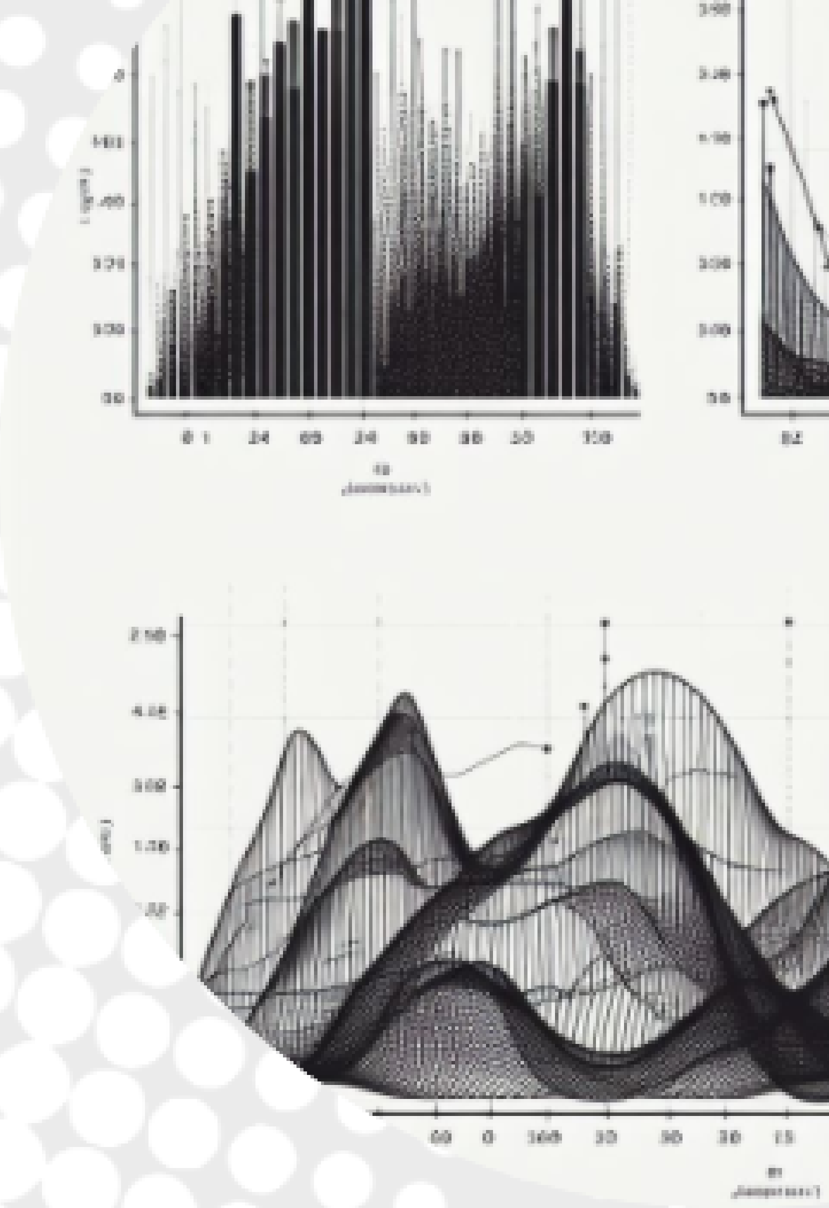
PREDICTING BANK TERM DEPOSIT SUBSCRIPTIONS



WHAT IMPACTS THE DECISION MORE ?

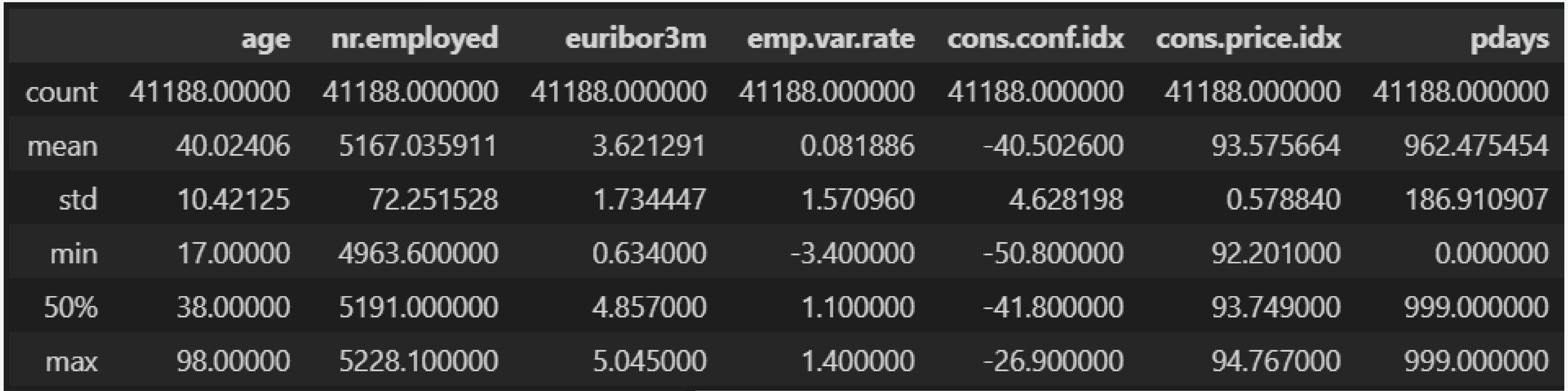
PROBLEM STATEMENT

- The goal is to **predict if a client will subscribe to a term deposit** based on their personal and financial information, as well as their response to previous marketing campaigns.
- This can help the bank to optimize its marketing strategy and increase its revenue by targeting the most potential customers.

