Higher Diploma in Science in Data Analytics

ASSESSMENT

Module Code: **B8IT160**

Module Description: **Applied Data Analytics**

Student Name: **Javier Ignacio Astorga Di Pauli**

Student ID: **20023978**

Table of Contents

1. Introduction

The focus of this report, therefore, is an astute analysis of a dataset pertaining to term deposit subscriptions within a banking institution. This dataset, laden with a wealth of information on client demographics, their banking behaviors, and the outcomes of marketing outreach, offers a fertile ground for extracting actionable insights through modern analytical methodologies, specifically through the lens of machine learning and data analysis.

The guiding purpose of this report is to harness the predictive potency inherent in this dataset, aiming to unveil patterns and factors that significantly influence a client's decision to subscribe to a term deposit. This objective is not only academically stimulating but bears immense practical importance for several reasons. Firstly, in an era where the financial sector is densely competitive, understanding the nuances that drive term deposit subscriptions can empower institutions to tailor their products, services, and marketing strategies more effectively, ensuring they resonate with the needs and preferences of their clientele.

The importance of this topic cannot be overstated. Term deposits, by their nature, represent a significant source of stable funding for banks, contributing to their liquidity and enabling them to extend more loans. From a client perspective, term deposits offer a risk-averse investment option, often with guaranteed returns. Therefore, dissecting the dynamics behind term deposit subscriptions can elucidate pathways for banks to optimize their offerings, enhancing customer retention, attracting new clients, and fostering a more robust financial ecosystem.

In terms of real-world application, the findings from this analysis are poised to serve as a robust foundation upon which banks can refine their marketing approaches and customer service paradigms. For instance, by identifying key demographic segments that are more inclined towards term deposit subscriptions or recognizing the times when clients are most receptive to such investment products, banks can orchestrate more personalized and timely marketing campaigns. Moreover, insights gleaned could inform product innovation, ensuring that term deposit features align closely with customer expectations and preferences.

In sum, this report endeavors to transcend mere academic exercise, aspiring to yield tangible, actionable insights that can be leveraged by banking institutions to fortify their market position, enhance customer satisfaction, and ultimately, contribute to a more vibrant, inclusive financial sector.

2. Exploratory Analysis

2.1 Dataset and Features Description

Overview of the Dataset:
The dataset originates from a study conducted by Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) in 2012, focusing on direct marketing campaigns of a Portuguese banking institution. These campaigns were executed via phone calls, aiming to ascertain whether clients would subscribe to a bank term deposit. The dataset primarily captures data spanning from May 2008 to November 2010, encompassing two main file variants: bank.csv, with all examples, and bank-test.csv, containing a 10% sample (4521 instances) of the full dataset, intended for testing more computationally demanding algorithms. For the purpose of this report, only the dataset from bankl.csv will be utilized.

The goal of the classification within this data is to predict if a client will subscribe to a term deposit, categorized as variable y.

Source:

- The dataset originates from the UCI Machine Learning Repository and can also be found on Kaggle.
- Dataset Size: 45211 instances in bank-full.csv
- Features: 16 input attributes plus one output attribute (y)

URL:

Bank Marketing Dataset on UCI Machine Learning Repository
Bank Marketing Dataset on Kaggle

Application Context:

- The dataset is instrumental in analysing the efficiency of bank marketing campaigns, particularly focusing on term deposits.
- Term deposits are defined as investments made with a financial institution that have a fixed interest rate and maturity date. These are beneficial for individuals seeking higher interest rates than standard deposit accounts. More about Term Deposits

Features Overview:

Feature No.	Name	Type	Description
1	age	Numeric	Age of the client
2	job	Categorical	Type of job (admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown)
3	marital	Categorical	Marital status (divorced, married, single; note: "divorced" includes widowed)

4	education	Categorical	Level of education (primary, secondary, tertiary, unknown)
5	default	Binary	Whether the client has credit in default (yes, no)
6	balance	Numeric	Average yearly balance, in euros
7	housing	Binary	Whether the client has housing loan (yes, no)
8	loan	Binary	Whether the client has personal loan (yes, no)
9	contact	Categorical	Type of contact communication (cellular, telephone, unknown)
10	day	Numeric	Last contact day of the month
11	month	Categorical	Last contact month of the year (jan, feb, mar, ..., nov, dec)
12	duration	Numeric	Last contact duration, in seconds
13	campaign	Numeric	Number of contacts performed during this campaign for this client (includes last contact)
14	pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted)
15	previous	Numeric	Number of contacts performed before this campaign for this client
16	poutcome	Categorical	Outcome of the previous marketing campaign (failure, other, success, unknown)

Output Variable

- y - has the client subscribed a term deposit? (Binary: "yes", "no")

Note: There are no missing attribute values within this dataset. This structured approach towards direct marketing campaigns provides a rich basis for analysis, targeting the prediction of term deposit subscriptions as a binary classification problem.

