Business Report – Customer Subscription Prediction Project

Table of Content

Data Tables

Charts & Figures

Problem Statement 1 Portuguese Bank Customer Subscription

1.1 Business Context

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

1.2 Objective

The classification goal is to predict if the client will subscribe a term deposit (variable y). For this, we will perform all the necessary steps such as data exploration, exploratory data analysis, feature engineering and selection, model building, performance evaluation, and final predictions.

1.3 Data Dictionary

- age - Age of the client (in years)
- job - Type of job (e.g., 'management', 'technician', 'blue-collar', etc.)
- marital - Marital status (e.g., 'single', 'married', 'divorced')
- education - Level of education ('primary', 'secondary', 'tertiary', or 'unknown')
- default - Whether the client has credit in default ('yes', 'no')
- balance - Average yearly balance in euros
- housing - Whether the client has a housing loan ('yes', 'no')
- loan - Whether the client has a personal loan ('yes', 'no')
- contact - Type of communication contact ('cellular', 'telephone', 'unknown')
- day - Last contact day of the month
- month - Last contact month of the year (e.g., 'jan', 'may', 'jul')
- duration - Last contact duration in seconds
- campaign - Number of contacts performed during this campaign for this client
- pdays - Number of days since the client was last contacted (-1 means not previously contacted)
- previous - Number of contacts performed before this campaign
- poutcome - Outcome of the previous marketing campaign ('success', 'failure', 'other', 'unknown')
- y - Target variable – has the client subscribed to a term deposit? ('yes', 'no')

1.4 Data Overview

1.4.1 How many rows and columns are present in the data?

- The dataset consists of 45,211 rows and 17 columns.

1.4.2 What are the data-types of the different columns in the dataset?

- Categorical - ['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'previous', 'poutcome']
- Numerical - ['age', 'balance', 'day', 'duration', 'campaign', 'pdays']
- Dependent - [‘y’]

1.4.3 Are there any missing values in the data?

- There are no missing values in any of the columns in the dataset.

1.4.4 Statistical summary of the data

TABLE 1. Statistical summary of the dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
age	45211	40.93	10.61	18	33	39	48	95
balance	45211	1362.27	3044.76	-8019	72	448	1428	102127
day	45211	15.80	8.32	1	8	16	21	31
duration	45211	258.16	257.52	0	103	180	319	4918
campaign	45211	2.76	3.09	1	1	2	3	63
pdays	45211	40.19	100.12	-1	-1	-1	-1	871
previous	45211	0.58	2.30	0	0	0	0	275

