Bank Marketing Dataset Classification

Miguel Cruzeiro, 50%
Machine Learning Fundamentals
Student 107660, DETI
Universidade de Aveiro
Aveiro, Portugal
miguelcruzeiro@ua.pt

Miguel Figueiredo, 50%
Machine Learning Fundamentals
Student 108287, DETI
Universidade de Aveiro
Aveiro, Portugal
miguel.belchior@ua.pt

Pétia Georgieva
Machine Learning Fundamentals
Course Instructor, DETI
Universidade de Aveiro
Aveiro, Portugal
petia@ua.pt

***Abstract*—This report presents a comprehensive analysis of the Bank Marketing Dataset. The main goal of this report is to evaluate and compare various machine learning models to determine their efficacy in predicting whether a client will subscribe to a term deposit after a marketing campaign. The study performs a detailed analysis on the dataset, exploring feature distributions and correlations to uncover key insights. The study emphasizes the importance of data preprocessing, parameter tuning and train-test splitting in optimizing the model's performance. Two distinct classification approaches and their results are evaluated: one utilizing all available features and the other excluding certain features to assess their impact on model performance.**

***Index Terms*—Bank Marketing Dataset, Machine Learning, Classification, Data Preprocessing, Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbors, Performance Metrics, Parameter Tuning, Accuracy, Feature Selection, Confusion Matrix**

I. Introduction

This report presents the first project undertaken as part of the Machine Learning Fundamentals course. As stated in the assignment description [1], "the goal of this project is to apply suitable machine learning algorithms learned in class or self-learned to solve a specific data science problem (classification, regression, clustering)". From the available topics proposed to complete the assignment, the bank marketing problem [2] was selected due to its relevance in understanding and predicting consumer decision-making and purchasing behavior.

The bank marketing problem involves analyzing data from a Portuguese banking institution's direct marketing campaigns, based on phone-calls, where the objective is to predict whether a client will subscribe to a term deposit. Therefore, this problem constitutes a binary classification problem. By applying machine learning techniques to the bank marketing dataset, insights can be gained into what factors most influence customer decisions, which is valuable to improve marketing strategies.

On this assignment, 4 machine learning algorithms were trained and tested to solve the classification problem: K-Nearest Neighbors, Logistic Regression, Random Forest Classifier and Support Vector Machine. Additionally, in order to evaluate the impact of feature selection on model performance, the models were trained and tested using both the complete feature set and a reduced subset of features. In both cases, all models were evaluated using appropriate metrics to assess their performance on the classification problem: accuracy, precision, recall and F1-score, as well as the corresponding confusion matrix.

II. State of the Art Analysis

The Bank Marketing Dataset has been widely utilized to develop models to predict client subscription behavior in banking campaigns. Researchers have adopted a wide range of techniques to analyse the data ranging from traditional models such as Logistic Regression or Decision Tree to more advanced approaches like ensemble methods.

The related works listed below focus primarily on either feature selection techniques and/or finding the best prediction model for the problem. With the analysis of the existing work, we aim to identify successful methodologies, effective strategies and contributions to understanding customer behavior in the context of financial marketing campaigns. The most proven strategies will be incorporated into the models developed for this assignment, aiming to enhance their effectiveness.

A. Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data

This study by Tuba Parlar and Songul Kakilli Acaravci [3] explores the application of data mining techniques to the Bank Marketing Dataset with the goal of improving the campaign's effectiveness through feature selection.

Feature Selection Methods

- **Information Gain (IG)**: Measures the contribution of each feature to prediction performance.
- **Chi-square**: Measures the lack of independence between a feature and a class.
- Four feature sizes (5, 8, 10, 15) were tested for the dataset with five-fold cross validation.

Machine Learning Classifier

- **Naive Bayes (NB)**: Used to evaluate the effectiveness of feature selection by comparing the results before and after applying the feature selection methods.

Experimental Results

The five most relevant features were `duration`, `poutcome`, `month`, `pdays`, and `contact`. The results show that reducing the feature set not only enhances the efficiency of the training process but also increases the classifier's performance.

With the reduced feature set, NB achieved an F-measure of 0.883, an improvement over the baseline (0.873).

Conclusion

This study shows that data mining and feature selection has a key role on achieving better predictive accuracy using classification algorithms.

B. A data-driven approach to predict the success of bank telemarketing

This study by Sérgio Moro, Paulo Cortez and Paulo Rita [4] investigates the use of data mining techniques to predict the success of telemarketing calls in selling long-term deposits. It introduces robust feature selection and modeling techniques to maximize prediction accuracy and practical relevance.

Dataset and Problem Context

- The dataset includes 52,944 telemarketing records and 150 initial features.
- The success rate for client subscription is low (12.38%), making the dataset unbalanced.

Feature Selection

A semi-automated feature selection process was utilized in this study. The initial dataset consisted of 150 features that were reduced to 22 relevant attributes through this process.

Machine Learning Classifiers

- **Logistic Regression (LR)**
- **Decision Trees (DT)**
- **Neural Networks (NN)**
- **Support Vector Machine (SVM)**

Experimental Results

- Among the tested models, Neural Networks outperformed the others, achieving an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.80 and an Area Under the Lift Cumulative CurveA (LIFT) of 0.67 in the rolling window evaluation.

Conclusion

The study highlights that feature selection significantly improves predictive accuracy and demonstrates how predictive modeling can enhance telemarketing campaigns, reduce costs and improve client targeting. They show that contacting only 50% of clients ranked by the predictive model, the bank could achieve 79% of successful outcomes.

