

**IMPROVING SUCCESS IN TELEMARKETING: AN APPLICATION OF STATISTICAL METHODS IN THE BANKING INDUSTRY**

by

Dimpho Mampuru (4142232)

Rofhiwa Bulannga(4427011)

Ongeziwe Kolweni(4329993)

A research project is submitted in partial fulfilment of the requirements for the Hons in Statistical Sciences

STA761

University of the Western Cape

2024

Supervisor : Dr Rechelle Jacobs

Date : 2 September 2024

Table of Contents

[Introduction 2](#_Toc176086079)

[Description and background 2](#_Toc176086080)

[Comparison of model performance 4](#_Toc176086081)

[Conclusion 5](#_Toc176086082)

[Data preparation 8](#_Toc176086083)

[Descriptive analysis 9](#_Toc176086084)

[Descriptive statistics: 9](#_Toc176086085)

[Distribution plots(numerical variables) 9](#_Toc176086086)

[1. Visualization of numerical variables: 9](#_Toc176086087)

[2. Visualization of categorical variables 10](#_Toc176086088)

[3. Correlation analysis 11](#_Toc176086089)

[4. Outliers 12](#_Toc176086090)

[Statistical analysis 13](#_Toc176086091)

[Summary of results 13](#_Toc176086092)

[1. Logistic regression: 13](#_Toc176086093)

[2. Decision tree: 14](#_Toc176086094)

[Recommendations 14](#_Toc176086095)

[Conclusion 15](#_Toc176086096)

[References 16](#_Toc176086097)

# Introduction

Understanding customer behaviour is crucial for effective marketing strategies, with financial institutions enhancing their productivity and earnings through innovative marketing (Nagami & Kishor, 2015). In the banking industry, customer attraction often relies on telemarketing, a direct marketing method involving phone calls to persuade potential customers to purchase products ([Palaniappan et al., 2017](#Palaniappan)). Telemarketing is favoured for its convenience and resource efficiency ([Yan et al., 2020](https://www.sciencedirect.com/science/article/abs/pii/S156849462030199X?via%3Dihub)), but contacting uninterested customers can be counterproductive and lead to employee burnout ([Hosseini, 2021](https://www.tandfonline.com/doi/full/10.1080/23270012.2021.1897956#d1e109); [Neysiani et al., 2016](https://ieeexplore.ieee.org/document/7436136)). Therefore, innovative telemarketing methods are essential for customer attraction and retention.

 Machine learning, a data mining technique, can identify key factors predicting positive customer responses ([Moro et al., 2014](https://www.sciencedirect.com/science/article/abs/pii/S016792361400061X?via%3Dihub); [Vajiramedhin & Suebsing, 2014](https://doi.org/10.12988/ams.2014.47222)). Machine learning has improved earnings by 6% and productivity by 5% in various sectors ([McAfee & Brynjolfsson, 2012](https://pubmed.ncbi.nlm.nih.gov/23074865/)). This research focuses on using classification methods in telemarketing to predict customers likely to make long-term deposits. According to Dincer ([2019](#DIncer)), this is important because banks are mostly sustained by customer deposits. In South Africa's competitive banking industry, a predictive model using machine learning can optimize telemarketing, making it more cost-effective and boosting employee morale.

This study addresses the following research questions:

1. ﻿﻿﻿With the aim of improving efficiency in telemarketing, does a logistic regression model and decision trees provide sufficient accuracy in predicting long-term deposits in a South African bank?
2. Is contact time a key factor in predicting long-term deposit uptake?

# Description and background

This study assumes that telemarketing effectiveness in the banking industry can be enhanced through statistical methods, leveraging machine learning techniques known to boost productivity across various sectors. It presumes that the provided data is accurate and comprehensive, enabling effective results. The project will focus on reviewing existing literature on the topic of enhancing telemarketing within banks. Followed by developing and validating a predictive model within a set timeline, using customer data from a South African bank, including personal, credit, and contact information. Programming language Python will be utilised for analysis and model development. The study aims to inform business decisions, enhancing marketing strategies and resource allocation through accurate predictions. However, inaccurate predictions could mislead strategies and waste resources, so rigorous evaluation using precision and recall metrics will be conducted. Objectives and deliverables

The following are the objectives of this research project.

* **Review of previous studies**: To identify gaps in existing knowledge on the use of machine learning in telemarketing strategies in the banking industry through a review of a maximum of ten articles.
* **Model Development:** To construct and implement logistic regression and decision tree algorithm on historical data to identify key factors in predicting long-term deposit uptake.
  + **Data Pre-processing:** To detect and correct errors in the data that might affect the accuracy of the analysis.
  + **Exploratory Data Analysis**: To gain understanding of customer behaviour relating to long-term deposit uptake through the identification of patterns in the predictor variables. This will be done with use of visualisations such as histograms, heatmaps of correlation matrices and pair plots.
  + **Model Evaluation**: To determine the suitability of each model in the stated problem by assessing the accuracy of the two classification models constructed using appropriate metrics, such as Accuracy, Precision, Recall, F1 Score, AUC-ROC, Root Mean Square etc.

The following are the expected deliverables of this research:

* **Literature review**: A comprehensive report of what has been accomplished by others on similar studies. The report will highlight the boundaries of the review, crucial findings, definitions of concepts, and a conclusion to the review indicative of areas that need further exploration.
* **Research Design and Methodology**: Outline of the research structure and all the statistical processes employed in this study.
* **Two Developed Classification Models**: Logistic regression and decision tree algorithm will be used to predict long-term deposit uptake. This deliverable includes the codes used to construct the models, as well as the validation metrics.
  + **Results**: This subsection will report on all data processing techniques, statistical analysis, in-depth interpretations of the graphical findings and an overall summary of the results obtained.
  + **Recommendations**: Based on findings, this subsection will recommend ways to improve a South African bank’s telemarketing strategy taking into consideration the key factors indicative of a customer’s long-term deposit uptake. Limitations in this research will also be discussed highlighting what can be done in future studies.