1.5 Exploratory Data Analysis (EDA)

FIG. 1 Count of dependent variable

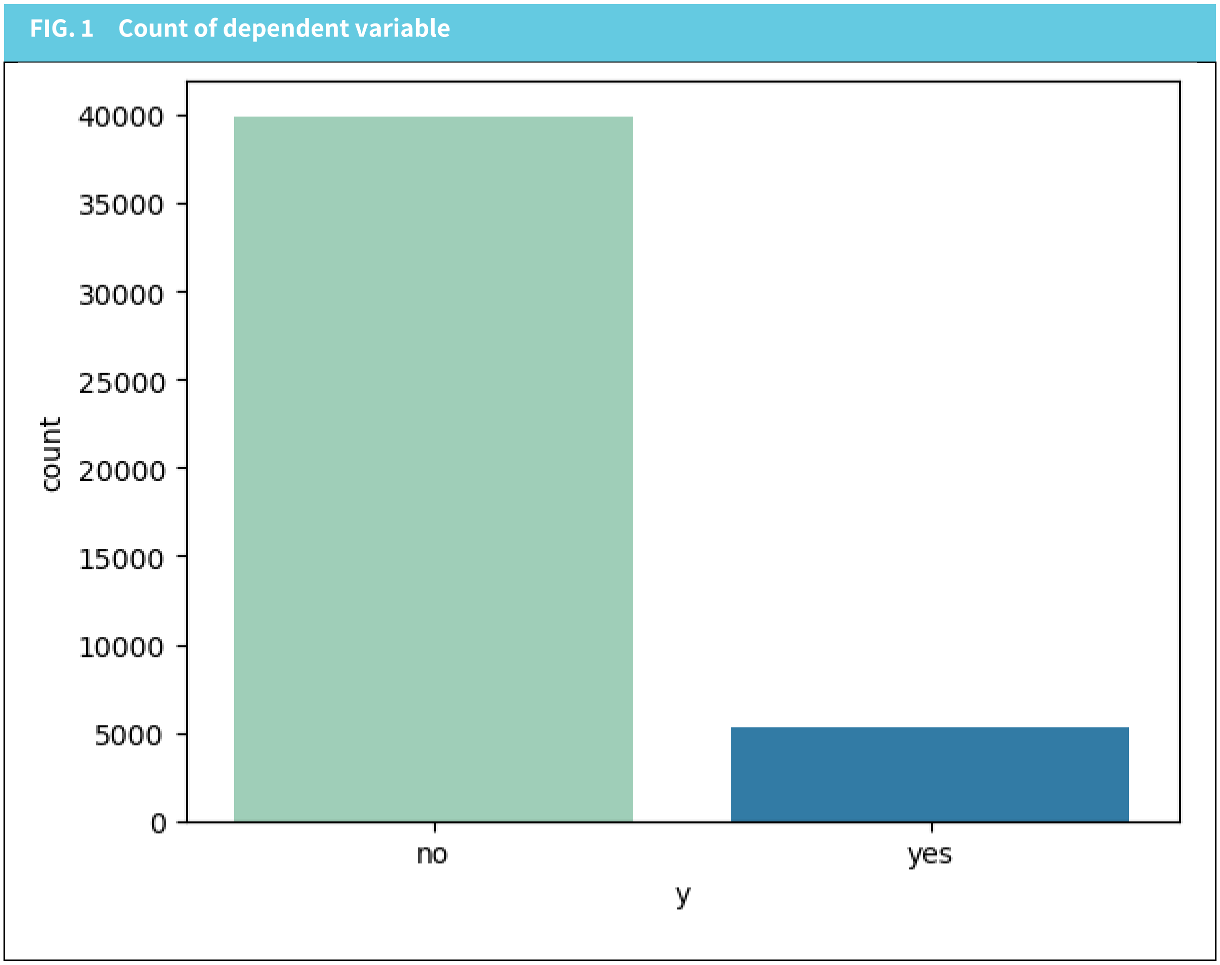


FIG. 2 Dependent variable vs Age

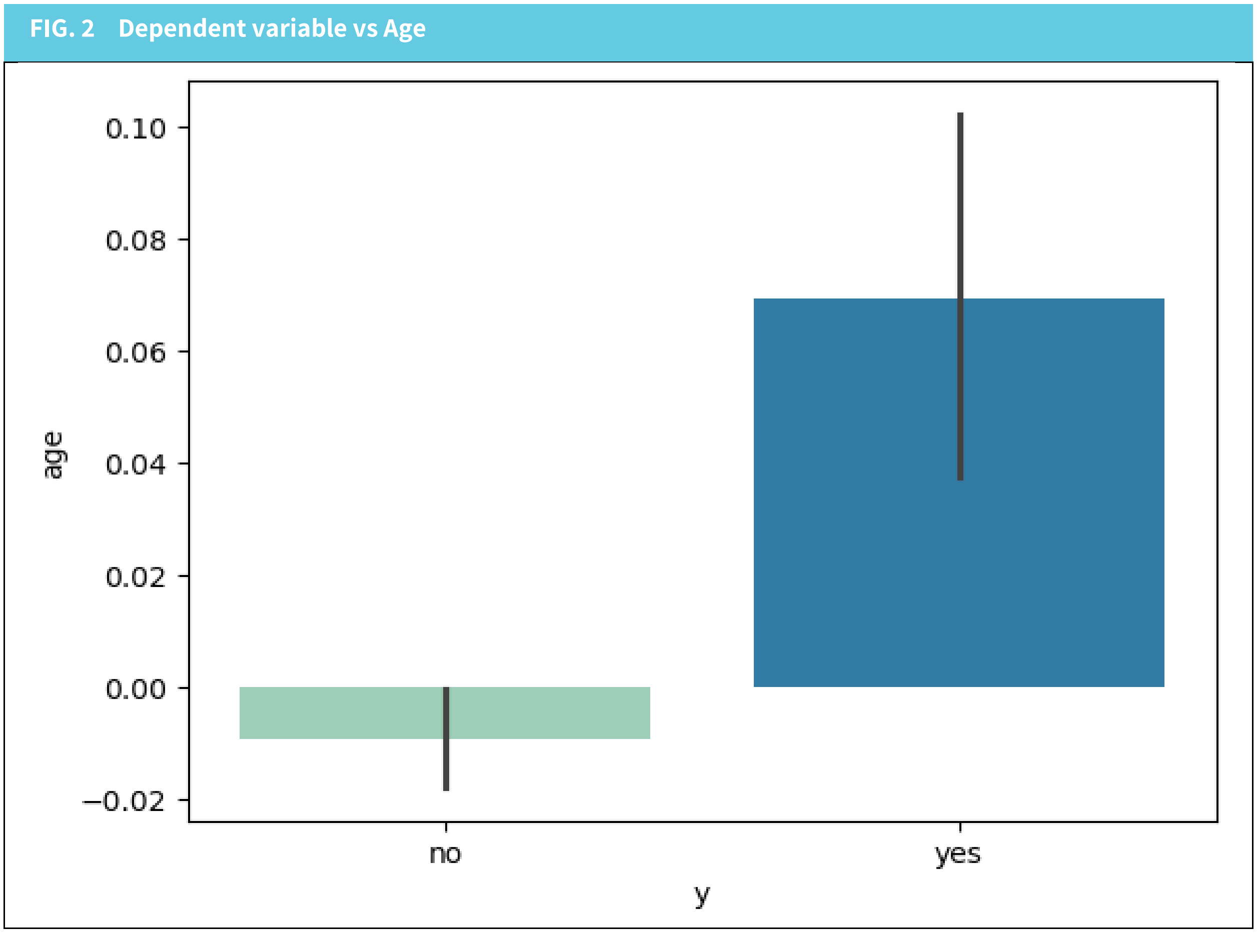


