Programming Exercise

MMA Marketing

1. **Introduction:**

Data-driven decision-making (DDDM) involves using data, facts, and metrics for strategic decision-making. In 2018, IDC found that 70% of organizations failed in their digital transformations due to a lack of a supportive data culture (Forth et al., 2020).

DDDM is vital for organizational growth and is used to create data-driven models, as seen in the Portuguese bank's telemarketing campaign. The dataset used here covers 2008-2013, and contains 4100 observations, an output target (y), and 19 input features. The dataset includes client information (e.g., age, job, marital status) and marketing campaign details.

Data consists of categorical (ordinal, nominal, binary) and numerical (continuous, discrete) variables. The goal is to build a model to predict client decisions on opening long-term deposit accounts.

1. **Methodology:**

It's crucial to thoroughly analyse and visualize data before model development. This involves assessing data quality, exploring variable relationships and relevance, and examining data distribution (Sanat, 2018).

1. **Exploratory Data Analysis (EDA):**
   1. **Initial observation:**

Both R and Python were used for creating histograms and scatterplots to display the data. Histograms show categorical variables' distribution, making it easy to interpret. Scatterplots visualize numerical values, displaying each data point and identifying outliers while showing data concentration.

All dataset-based plots are in the appendices. Notable plots follow.

A graph of blue bars

Description automatically generatedFigure 1 illustrates the last contact month with May having the most clients contacted, and December the least. The month may impact call outcomes.

Figure 1: This figure displays the formation of which month the clients were last contacted.

Figure 2 displays client contact frequency in the previous campaign. The sharp drop after one or two contacts could affect their willingness to open long-term accounts. Research suggests buyers may say no up to 4 times before saying yes, but few clients were contacted 5 times (acumen.sg, 2022).

A graph with blue squares

Description automatically generated

Figure 2: The number of times the client was contacted by the previous campaign.

In Figure 3, most clients are married. Marital status's impact on long-term deposit account willingness is uncertain. Financial stability, goals like property purchase, and family planning are more likely indicators for account opening (Husejinovic et al., 2020).

A graph of a number of individuals

Description automatically generated

Figure 3: Marital status of clients.

Figure 4 shows client job distribution, with admin and blue-collar jobs being the most common. Students are the least represented group. Offering more benefits to students could attract more of them as clients.

A graph of a number of people in a job distribution

Description automatically generated

Figure 4: Job distribution of clients.

Euribor is the rate at which European banks lend money to each other. This bank's Euribor 3-month rate, comprising the current figure, varies among clients due to factors like credit score, loan details, and their relationship with the bank. Most clients have around a 5% interest rate, with a significant group falling between <1% and 2%.

A graph of a graph showing the number of indicators

Description automatically generated with medium confidence

Figure 5: Euribor 3-month rate of clients.

* 1. **Correlations:**

To ensure relevant variables and avoid overfitting, variable selection is essential (Hastie et al., 2009; Harrell, 2015). Assessing correlations aids in this process (Kuhn, 2013). Strong correlations (±0.7-1.0) suggest redundancy; weaker correlations (±0-0.3) indicate unique information (Hair et al., 2019).

Categorical to numeric conversion and renaming were performed to create a correlation matrix. A heatmap (Figure 6) used thresholds:

Strong: ±0.7-1.0

Moderate: ±0.3-0.7

Weak: ±0-0.3 (Ratner, 2009).

A colorful grid with white and red squares

Description automatically generated with medium confidence

Figure 6: Heatmap for correlation coefficient of the numerical values in the MMA dataset.

The heatmap indicates mainly weak correlations among variables. Notable strong correlations include:

* Euribor3m and emp.var.rate: Very strong positive (0.97).
* Nr.employed and emp.var.rate: Strong (0.9).
* Cons.price.idx and emp.var.rate: Moderate (0.76).
* Poutcome and previous: Strong (0.85).
* Poutcome and pdays: Strong negative (-0.74).
  1. **Assessing the relevance of variables:**

Initially, emp.var.rate and nr.employed were excluded as they pertain to employees, not clients. Additionally, poutcome will be excluded due to numerous non-existent values (Figure 7).

A graph showing the results of a campaign

Description automatically generated

Figure 7: Demonstrates the outcome of the previous marketing campaign.