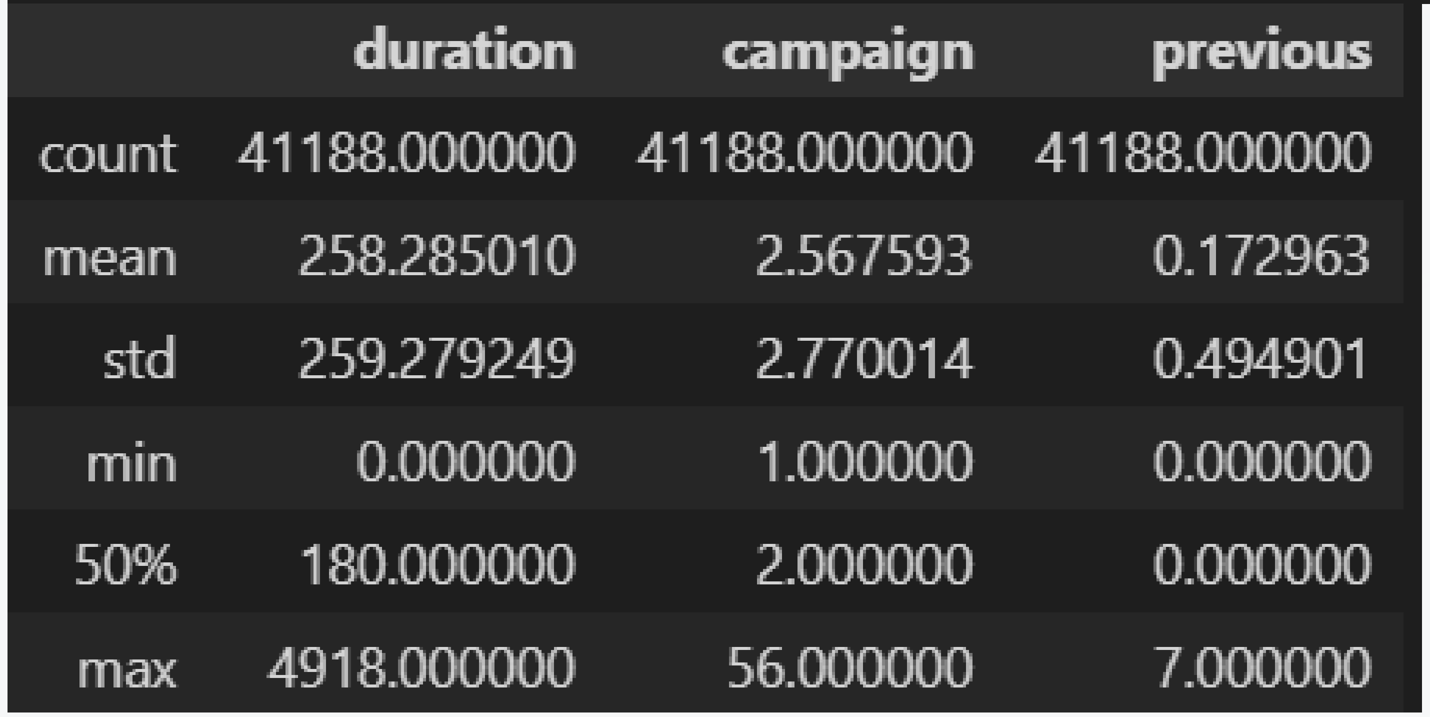
DATASET

- The data is sourced from the UCI Machine Learning Repository1, which contains data from direct marketing campaigns (phone calls) of a Portuguese banking institution.
- The data set has **41,188** observations and **21** features, including the target variable y (yes/no).
- The features include **demographic**, **economic**, **social**, and **behavioral** attributes of the clients, as well as information about the contact, duration, and outcome of the previous and current campaigns.