FIG. 3 Marital Status vs Age

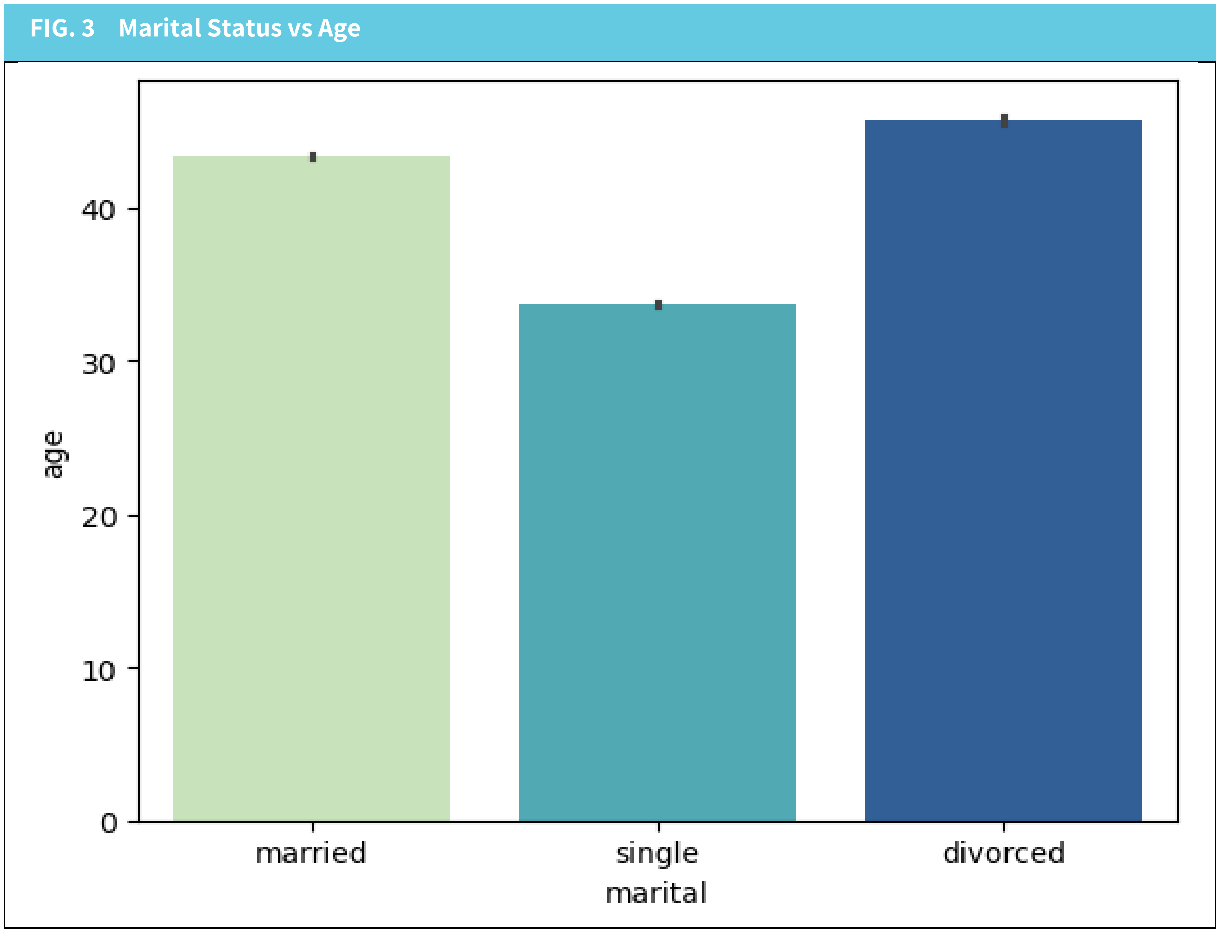