Day\_of\_week values are evenly distributed across the working week, showing little variation in the day clients were last contacted. Consequently, this variable is unlikely to impact the final client outcome and will be excluded from the model.

A graph showing the number of days of the week

Description automatically generated

Figure 8: A plot showing the day of the week a client was last contacted.

Credit default (Figure 9) is noteworthy. A vast majority of clients report no credit default. Since almost none in this dataset have credit default, it's a constant/near-constant variable, offering no useful model information (Hair et al., 2019)

A graph showing a number of credit default distribution

Description automatically generated

Figure 9: Plot demonstrates client credit default status.

In conclusion, the model will use the following variables: Age, job, marital, education, housing, loan, contact, month, duration, campaign, cons.conf.idx, cons.price.idx, outcome, Euribor3m.

1. **Model**

Choosing the right model is crucial as different models address specific problems, impacting accuracy and performance. The model choice depends on factors like interpretability, scalability, complexity, data type, and business objectives (Müller & Guido, 2017).

* 1. **Determining the model type**

Regression analysis predicts a variable (e.g., client's decision on a long-term deposit account) based on selected variables. The choice between simple and multiple regression was straightforward. With multiple variables, simple regression wasn't suitable due to data variety. Multiple regression accommodates quantitative and qualitative variables. Among multiple regression types, logistic regression is chosen for binary (yes/no) output, aligning with report’s requirements (Mitchell, 2019).

* 1. **Logistic model**

Two logistic regression models were constructed to assess the impact of excluded variables on the model's suitability. The first model includes all variables, serving as the initial reference model.

Logistic regression model 1:

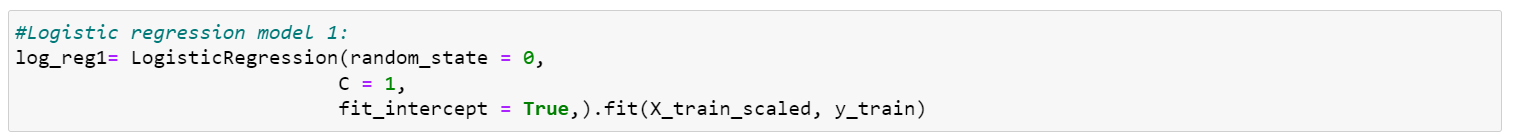


Figure 10: Screenshot of logistic regression model 1: containing all variables.

Train data accuracy score: 0.9118466898954704.

Test data accuracy score: 0.9252032520325203

91.2% of the train data’s output was accurate, whilst 92.5% (of the test data output was accurate. The output of the train and test data set are similar, suggesting good accuracy of the model.

Logistic regression model 2:

This (Figure 11) had all the variables removed and used only the ones that were highlighted in section 3.3.

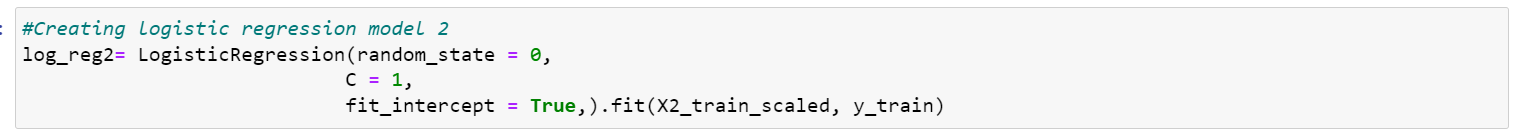


Figure 11: Logistic regression model 2, specified variables have been removed.

Train data set accuracy score: 0.9097560975609756

Test data set accuracy score: 0.9219512195121952

After excluding specific variables, minor adjustments occurred in the train and test data outputs. The training dataset's accuracy decreased marginally to 91.0%, and the test dataset's accuracy reduced slightly to 92.2%. This minimal accuracy change indicates that the removed variables exert little to no influence on the outcome.

* 1. **Model evaluation**

To evaluate the performance of the model, both a classification report and a ROC/AUC plot were done.

Logistic regression model 1:

Classification report:

A number of numbers on a white background

Description automatically generated

Table 1: Table showing classification report of logistic regression model one. All variables included and Class\_weight imbalanced.

The classification report showed a high f1 score for '0' (no) and a low f1 score for '1' (yes) due to data imbalance. Changing the threshold post-learning might introduce bias and overfitting. Balancing class weights can mitigate the imbalance issue and improve the f1-score.