These objectives and deliverables aim to provide a thorough action plan for the investigation and offer practical recommendations that can be applied to the South African banking industry.

Literature review

Long-term deposits are a bank’s lifeline, and predicting these deposits has become a topic of interest over the years ([Lu et al., 2016](https://doi.org/10.1007/s10844-016-0399-2)). With the growing concern of wasting resources through telemarketing, statistical methods and machine learning algorithms have emerged as potential solutions for defining the target market by narrowing it down based on key factors ([Vajiramedhin et al., 2014](https://doi.org/10.12988/ams.2014.47222)). This section critically reviews literature from the past 10 years on improving bank telemarketing strategies using statistical methods, with a key emphasis on the performance of predictive models. The key discoveries, challenges, and limitations of each study will be highlighted, focusing on how they can be improved and utilized for predicting long-term deposit uptake in a South African bank.

### Comparison of model performance

Statistical models work by running raw data through an algorithm to identify patterns and factors with significant predictive power ([Witten et al., 2002](https://doi.org/10.1145/507338.507355)). By comparing predictive power among algorithms, the most suitable model can be identified. Across seven classifiers: k-nearest neighbours (KNN), decision tree (DT), neural networks (NN), logistic regression (LR), naïve bayes (NB), and support vector machine (SVM), Ilham ([2019](https://doi.org/10.1088/1742-6596/1175/1/012035)) concluded that SVM, with the highest accuracy score of 91.07%, was best for predicting long-term deposit uptake. This study used a Portugal bank dataset with 4120 records. However, Tekouabou ([2019](https://doi.org/10.1145/3320326.3320389)) found that DT and LR outperformed SVM with accuracies of 100% and 98.61%, respectively, using a similar dataset with 4118 records. Moro ([2014](https://doi.org/10.1016/j.dss.2014.03.001)) found that NN with an (area under the curve) AUC of 0.90 was the best model among LR, SVM, and DT using a larger dataset of 52944 records. These differences highlight the importance of data quality and the nature of datasets in producing consistent results.

Previously identified key factors

According to Tekouabou ([2019](https://doi.org/10.1145/3320326.3320389)), the most significant factors in predicting long-term deposit uptake are age, job, education, and housing. Moro([2014](https://doi.org/10.1016/j.dss.2014.03.001)) suggested that gender, phone call context, and client-bank relationship also play a role in client response to long-term deposit sales calls.

Challenges of applying statistical models in telemarketing

The use of advanced statistical techniques can improve marketing efforts but comes with various challenges. Regulatory constraints emphasize responsible data use, leading to stricter access to data and complicating the implementation of statistical techniques ([Mehrabi et al., 2021](https://dl.acm.org/doi/10.1145/3457607)). Data quality issues, such as inconsistencies and missing entries, severely affect the accuracy and reliability of statistical models ([Moro et al., 2014](https://doi.org/10.1016/j.dss.2014.03.001)). Model interpretability is another challenge, as complex methods can be difficult for decision-makers to understand, impacting trust and accountability ([Guidotti et al. 2018](https://dl.acm.org/doi/10.1145/3236009)

Approaches to address these challenges

To address regulatory issues, fair modelling techniques, such as differential privacy, protect customer privacy while allowing statistical analysis ([Bonisch et al., 2019](https://www.mi.fu-berlin.de/inf/groups/ag-idm/theseses/2019_Boenisch_MSc.pdf)). Bias detection strategies are increasingly integrated into model building to prevent discrimination ([Mehrabi et al., 2021](https://dl.acm.org/doi/10.1145/3457607)). Moro ([2014](https://doi.org/10.1016/j.dss.2014.03.001)) suggest that data pre-processing techniques, such as data imputation, outlier detection, and normalization, are crucial for ensuring data. Simpler techniques, like DT can improve interpretability without sacrificing accuracy ([Guidotti et al. 2018](https://dl.acm.org/doi/10.1145/3236009)). To aid interpretability of other complex models, additional tools have been proposed, i.e., (local interpretable model-agnostic explanations) LIME a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner ([Ribeiro, M.T. et al., 2016](https://arxiv.org/abs/1602.04938)).

### Conclusion

The review highlights that predictive models help narrow down target markets for financial products such as long-term deposit uptake, make the best use of resources, and improve the process of direct marketing. However, there are several challenges associated with their application including data quality issues, regulatory restrictions as well as the need for model interpretability. Data of poor quality in terms of errors, missing records or unbalanced datasets can cause inaccuracies and reduce the reliability of predictive models. There are a few emerging methodologies regarding the responsive usage of information for differential privacy as well as identifying bias in modeling such as fair modeling checks and diagnostic tests. Data preprocessing methods such as imputation, outlier detection and normalization are essential tools for improving data quality. Additionally, more simplified approaches such as rule-based techniques should be considered to ensure that the model remains interpretable without sacrificing the accuracy of predictions. By utilizing these precautionary methods, the effectiveness of bank telemarketing can be improved, with a special focus on the unexplored South African context. With the key factors identified in existing studies, the effect of contact time in telemarketing still calls for further exploration. In the upcoming sections classification models, which have proven to be interpretable and fairly accurate, such as Logistic Regression and Decision Tree, will be utilized to determine this effect. Moreover, the models will be compared to evaluate their performance against each other and those in existing research.

Research design and methodology

This study employs a systematic approach to develop and evaluate predictive models aimed at enhancing the success of telemarketing efforts within the banking industry. All the processes were carried out using the programming language Python 3.10 in a Jupyter notebook. The research design is structured as follows:

Figure 1: Crisp methodology

**1. Business understanding**

The research focuses on identifying customer segments that are most likely to respond positively to telemarketing campaigns. The insights gained from this study will guide the development of predictive models and align them with the bank’s strategic goals.

**2. Analytic approach**

The research utilizes logistic regression and decision tree models to predict customer responses to telemarketing calls. Logistic regression is chosen for its ability to handle binary outcomes and provide insights into the relationships between predictor variables and the target outcome. Decision trees are selected for their interpretability, allowing for the identification of key factors that influence customer decisions.