FIG. 4 Count of education variable

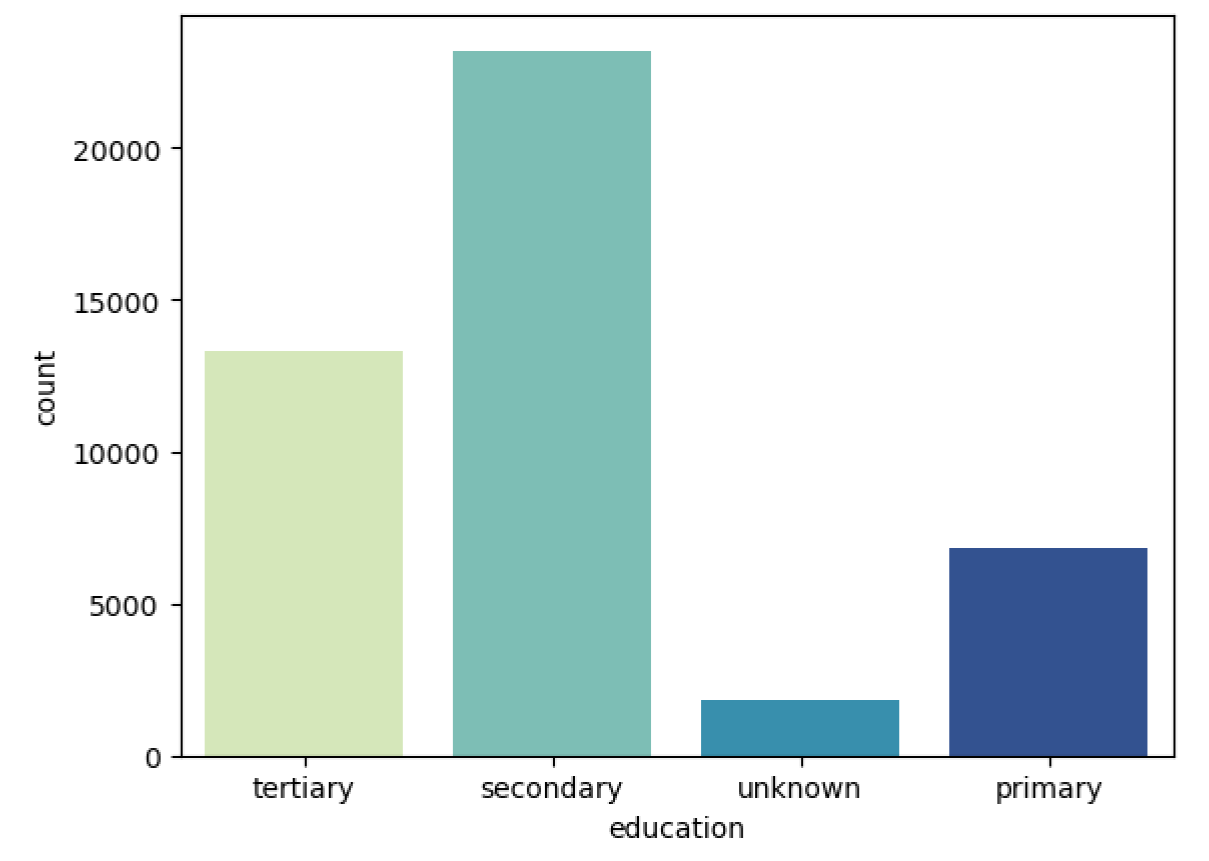