III. Dataset Information

The dataset used for this assignment contains data related to direct marketing campaigns, based on phone calls, of a Portuguese banking institution. Often, multiple contacts with the same client were necessary to determine whether they would subscribe to the product (a bank term deposit), with the outcome recorded as either 'yes' or 'no'. The dataset includes information from 11162 individuals. Before delving into data distribution and detailed analysis, a comprehensive list of the datasets features, extracted from [5], can be found in Table I.

TABLE I
Dataset Descriptions

Attribute	Type	Description
age	Integer	Age of the client.
job	Categorical	Type of client's job: 'admin.','blue-collar','entrepreneur', etc.
marital	Categorical	Marital Status: 'divorced','married','single', etc.
education	Categorical	Education Level: primary, secondary, tertiary, etc.
default	Binary	Whether the client has credit in default: yes, no.
balance	Integer	Average yearly balance
housing	Binary	Whether the client has a housing loan: yes, no.
loan	Binary	Whether the client has a personal loan: yes, no.
contact	Categorical	Contact Communication Type: 'cellular','telephone'.
day	Date	The day of the week the last contact was made.
month	Date	The month of the last contact: 'jan', 'feb', 'mar', ..., 'nov', 'dec'.
duration	Integer	Duration of the last contact in seconds.
campaign	Integer	Number of contacts performed during this campaign for the client.
pdays	Integer	Number of days since the client was last contacted in a previous campaign; -1 means the client was not previously contacted.
previous	Integer	Number of contacts performed before this campaign for the client.
poutcome	Categorical	Outcome of the previous marketing campaign: 'failure','nonexistent','success'.
y - deposit	Binary	Whether the client subscribed to a term deposit: yes, no.

A. Deposit subscriptions

The target variable in the dataset is deposit, which indicates whether a client subscribed to the bank term deposit. Since this is a binary classification problem, the deposit variable has two possible outcomes: 'yes' or 'no'. The Figure 1 shows that the counts for both outcomes are relatively balanced. As a result, accuracy is a suitable metric for evaluating the performance of the model.

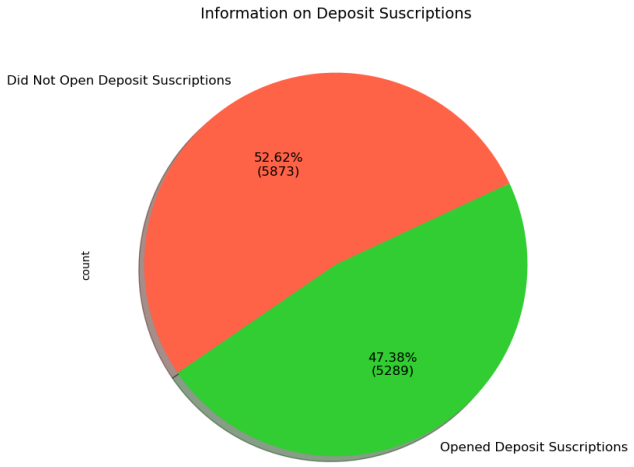

Fig. 1. Deposit subscriptions

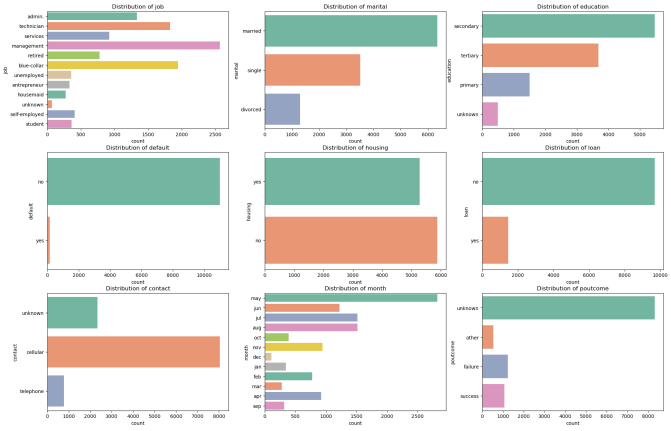

Fig. 2. Distribution of Categorical Features

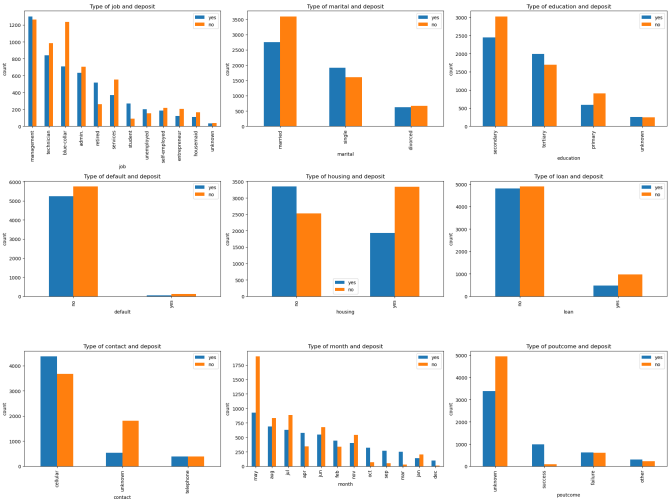

Fig. 3. Distribution of Categorical Features related to the term deposit

B. Feature Distribution

Understanding the feature distribution in the dataset is crucial for identifying patterns and gain additional insights on the data. It can also help on identifying outliers and potential pre-processing needs. This section will dive deeper on the distributions of both the categorical and numerical features of our dataset, related to the target variable.