2.2 Analysis of Continuous Variables

The analysis of continuous variables in the dataset reveals several key insights into the distribution and spread of each variable. Below are the central measures and variation measures for each continuous variable in the dataset:

- **Age:**
    - Count: 11,162 (Total number of entries)

    - Mean (Average Age): 41.23 years
    - Standard Deviation (Variability in Age): 11.91
    - Minimum Age: 18 years
    - 25% of the population is younger than 32 years (1st quartile)
    - Median (Middle value): 39 years
    - 75% of the population is younger than 49 years (3rd quartile)
    - Maximum Age: 95 years
- **Balance:**
    - Mean (Average Balance): 1,528.54
    - Standard Deviation (Variability in Balance): 3,225.41
    - Minimum Balance: -6,847
    - 25th Percentile: 122
    - Median Balance: 550
    - 75th Percentile: 1,708
    - Maximum Balance: 81,204
- **Day:**
    - Mean: 15.66
    - Standard Deviation: 8.42
    - Minimum: 1
    - 25th Percentile: 8
    - Median: 15
    - 75th Percentile: 22
    - Maximum: 31
- **Duration:**
    - Mean: 371.99 (seconds)
    - Standard Deviation: 347.13
    - Minimum Duration: 2 seconds
    - 25th Percentile: 138 seconds
    - Median: 255 seconds
    - 75th Percentile: 496 seconds
    - Maximum Duration: 3,881 seconds
- **Campaign:**
    - Mean: 2.51 (contacts during this campaign)
    - Standard Deviation: 2.72
    - Minimum: 1 contact
    - 25th Percentile: 1
    - Median: 2 contacts
    - 75th Percentile: 3
    - Maximum: 63 contacts
- **Pdays:**
    - Mean: 51.33
    - Standard Deviation: 108.76
    - Minimum: -1 (indicates client was not previously contacted)
    - 25th Percentile: -1

    - Median: -1
    - 75th Percentile: 20.75
    - Maximum: 854
- **Previous:**
    - Mean: 0.83 (contacts before this campaign)
    - Standard Deviation: 2.29
    - Minimum: 0 contacts
    - 25th Percentile: 0
    - Median: 0 contacts
    - 75th Percentile: 1
    - Maximum: 58 contacts

These statistics provide a comprehensive view of each continuous variable's central tendency and variability. It's noted that some variables, like 'balance', 'duration', and 'campaign', exhibit high variability as indicated by their standard deviations relative to their means. This suggests a wide spread of values which can be further investigated with histograms, box plots, and density plots to visualize the distribution of each variable.

2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical first step in the data science process, crucial for understanding the data we are working with. By conducting EDA, we gain valuable insights that guide subsequent data processing and model building decisions. Here's a breakdown of each part of the EDA process mentioned and an explanation of its importance:

1. **Find Unwanted Columns:** Identifying columns that might not be useful for analysis helps in focusing on relevant data. Unwanted columns could be those that do not add value, are redundant, or are not applicable to the problem at hand. Removing them simplifies the dataset.
2. **Find Missing Values:** Discovering and handling missing values is fundamental to prevent data bias and inaccuracies in analysis. It is crucial to identify these gaps in data to decide whether to impute, remove, or otherwise address them.
3. **Find Features with one value:** Features that have a single value (or very little variance) across the dataset are not useful for predictive models since they do not contribute to the model's ability to distinguish between different observations. Identifying and removing these features can improve model performance.
4. **Explore the Categorical Features:** Understanding how categories are distributed and their characteristics can provide insights into the data's structure and suggest how these features might be used in analysis or preprocessed.
5. **Find Categorical Feature Distribution:** This involves examining how often each category within a feature appears in the dataset. It helps in identifying dominant categories and can influence encoding strategies and imbalance handling.