Below is the classification report of model 1 after introducing the class weight as balanced into the regression model. The f1 score for 1 is significantly improved, however, the recall score has decreased. Improving precision will reduce recall scores and vice versa.

A number of numbers in a row

Description automatically generated with medium confidence

Table 2: Is a classification report of Model 1. All variables included and Class\_weight balanced.

The f1 score is a combination of the precision and recall score and reflects the model's overall ability to make correct positive predictions whilst minimizing false positives and false negatives.

Additionally, the accuracy score of the train data has dropped to 85%, and 86% for the test data.

Logistic regression Model 2:

Classification report model:

A number of numbers in a row

Description automatically generated with medium confidence.

Table 3: Classification report of model 2. Irrelevant variables were removed and class\_weight was balanced.

ROC curve:

Figure 12 displays the Logistic regression models' performance. Both models had balanced class weights, as altering class weights affected the AUC. Comparing the models reveals that excluding select variables had a big impact. Model 1 had an AUC of 0.71 whilst model 2’s AUC is 1.0.

Possible reasons for improved AUC:

* Removed variables that were noisy/irrelevant
* High correlation existed between removed and retained variables.
* The rebalance of the dataset
* Modules robustness was increased after removing variables introducing sensitivity.

A combination of these factors has resulted in an improved AUC, ROC and metrics.

A graph of a logistic regression model

Description automatically generated

Figure 12: ROC curve of both logistic regression models 1 and 2 with the weight class balanced. In orange is the first logistic regression model, containing all variables. In blue is the second logistic regression model with the removed variables.

A ROC curve positioned closer to the upper-left corner signifies a higher True Positive Rate (TPR) and a lower False Positive Rate (FPR), indicating stronger discriminatory power. While this model exhibits a relatively higher TPR and a relatively lower FPR, there's still room for enhancement in its performance.

The AUC, serving as a summary measure of the ROC curve's performance, quantifies the model's ability to distinguish between classes. A value of 0.5 suggests random guessing, while 1.0 represents perfect classification. Model 2 has a perfect classification. The first model is moderate to good discriminatory power with an AUC score of 0.71. The improved model (Model 2) makes accurate predictions with a TPR of 1 and a (False positive Rate) TPR of 0.

Since 1.0 is a perfect AUC score, no further changes are required to be made for the model.

1. **Conclusion**

This programming exercise aimed to build an accurate model for predicting clients' decisions on opening long-term deposit accounts. Data exploration raised questions about variable relevance and redundancy. After careful consideration, relevant variables were selected. To validate this choice, two regression models were created and compared. The results indicated that the excluded variables had little to no impact on the model's robustness and accuracy. This was affirmed through the classification report and ROC/AUC analysis.

Collectively, these assessments suggest two key findings: firstly, the removed variables that were irrelevant improved the model's performance greatly, creating a robust prediction model. Secondly, the discarded variables had no effect on the clients' decisions to open long-term deposit accounts. The prior analysis aided in the identification of these variables, highlighting the importance of removing variables that can potentially affect accuracy.

1. **References:**

‌ Acumen.sg. (2022). *60 Percent of Customers Say ‘No’ Four Times Before Saying ‘Yes’ -*. [online] Available at: https://acumen.sg/60-percent-of-customers-say-no-four-times-before-saying-yes/ [Accessed 1 Oct. 2023].

‌Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

Forth, P., Reichert, T., de Laubier, R. and Chakraborty, S. (2020). Flipping the Odds of Digital Transformation Success. *BCG*. [online] 29 Oct. Available at: <https://www.bcg.com/publications/2020/increasing-odds-of-success-in-digital-transformation>.

Hair, J.F., Black, W.C. and Babin, B.J. (2019). *Multivariate Data Analysis*. Andover, Hampshire, United Kingdom: Cengage Learning Emea. Copyright.

Harrell, F. (2016). *REGRESSION MODELING STRATEGIES : with applications to linear models, logistic and ordinal ... regression, and survival analysis.* Springer.

Husejinovic, Admel and Efendic Ph.D., Velid, Investigation of Factors Effecting Clients’ Decision on Engagement of Bank Deposits (December 31, 2020). Available at SSRN: https://ssrn.com/abstract=3963034 or http://dx.doi.org/10.2139/ssrn.3963034

Kuhn, M., Johnson, K. and Springer Science+Business Media (2016). *Applied predictive modeling*. New York: Springer.

