Statistical and Machine Learning Approaches for Evaluating Customer Behavior in the Banking Industry

Lola Akinsehinwa
Department of Data Science
Carolina University
Winston-Salem, USA
akinsehinwak@carolinau.edu

Abstract— Marketing campaigns use a variety of strategies to promote product and service offerings. In the banking industry, these strategies range from direct marketing to digital. Direct marketing is a traditional strategy that includes telemarketing, mail, and email. This study focuses on telemarketing, a long-standing method that involves phone contact with existing and potential customers. Using data from the University of California, Irvine (UCI) data repository, the effectiveness of a telemarketing campaign conducted by a Portuguese banking institution to promote term deposit subscriptions was analyzed. This study applied various statistical techniques, including exploratory data analysis, hypothesis testing, Bayesian Inference, A/B testing, and Logistic Regression, to understand the factors influencing subscription behavior. The logistic regression model performed well in predicting customer subscription outcomes, achieving an accuracy of 92%, a precision of (94%), a recall of (90%), and an F1-score of (92%). These results highlight the effectiveness of predictive modeling, which can support key marketing strategies.

Keywords—Bank Marketing, Term Deposit, Statistical Analysis, Hypothesis Testing, Probability Distribution

I. Introduction

Marketing campaigns are commonly used in several industries. For instance, the banking industry relies on campaigns to promote its products and services [1]. The effectiveness of these campaigns depends on the strategies used and channels through which they are conducted. These channels range from traditional approaches, such as print and telemarketing, to digital channels, including social media and email. Telemarking is a form of direct marketing through phone channels that offers several advantages, including cost efficiency, a high rate of outreach, and the opportunity for real-time interactions [1]. Real-time engagement can help build rapport, clarify questions, and allow representatives to collect feedback [1], especially when customers choose not to subscribe to a product. Although there are several advantages to telemarketing, they also present limitations. For example, customers may provide rushed responses, especially if contacted at an inconvenient time. Furthermore, customers may feel pressured to subscribe because of concerns about how the representative perceives them [1].

The Bank Marketing dataset from the UCI data repository was used to analyze the effectiveness of a telemarketing campaign conducted by a Portuguese bank institute to promote term deposit subscriptions. This study applies statistical analysis through exploratory data analysis, hypothesis testing, A/B testing, and Bayesian inference to derive insights into the campaign. In addition, a logistic regression model is developed to predict customer subscription behavior based on various input features such as marital status, age, and average balance.

II. Methodology

A. Data Collection

The bank_full.csv dataset was sourced from the UCI data repository and consists of 45,211 complete records with 17 numerical and categorical features, including age (client age), job (job type), marital (marital status), education (education level), default (whether a customer has credit in default), balance (average yearly balance), housing (whether a customer has a housing loan), loan (whether a customer has a home loan), contact (contact communication type), day (last contact day of the week), month (last contact month of the year), duration (last contact duration, in seconds, campaign (number of contacts performed during the campaign), pdays (number of days that passed by after a client was last contacted from a previous campaign), previous (number of contacts performed before the campaign), outcome (the outcome of the previous marketing campaign), and the target variable y (whether a client subscribed to a term deposit).

B. Data Preprocessing

The steps applied to prepare the dataset for analysis and the logistic regression model include enhancing the readability of the column names. For instance, the target variable “y” was renamed to subscribed, “pdays” to days_since_last_contact, and “previous” to num_contacts_before_ campaign. Categorical features were encoded into a numerical format for optimal model performance. For example, ordinal encoding was applied to the education column, with levels defined as primary, secondary, tertiary, and unknown. This analysis assumed that “unknown” education is below "primary education". Binary encoding was applied to the following features: default, housing, loan, and subscribed, as they have two outcomes (e.g., yes, no). Dummy encoding was applied to the multi-class categorical features, including job, marital, contact, and outcome; this method helps to avoid multicollinearity as it drops one category from each feature. The target variable was imbalanced, with more “no” than “yes” responses. To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to increase the representation of the minority class, which improved the logistic regression model's ability to learn from customers who subscribed and those who did not.