1) Categorical Features: A categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values. The distributions of the categorical variables on our dataset are presented in Figure 2, leading to the following observations:

- **Job** - The most common occupations are management, blue-collar, and technician, while the least common are unknown, housemaid, and entrepreneur.
- **Marital** - The majority of individuals in the dataset are married, followed by single and divorced individuals.
- **Education** - The most common education level is secondary, followed by tertiary and primary, with a small portion categorized as unknown.
- **Default** - There is a significant imbalance between individuals who do not have credit in default and those who do, with the majority of individuals falling into the ”no default” category.
- **Housing** - The distribution between those with and without housing loans is fairly even.
- **Loan** - However, most people do not have a personal loan, with only a small portion marked as “yes”.
- **Contact** - The most frequent contact method is cellular, followed by unknown and telephone.
- **Month** - The majority of last contacts occurred in May, followed by July and August. December had the fewest contacts.
- **Poutcome** - The outcome of most previous marketing campaigns is unknown. Among the known outcomes, failure is the most common.

With this in mind, a more valuable analysis involves exploring the relationship between the possible values of each categorical feature and the target variable, deposit. Figure 3 illustrates this relationship and may provide insights into how these feature’s values influence the likelihood of a term deposit subscription. Overall, the features do not exhibit a clear distinction between the target values. However, there are several observations worth noting:

- **Job** - Customers with blue-collar and services jobs are less likely to subscribe for a term deposit.
- **Marital** - Married customers are more probable not to subscribe to a term deposit, whereas single ones are more probable to do so. Divorced individuals show an even balance between subscribing and not subscribing to a term deposit.
- **Education** - Clients with secondary and primary levels of education are more probable not to subscribe for a term deposit, whereas those with tertiary level of education are more likely to do so.
- **Default** - People are generally unlikely to subscribe to

a term deposit, regardless of whether they have credit in default.

- **Housing** - Clients without housing loans are more likely to subscribe to a term deposit, whereas those with housing loans are less likely to do so.
- **Loan** - Individuals without personal loans display an evenly distributed likelihood of subscribing or not to a term deposit. In contrast, those with personal loans are more inclined not to subscribe to a term deposit.
- **Contact** - Most contacts were made via cellular resulting in a higher number of positive outcomes. When contacts were made via telephone, the distribution for subscribing to a term deposit is fairly even.
- **Month** - The majority of contacts occurred in May, with negative outcomes being the most prevalent. Contacts made in February, March, April, September, October and December were more likely to result in term subscriptions.
- **Poutcome** - A previously successful marketing campaign has led to more positive outcomes in the current campaign. However, contrary to what would be expected as of this latest remark, an unsuccessful marketing campaign has led to a balanced distribution of both positive and negative term subscriptions.

To summarize, the findings indicate that there isn't a clear distinction of the target variable according to the features values. However, in same cases its possible to identify more likely outcomes with the most notable being an higher tendency to adhere to a term deposit after a previously successful campaign.

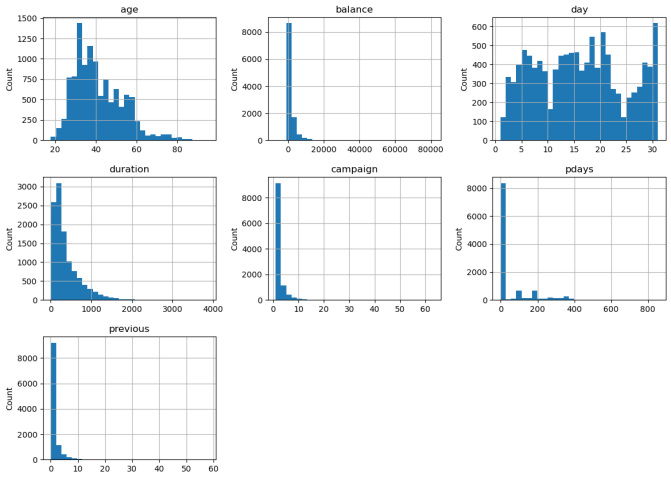

Fig. 4. Distribution of Numeric Features

2) Numerical Features: Numerical Features refer to continuous numeric values. The distributions of the numerical variables on the dataset are presented in Figure 4 leading to the following observations:

- **Age**: Most people are between 20 and 40 years old, with a peak around 30. The distribution skews slightly right, with a few individuals up to around 90 years.

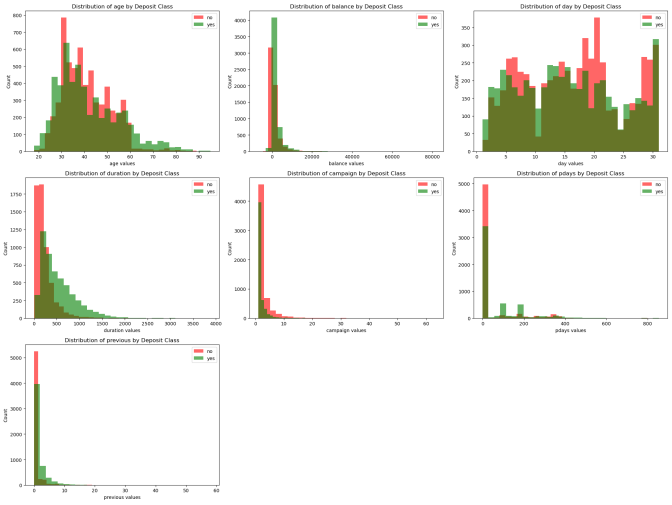

Fig. 5. Distribution of Numeric Features related to term deposit

- **Balance**: Heavily skewed towards low values, with most clients having minimal or no balance.
- **Day**: Has a uniform distribution, meaning that the calls were distributed evenly throughout the month.
- **Duration**: Exhibits a right-skewed distribution heavily concentrated around 0, indicating that most contacts were of short duration.
- **Campaign**: Most clients have been contacted less than 5 times, and it is unusual for a person to be contacted more than 10 times.
- **Pdays**: The `pdays` variable, indicating the number of days since a person was last contacted in a previous campaign, indicates that most clients were either never contacted or contacted recently, with the majority of values clustered around 0.
- **Previous**: The `previous` variable is also concentrated around 0, with a rapid decrease as the number of prior contacts rises, meaning that most clients had not been contacted before.