Mitchell, M. (2019). *Selecting the Correct Predictive Modeling Technique*. [online] Medium. Available at: https://towardsdatascience.com/selecting-the-correct-predictive-modeling-technique-ba459c370d59.

Müller, A.C. and Guido, S. (2017). Introduction to machine learning with Python : a guide for data scientists. Beijing: O’reilly.

Ratner, B. (2009). The correlation coefficient: Its values range between +1/−1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing*, [online] 17(2), pp.139–142. doi:https://doi.org/10.1057/jt.2009.5.

Sanat (2019). *Data Visualization for Machine Learning and Data science:* [online] Medium. Available at: https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7.

1. **Appendices**

Appendix 1: Figure demonstrating an outcome of clients choosing a long-term deposit account.

A graph with blue squares

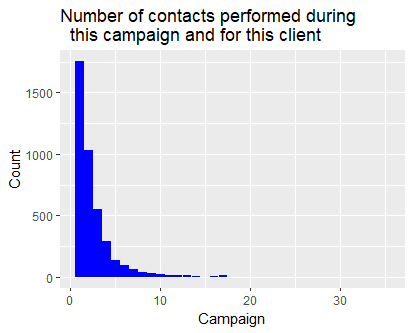
Description automatically generated

Appendix 2: Figure demonstrating the distribution of age of all clients.

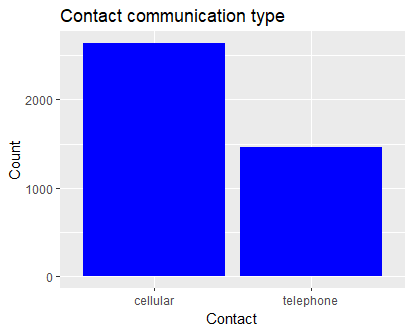
A graph of age distribution

Description automatically generated

Appendix 3: Figure demonstrating a number of times a client was contacted in this current campaign.



Appendix 4: The figure demonstrates the contact communication type with clients.

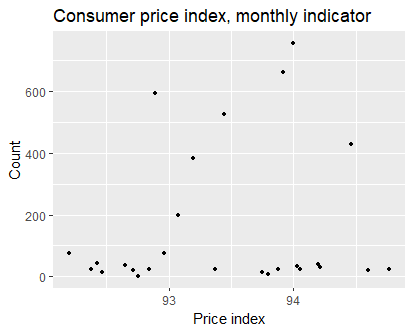


Appendix 5: Consumer confidence index of clients.

A graph with black dots

Description automatically generated

Appendix 6: Consumer price index of clients.

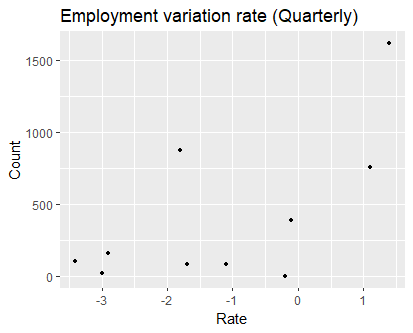


Appendix 7: Distribution of education of the clients.

A graph of a graph with blue bars

Description automatically generated with medium confidence

Appendix 8: Figure shows employment variation rate.



Appendix 9: Distribution of clients who do and don’t have housing loans. There is an almost equal distribution between those who do and don’t have a loan.

A graph of blue squares

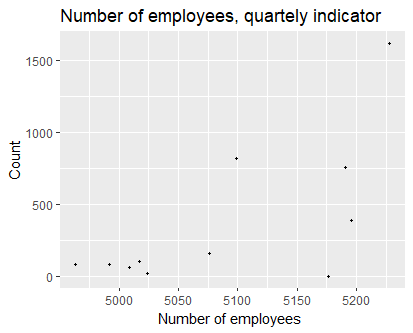
Description automatically generated

Appendix 10: This demonstrates the personal loan distribution of clients.

A graph with blue squares

Description automatically generated

Appendix 11: Figure shows distribution of the number of employees.



Appendix 12: This figure shows the number of days that have passed since a client was contacted by the previous campaign.

.