III. Exploratory Data Analysis

Exploratory Data Analysis (EDA) provides insights into a dataset and helps uncover patterns, trends, and potential relationships among variables. This process includes but is not limited to, univariate, bivariate, and multivariate analyses to generate statistical summaries and visual insights.

a) Univariate analysis refers to the analysis of a single variable; in this study, descriptive statistical analysis was used to gain insight into each variable. The data showed that customer ages range from 18 to 95 years, with an average of approximately 40 years and a standard deviation of 10. The balance average was 1,362 euros, with a standard deviation of 3,044 euros, indicating a widespread and skewed distribution. The average call duration was approximately 258 seconds, with outliers reaching 4,918 seconds. The dataset contains 10 unique job categories, with “blue-collar” being the most common. The target variable, "subscribe," represents whether a customer subscribed to a term deposit, with most customers not subscribing.

b) Bivariate analysis investigates the relationship between two variables. Analyzing the relationship between age and balance revealed a weak positive relationship (r=0.098), suggesting little to no influence of age on customer balance (Fig.1). Further analysis revealed a weak negative relationship between duration and age (r = -0.0046), indicating that as age increases, the duration of calls tends to be shorter (Fig. 2). However, some customers over 60 had calls lasting more than 4000 seconds (66.6 minutes). Additionally, the relationship between the campaign and "Pdays" was investigated. Both variables show a weak negative relationship, which suggests that as the number of days since the last contact(days) increases, the number of contacts made during the current campaign tends to decrease (Fig. 3).

Finally, as shown in Figure 4, a moderately positive relationship exists between Pdays and previous (r= 0.45)which suggest that customers who have had more days since being contacted tend to have more prior interactions.

c) Multivariate Analysis: Multivariate analysis analyzes the relationships among multiple variables. As shown in Figure 1.5, the correlation analysis supports the findings from the bivariate analysis. A moderate positive correlation (r = 0.45) exists between "previous" and "pdays," while combinations of other variables show a weak or no linear relationship.