To extend this analysis, once again it is important to examine the relationship between the numerical features and the target variable, deposit. Figure 5 displays, for each feature, separate distributions related to the positive and negative outcomes of the output class, on the same plot. Neither of the features reveals a distinct, non-overlapping distribution between the two outcome classes. However, a few insights can still be drawn from these distributions:

- **Age** - clients who are slightly older tend to subscribe to term deposits.
- **Balance** - higher account balances are related to an increased likelihood of subscribing to term deposits.
- **Duration** - call duration is strongly correlated with subscribing to a term deposit, with longer durations indicating a higher probability of a positive outcome.
- **Campaign** - A higher number of contacts during this campaign is associated with a lower likelihood of the

client subscribing to a term deposit.

- **Pdays** - The distribution of the number of days since the client was last contacted from a previous campaign is similar for both outcomes, with a slight tendency toward the positive outcome. Once again, the spike surrounding 0 indicates no previous contact for a vast majority of the clients.
- **Previous** - Clients that have been contacted more times in previous campaigns show a higher probability to subscribe to a term deposit. Once again, the spike surrouding 0 indicates no previous contact for a vast majority of the clients.

C. Occupation and Financial Demographics

In this section, we explore the relationship between age, occupation, and financial standing within the dataset. Specifically, we analyze the age distribution across different job occupations in Figure 6 and examine how these occupations correlate with mean balance in accounts in Figure 7.

As expected, Figure 6 shows that students represent the youngest demographic, with ages clustered between 15 and 25, exhibiting little overlap with other occupations. Additionally, the figure reveals that retired individuals have the highest median age, which is also to be expected.

Interestingly, Figure 6 also reveals an unexpected detail: self-employed individuals have the second lowest median age, following students. This suggests that younger people are increasingly opting for self-employment as a career path.

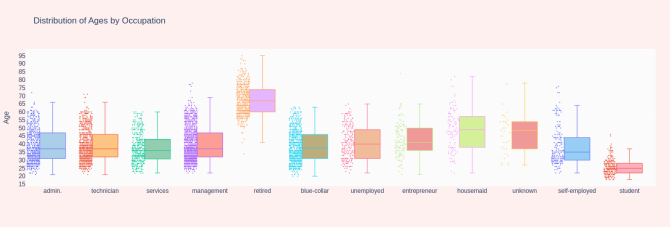

Fig. 6. Distribution of Ages by Occupation

Figure 7 displays the relation between the job occupation of a person and the respective balance on their account (mean balance). As anticipated, retirees and entrepreneurs have the highest mean balances, while students tend to have the lowest. Occupations such as management and self-employed roles also exhibit substantial mean balances in the ”high” category, suggesting financial stability.

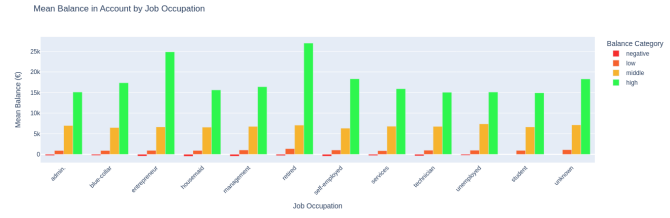

Fig. 7. Mean Balance in Account by Job Occupation

D. Clustering Marital Status and Education Level

The analysis of these features provides insight of how marital status and education level interact to influence the balance of an individual. The chart in Figure 8 exhibits the median balance by each group of marital status and education level.

From the chart,the following conclusions can be drawn:

- **Divorce significantly impacts an individual's overall balance** - with divorced individuals typically exhibiting lower median balances compared to both their married and single counterparts (the exception being the divorced/primary and single/primary pair).
- **Balance is strongly influenced by education level** - higher education levels correlate with higher median balances (the exception being divorced/primary cluster again).