DATA EXPLORATION

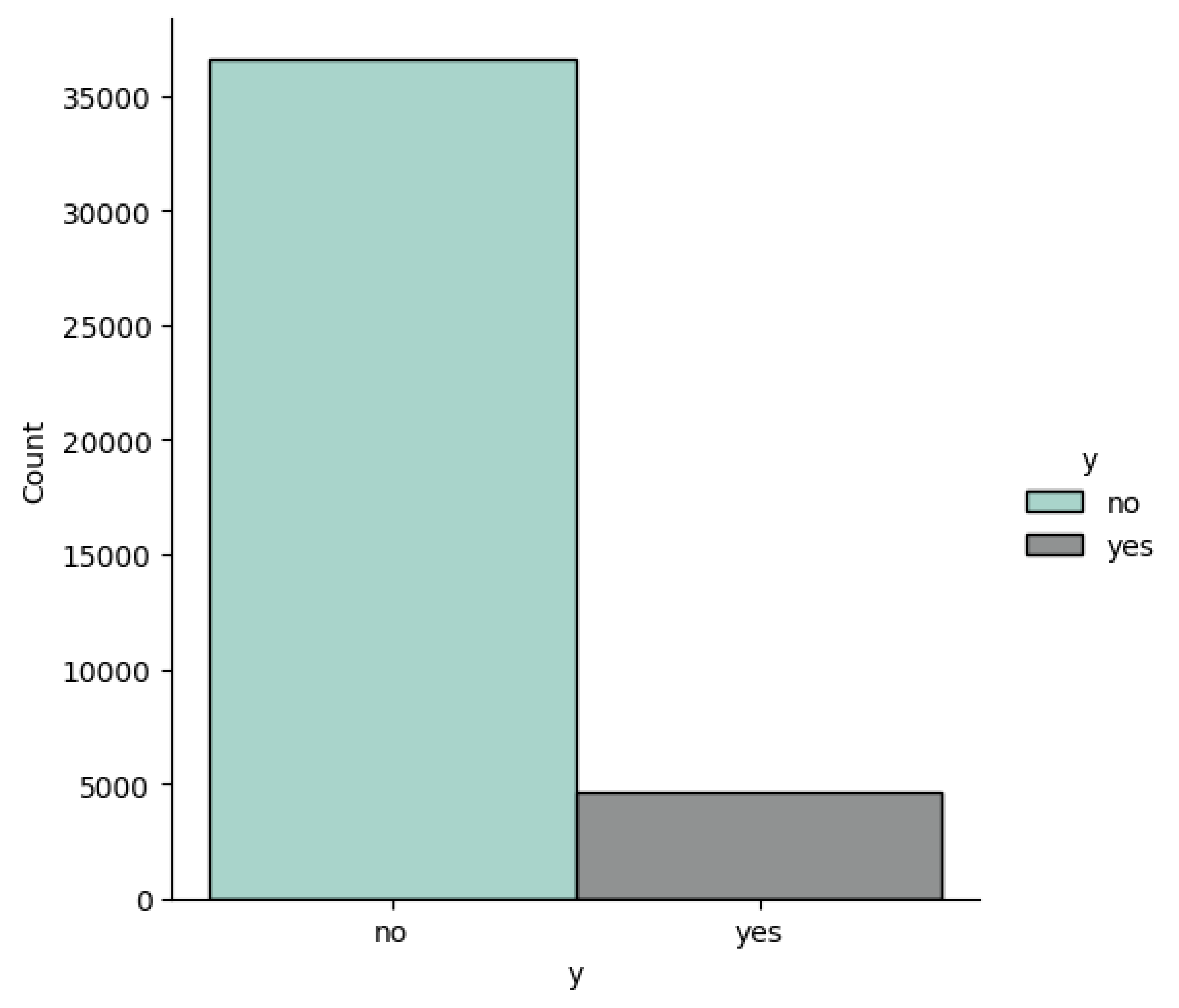


	age	nr.employed	euribor3m	emp.var.rate	cons.conf.idx	cons.price.idx	pdays
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	5167.035911	3.621291	0.081886	-40.502600	93.575664	962.475454
std	10.42125	72.251528	1.734447	1.570960	4.628198	0.578840	186.910907
min	17.00000	4963.600000	0.634000	-3.400000	-50.800000	92.201000	0.000000
50%	38.00000	5191.000000	4.857000	1.100000	-41.800000	93.749000	999.000000
max	98.00000	5228.100000	5.045000	1.400000	-26.900000	94.767000	999.000000