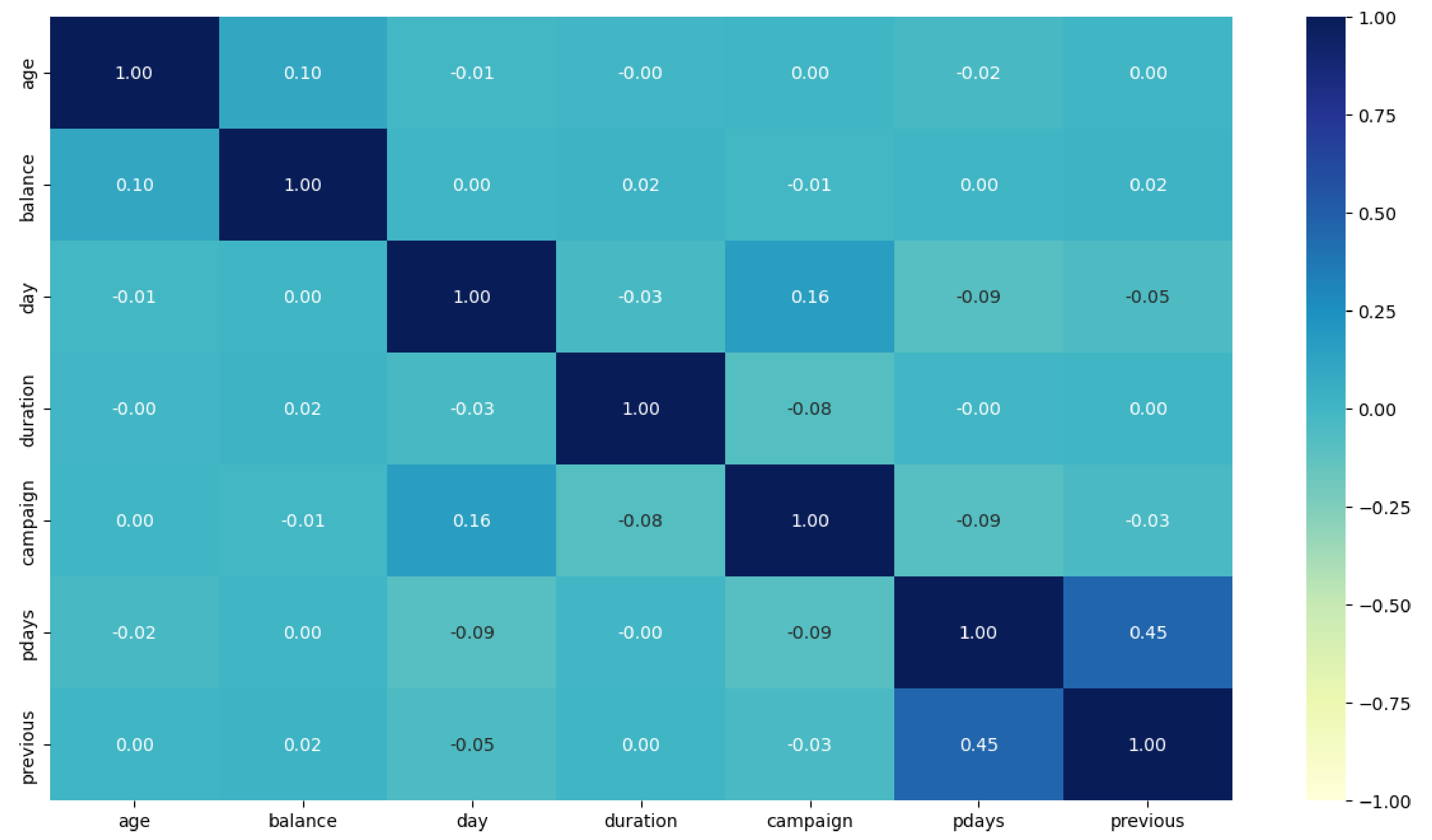
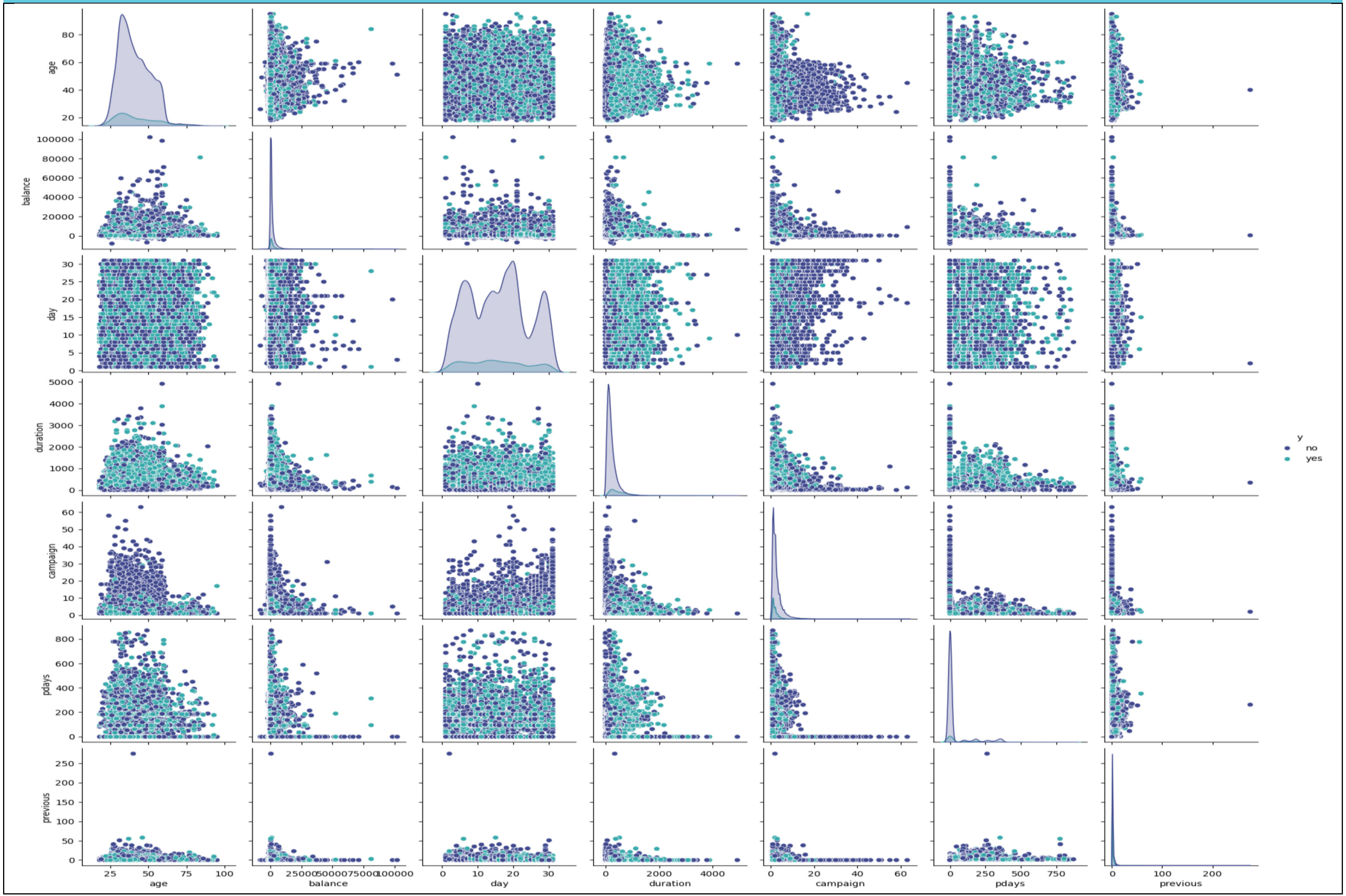


