Methodology.

1. Overview of the document's methodology

The objective of this project is to analyze and model the given data to come up with insights to answer the research question and come up with recommendations. To achieve this goal, the following methodology was employed:

The data analysis for this study involved several steps, including data preprocessing, descriptive and frequency analysis, exploring differences and correlation in two variables, logistic regression, and ensemble methods.

Data preprocessing involved handling missing data and reducing noise. The Md.patern function was used to detect missing values, and the balance and duration variables were subjected to noise reduction. The duration variable was also converted to minutes for further analysis.

Descriptive and frequency analysis were conducted for both numerical and categorical demographic variables. Visualization of three numerical variables, age, balance, and duration, was done using a visualization chart. Statistical analysis for these variables was presented in Table 1. For categorical demographic variables, pie charts and bar charts were used to present the marital, job, education level, and dependent Y variables.

Exploring differences and correlation in two variables involved T-test and Pearson correlation testing between the dependent Y variable and default variable.

Logistic regression was used to build a model, and the likelihood ratio (deviance) test was conducted to evaluate the model. Visualization of the model was done using a coefficient plot.

Ensemble methods were used to develop four models, including a single classification tree, bagging of classification tree, random forest for classification trees, and random forest with boosting. Model comparisons were done using accuracy, ROC curve, and F-beta score.

The best measure to select the model was the F-beta score, which was better than the F1 score. Model 0 involved a single classification tree, model 1 involved bagging of classification tree, model 2 involved random forest for classification trees, and model 3 involved random forest with boosting. Evaluation of these models was done using accuracy, ROC curve, and F-beta score.

Finally, model evaluation and deployment were conducted to assess the performance of the models and deploy them in practical applications. Overall, the data analysis involved several steps that allowed for a comprehensive evaluation of the data and the development of effective models for predicting bank deposits.

2. Selecting the model measure

For the classification test, prediction accuracy on the entire data set is popular. Hence, the most common metric is accuracy, so as the document.

The following formula is utilized to calculate the accuracy:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Whereas:

- TP stands for True positive.
- TN stands for True negative.
- FP is False positive.
- FN is False negative.

However, data imbalance is a common challenge in machine learning, and it can affect the performance of models trained on such datasets. One way to evaluate the performance of these models is by using evaluation metrics such as accuracy or F-beta score. However, accuracy may not be reflective of the true performance of the model in imbalanced datasets. Instead, studies suggest that the F-beta score is a better evaluation metric for imbalanced datasets [8]. F-score considers both precision and recall, which are critical measures of performance in such scenarios. These studies argue that F-score is a more robust evaluation metric than accuracy for imbalanced datasets, especially in situations where the cost of false negatives is high. Therefore, when dealing with imbalanced datasets, it is recommended to use the F-beta score instead of accuracy to compare between models. The F-beta score is calculated via precision and recall figures, with the following formulas:

$$Precision = \frac{\mathrm{TP}}{\text{Total predicted positive}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$Recall = \frac{\mathrm{TP}}{\text{Total actual positive}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Initially, it can be challenging to determine which of these metrics, precision or recall, is more important. In addition to f1-score and accuracy, there are other more advanced metrics that can be used to evaluate classification power, such as AUC, Gini Index, and Cohen's Kappa.

$$f_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

However, in my opinion, accuracy and f1-score are still the two most fundamental metrics for classification problems. Optimally, F-beta is the more general case of when we consider the importance of recall equal to beta times precision.

$$f_{ß} = \frac{1 \ + \ ß^2}{\frac{ß^2}{Recall} + \frac{1}{Precision}} \ = \ \frac{(1 \ + \ ß^2) \ x \ precision \ x \ recall}{ß^2 \ x \ precision \ + \ recall}$$

$$= \frac{(1 \ + \ ß^2) \ x \ \mathrm{TP}}{(1 \ + \ ß^2) \ x \ \mathrm{TP} \ + \ ß^2 \ \mathrm{x} \ \mathrm{FN} \ + \ \mathrm{FP}}$$