6. **Relationship between Categorical Features and Label:** Analyzing how categorical features relate to the target variable enables the identification of potential predictors and can guide feature engineering efforts.
7. **Explore the Numerical Features:** Investigating the numerical features helps in understanding their distribution, scale, and variance, providing insights into how they might be normalized or standardized.
8. **Find Discrete Numerical Features:** These are numerical features that represent categories or countable measurements. Understanding them can lead to better handling, such as possibly treating them as categorical data.
9. **Relation between Discrete Numerical Features and Labels:** Exploring how these features relate to the target can highlight their predictive value or inform transformations to improve their utility.
10. **Find Continuous Numerical Features:** Continuous features can have any value within a range. Understanding their distribution is key to identifying patterns or trends and deciding on scaling or transformation strategies.
11. **Distribution of Continuous Numerical Features:** Analyzing the distribution helps in recognizing normal, skewed, or bimodal trends, which is vital for selecting appropriate preprocessing steps.
12. **Relation between Continuous Numerical Features and Labels:** This analysis identifies potential relationships and patterns between features and the target, which is key to selecting features and modelling strategies.
13. **Find Outliers in Numerical Features:** Outliers can significantly affect model performance. Identifying and addressing outliers ensures that models are not unduly influenced by these atypical data points.
14. **Explore the Correlation between Numerical Features:** Understanding how features relate to each other helps in identifying redundancy and potential multicollinearity issues, guiding feature selection and dimensionality reduction efforts.
15. **Find Pair Plot:** This is a visual exploration tool to understand the bivariate relationships between each pair of features, helping in identifying patterns, trends, and possible interactions between features.
16. **Check the Dataset is Balanced or Not Based on Target Values in Classification:** Imbalanced data can result in biased models favoring the majority class. Identifying balance helps in deciding if resampling or other techniques are needed to address imbalance.

2.4 Relationships Between Features

In the Exploratory Data Analysis section of the code, a comprehensive set of visualisation techniques, including scatter plots, pair plots, heatmaps, distplots, and boxplots, is utilized to explore the relationships between features. These plots serve as indispensable tools for gaining insights into the dataset's underlying structure, correlation among variables, distribution characteristics, and potential outliers.

1. **Scatter Plots and Pair Plots:** These plots illustrate the relationship between two continuous variables. They are instrumental in identifying whether there exists a linear, nonlinear, or no relationship between pairs of variables, by plotting them on the X and Y axes.
2. **Heatmaps:** Serving as a visual representation of correlation coefficients, heatmaps clearly display the strength and direction of the relationship between variables. Correlation coefficients range from -1 to 1, where values near 1 indicate a strong positive correlation, values near -1 denote a strong negative correlation, and values around 0 suggest no linear correlation.
3. **Distribution Plots:** Distplots, or distribution plots, are used to observe the distribution of a single continuous variable. They combine the histogram and kernel density estimate (KDE) plots to provide a clear view of the data distribution, revealing any skewness, bimodality, or normality. This is vital for understanding how data spread and potential transformations that may be needed for modeling.
4. **Boxplots:** Boxplots are crucial for identifying outliers within the dataset as well as understanding the distribution characteristics of a variable, such as its quartiles, median, and range. They visually outline the spread and central tendency of the data, making it easier to spot outliers represented by points outside the typical range (whiskers).

Incorporating histplots and boxplots, alongside scatter plots, pair plots, and heatmaps, in the EDA process offers a rounded comprehension of the relations between features. Distplots allow for a granular view of individual variables' distributions, revealing essential distribution characteristics. Simultaneously, boxplots provide an efficient means of detecting outliers and understanding the distribution properties of the data. Altogether, the utilisation of these diverse visualisation techniques in the code engenders a holistic understanding of the dataset. This knowledge is invaluable for making informed preprocessing and modeling decisions, ultimately leading to the development of more precise and robust predictive models.

3. Data Preprocessing

The data preprocessing phase is a foundational step in preparing the dataset for machine learning model development. It involves several critical tasks aimed at optimizing the data for better model performance. Below is an exploration of the specific preprocessing actions taken, as detailed in the provided code snippet:

1. **Handling Imbalanced Dataset:** Balancing the dataset is essential to avoid model bias toward the majority class, which can adversely affect its performance, especially on the minority class. Although the specific technique employed for balancing the dataset is not explicitly mentioned in the code, addressing imbalance is crucial for ensuring that predictive models perform well across all classes of the data.
2. **Categorical Variables:** The conversion of categorical variables into dummy variables was undertaken for variables such as 'job', 'marital', 'education', 'contact', 'month', and 'poutcome'. This step is necessary because machine learning algorithms require numerical input, and categorical data must be transformed into a format that these

algorithms can process. Boolean variables like 'housing', 'loan', and 'deposit' were converted into numeric values (0 and 1), enabling their use in the modeling process. This encoding not only facilitates the use of categorical data in predictive models but also helps in uncovering patterns within these variables.