**3. Data requirements**

The analysis uses a dataset named "Bank-Data," which initially contains 20 columns and 45,220 rows. This dataset includes customer demographic information, contact history, and financial data. The target variable is whether the customer agrees to a deposit-take-up or not.

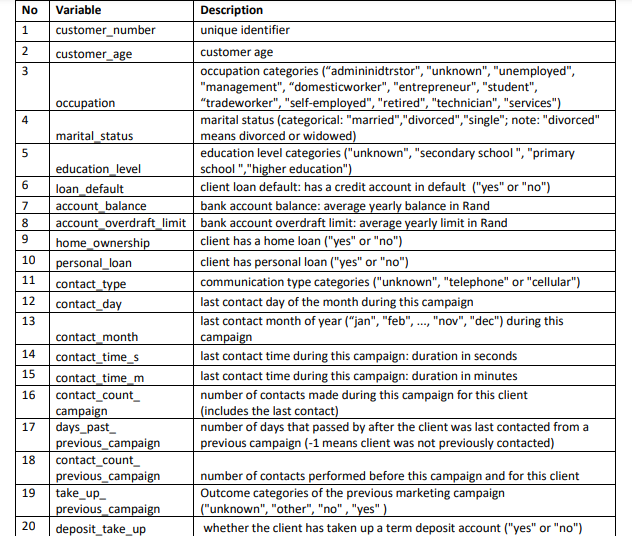


Table 1: Variable - description

**4. Data collection**

The data was pre-collected and provided by the University of Western Cape (Statistics department). It will undergo rigorous quality checks to ensure its accuracy and completeness. This includes verifying the consistency of data entries, checking for duplicates, and ensuring the correct formatting of variables.

**5. Data exploration**

Exploratory data analysis (EDA) will be conducted to understand the distributions of the data, identify anomalies, and explore relationships between variables. Multicollinearity and the risk of redundancy of the variables will be checked with a correlation heatmap. A combined distribution plot and histogram and normal curve will be utilized to assess deviation from normality, the spread of continuous values and identify patterns. To compare categorical variables and display differences in frequency/counts, a bar plot will be utilized. Scatterplots are to be generated to examine the existence of potential patterns between pairs of continuous variables. The data will be checked for other anomalies, and they will be handled accordingly. The libraries which will be utilized in this step are Pandas, Matplotlib, Seaborn and NumPy, all belonging to Python version 3.10.

**6. Data cleaning**

Data cleaning is a crucial step in this study. Missing values will be handled using imputation techniques, with numerical features imputed using the median and categorical features using the mode. Features with more than 20% missing data will be considered for removal as their inclusion can lead to unreliable results. Identification of outliers will do with use of boxplots and other metrics such as the interquartile range, to highlight values significantly deviating from the median. Numerical features are to be normalized or standardized to ensure consistent scaling. Categorical variables will be encoded using a one-hot encoder and ordinal encoding where appropriate, to make it easy for the models to read the data. These processes will be carried out using Python’s Pandas, NumPy, and Scikit-learn libraries.

**7. Model building**

Before the models are developed, the data will first be split for training and testing, with 70% allocated for training and 30% for testing. Two models will be developed, a logistic regression and decision tree model. With aim of predicting the likelihood of a customer agreeing to a long-term deposit.

*Logistic regression* is used to estimate the probability of a positive response to telemarketing, using the formula:

where:

* is the probability of a positive response
* are the predictor variables.

*The decision tree model*, which is non-parametric, classifies data by recursively splitting the dataset based on feature values to minimize impurity, which is the mixture of classes in each node. The two measures that the decision tree employ to split the tree are:

1. Gini=

* is the probability of a randomly chosen element being classified into class

1. Entropy=−

* The aim is to have the purest nodes, by selecting splits that result in the highest reduction in impurity and leading to better classification performance.

**8. Model validation**

* K-fold cross-validation: To validate the generalizability of the model, K-fold cross-validation will be employed, dividing the data into k folds of the same size, where . The models will undergo training on four of the folds except one, on which the models will be tested. This approach, utilizing the ‘cross\_val\_score’ function in the scikit-learn library, aims to decrease overfitting and overall model fit.

**9. Model evaluation**

The model performance will be evaluated using the following key metrics:

* **Accuracy**: Measures the proportion of correctly predicted instances in a dataset relative to the total number of instances. The formula is as follows

The ‘accuracy\_score’ function from the scikit-learn library was employed to calculate this metric

* **Precision**: The proportion of accurate positive predictions to all positive predictions.
* **Recall**: Measures the model's ability to capture actual positive cases. This is the proportion of correctly predicted positive cases to the total number of real positive cases.
* **F1 Score**: The harmonic mean of precision and recall, providing a balanced measure for imbalanced datasets.
* **ROC index**: Evaluates the model's ability to distinguish between the positive and negative classes. This is done by measuring the connection between sensitivity (true positive rate) and 1-specificity (false positive rate) at various thresholds. The AUC score, ranges from 0 to 1, where a score of 0.5 indicates random guessing, while a value approaching 1 indicates exceptional performance. The ‘roc\_auc\_score’ function from the scikit-learn library was used for this

**10. Deployment**

The best performing model will be deployed into the bank’s telemarketing systems, integrating with existing customer relationship management (CRM) tools. This integration shall allow the model to guide telemarketing strategies, helping the bank focus on customers more likely to respond positively. Dashboards and reports will be generated using visualization libraries like Plotly or Dash to provide real-time insights into campaign performance.

Results

This section provides insightful details on the data preparation, descriptive analysis and statistical analysis processes. Additionally, it includes a discussion of findings and recommendations for future research.

### Data preparation

This part of the research outlines the steps taken in preparing the data for analysis. Data preparation plays an important role in ensuring that the analysis and modelling procedure produces accurate and valid results. This process began with data cleaning. There were no missing detected. The presence of negative values in "days past previous campaign” has been set to zero (days can't be negative) indicating no previous contact. Outliers detected were remedied with the interquartile range (IQR) method followed by cross-validation. Moreover, the customer number has been removed from the data as it is a unique identifier and does not contribute anything to the analysis.