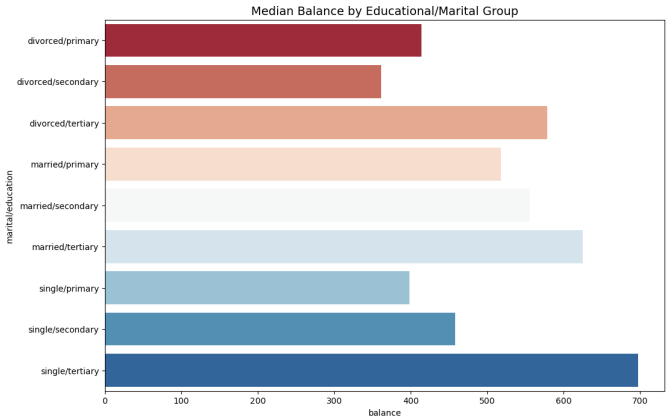

Fig. 8. Median Balance by Educational/Marital Group

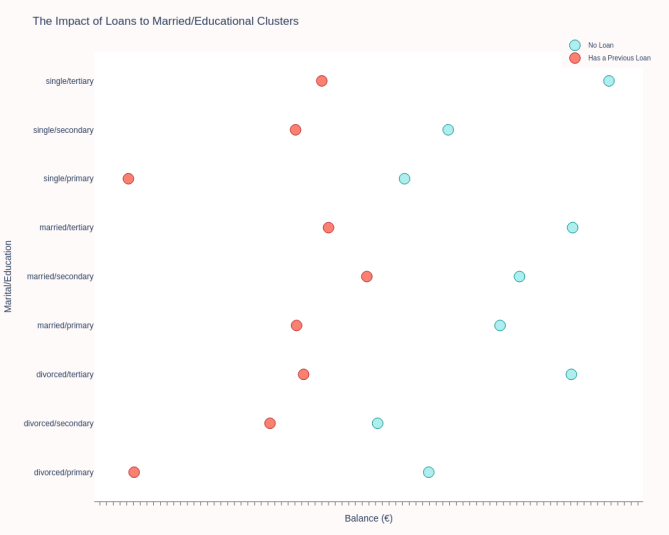

Fig. 9. Personal loans on median balance of marital and education clusters

The chart in Figure 9 illustrates the impact of personal loans on the median balances across various marital and educational clusters, comparing the balances for each cluster with and without a loan.

From its analysis, the following can be concluded:

- **Individual's balance is highly influenced by whether or not he/she has had a previous loan** - as the median balance for any given combination of marital status and education level is consistently lower for individuals who have had a loan.

E. Analyzing factors influencing term deposit subscriptions

This analysis examines three key factors that influence the subscription rate for term deposits: campaign duration, the number of contacts, and account balance.

From the plots in Figure 10 we can observe the following:

- **The duration of the campaign has a significant impact on the subscription rate** - As the campaign duration increases, the subscription rate steadily improves. However, after a certain threshold (around 900 seconds), the subscription rate begins to stabilize, reaching a peak of around 80-90%. This indicates that while longer campaign durations are beneficial, excessively long interactions may not produce significant additional gains.
- **The number of contacts has a negative correlation with subscription rates** - It decreases linearly as the number of contacts increases.
- **The account balance also has a significant impact on the subscription rate** - The percentage of subscriptions increases rapidly to around 60% as the account balance increases from 0 to 4000 euros. After this point, the subscription rate starts decreasing, meaning that very high account balances do not necessarily lead to higher subscription rates.

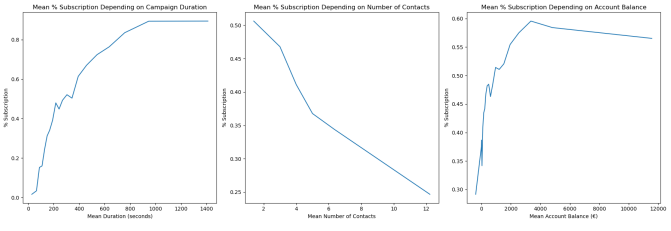

Fig. 10. Subscription rate vs duration, contacts and balance

F. Analysis of the correlation matrix

The correlation matrix reveals key relationships between numerical features and their impact on the target variable (deposit). From the matrix in Figure 11, the following observations can be made:

- **Strong Correlation Between Call Duration and Deposit Outcome** - The most significant positive correlation with the target variable is the duration (0.45). This aligns with previous observations (Figure 10 and Figure 2) that campaign duration plays a crucial role in influencing customer decisions. Therefore, duration presents itself as the most dominant feature in our dataset.
- **Minimal Correlation of some Variables with Deposit Outcome** - Variables such as age and day, show little to no correlation with the deposit outcome, indicating they may have minimal predictive value for determining subscription probability.

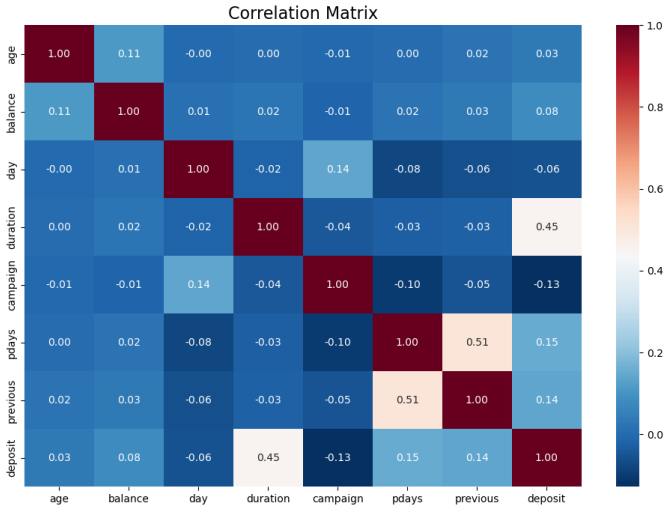

Fig. 11. Correlation matrix

IV. Methodology

In the following section, we will outline the methodology employed to solve the bank marketing dataset problem, which includes a detailed analysis of the following steps: data preprocessing, machine learning models implementation, hyperparameter tuning and the data splitting strategy.

A. Data Preprocessing

Effective preprocessing is crucial to prepare the dataset for modeling and ensure optimal performance of the machine learning algorithms.

1) Handling missing values: The dataset was inspected using `df.isnull().sum()` to check for missing values. However, no missing values were found. Therefore, there is no need to fill any missing values with the median, mean or mode.

2) Target Variable Encoding: The target variable `deposit` possessed two possible string values: `yes` and `no`. These were converted to a binary numeric format using `LabelEncoder` [6], where `yes` was encoded as 1 and `no` was encoded as 0. A similar approach was applied for all other features with possible values `yes` and `no`.

3) Convert categorical variable into dummy/indicator variables: The `pandas.get_dummies` function was used to transform categorical variables into binary values. Each categorical variable is converted in as many 0/1 variables as there are different values. For example, `month` class is transformed into `month_jan`, `month_fev`, etc. This approach

was applied to categorical features with more than two possible values.