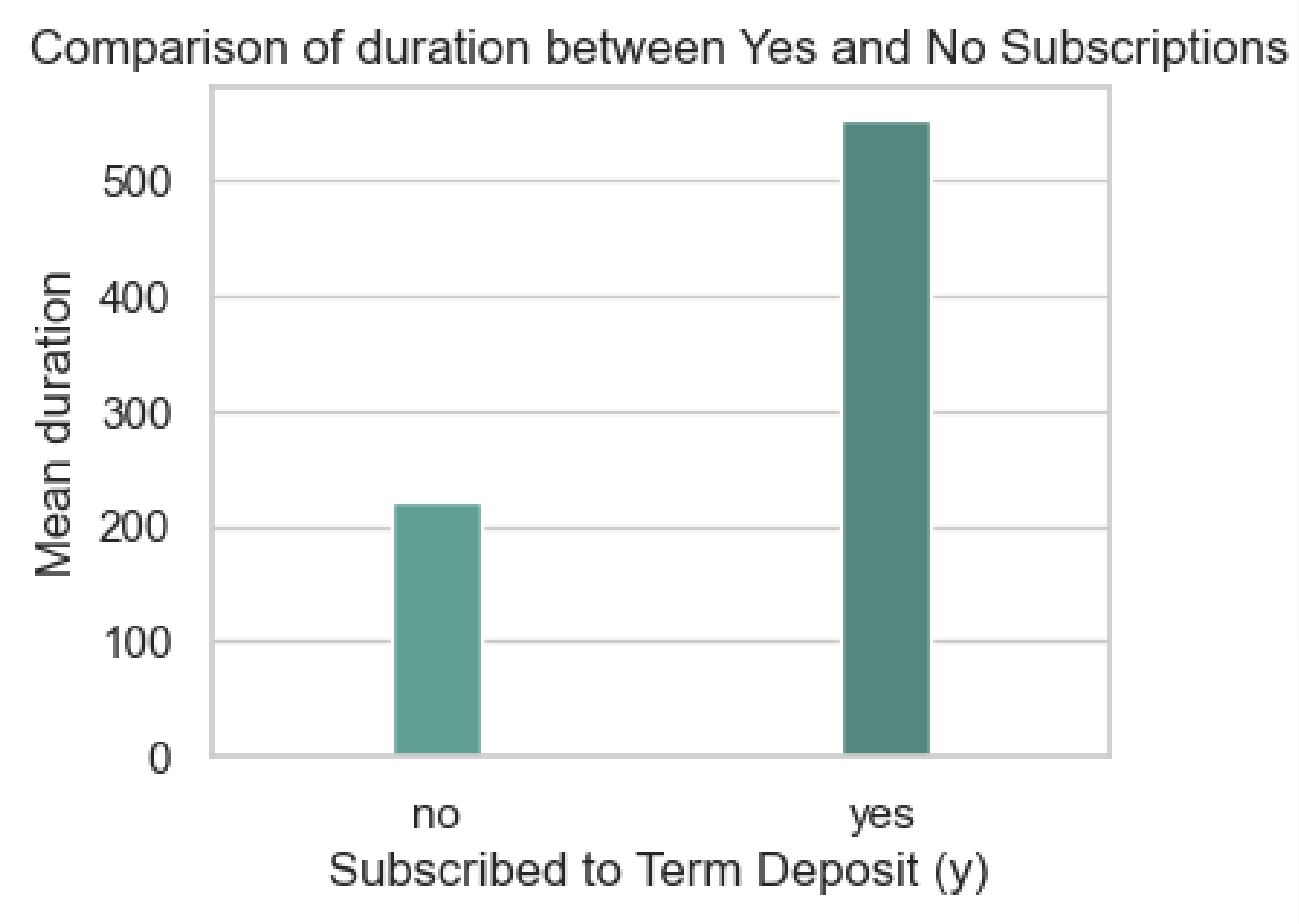
	duration	campaign	previous
count	41188.000000	41188.000000	41188.000000
mean	258.285010	2.567593	0.172963
std	259.279249	2.770014	0.494901
min	0.000000	1.000000	0.000000
50%	180.000000	2.000000	0.000000
max	4918.000000	56.000000	7.000000

DATA VISUALISATION



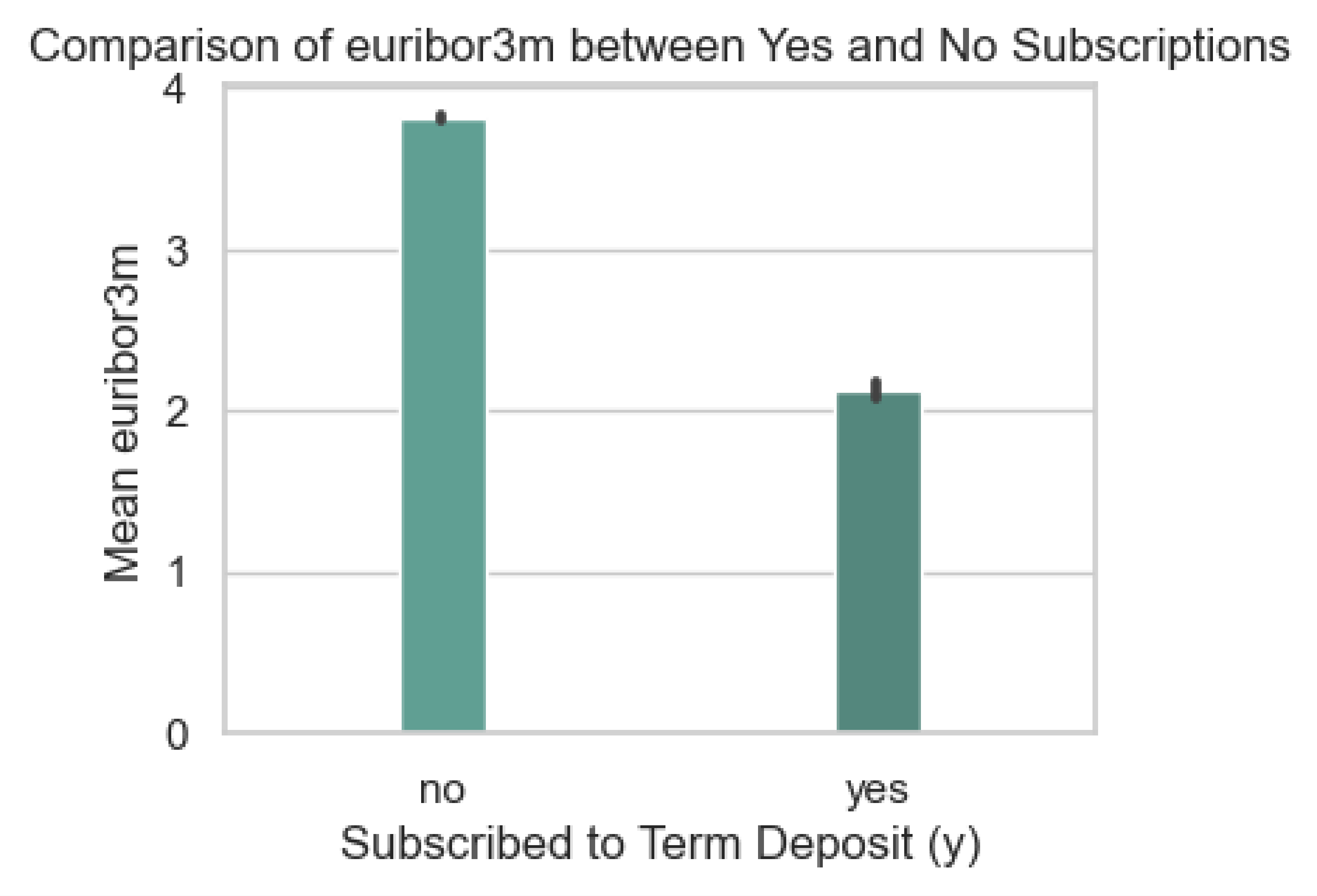
- Dataset is Biased towards ‘no’ outcome class