### Descriptive analysis

This section describes the data, identifies patterns in the dataset with the aid of summary statistics, and exploratory data analysis (visualizations).

### Descriptive statistics:

The following summarizes key statistics and provides insight into the customers’ account details, demographics and contact frequency.

* The average customer age is around 41 years, with a standard deviation of approximately 10.6 years, indicating moderate variation in age.
* The account balance varies widely, with an average of 1,362 and a large standard deviation of 3,044, reflecting significant disparities among customers. Notably, there are extreme values, with a minimum balance of - 8,019 (indicating possible debt) and a maximum of 102,127.
* The overdraft limits and the number of contacts during campaigns also show considerable variation, with most customers having a limit of zero and the majority being contacted once or twice.
* The data on the previous campaign indicates that many customers were not contacted (as shown by the minimum and median values being zero), though the maximum suggests some outliers with many contacts. The presence of negative values in "days past previous campaign" suggests a data entry issue, potentially requiring further investigation or cleaning.

# Visualizations:

#### Visualization of numerical variables:

Numerical variables are visually represented with a combined histogram and distribution curve to gain insight into their distributions and identify possible anomalies in the data.

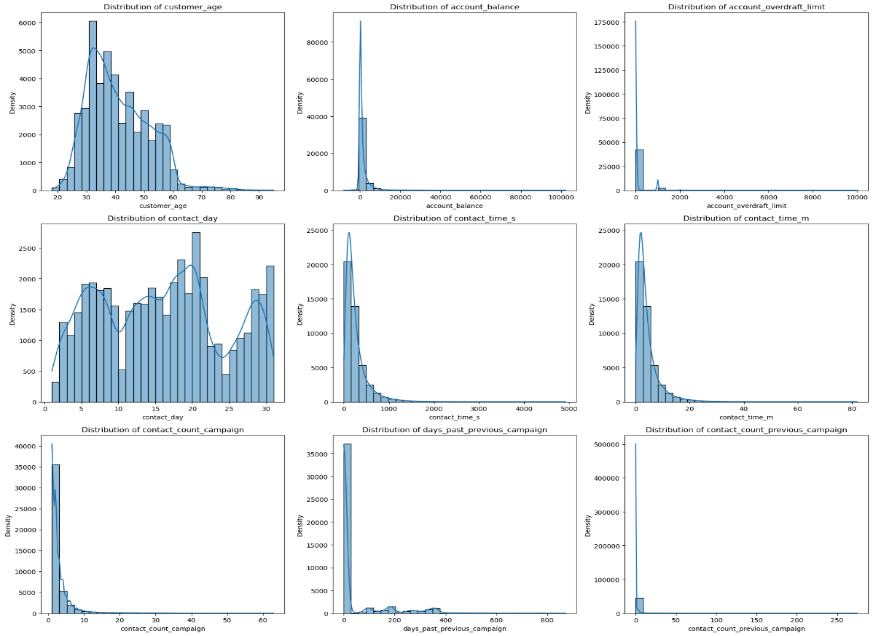


Figure 2: Histogram

**Interpretation**

* Age: The age distribution is right-skewed, with most customers aged 20 to 60, peaking around 30-35, and fewer customers over 60.
* Account Balance: The distribution is highly right-skewed, with most customers having low balances and a few with significantly higher balances, creating a long right tail.
* Account Overdraft Limit: The overdraft limit is also right skewed, with most customers having minimal limits and a few with much higher ones.
* Contact Day: Contact days are nearly uniformly distributed across the month, with slight peaks on specific days, indicating potential contact patterns.
* Contact Time (Seconds and Minutes): Both in seconds and minutes, contact time is sharply right skewed, with most contacts being brief and only a few lasting longer.
* Contact Count per Campaign: This variable shows right-skewness, with most customers contacted only a few times and very few contacted multiple times.
* Days Since Previous Campaign: The distribution is right-skewed, with most customers having a short interval since the last campaign and a few with much longer gaps.
* Previous Campaign Contact Count: Extremely right skewed, with most customers having few or no previous contacts, while a small number have been contacted many times.

#### Visualization of categorical variables

* 1. Bar chart

The following visualizations offer a detailed graphical representation of categorical features, providing information on the counts of the variables and aiding in uncovering informative patterns.