Fig. 1 Age vs. Balance Scatterplot

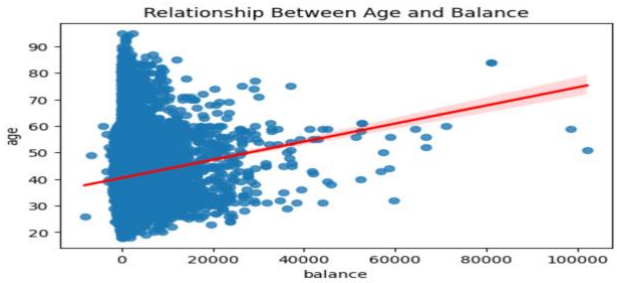


Fig.2 Age v. Duration Scatterplot

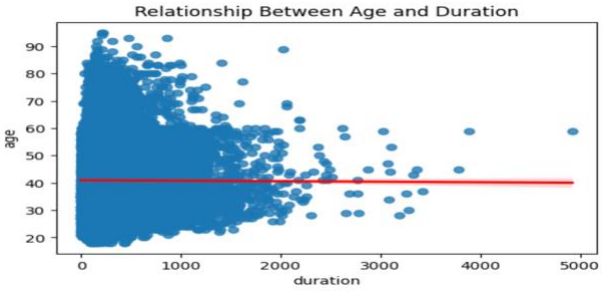


Fig. 3 Campaign v. PDays Scatterplot

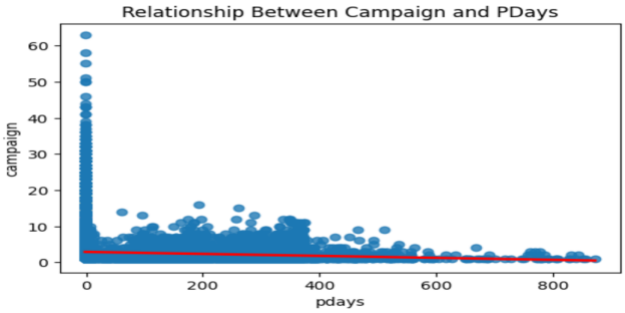


Fig. 4 PDays v. Previous Scatterplot

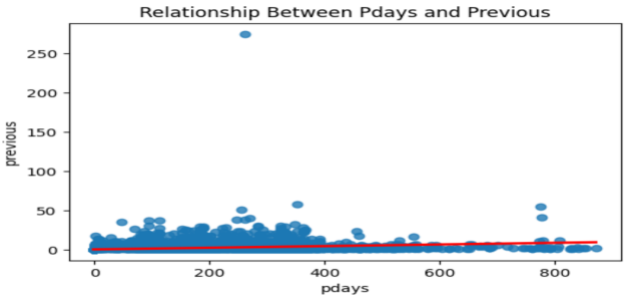
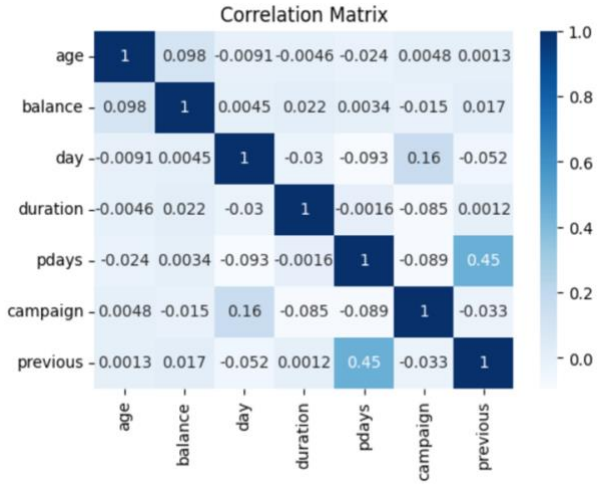


Fig.5 Correlation Matrix of Variables



IV. Probability and Distributions

Conditional, joint, and marginal probabilities were computed to explore how education level influences subscription behavior.

a) The conditional probability is the probability of observing an event given that another event has already occurred [2]. This study computes the conditional probabilities for subscription decisions based on education levels. The results showed that the probability of subscribing to a term deposit for customers with a tertiary education was 0.15. Customers with secondary education have a lower subscription rate, with a probability of 0.11. For those with primary education, the subscription probability was lower at 0.09, with no subscription at 0.91. The subscription rate for customers with unknown education is 0.14. Considering these insights, marketing strategies targeting customers with tertiary and secondary education levels can be developed to gain additional term deposit subscriptions.

b) Joint probability represents the likelihood of two events occurring simultaneously [3]. In this study, the joint probabilities for education level and subscription results are as follows: for primary education, the probability of not subscribing is 0.138, and the probability of subscribing is 0.013; for secondary education, the probabilities are 0.459 (did not subscribe) and 0.054 (subscribed); for tertiary education, the probabilities are 0.250 (did not subscribe) and 0.044 (subscribed); and for customers with an unknown education level, the probabilities are 0.036 (did not subscribe) and 0.006 (subscribed).

c) Marginal probability refers to the probability of a single event occurring without considering other events [4]. The analysis revealed that the marginal probabilities for education level included primary (0.152), secondary (0.513), tertiary (0.294), and unknown (0.041). For subscription status, the marginal probability of subscribing was 0.117, whereas the probability of not subscribing was 0.883. Overall, secondary education is the most common education level among customers of this financial institution.

B. Fitting Probability Distributions

Analysis of the age distribution revealed that the variable closely follows a normal distribution (Fig.6). According to [5], a perfectly normal distribution has a skewness value of zero. The age distribution skewness value is 0.68, which falls within the generally accepted normal range of -1 to + 1 [5]. The subscribed feature fits a Bernoulli distribution (Fig. 7), representing binary outcomes (yes or no). The Bernoulli distribution is also appropriate for this variable as the probability of success is assumed to be the same for each trial, and the outcome of one trial does not affect the outcome of any other trial [6].

Fig.6 Age Distribution

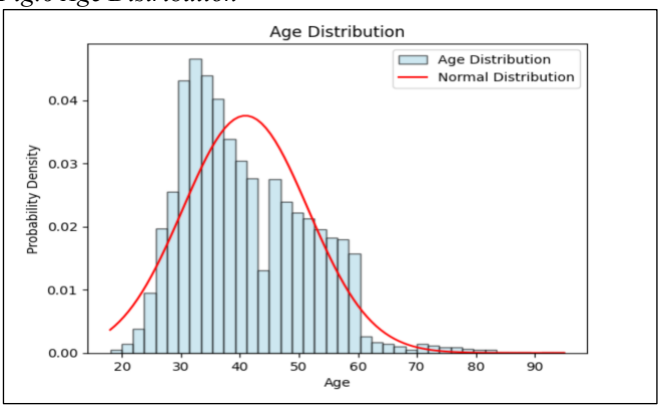
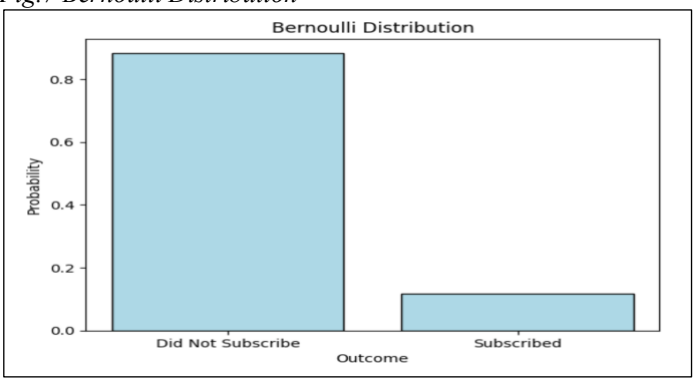