4) Standardization of Numerical Features: Numerical features were scaled with `StandardScaler` [7] to ensure that all features had a mean of 0 and a standard deviation of 1. This ensures that all numerical features are on a similar scale and can improve performance for some of our models, such as K-Nearest Neighbors.

B. Splitting the dataset

The dataset was split using stratified sampling to ensure that the distribution of the target value (deposit) remains consistent between the two following subsets:

- **Training Set** - 80% of the data was allocated to the training set, that is used to train and optimize the machine learning models. This set will be used by `GridSearchCV` [8] function to conduct a k-fold cross validation to find the optimal hyperparameter grid.
- **Test Set** - 20% of the data was allocated to the test set, used to evaluate the final models performance.

C. Model Selection

Four different models were trained and tested, chosen for their suitability for the classification task and their alignment with what has been learned in the machine learning fundamentals' course.

- **K-Nearest Neighbors (KNN)** - The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point [9]. The proximity is measured using distance metrics like Euclidean, Manhattan, or Minkowski. Thus, each data point is classified based on the majority class of its k-nearest neighbors (in the feature space).
- **Logistic Regression** - The Logistic Regression model estimates the probability of a data point belonging to a certain class using a sigmoid (logistic) function. The output of the model is a value between 0 and 1, which is interpreted as a probability. Based on this probability, the model makes a prediction: if the probability exceeds 0.5, the data point is classified as class 1. Otherwise, it is classified as class 0.
- **Random Forest** - The Random Forest model is an ensemble model that combines the output of multiple decision trees to reach a single result [10]. Being this a classification problem, the final prediction is determined by majority voting among the various decision tree classifiers. However, this comes at a cost as it's very computational expensive, especially for large datasets or when the number of trees and features is high, as it requires building and evaluating multiple trees.
- **Support Vector Machine (SVM)** - A Support Vector Machine is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space [11]. SVM can also handle non-linear separable data by using kernel functions to transform the input space into a higher-dimensional feature space.

D. Parameter Tuning

Hyperparameter tuning involves systematically searching through a set of hyperparameters and selecting the combination that yields the best results on the given data. In order to perform this step, the function `GridSearchCV` was used, which performs an exhaustive search over the specified parameter values for either of the following estimators: KNeighborsClassifier [12], LogisticRegression [13], RandomForestClassifier [14] or SVC [15]. The `GridSearchCV` function requires a scoring function to optimize based on the given parameters - the scoring metric used was accuracy, since, as previously mentioned, the dataset is fairly balanced. The tuned hyperparameters for each of the previous models were as follows:

- **K-Nearest Neighbors (KNN)** - A range of values from 1 to 19 were considered for the `n_neighbors` parameter in order to assess the impact of different neighborhood sizes on model performance. The weighting function (`weight` parameter) applied to the neighbors was also varied between 'uniform' - all points in each neighborhood are weighted equally - and 'distance' - weight points corresponding to the inverse of their distance giving greater influence to the closest neighbors - values. In addition, different metrics for the distance computation were tested: 'minkowski', 'euclidean' and 'manhattan'.
- **Logistic Regression** - The regularization parameter `C` was varied between the following values [0.001, 0.01, 0.1, 1, 10, 100, 1000]. `C` corresponds to the inverse of the regularization strength λ. Therefore, a small value of `C` (like 0.001) leads to a stronger regularization and a simpler model, whereas a large value of `C` (like 1000) leads to weaker regularization and a more complex model - really small and really large values of `C` may lead to underfitting or overfitting, respectively. Furthermore, different optimization algorithms (`solver` parameter) were used between 'liblinear' and 'saga'. The type of regularization (`penalty` parameter) applied to the model was also differed between Lasso regularization ('l1' value) and Ridge regularization ('l2' value). By varying all these parameters, it's possible to find an optimal balance between bias and variance in the model.
- **Random Forest** - The following values [100, 200, 400] were experimented for the `n_estimators` which determines the number of trees in the forest. Additionally, the `max_depth` of the tree was changed between [10, 15, 20, 30, None]. The parameter `min_samples_split` (the minimum number of samples required to split an internal node) has also been varied between the values [1, 2, 4]. This way, the goal was to find a combination of parameters that improves model performance, while balancing computational efficiency (as it is computationally expensive due to being an ensemble model) and avoid overfitting.

- **Support Vector Machine (SVM)** - Similarly to the logistic regression model, the regularization parameter `C` was varied between the values [0.1, 1, 10, 100, 1000]. Different values for the kernel coefficient `gamma` were also applied - [1, 0.1, 0.01, 0.001, 0.0001]. Different kernel functions weren't applied due to the performance impact of doing so in the grid search method.

Therefore, `GridSearchCV` conducted k-fold cross-validation while systematically exploring the parameter grids for each model, with k being 10 for all models excluding SVC (too computationally expensive for a bigger fold value). The goal was to identify the optimal hyperparameters that would enhance model performance and improve its ability to generalize to unseen data. The resulting hyperparameters of this exhaustive search are presented in Table II, along with the corresponding best accuracy value achieved for the best parameter configuration on the training data.

Model Name	Best Hyperparameters	Best Score
K-Nearest Neighbors (KNN)	`metric: 'minkowski', n_neighbors: 16, p: 2, weights: 'distance'`	0.818
Logistic Regression	`C: 0.1, solver: 'saga', penalty: 'l1'`	0.828
Random Forest	`max_depth: 30, min_samples_leaf: 2, n_estimators: 400`	0.855
Support Vector Classifier	`C: 100, gamma: 0.01`	0.851

TABLE II
BEST HYPERPARAMETERS FOR EACH MODEL AFTER TUNING.