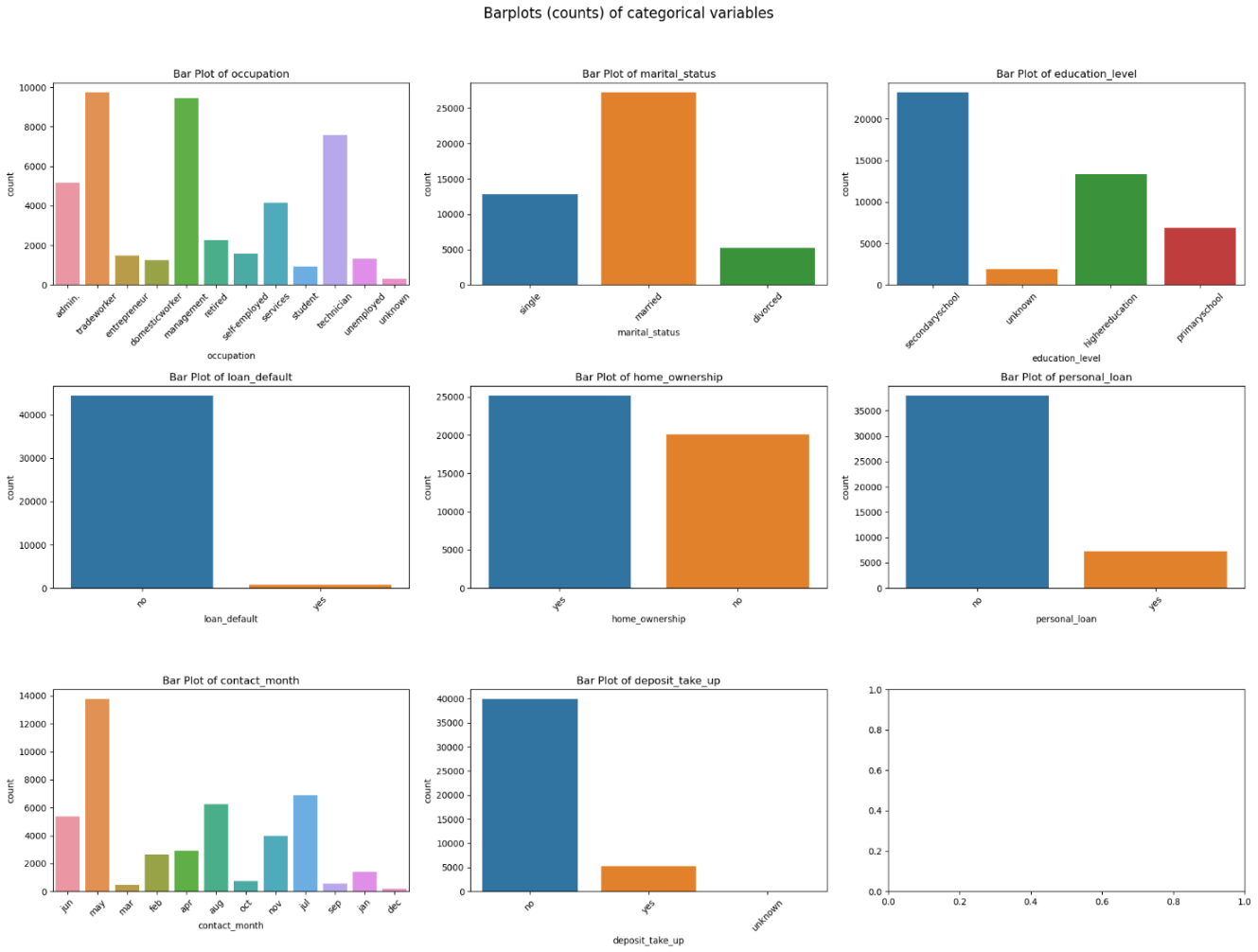


Figure 3: Bar plots

The bar charts illustrate that most of the bank's customers are married, employed as technicians or managers, and have attained a secondary education level. The vast majority do not have personal debts or a history of loan defaults; however, property ownership is somewhat more common. The bank reaches out to the highest number of customers in May, yet only a few accept the deposit offers. Overall, more individuals decline the long-term deposit proposals than agree to the sales calls.

* 1. *Stacked bar chart*

The stacked bar chart is used to uncover the relationship between customer age and deposit-take-up.

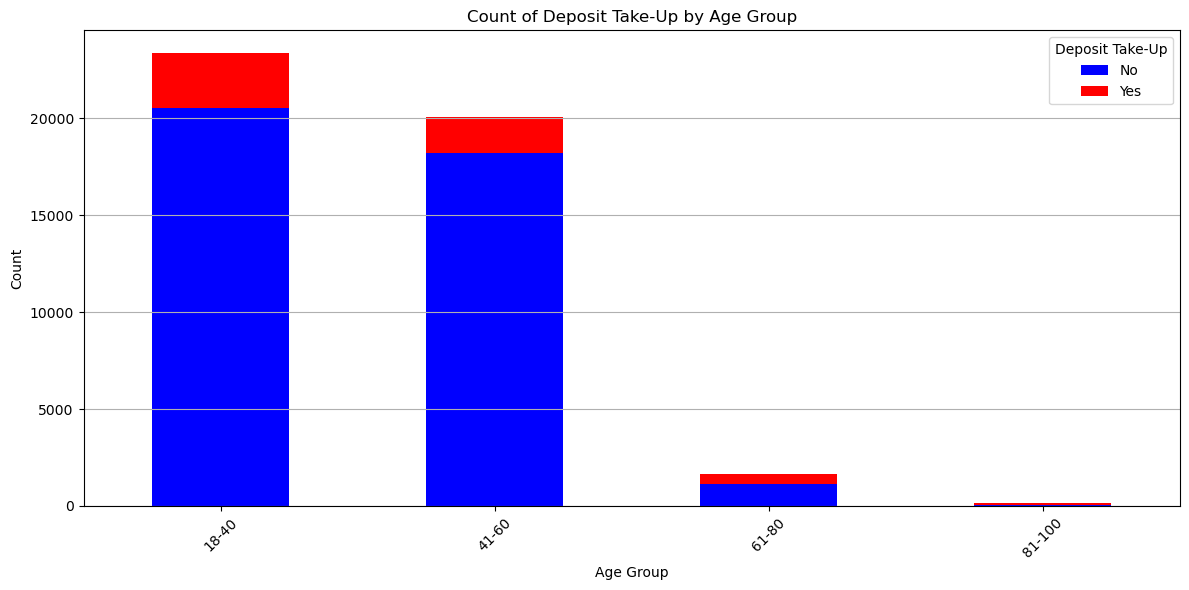


Figure 4: stacked bar chart

The customer age variable had an unrealistic age range, of 0 to 150 years, which was adjusted to a more reasonable age range of 18 to 100 years, by filtering out the other ages. The age of the customer was then divided into four distinct groups. The age ranges with the highest number of deposit take-ups were 18–40 and 41–60. However, within these ranges, most individuals declined the deposit offer, as indicated by the blue bars. Across all age groups, only a small percentage of people accepted the deposit offer, represented by the red bars. The 681–80 and 81–100 age groups exhibited both a low overall population and a low rate of deposit take-up.

#### Correlation analysis

This part of exploratory data analysis (EDA) examines the correlation between the variables, checks for multicollinearity and identifies potential redundant variables to be dealt with.