DATA VISUALISATION



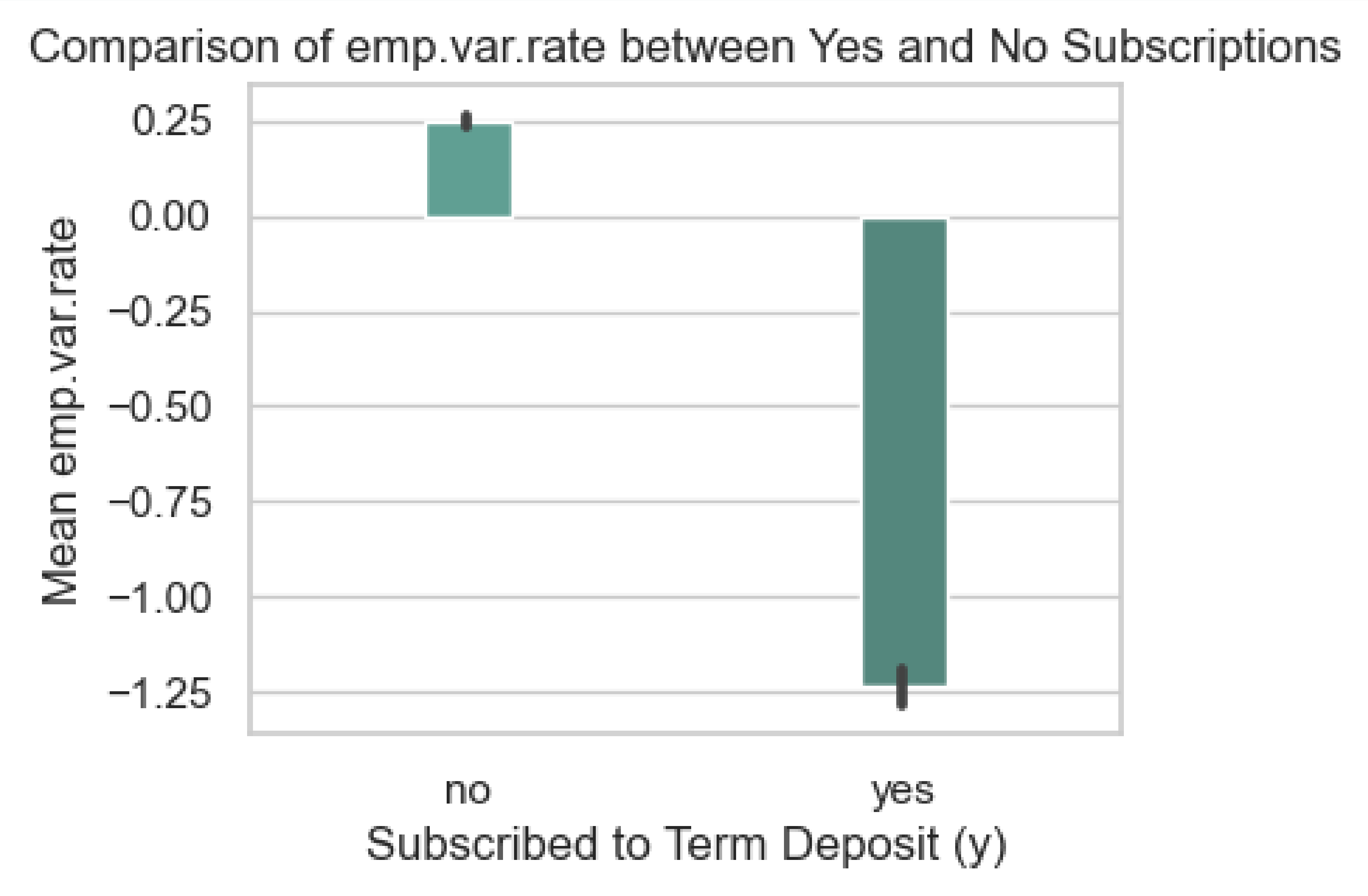
- If the duration is long, the outcome will be 'yes.' However, this information is only determined after the call, at which point the customer has already subscribed.