Fig.7 Bernoulli Distribution



V. Inference and Hypothesis Testing

Statistical analysis based on sample data was used to make inferences about the population. This study applies hypothesis testing techniques, including a two-sample t-test, chi-square test, and confidence interval estimation.

a) The two-sample t-test is a statistical test used to analyze whether the means of two independent groups are significantly different [2]. According to [7], the alternative hypothesis (H1) suggests that the means of the dependent variables for the two groups are different. In contrast, the null hypothesis (H0) implies that the means of the dependent variable for the two groups are the same [7]. Administrators and technicians were compared among the ten job types defined in the dataset. The null hypothesis (H0) states that there is no significant difference in the average subscription rates between admins and technicians. In contrast, the alternative hypothesis

(H1) suggests a significant difference in the average subscription rates between the two groups. The results returned a t-statistic of 1.991 with a p-value of 0.047. The null hypothesis is rejected since the p-value is less than the significance level ($\alpha = 0.05$). This finding suggests a significant difference in subscription rates between administrative staff and technicians, with administrative staff showing a higher chance of subscribing to term deposits.

b) Chi-squre Goodness-of-Fit Test: The chi-square test is a nonparametric statistical test used to determine whether the observed distribution of categorical data matches an expected distribution (i.e., a test of goodness-of-fit) [8]. This test helps determine whether there is a significant difference between the observed and expected frequencies [8]. A chi-squared test was conducted to determine whether job types follow a uniform distribution. The null hypothesis (H0) suggests that job categories follow a uniform distribution (i.e., all job types have equal representation). In contrast, the alternative hypothesis (H1) suggests that job types do not follow a uniform distribution. The analysis returned a chi-square statistic of 34,389 and p-value of 0.0. The null hypothesis is rejected based on the p-value (significance level $\alpha = 0.05$). This indicates that the distribution of job categories is not uniform and that certain job types (e.g., admin) are more common than others (e.g., students).

c) The credible interval is a range of values calculated from sample data that are likely to contain the true population parameter [2]. Given the observed data, a 95% credible interval suggests a 95% probability that the true parameter value lies within this interval [2]. This study's 95% credible interval for the campaign success rate ranges from 0.1140 to 0.1199, indicating a 95% probability that the true subscription rate falls between 11.4% and 11.9%.

VI. Bayesian Analysis

Bayesian inference is a statistical method used to calculate the probability of an event based on prior knowledge and observed data [2]. In Bayesian analysis, a prior distribution (i.e., beta distribution) is initially assumed, and as information is accumulated, the prior belief is updated [2]. This study's prior belief was that of a uniform distribution, Beta (1,1). A uninformed prior suggests that the possibility of a customer subscribing to term deposits is equally likely. After observing the data, the estimated subscription probability is 0.117, indicating that approximately 11.7% of customers are expected to subscribe to a term deposit. Based on the observed data and prior beliefs, the 95% credible interval ranges from 11.4% to 12.0%, which means there is a 95% probability that the true subscription rate falls between these values.

VII. A/B Testing

According to [2], A/B testing is a statistical method used to compare two versions of a variable, such as a marketing strategy, to determine which one performs better. This study compared two marketing strategies–email and phone calls–to evaluate their effectiveness in driving client subscriptions to term deposits. The null hypothesis (H0) states no significant difference in subscription rates between customers contacted by email and those contacted by phone. In contrast, the alternative hypothesis (H1) states that there is a difference between these two strategies. The Chi-square statistic was 0.2177 with a p-value of 0.6408. The null hypothesis cannot be rejected because the p-value is greater than the significance level of 0.05. These results indicate that there is no statistically significant difference in the effectiveness of either strategy.