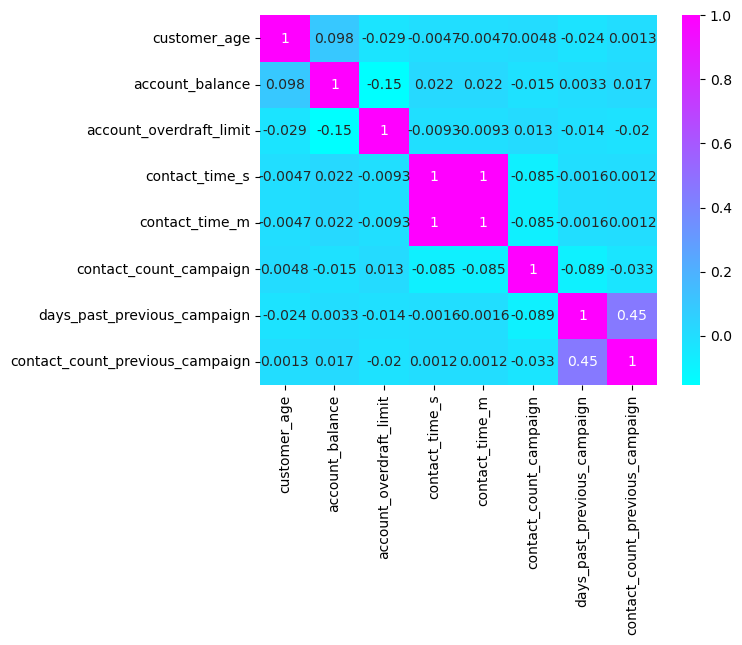


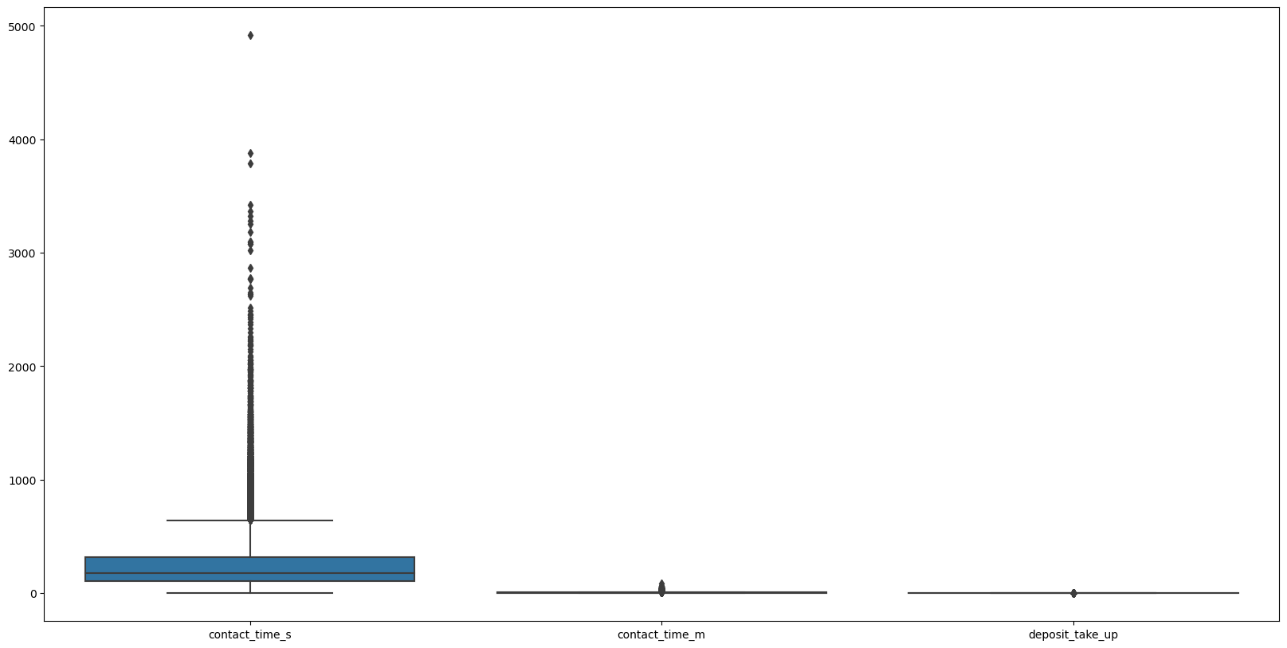
Figure 5: Correlation matrix

There is a perfect correlation of 1 between contact time (seconds and minutes). Since they provide the same information about the data, it would be useful to reduce it to one variable to avoid redundancy. The rest of the variables exhibit weak correlations, with the days past the previous campaign and contact count in the previous campaign being moderately positive with a correlation of 0.45. This indicates that customers who have been previously contacted often have had a greater number of days since the last campaign.

#### Outliers

This section identifies outliers in the dataset and assesses their impact on the analysis, followed by appropriate treatment.

Figure 6: Box plots



1. **Interpretation of findings and rationale for retaining outliers**

The initial dataset comprised 45,211 observations and 34 variables, including 8 numerical columns and 26 categorical columns that were encoded numerically. When outliers were removed, the number of observations was reduced to 26,530, which signifies a substantial data reduction of nearly 50%.

1. **Outlier treatment:**

Detection**:** Outliers were identified using standard statistical techniques, such as the interquartile range (IQR) and visualizations (e.g., box plots), to highlight values significantly deviating from the median.

1. **Rationale for not removing outliers**

Significant Data Loss: The removal of outliers resulted in a loss of nearly 50% of the data, which could lead to a significant reduction in the richness and variability of the dataset. This loss compromises the statistical power of the analysis and limits the ability to generalize the findings to the broader customer base.