DATA VISUALISATION



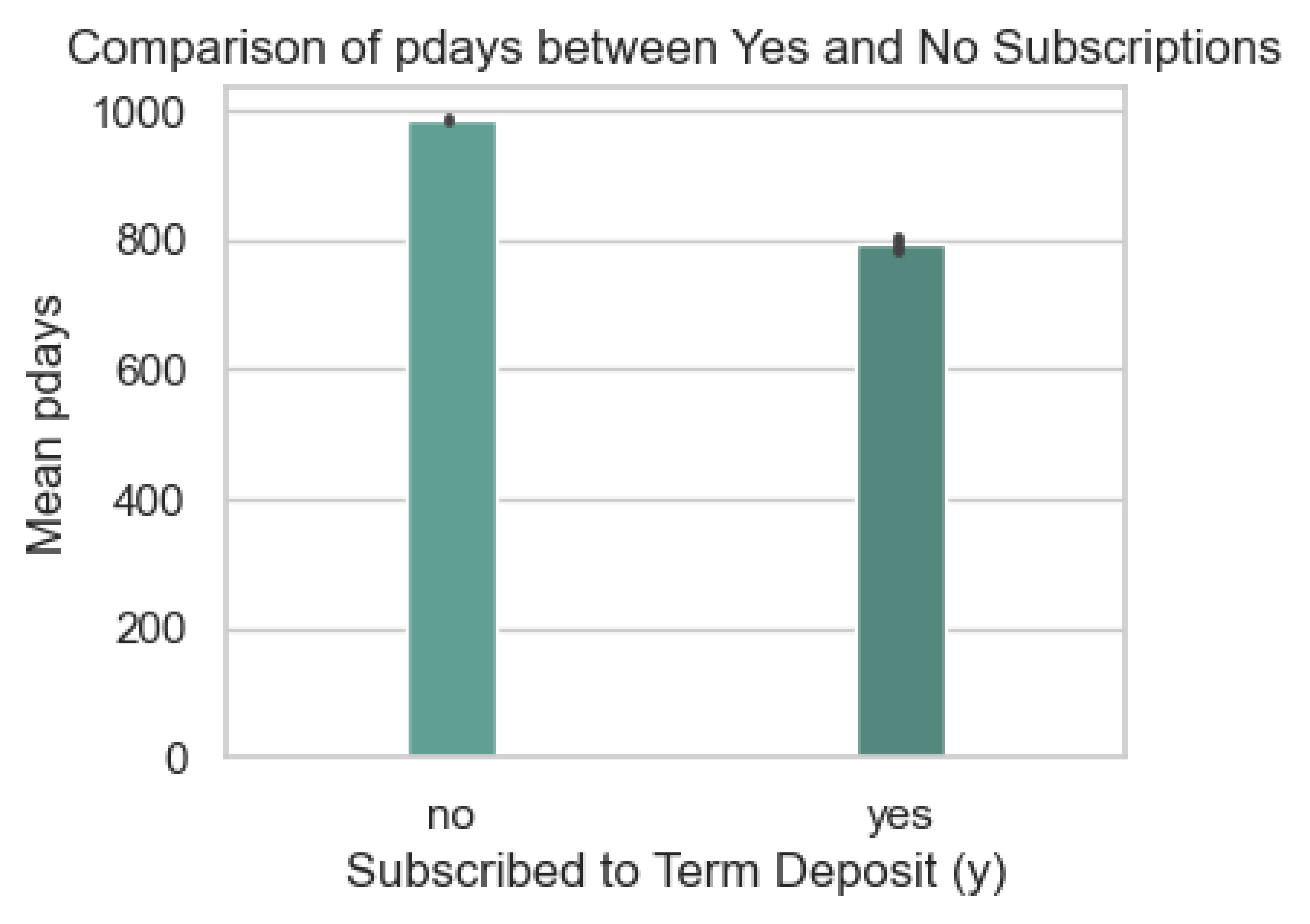
- Euribor interest rate for 3 months have a negative effect on subscription,

DATA VISUALISATION



- The employment variation rate also has a negative effect on subscription.

DATA VISUALISATION



- Lesser the number of days after last contact, subscription probability is high.