3. **Missing Data:** The approach to managing missing data involved the removal of features deemed not significant or having a substantial amount of missing values. For example, the feature 'pdays' was dropped due to its high proportion of '-1' values, representing missing or irrelevant data. This strategy helps in streamlining the dataset and focusing on more meaningful and impactful features for the prediction task.
4. **Outliers:** Instead of applying Chebyshev's rule or box plot techniques traditionally used for outlier detection, the decision on handling outliers was made based on an analysis that considered the business and statistical significance of data points. For features like 'age', 'balance', and 'duration', outliers were assessed with the understanding that certain outlier values, especially in 'balance' and 'duration', indicate a higher interest in deposits. This reflects a more nuanced approach to dealing with outliers, emphasizing the importance of domain knowledge in preprocessing decisions.
5. **Normalization:** Although the specifics of normalization are not detailed in the code snippet, normalization is recognized as an essential step for ensuring that numeric features have a consistent scale. This is especially important for models sensitive to input scales or when features span varying units of measurement. Through normalization, all features can be brought to a uniform range, contributing equally to the model's accuracy and effectiveness.

In conclusion, the preprocessing steps, including the conversion of categorical variables, management of missing data, and thoughtful consideration of outliers, were critical in preparing the dataset for model building. These steps ensure the development of a robust machine learning model capable of delivering accurate and reliable predictions.

Dataset Splitting

Splitting the dataset into training and testing sets is a foundational practice in machine learning that ensures the developed models are robust, generalizable, and perform well on unseen data. In the provided code, the dataset is divided into training (80%) and testing (20%) sets, a common split ratio that balances between having enough data for training the model while retaining a significant portion for unbiased evaluation.

Why Splitting is Necessary:

1. **Model Training:** The training set is used to train the machine learning model, allowing it to learn the relationship between the input features (X) and the target variable (y). An adequately large training set enables the model to explore a wide variety of patterns, trends, and associations within the data, which is crucial for building a model that captures the underlying structure of the dataset effectively.
2. **Model Evaluation:** The testing set acts as new, unseen data for the model. It is essential for evaluating the model's performance metrics, such as accuracy, precision,

recall, and more, depending on the problem type (classification, regression, etc.). This evaluation provides an unbiased assessment of the model's ability to generalize its learned patterns to new data, which is vital for understanding how the model will perform in real-world scenarios.

3. **Preventing Overfitting:** Overfitting occurs when a model learns the training data too well, including its noise and outliers, to the detriment of its performance on new data. By keeping a portion of the dataset unseen during the training phase and using it solely for testing, it's possible to detect overfitting. If a model shows high performance on the training set but low performance on the testing set, this is a clear indication of overfitting.
4. **Tuning Model Parameters:** The separation of data into training and testing sets also facilitates the process of hyperparameter tuning. Models can be adjusted and fine-tuned using the training set, with the testing set serving as the final evaluator of the tuning process's effectiveness.

In the code segment, the train_test_split function from sklearn.model_selection is utilized to partition the dataset, ensuring that 8921 instances are used for training and 2231 for testing. This split, alongside the specified random_state for reproducibility, lays the groundwork for a reliable machine learning workflow, where models are trained effectively and evaluated accurately.

4. Predictive Analysis

The predictive analysis phase of this project focuses on utilising machine learning models to predict whether a client will subscribe to a term deposit. This section details the steps taken, from selecting input and output variables to model fitting, interpretation, and performance analysis.

4.1 Input and Output Variables

For the predictive model, the input variables (features) comprise the entire dataset excluding the target variable 'deposit_new', which represents whether a client subscribes to a term deposit. The output variable (target) is 'deposit_new', indicating the binary outcome of subscription (1) or non-subscription (0).

4.2 Regression Model Fitting

Two models were initially considered for the task: RandomForestClassifier and XGBoost Classifier (XGBClassifier). Cross-validation scores were computed for both models, leading to the selection of the XGBClassifier based on its superior performance.

The XGBoost model was configured with the following parameters: objective='binary:logistic', learning_rate=0.1, max_depth=10, and n_estimators=100. These parameters were determined

via GridSearchCV, which optimised the model by searching through a predefined grid of parameters to find the combination that yielded the best cross-validation score. The final model was then fitted to the training dataset.