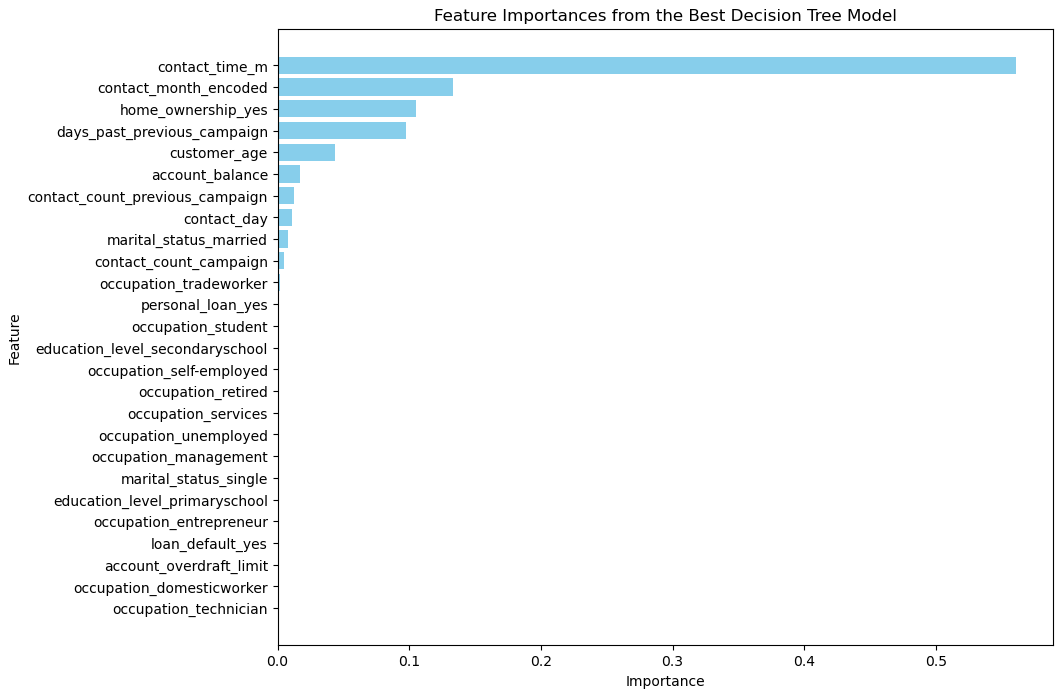
### Statistical analysis

The following showcases the performance of the logistic regression model and the decision tree model, as well as the contribution of each variable in predicting deposit-take-up.

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic Regression | | Decision Tree | |
| Overall Accuracy | 88.84% |  |  |
| CV Training Accuracy | 88.92% | CV Training Accuracy | 90.4% |
| CV Test Accuracy | 90.17% | CV Test Accuracy | 90.4% |
| ROC AUC score | 0.88 | ROC AUC score | 0.86 |
| Classification Report | | Classification Report | |
| Class 0 Performance | Class 1 Performance | Class 0 performance | Class 1 performance |
| 90% precision | 54% precision | 92% precision | 64% precision |
| 98% recall | 22% recall | 97% recall | 39% recall |

Table 2: Results summary

Figure 7: Significant Variables



### 

### Summary of results

#### Logistic regression:

The logistic regression model shows a high overall accuracy of 88.84% and a good ROC AUC score of 0.88. However, the performance result for class 1 reveals significant weaknesses. The precision of 54% and recall of 22% for class 1 indicate that the model struggles to correctly identify and predict the positive class. This poor performance is likely due to class imbalance, where class 0 significantly outnumbers class 1 in the dataset. This is evident from support value of 12022 for class 0 and 1542 for class 1.

The model's high accuracy is largely driven by its ability to correctly predict the majority class (class 0). However, the low precision and recall for class 1 highlight the model's limitations in applications where accurately identifying the minority class is critical. The cross-validation results support the model's stability, with a consistent accuracy around 89% across validation folds, training, and test sets. However, the poor recall for the minority class suggests that the model may not be suitable for situations where identifying positive instances is essential.

At significance level of 0.05, using the p-values, the logistic regression model suggests the following to be true in predicting deposit take-ups: successful previous campaigns and retirement status contribute positively, while factors such as larger overdraft limits, older age, and being a domestic worker or entrepreneur influence deposit-take-up negatively. Additionally, property ownership and higher balances on personal loans also decrease the likelihood of making deposits. The factors identified to not be significant are telephone communication, being a student and loan defaults.

#### Decision tree:

The Decision Tree model's performance before cross-validation indicated overfitting, as evidenced by the perfect training accuracy of 1.0 and lower test accuracy of 0.872. This overfitting suggests that the model was too complex, capturing noise in the training data rather than generalizing well to unseen data. After cross-validation, the model's training accuracy decreased to 90.4%, while the test accuracy slightly increased to 90.4%. This adjustment demonstrates improved generalization, with the model now better balancing its performance between the training and test datasets. The slight increase in test accuracy and decrease in training accuracy indicate that the model is now less prone to overfitting, making it more reliable for predicting unseen data.

The ROC AUC score of 0.86 further supports the model’s ability to distinguish between the two classes. However, the classification report reveals that the model performs significantly better for class 0 (the majority class) than for class 1. The lower precision of 0.64 and recall of 0.39 for class 1 suggest that the model struggles to accurately identify and predict the minority class, which could be problematic in applications where detecting the positive class is crucial.

The feature importance analysis highlights that contact time, contact month, and days past the previous campaign are key predictors in the model. These findings suggest that customer interactions and timing play significant roles in predicting deposit take-up. The moderate importance of home ownership and customer age also aligns with expectations, as these factors likely influence financial decisions.

***Model comparison:***

The logistic regression model's poor performance in predicting the minority class suggests that it may not be suitable for applications requiring high sensitivity to the positive class. The Decision Tree model initially exhibited overfitting, but cross-validation significantly improved its generalization ability. While the overall accuracy and ROC AUC score indicate good model performance, the disparity in precision and recall between classes suggests a need for further refinement. The feature importance analysis by the decision tree coincides with the logistic regression, showing that home ownership and customer interaction metrics, such as contact time and contact month are crucial for predicting deposit take-up. However, the logistic regression model, also indicates that previous campaign success and occupation offer critical influence. The decision tree model performs better in this context. This coincides with previous studies where the decision tree model outperformed the logistic regression model, Moro ([2014](https://doi.org/10.1016/j.dss.2014.03.001)).

# Recommendations

**Optimize Contact Timing**: The study findings reveal that contact time significantly influences the success of telemarketing efforts. Banks should leverage predictive analytics to identify the optimal times for contacting customers, thereby maximizing the likelihood of a positive response. Implementing a data-driven approach to contact timing will enhance the efficiency of telemarketing operations.