FIG. 5 Heatmap



	age	balance	day	duration	campaign	pdays	previous
age	1.00	0.10	-0.01	-0.00	0.00	-0.02	0.00
balance	0.10	1.00	0.00	0.02	-0.01	0.00	0.02
day	-0.01	0.00	1.00	-0.03	0.16	-0.09	-0.05
duration	-0.00	0.02	-0.03	1.00	-0.08	-0.00	0.00
campaign	0.00	-0.01	0.16	-0.08	1.00	-0.09	-0.03
pdays	-0.02	0.00	-0.09	-0.00	-0.09	1.00	0.45
previous	0.00	0.02	-0.05	0.00	-0.03	0.45	1.00

FIG. 6 Pairplot



1.5.1 Observations from the exploratory data analysis

- Approximately 85-90% of the observations in the dependent variable are non-subscribers.
- Among the job categories, blue-collar words account for the highest number of clients in the dataset.
- Customers contacted through cellular medium have a higher tendency to subscribe a term deposit as compared to other contact mediums.
- The month of may witnesses the highest number of customer interactions, in comparison to any other month of the year.
- The ‘previous’ and ‘pdays’ variables have the only highest correlation of 0.45, whereas the other variables all have a correlation with each other of less than 0.2

1.6 Feature Engineering and Selection

- There are total 10 categorical,6 numerical, and 1 dependent variable in the dataset.
- One hot encoding was performed on the categorical variables to ensure scalability for model fitting and predictions.
- Additionally, a standardscaler() function was used to perform scaling on the numerical variables to ensure optimal data preparation.
- The data was then split into training and testing sets with a test size ratio of 0.3. The training data had a total of 31,647 rows and 14 features, whereas the testing data had a total of 13,564 rows and 14 features.
- After the scaling and one-hot encoding the dataset had a total 81 features.
- To further shrink the data and select only important and useful features, we performed feature selection using a Lasso regression model.
- Total 14 features were identified as important by the Lasso regression model, where as 67 other features had a coefficient of 0 and were dropped.
- The list of selected features includes ['balance', 'duration', 'campaign', 'pdays', 'marital_married', 'marital_single', 'education_tertiary', 'housing_yes', 'loan_yes', 'contact_unknown', 'month_jul', 'month_may', 'poutcome_success', 'poutcome_unknown'].