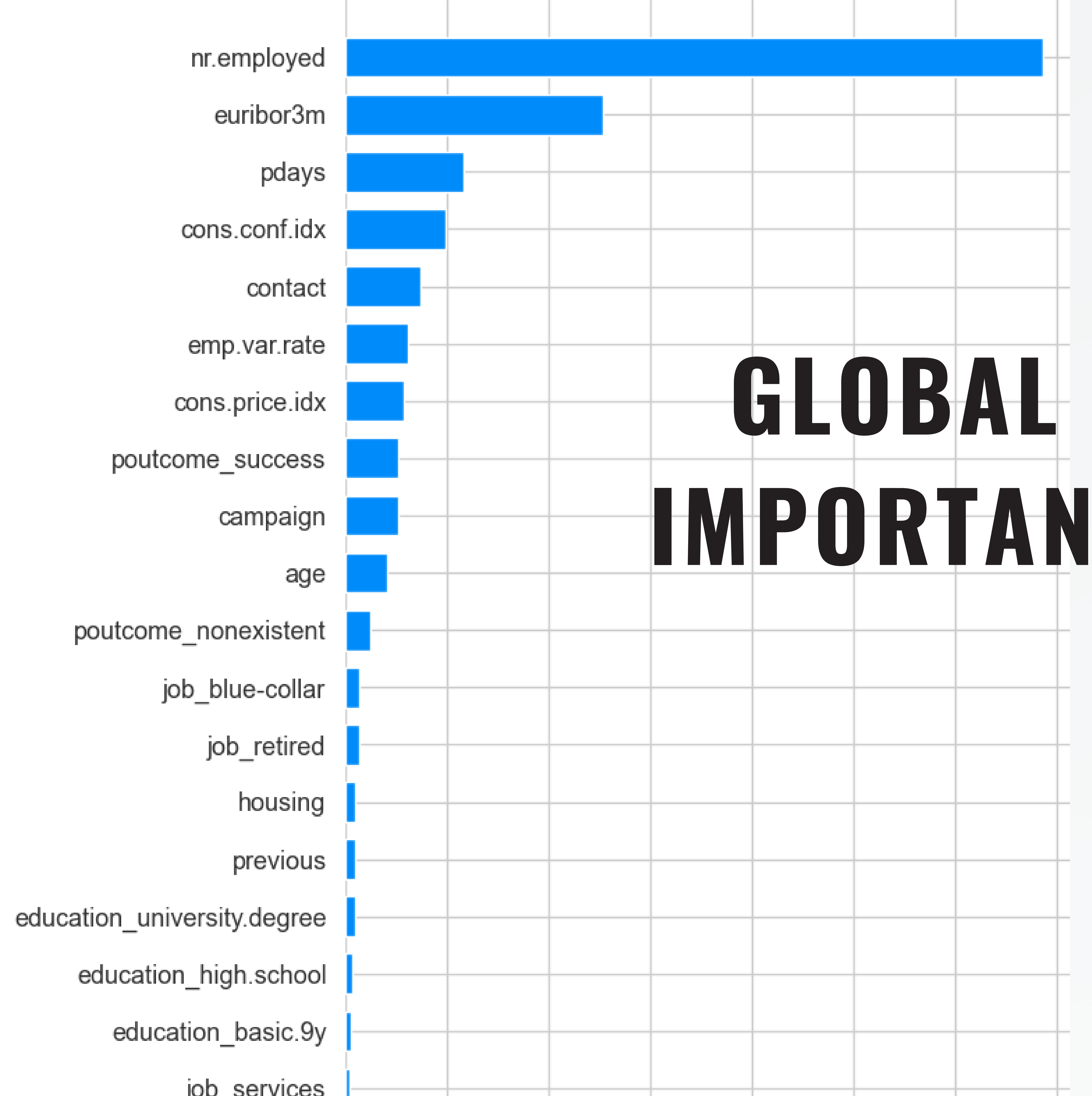
DATA PREPROCESSING

- Nulls were replaced with mode and mean values.
- Categorical fields except job and education, having unique values greater than 3 were discarded from the dataset.
- 'job','marital','poutcome','education'
- Columns 'job','marital','poutcome', and 'education' were one hot encoded and 'default','housing','loan', and 'contact' were label encoded.
- Numerical columns were scaled using Standard Scaler.
- Column ‘duration’ was dropped from the dataset since it is an information we derive only along with target variable ‘y’.
- **As the dataset exhibits bias towards one outcome class, it has been resized to ensure a balanced representation of both outcome classes, with each class contributing 50% to the reduced dataset length.**

MODEL BUILDING

- Employed logistic regression and decision tree classifier models for initial analysis.
- Utilized ensemble models including gradient boosting, bagging, and random forest to enhance predictive capabilities.
- Conducted hyperparameter tuning specifically on the gradient boosting classifier to optimize for superior metric values.
- Aligned our model selection with the strategic goal of resource efficiency in marketing efforts.
- Given our objective of retaining potential customers while efficiently allocating resources, we aim for a model with a high F1 score. This metric balances the trade-off between minimizing false negatives (missing potential subscribers) and false positives (allocating resources to unlikely subscribers).