**Continuous Monitoring and Updating of Models**: As customer behaviour and market dynamics evolve, it is essential to regularly monitor the performance of predictive models and update them with new data. Continuous model validation and recalibration will help maintain the accuracy and relevance of predictions, ensuring sustained improvements in telemarketing success.

**Balance Model Complexity and Interpretability**: While more complex models may offer superior predictive accuracy, their interpretability is crucial for practical application in business decision-making. The use of decision tree models, which balance accuracy with interpretability, is recommended. For more complex models, interpretability tools such as SHAP (Shapley Additive Explanations) should be utilized to explain model outputs effectively.

# Conclusion

By applying logistic regression and decision tree models, this research aimed to enhance customer targeting and optimize telemarketing strategies. The analysis revealed that factors such as customer age, occupation, education level, and contact timing significantly influence the likelihood of a positive response to telemarketing calls. Both models achieved high overall accuracy in predicting customer behaviour, although class imbalance issues affected the precision and recall for predicting positive deposit uptake.

The findings highlight the importance of a targeted telemarketing approach. Focusing on customer segments more likely to respond positively can improve engagement, optimize resource allocation, and boost conversion rates. The use of predictive modelling enables more precise targeting, which enhances the efficiency and effectiveness of marketing campaigns.

Integrating these predictive models into the bank's Customer Relationship Management (CRM) systems will facilitate real-time, data-driven decision-making. This integration is expected to improve customer satisfaction, reduce unnecessary contact attempts, and increase the uptake of long-term deposits. Additionally, developing dashboards to monitor key performance indicators will support ongoing adjustments to telemarketing strategies based on real-time insights.

The study also identified significant right-skewness in variables like account balance and contact time, with most customers having lower balances and shorter contact durations. Moreover, anomalies such as unrealistic age values and errors in the "days since the previous campaign" variable point to potential data entry issues, emphasizing the need for robust data management and validation practices.

**Recommendations for Business Decision-Making:**

1. **Optimize Contact Timing Using Predictive Insights:** Leverage data-driven insights to identify optimal contact times, thereby enhancing customer engagement and response rates.
2. **Update Models Regularly:** Continuously monitor and update predictive models to maintain their relevance and accuracy in the face of changing customer behaviour and market conditions.

# References

* Dincer, H., Hacioglu, U., Tatoglu, E., & Delen, D. (2019). Developing a hybrid analytics approach to measure the efficiency of deposit banks. Journal of Business Research, 104. <https://www.sciencedirect.com/science/article/abs/pii/S0148296319303959?via%3Dihub>
* McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution - Harvard Business Review. Harvard Business Review, October. <https://pubmed.ncbi.nlm.nih.gov/23074865/>
* Neysiani, B. S., Soltani, N., & Ghezelbash, S. (2016). A framework for improving find best marketing targets using a hybrid genetic algorithm and neural networks. Conference Proceedings of 2015 2nd International Conference on Knowledge-Based Engineering and Innovation, KBEI 2015. <https://ieeexplore.ieee.org/document/7436136>
* Palaniappan, S., Mustapha, A., Foozy, C. F. M., & Atan, R. (2017). Customer profiling using classification approach for bank telemarketing. International Journal on Informatics Visualization, 1(4–2). <https://doi.org/10.30630/joiv.1.4-2.68>
* Yan, C., Li, M., & Liu, W. (2020). Prediction of bank telephone marketing results based on improved whale algorithms optimizing S\_Kohonen network. Applied Soft Computing Journal, 92. <https://www.sciencedirect.com/science/article/abs/pii/S156849462030199X?via%3Dihub>
* Hosseni, S. (2021). A decision support system based on machined learned Bayesian network for predicting successful direct sales marketing. <https://www.tandfonline.com/doi/abs/10.1080/23270012.2021.1897956>
* Witten, I. H., Frank, E., & Geller, J. (2002). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. *SIGMOD Record*, *31*(1). <https://doi.org/10.1145/507338.507355>
* Ilham, A., Khikmah, L., Indra, Ulumuddin, & Bagus Ary Indra Iswara, I. (2019). Long-term deposits prediction: A comparative framework of classification model for predict the success of bank telemarketing. Journal of Physics: Conference Series, 1175(1). <https://doi.org/10.1088/1742-6596/1175/1/012035>
* Lu, X. Y., Chu, X. Q., Chen, M. H., Chang, P. C., & Chen, S. H. (2016). Artificial immune network with feature selection for bank term deposit recommendation. Journal of Intelligent Information Systems, 47(2). <https://doi.org/10.1007/s10844-016-0399-2>
* Tekouabou, S. C. K., Cherif, W., & Silkan, H. (2019). A data modeling approach for classification problems: Application to bank telemarketing prediction. ACM International Conference Proceeding Series, Part F148154. <https://doi.org/10.1145/3320326.3320389>
* Vajiramedhin, C., & Suebsing, A. (2014). Feature selection with data balancing for prediction of bank telemarketing. Applied Mathematical Sciences, 8(113–116). <https://doi.org/10.12988/ams.2014.47222>
* Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6). A Survey on Bias and Fairness in Machine Learning.<https://dl.acm.org/doi/10.1145/3457607>
* Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62. A data-driven approach to predict the success of bank telemarketing – ScienceDirect.<https://www.sciencedirect.com/science/article/abs/pii/S016792361400061X>
* Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black-box models. ACM Computing Surveys, 51(5). A Survey of Methods for Explaining Black Box Models. <https://dl.acm.org/doi/10.1145/3236009>
* Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1602.04938.<https://arxiv.org/abs/1602.04938>
* Boenisch, F. (2019). Differential Privacy: General Survey and Analysis of Practicability in the Context of Machine Learning. 2019\_Boenisch\_MSc.pdf.<https://www.mi.fu-berlin.de/inf/groups/ag-idm/theseses/2019_Boenisch_MSc.pdf>