The `GridSearchCV` function also refits automatically (`refit` parameter set to true by default) each estimator using the best found parameters on the whole training set. Thus, to evaluate the model's true performance, we simply need to make predictions on the test data and compute the relevant metrics. This analysis will be presented in the next section.

V. RESULTS AND ANALYSIS

In this section, the performance metrics of each model are presented based on their evaluation on the test set. For each model, the best hyperparameters found in the Parameter Tuning section were used. This metrics reflect how well each of the models are expected to generalize with unseen data, and can be found in Table III.

Model Name	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (KNN)	0.827	0.827	0.825	0.826
Logistic Regression	0.829	0.828	0.828	0.828
Random Forest	0.858	0.860	0.860	0.858
Support Vector Classifier	0.861	0.861	0.862	0.861

TABLE III
EVALUATION METRICS FOR TUNED MODELS ON THE TEST SET.

Firstly, it is important to note that the accuracy values for the training and testing set, as shown in Tables II and III, are similar, which suggests that the models are well generalized - the model is neither overfitting nor underfitting.

Based on the results of the Table III, Random Forest and Support Vector Classifier emerged as the top-performing models with an accuracy of approximately 0.86, whereas K-Nearest Neighbors and Logistic Regression show a slightly lower accuracy.

Overall, all models maintained consistent values across all the precision, recall and f1-score metrics which means that the models are not biased towards one class over the other, with the Random Forest and Support Vector Classifier remaining the top performers. This consistency suggests that the models effectively differentiate between positive and negative classes.

This results indicate that if robustness and accuracy is needed, Random Forest and SVC are the recommended models to use. However, if speed and computational cost is a concern, Logistic Regression will be a great choice providing a balance between efficiency and performance.

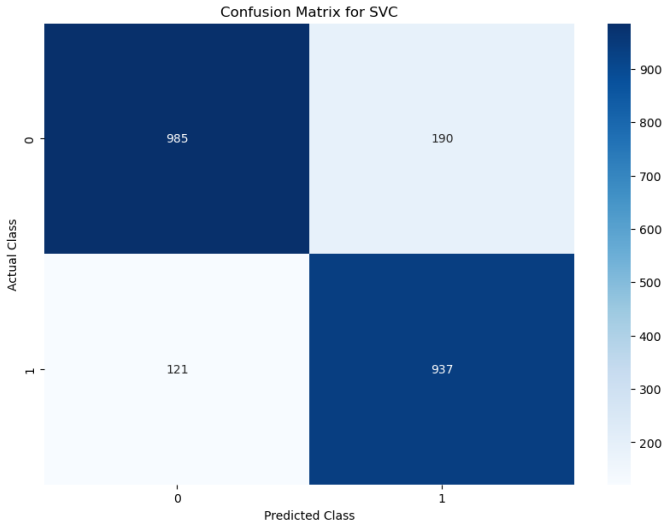

Fig. 12. Confusion Matrix for SVC

The Figure 12 displays the confusion matrix for the top performer model: Support Vector Classifier. Similar confusion matrices were plotted for the rest of the models which can be found in the source code. In Figure 12, the target classes are represented as actual versus predicted values for the binary classification problem. Correct classifications, where predictions align with the actual values, are reflected in the diagonal cells with higher counts. The model is showing a much higher number of True positives and True Negatives correspondent to the correct classification of 'deposit' and 'no deposit' classes, which is high for both classes. This indicates that the SVC model is effective on identifying both classes.

A. Classification with feature removal

To optimize the performance of the machine learning models and reduce dimensionality, feature selection was performed using Random Forest Classifier. The Model was trained on the dataset after encoding and converting categorical data to dummy values. The model's `feature_importances_` [14] attribute was used to obtain the most important features, ranked based on their contribution to the model's predictive power.

After that, to limit the dataset to the most relevant features, only the top 40 features were selected. The procedure remained consistent with what was presented on the rest of the report, using an 80%-20% train-test split, followed by a grid search applied to the training set. The evaluation metrics results on the test set for each of the models after the removal of the 8 least important features is shown in Table IV.

Model Name	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (KNN)	0.830	0.830	0.828	0.829
Logistic Regression	0.830	0.829	0.829	0.829
Random Forest	0.858	0.859	0.860	0.858
Support Vector Classifier	0.861	0.861	0.862	0.861

TABLE IV
EVALUATION METRICS FOR TUNED MODELS ON THE TEST SET WITH FEATURES REMOVED

Upon further inspection of Table IV, it can be concluded that the values for the accuracy, precision, recall and F1-score metrics either improved or remained consistent after the feature removal - with the exception being the Random Forest's precision which saw a really small decrease. Therefore, it is recommended to decrease the feature's scope, as this simplification appears to enhance the model's performance or at least maintain it, while reducing their computational complexity and without compromising their predictive power. Similarly to before, the Figure 13 displays the confusion matrix for the top performer model: Support Vector Classifier. Similar confusion matrices were plotted for the rest of the models which can be found in the source code.

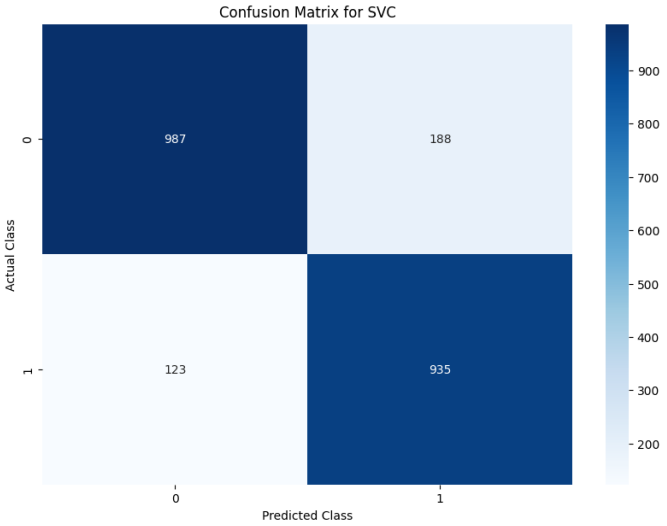

Fig. 13. Confusion Matrix for SVC with features removed

The Figure 13 shows a similar distribution of True Positives, True Negatives, False Positives, and False Negatives, with only minor variations. Based on this the previous recommendation still holds: decrease the feature's scope, as this simplification appears to enhance the model's performance or at least maintain it, while reducing their computational complexity and without compromising their predictive power.