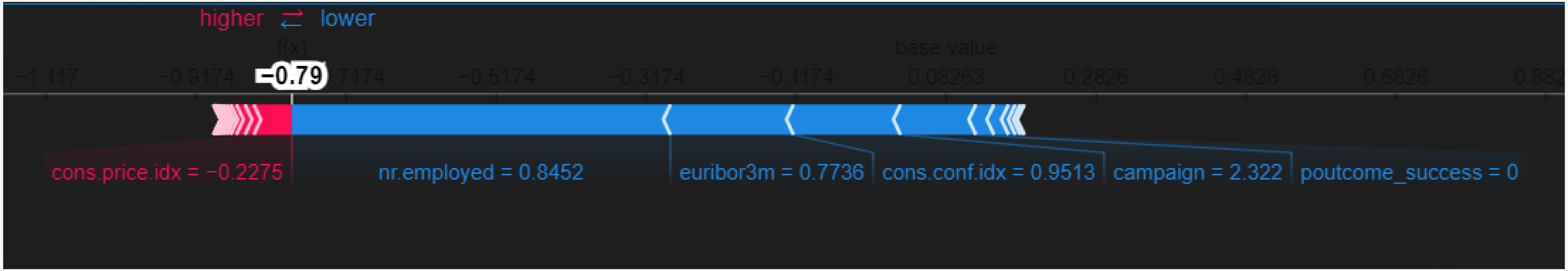
GLOBAL FEATURE IMPORTANCE (SHAP)



Features such as **‘nr.employed’,’euribor3m’,’pdays’,’cons.conf.idx’,’contact’,’emp.var.rate’, and ’cons.price.idx’** influences the outcome most

LOCAL FEATURE IMPORTANCE (SHAP)

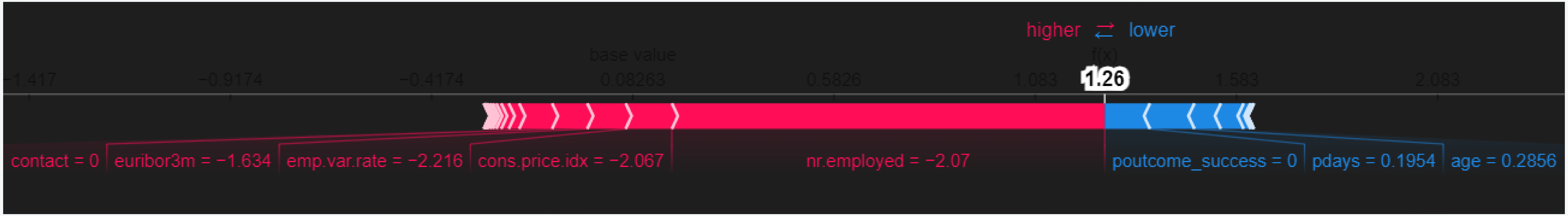
# Observation 4



In this observation, it is noteworthy that the features 'nr.employed' and 'euribor3m' contribute the most to the predicted value.

LOCAL FEATURE IMPORTANCE (SHAP)

# Observation 20



Here also ‘nr.employed’ contributes the most to the predicted value.

MODEL EVALUATION

Model Name	Training Accuracy %	Testing Accuracy %	Precision %	Recall %	F1 Score
Logistic Regression	73	73	78	65	71
Decision Tree Classifier	74	73	86	54	66
Gradient Boosting	80	73	86	54	66
Bagging	95	73	78	64	70
Random Forest	99.6	72	76	65	70
Gradient Boosting Hypertuned	75	74	81	64	72

DECIDING THE MODEL

- Following extensive training of diverse models, the **hyper-tuned Gradient Boosting Classifier** emerged as the top performer, attaining an impressive **F1 score of 72**.
- We select this model with confidence, recognizing its robust performance aligns seamlessly with our precise objectives.

CONCLUSION AND INSIGHTS

- In the hyper-tuned model, 'nr.employed' (number of employees) emerges as the most influential feature, with a negative impact on the outcome—indicating higher employee numbers correlate with lower subscription probability.
- Following closely is 'euribor3m' (3-month Euribor interest rate), showcasing a negative effect where higher interest rates correspond to lower subscription likelihood.
- Other key contributors include 'pdays,' 'cons.conf.idx,' 'poutcome,' 'emp.var.rate,' 'cons.price.idx,' and 'contact,' encapsulating factors such as previous contacts, consumer confidence, price indices, employment variation, and contact type.
- Feature selection methods efficiently reduce the feature set from 20 to 10, streamlining the model for improved performance and simplicity.

FUTURE WORK

- Investigate the influence of dataset bias on model performance.
- Explore the potential benefits of utilizing the entire dataset for training.
- Assess how eliminating bias and utilizing the complete dataset can contribute to achieving an even higher F1 score.
- Consider strategies for refining the model in future iterations to optimize performance in unbiased scenarios.

THANK YOU

Farzeena P A