4.3 Model Parameters Interpretation

The significance of the model's parameters lies in their contribution to handling the predictive task:

- **Objective:** Specifies that the model's goal is binary classification.
- **Learning rate (0.1):** Controls the model's adaptation rate to new patterns, with a lower value helping prevent overfitting.
- **Max depth (10):** Indicates the maximum tree depth, allowing the model to learn complex patterns without becoming too specific to the training data.
- **N_estimators (100):** The number of trees in the forest, providing a balance between learning capability and computational efficiency.

Ultimately, these parameters underscore the model's capacity to efficiently learn from data whilst minimising overfitting, ensuring robust predictive performance.

4.4 Predictions and Model Performance

The model’s performance was evaluated on the test data, resulting in an accuracy score of approximately 85.84%. To further assess the model, a confusion matrix was generated, revealing the distribution of true positives, false positives, true negatives, and false negatives. The confusion matrix visualisation and the computed accuracy highlight the model's effectiveness in predicting term deposit subscriptions, with a higher number of correct predictions as compared to incorrect ones.

Further, feature importance analysis was conducted, revealing the significance of various features in predictions. This analysis, depicted through a bar chart, provides insights into which factors most significantly influence a client's decision to subscribe to a term deposit, underlining the value of specific input variables in the predictive process.

In conclusion, the predictive analysis stage encapsulated the model selection, fitting, and evaluation processes, substantiated by the robust performance of the XGBoost Classifier. The model's ability to accurately predict term deposit subscriptions showcases the potential for machine learning to inform strategic decision-making in banking services.

5. Conclusion

The thorough analysis undertaken in this report, aimed at dissecting the intricacies of term deposit subscriptions within a banking institution, has culminated in several enlightening findings. Through a careful examination of the dataset provided, encompassing client demographics, financial behaviours, and responses to marketing campaigns, we have identified key factors that significantly influence the likelihood of a client subscribing to a term deposit. These factors include demographic attributes such as age, job type, marital status, and educational background, as well as transactional and contact-related variables like the average yearly balance, contact communication type, and the outcome of previous marketing campaigns.

The implications of these findings are manifold and offer valuable insights for both academic and practical applications. From a practical standpoint, banking institutions can leverage this information to refine their marketing strategies, tailoring their outreach efforts to target demographics more likely to engage with term deposit offerings. Moreover, understanding the variables that play a pivotal role in decision-making processes enables banks to design more appealing term deposit products, potentially boosting subscription rates and fostering long-term customer loyalty.

Despite the robust analysis and the considerable insights generated, this study is not without its limitations. One such limitation is the reliance on historical data, which, while comprehensive, may not fully capture the dynamism of customer preferences and market conditions over time. Furthermore, the dataset's geographic focus on a Portuguese banking institution may limit the generalizability of the findings across different cultural and economic contexts.

Recognizing these limitations, future research directions could include a broader analysis incorporating datasets from banking institutions across various countries and contexts, to assess the universality of the identified factors. Additionally, longitudinal studies could offer insights into the evolving nature of customer behaviour and preferences, enabling a more dynamic understanding of factors influencing term deposit subscription. Lastly, the integration of machine learning and artificial intelligence techniques could further enhance the predictive accuracy and granularity of insights, paving the way for more nuanced and effective banking strategies.

In conclusion, this report not only sheds light on the essential factors influencing term deposit subscriptions but also underscores the significance of data-driven insights in shaping banking strategies. Moving forward, embracing a more holistic and dynamic approach to research could substantially contribute to the evolution of banking services and customer satisfaction.

6. References

The analysis presented in this report draws upon a range of sources, datasets, and tools which have been instrumental in shaping the research and findings. Notably, the dataset used, originating from the UCI Machine Learning Repository and also found on Kaggle, captures client demographics, financial behaviours, and responses to marketing campaigns for a Portuguese banking institution from May 2008 to November 2010[1]. This rich dataset, provided by Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL), serves as the foundation for our predictive analysis of term deposit subscriptions.

Key references include:

- Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62, 22-31.
- UCI Machine Learning Repository: Bank Marketing Dataset. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/bank+marketing
- Kaggle: Bank Marketing Dataset. [Online]. Available: https://www.kaggle.com/rouseguy/bankbalanced

Additionally, the Python programming language, along with libraries such as Pandas, NumPy, Matplotlib, and Scikit-learn, played a crucial role in data processing, exploratory analysis, and model development stages of this report.