1.7 Model Building

TABLE 2. Model performances on training data

Model	Accuracy Score
XGBClassifier	92.04%
LGBMClassifier	92.01%
K-Nearest Neighbors Classifier	91.99%
Random Forest Classifier	90.97%
Support Vector Machine	90.42%
Decision Tree Classifier	90.13%
Logistic Regression	90.08%
Naive Bayes Classifier	87.48%

- A total of 8 classification models were trained, for which the accuracy scores are given below.
- Hypertuning was performed for the Decision Tree, Random Forest, XGB Classifier, and LGBM Classifier models using Randomized Search Cross Validation.
- XGBoost classifier had the highest accuracy of 92.04% on the training data, and was selected as the final model for predictions on the test data.

1.8 Test Data Predictions

TABLE 3. Prediction on the test data

Model	Accuracy Score
XGBClassifier	90.12

TABLE 4. Confusion matrix for test predictions

	Predicted Positive	Predicted Negative
Actual Positive	580	1007
Actual Negative	333	11644

- The XGB classifier model attained an accuracy of 90.12 % on the test data, indicating that the model generalized exceptionally well.

1.9 Conclusions and Recommendations

- Prioritize outreach to customers via cellular contact methods, as this medium has shown to be significantly more effective in converting prospects into subscribers compared to others.
- Launch major marketing campaigns outside the month of May, as customer interactions peak in May, suggesting potential diminishing returns if all efforts are concentrated in a single month.
- Develop personalized offers for blue-collar customers, since they form the largest job category in the dataset and may respond well to tailored term deposit solutions.
- Target single and married individuals differently, as both marital statuses showed up in important features, indicating that tailored messaging based on life stage could increase engagement.
- Educate tertiary-educated customers about the benefits of term deposits, as this group is likely financially literate and may respond positively to data-driven investment products.
- Focus retention efforts on customers who previously interacted with the bank, especially those with prior outcomes labeled as ‘success’, a key feature in predicting subscription.
- Exclude or deprioritize customers with missing or unknown outcomes, as 'poutcome_unknown' appeared as a significant feature, suggesting some predictive value but potentially less ROI than known segments.
- Increase interaction duration strategically, since the 'duration' variable was one of the most predictive features, indicating longer conversations positively influence subscription likelihood.
- Reevaluate the number of contacts per campaign, because the ‘campaign’ variable is significant—optimizing follow-up frequency could balance between conversion and customer fatigue.
- Improve customer experience with unknown or uncertain contact types, as 'contact_unknown' is still a relevant predictor and addressing its ambiguity could improve model performance and real-world outcomes.
- Utilize the Lasso-selected features to develop customer segmentation strategies, allowing for smarter allocation of marketing budgets based on proven influencing factors.
- Implement proactive campaigns during months like July, since this month was selected as predictive, indicating it’s an effective time for conversions outside peak periods.
- Use the 'pdays' variable to reengage clients who were previously contacted, as this variable’s strong correlation with 'previous' shows a clear opportunity in timing re-contact efforts.
- Introduce interest-rate incentives or tiered benefits for clients with higher balances, since ‘balance’ is a strong predictor—financial incentives might push this group toward subscribing.

- Reassess communication strategies for customers with existing housing and personal loans, given ‘housing_yes’ and ‘loan_yes’ are predictive—offering bundled financial products could increase subscription likelihood.
- Use predictive modeling outputs to prioritize high-probability leads, enabling the sales and outreach teams to focus time and resources on customers more likely to convert.
- Avoid excessive re-targeting of clients from prior unsuccessful campaigns, since prior contact history is crucial; strategic restraint might preserve long-term engagement potential.
- Educate internal teams on interpreting model outputs and feature importance, so they can tailor pitches or customer service interventions based on statistically relevant behaviors.
- Develop and test new contact scripts aimed at high-impact variables, such as ‘duration’ and ‘pdays’, to better guide conversations toward closing term deposit deals.