VI. COMPARING WITH SOLUTIONS FROM OTHER REFERENCES

This section does not include comparisons with the works referenced in the State of the Art Analysis section as they either used different datasets (similar datasets with a larger number of records) or did not focus on model development. As a result, the comparisons are based on models and related Jupyter notebooks publicly accessible in kaggle [16]. Special credit is given to both Janio Martinez Bachmann [17] and Aleksandra Deis [18], whose work has served as both a valuable comparative reference and a source of inspiration for this report, specially regarding the data analysis section.

In the work of Janio Martinez Bachmann [17], the same classifiers were used, but all of their models are underperforming compared to ours. The most notable difference is observed with K-Nearest Neighbors (KNN), where the accuracy difference is 0.0255 (approximately 2.5%). This could be because the models in their study may not have undergone hyperparameter tuning, using the default hyperparameters for each model. In contrast, our models benefited from an extensive search for the best hyperparameters, using `GridSearchCV` to optimize the performance of each classifier.

The Rhythm Shah work [19] utilizes RandomForestClassifier and SVC. RandomForest achieved a score very close to ours (0.85), while for SVC, their accuracy is 0.76 while our is 0.86, a significant improvement of about 10%. This could be because of data normalization and encoding as he only uses `LabelEncoder` for all the features and doesn't normalize the data.

Aleksandra Deis [18] uses XGBClassifier, achieving a test accuracy of 0.848. It outperforms two of our models. This is expected, as XGBoost is an advanced ensemble method, making it generally more robust than simpler models like KNN and Logistic Regression.

VII. CONCLUSION

This report provides a comprehensive analysis of the bank marketing dataset, covering everything from an in-depth examination of the dataset to the application of various models for addressing the classification problem. It also encompasses the full pipeline to train, test and evaluate each of the proposed models as well as the impact of data preprocessing and feature selection on the model's performance. More specifically by comparing results from models trained on the complete feature set against those trained on a reduced subset, we have demonstrated the potential for optimizing predictive accuracy while simultaneously enhancing computational efficiency.

REFERENCES

[1] P. Georgieva, "PROJECT 1 2024/2025 – Instructions," [Online]. Available: https://elearning.ua.pt/mod/resource/view.php?id=996489. [Accessed: Nov. 23, 2024].

[2] J. Bachmann, "Bank Marketing Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset. [Accessed: Nov. 27, 2024]

[3] Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data. *Online*, Available: https://dergipark.org.tr/en/download/article-file/365990

[4] A data-driven approach to predict the success of bank telemarketing *Online*, Available: https://www.sciencedirect.com/science/article/abs/pii/S016792361400061X?fr=RR-2&ref=pdf_download&rr=8e72ac1ad8595bd6

[5] UCI Machine Learning Repository, "Bank Marketing Dataset," 2014. [Online]. Available: https://archive.ics.uci.edu/dataset/222/bank+marketing. [Accessed: Nov. 23, 2024].

[6] Scikit-learn, "LabelEncoder," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html. [Accessed: Nov. 27, 2024].

[7] Scikit-learn, "StandardScaler," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. [Accessed: Nov. 27, 2024].

[8] Scikit-learn, "GridSearchCV," [Online]. Available: https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html. [Accessed: Nov. 27, 2024].

[9] IBM, "K-Nearest Neighbors (KNN)," available at: https://www.ibm.com/topics/knn, accessed Nov. 26, 2024.

[10] IBM, "Random Forest," [Online]. Available: https://www.ibm.com/topics/random-forest. [Accessed: Nov. 27, 2024].

[11] IBM, "Support Vector Machine," [Online]. Available: https://www.ibm.com/topics/support-vector-machine. [Accessed: Nov. 27, 2024].

[12] Scikit-learn, "KNeighborsClassifier," [Online]. Available: https://scikit-learn.org/dev/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. [Accessed: Nov. 27, 2024].

[13] Scikit-learn, "LogisticRegression," [Online]. Available: https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html. [Accessed: Nov. 27, 2024].

[14] Scikit-learn, "RandomForestClassifier," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed: Nov. 27, 2024].

[15] Scikit-learn, "SVC," [Online]. Available: https://scikit-learn.org/dev/modules/generated/sklearn.svm.SVC.html. [Accessed: Nov. 27, 2024].

[16] Kaggle, "Kaggle: Your Home for Data Science," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/. [Accessed: Nov. 27, 2024].

[17] J. Bachmann, "Bank Marketing Campaign: Opening a Term Deposit," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/code/janiobachmann/bank-marketing-campaign-opening-a-term-deposit. [Accessed: Nov. 27, 2024].

[18] A. Deis, "Bank Marketing Analysis," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/code/aleksandradeis/bank-marketing-analysis. [Accessed: Nov. 27, 2024].

[19] R. Shah, "Bank Customer Segmentation," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/code/rhythmshah11/bank-customer-segmentation#Random-Forest. [Accessed: Nov. 27, 2024].