VIII. Logistic Regression

Logistic Regression is a supervised machine learning model commonly used for binary classification tasks such as predicting whether a customer will subscribe to a term deposit. The logistic regression model estimates the probability of an outcome using a sigmoid function defined as $\sigma(t) = 1/(1+\exp(-t))$ [2][10]. If the predicted probability is greater than 0.5, the observation is classified as 1, indicating it belongs to the positive class. If the probability is less than or equal to 0.5, it is classified as zero, indicating that it belongs to the negative class[2][10]. This study applied logistic regression to predict customer subscription to term deposits based on several input features, including but not limited to age, duration, education, and job type..

IX. Results

The results suggest that the logistic regression model effectively classifies customers who subscribed to term deposits. The metrics used to assess the performance of the models include accuracy, precision, recall, and F1-score. Accuracy is a metric that measures the proportion of correct classifications [11]. The recall metric represents the proportion of actual positives identified as positives, and the F1-Score is the harmonic mean of precision and recall, indicating how well the model balances these two metrics[11]. Table 1 summarizes the performance of Class 1(subscribed).

Table 1. Logistic Regression Classification Report

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.92	0.94	0.90	0.92

X. Conclusion

This study analyzed the effectiveness of a telemarketing campaign conducted by a Portuguese banking institution using data from the UCI data repository. Insights about term deposit subscriptions were gained through statistical techniques, including exploratory data analysis, hypothesis testing, Bayesian inference, A/B testing, and logistic regression modeling. The results show that the subscription rate, approximately 11.7%, has room for improvement. The logistic regression model performed well across all metrics, effectively predicting term deposit subscriptions. With no difference between email and phone marketing strategies, the retail bank can target customers based on other key factors such as job type. Overall, this study provides insights into customer behavior and some factors that increase the likelihood of subscribing to term deposits. Future work could include analyzing additional features and hyperparameter tuning for enhanced model performance results.

REFERENCES

[1] R. B. Rasool Basha, "A Study on the Effectiveness of Telemarketing in the Banking Industry," *Shanlax International Journal of Management*, vol. 11, no. S1-Mar, pp. 134–143, Mar. 2024, doi: https://doi.org/10.34293/management.v11is1-mar.8101.

[2] P. C. Bruce, A. Bruce, and P. Gedeck, Practical statistics for data scientists : 50+ essential concepts using R and Python. Sebastopol, Ca: O'reilly Media, Inc, 2020.

[3] J. A. and J. Hu, *Probability and Bayesian Modeling*. Accessed: Jul. 12, 2021. [Online]. Available: https://bayesball.github.io/BOOK/probability-a-measurement-of-uncertainty.html

[4] GeeksforGeeks, "Marginal Probability," *GeeksforGeeks*, Aug. 26, 2024. https://www.geeksforgeeks.org/marginal-probability/

[5] H. Y. Kim, "Statistical Notes for Clinical Researchers: Assessing Normal Distribution (2) Using Skewness and Kurtosis," *Restorative Dentistry & Endodontics*, vol. 38, no. 1, pp. 52–54, 2013, doi: https://doi.org/10.5395/rde.2013.38.1.52.

[6] "Bernoulli Trials & Binomial Distribution: Fundamentals of Probability," *GeeksforGeeks*, Dec. 31, 2020. https://www.geeksforgeeks.org/bernoulli-trials-binomial-distribution/

[7] S. W. Lee, "Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee," *Life Cycle*, vol. 2, no. 1, Jan. 2022, doi: https://doi.org/10.54724/lc.2022.e1.

[8] R. Singhal and R. Rana, "Chi-square Test and Its Application in Hypothesis Testing," *Journal of the Practice of Cardiovascular Sciences*, vol. 1, no. 1, pp. 69–71, 2015, doi: https://doi.org/10.4103/2395-5414.157577.

[9] W. Zhang, "Confidence intervals: Concepts, fallacies, criticisms, solutions and beyond," vol. 12, no. 3, pp. 97–115, Jun. 2022, Available: https://www.researchgate.net/publication/361266357_Confidence_intervals_Concepts_fallacies_criticisms_solutions_and_beyond

[10] A. Géron, Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems, 2nd ed. O'Reilly Media, Inc., 2019.

[11] google, "Classification: Accuracy, recall, precision, and related metrics," *Google for Developers*, 2024. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall