-world bank marketing scenarios. It allows banks to improve targeting, potentially increasing response rates while reducing marketing costs.

Future Work:

In future studies, we aim to improve the algorithm's efficiency by incorporating a wider variety of financial instruments and comparing LR with other advanced techniques, such as domain knowledge-based models and deep learning approaches. These enhancements could further boost the model's accuracy and predictive power, contributing to a more refined and robust telemarketing strategy.

Reference

[1] Khan, S.U.R., Asif, S., Bilal, O. et al. Lead-cnn: lightweight enhanced dimension reduction convolutional neural network for brain tumor classification. Int. J. Mach. Learn. & Cyber. (2025). https://doi.org/10.1007/s13042-025-02637-6.

[2] Khan, S. U. R., Asim, M. N., Vollmer, S., & Dengel, A. (2025). Robust & Precise Knowledge Distillation-based Novel Context-Aware Predictor for Disease Detection in Brain and Gastrointestinal. arXiv preprint arXiv:2505.06381..

[3] Hekmat, A., et al., Brain tumor diagnosis redefined: Leveraging image fusion for MRI enhancement classification. Biomedical Signal Processing and Control, 2025. 109: p. 108040.

[4] Khan, Z., Hossain, M. Z., Mayumu, N., Yasmin, F., & Aziz, Y. (2024, November). Boosting the Prediction of Brain Tumor Using Two Stage BiGait Architecture. In 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (pp. 411-418). IEEE.

[5] Khan, S. U. R., Raza, A., Shahzad, I., & Ali, G. (2024). Enhancing concrete and pavement crack prediction through hierarchical feature integration with VGG16 and triple classifier ensemble. In 2024 Horizons of Information Technology and Engineering (HITE)(pp. 1-6). IEEE https://doi. org/10.1109/HITE63532.

[6] Khan, S.U.R., Zhao, M. & Li, Y. Detection of MRI brain tumor using residual skip block based modified MobileNet model. Cluster Comput 28, 248 (2025). https://doi.org/10.1007/s10586-024-04940-3

[7] Khan, U. S., & Khan, S. U. R. (2024). Boost diagnostic performance in retinal disease classification utilizing deep ensemble classifiers based on OCT. Multimedia Tools and Applications, 1-21.

[8] Raza, A., & Meeran, M. T. (2019). Routine of encryption in cognitive radio network. Mehran University Research Journal of Engineering & Technology, 38(3), 609-618.

[9] Al-Khasawneh, M. A., Raza, A., Khan, S. U. R., & Khan, Z. (2024). Stock Market Trend Prediction Using Deep Learning Approach. Computational Economics, 1-32.

[10] Khan, U. S., Ishfaque, M., Khan, S. U. R., Xu, F., Chen, L., & Lei, Y. (2024). Comparative analysis of twelve transfer learning models for the prediction and crack detection in concrete dams, based on borehole

images. Frontiers of Structural and Civil Engineering, 1-17.

[11] Khan, S. U. R., & Asif, S. (2024). Oral cancer detection using feature-level fusion and novel self-attention mechanisms. Biomedical Signal Processing and Control, 95, 106437.

[12] Farooq, M. U., Khan, S. U. R., & Beg, M. O. (2019, November). Melta: A method level energy estimation technique for android development. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-10). IEEE.

[13] Waqas, M., Tahir, M. A., & Khan, S. A. (2023). Robust bag classification approach for multi-instance learning via subspace fuzzy clustering. Expert Systems with Applications, 214, 119113.

[14] Asim, M. N., Ibrahim, M. A., Malik, M. I., Dengel, A., & Ahmed, S. (2020). Enhancer-dsnet: a supervisedly prepared enriched sequence representation for the identification of enhancers and their strength. In Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part III 27 (pp. 38-48). Springer International Publishing.

[15] Meeran, M. T., Raza, A., & Din, M. (2018). Advancement in GSM Network to Access Cloud Services. Pakistan Journal of Engineering, Technology & Science [ISSN: 2224-2333], 7(1).

[16] Raza, A.; Meeran, M.T.; Bilhaj, U. Enhancing Breast Cancer Detection through Thermal Imaging and Customized 2D CNN Classifiers. VFAST Trans. Softw. Eng. 2023, 11, 80–92.

[17] Dai, Q., Ishfaque, M., Khan, S. U. R., Luo, Y. L., Lei, Y., Zhang, B., & Zhou, W. (2024). Image classification for sub-surface crack identification in concrete dam based on borehole CCTV images using deep dense hybrid model. Stochastic Environmental Research and Risk Assessment, 1-18.

[18] Muhammad, N. A., Rehman, A., & Shoaib, U. (2017). Accuracy based feature ranking metric for multi-label text classification. International Journal of Advanced Computer Science and Applications, 8(10).

[19] Mehmood, F., Ghafoor, H., Asim, M. N., Ghani, M. U., Mahmood, W., & Dengel, A. (2024). Passion-net: a robust precise and explainable predictor for hate speech detection in roman urdu text. Neural Computing and Applications, 36(6), 3077-3100.

[20] Waqas, M., Tahir, M. A., Al-Maadeed, S., Bouridane, A., & Wu, J. (2024). Simultaneous instance pooling and bag representation selection approach for multiple-instance learning (MIL) using vision transformer. Neural Computing and Applications, 36(12), 6659-6680.

[21] Khan, S.U.R.; Asif, S.; Bilal, O.; Ali, S. Deep hybrid model for Mpox disease diagnosis from skin lesion images. Int. J. Imaging Syst. Technol. 2024, 34, e23044.

[22] Mahmood, F., Abbas, K., Raza, A., Khan,M.A., & Khan, P.W. (2019). Three Dimensional Agricultural Land Modeling using Unmanned Aerial System (UAS). International Journal of Advanced Computer Science and Applications (IJACSA) [p-ISSN : 2158-107X, e-ISSN : 2156-5570], 10(1).

[23] Waqas, M., Tahir, M. A., & Qureshi, R. (2023). Deep Gaussian mixture model

based instance relevance estimation for multiple instance learning applications. Applied intelligence, 53(9), 10310-10325.

[24] Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X.; Zhu, Y. GLNET: Global–local CNN's-based informed model for detection of breast cancer categories from histopathological slides. J. Supercomput. 2023, 80, 7316–7348.

[25] Saleem, S., Asim, M. N., Van Elst, L., & Dengel, A. (2023). FNReq-Net: A hybrid computational framework for functional and non-functional requirements classification. Journal of King Saud University-Computer and Information Sciences, 35(8), 101665.

[26] Hekmat, Arash, Zuping Zhang, Saif Ur Rehman Khan, Ifza Shad, and Omair Bilal. "An attention-fused architecture for brain tumor diagnosis." Biomedical Signal Processing and Control 101 (2025): 107221.

[27] Khan, S.U.R.; Zhao, M.; Asif, S.; Chen, X. Hybrid-NET: A fusion of DenseNet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. Int. J. Imaging Syst. Technol. 2024, 34, e22975.

[28] Waqas, M., & Khan, M. A. (2018). JSOPT: A framework for optimization of JavaScript on web browsers. Mehran University Research Journal of Engineering & Technology, 37(1), 95-104.

[29] HUSSAIN, S., Raza, A., MEERAN, M. T., IJAZ, H. M., & JAMALI, S. (2020). Domain Ontology Based Similarity and Analysis in Higher Education. IEEEP New Horizons Journal, 102(1), 11-16.

[30] Khan, S.U.R.; Raza, A.;Waqas, M.; Zia, M.A.R. Efficient and Accurate Image Classification Via Spatial Pyramid Matching and SURF Sparse Coding. Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol. 2023, 7, 10–23.

[31] Farooq, M.U.; Beg, M.O. Bigdata analysis of stack overflow for energy consumption of android framework. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 1–2 November 2019; pp. 1–9.

[32] Waqas, M., Ahmed, S. U., Tahir, M. A., Wu, J., & Qureshi, R. (2024). Exploring Multiple Instance Learning (MIL): A brief survey. Expert Systems with Applications, 123893.

[33] Shahzad, I., Khan, S. U. R., Waseem, A., Abideen, Z. U., & Liu, J. (2024). Enhancing ASD classification through hybrid attention-based learning of facial features. Signal, Image and Video Processing, 1-14.

[34] Khan, S. R., Raza, A., Shahzad, I., & Ijaz, H. M. (2024). Deep transfer CNNs models performance evaluation using unbalanced histopathological breast cancer dataset. Lahore Garrison University Research Journal of Computer Science and Information Technology, 8(1).

[35] Bilal, Omair, Asif Raza, and Ghazanfar Ali. "A Contemporary Secure Microservices Discovery Architecture with Service Tags for Smart City Infrastructures." VFAST Transactions on Software Engineering 12, no. 1 (2024): 79-92.

[36] Asif Raza, Inzamam Shahzad, Ghazanfar Ali, and Muhammad Hanif Soomro. "Use Transfer Learning VGG16, Inception, and Reset50 to Classify IoT Challenge in Security Domain via Dataset Bench Mark."

Journal of Innovative Computing and Emerging Technologies 5, no. 1 (2025).

[37] Khan, S. U. R., Asif, S., Zhao, M., Zou, W., Li, Y., & Li, X. (2025). Optimized deep learning model for comprehensive medical image analysis across multiple modalities. Neurocomputing, 619, 129182.

[38] M. Waqas, Z. Khan, S. U. Ahmed and A. Raza, "MIL-Mixer: A Robust Bag Encoding Strategy for Multiple Instance Learning (MIL) using MLP-Mixer," 2023 18th International Conference on Emerging Technologies (ICET), Peshawar, Pakistan, 2023, pp. 22-26.

[39] Khan, S. U. R., Asif, S., Zhao, M., Zou, W., & Li, Y. (2025). Optimize brain tumor multiclass classification with manta ray foraging and improved residual block techniques. Multimedia Systems, 31(1), 1-27.

[40] Khan, S. U. R., Asim, M. N., Vollmer, S., & Dengel, A. (2025). AI-Driven Diabetic Retinopathy Diagnosis Enhancement through Image Processing and Salp Swarm Algorithm-Optimized Ensemble Network. arXiv preprint arXiv:2503.14209.

[41] Raza, A., Salahuddin, & Inzamam Shahzad. (2024). Residual Learning Model-Based Classification of COVID-19 Using Chest Radiographs. Spectrum of Engineering Sciences, 2(3), 367–396.

[42] Shahzad, Inzamam, Asif Raza, and Muhammad Waqas. "Medical Image Retrieval using Hybrid Features and Advanced Computational Intelligence Techniques." Spectrum of engineering sciences 3, no. 1 (2025): 22-65.

[43] Khan, Z., Khan, S. U. R., Bilal, O., Raza, A., & Ali, G. (2025, February). Optimizing Cervical Lesion Detection Using Deep Learning with Particle Swarm Optimization. In 2025 6th International Conference on Advancements in Computational Sciences (ICACS) (pp. 1-7). IEEE.

[44] M. Wajid, M. K. Abid, A. Asif Raza, M. Haroon, and A. Q. Mudasar, "Flood Prediction System Using IOT & Artificial Neural Network", VFAST trans. softw. eng., vol. 12, no. 1, pp. 210–224, Mar. 2024.

[45] Khan, S.U.R., Raza, A., Shahzad, I., Khan, S. (2025). Subcellular Structures Classification in Fluorescence Microscopic Images. In: Arif, M., Jaffar, A., Geman, O. (eds) Computing and Emerging Technologies. ICCET 2023. Communications in Computer and Information Science, vol 2056. Springer, Cham. https://doi.org/10.1007/978-3-031-77620-5_20

[46] Hekmat, A., Zuping, Z., Bilal, O., & Khan, S. U. R. (2025). Differential evolution-driven optimized ensemble network for brain tumor detection. International Journal of Machine Learning and Cybernetics, 1-26.

[47] Khan, S. U. R. (2025). Multi-level feature fusion network for kidney disease detection. Computers in Biology and Medicine, 191, 110214.

[48] Raza, A., Soomro, M. H., Shahzad, I., & Batool, S. (2024). Abstractive Text Summarization for Urdu Language. Journal of Computing & Biomedical Informatics, 7(02).

[49] Khan, S. U. R., Asif, S., & Bilal, O. (2025). Ensemble Architecture of Vision Transformer and CNNs for Breast Cancer Tumor Detection From Mammograms. International Journal of Imaging Systems and Technology, 35(3), e70090.

[50] Khan, S. U. R., & Khan, Z. (2025). Detection of Abnormal Cardiac Rhythms Using Feature Fusion Technique with Heart Sound Spectrograms. *Journal of Bionic Engineering*, 1-20.

51] Khan, M.A., Khan, S.U.R. & Lin, D. Shortening surgical time in high myopia treatment: a randomized controlled trial comparing non-OVD and OVD techniques in ICL implantation. BMC Ophthalmol 25, 303 (2025). https://doi.org/10.1186/s12886-025-04135-3.