Business Case

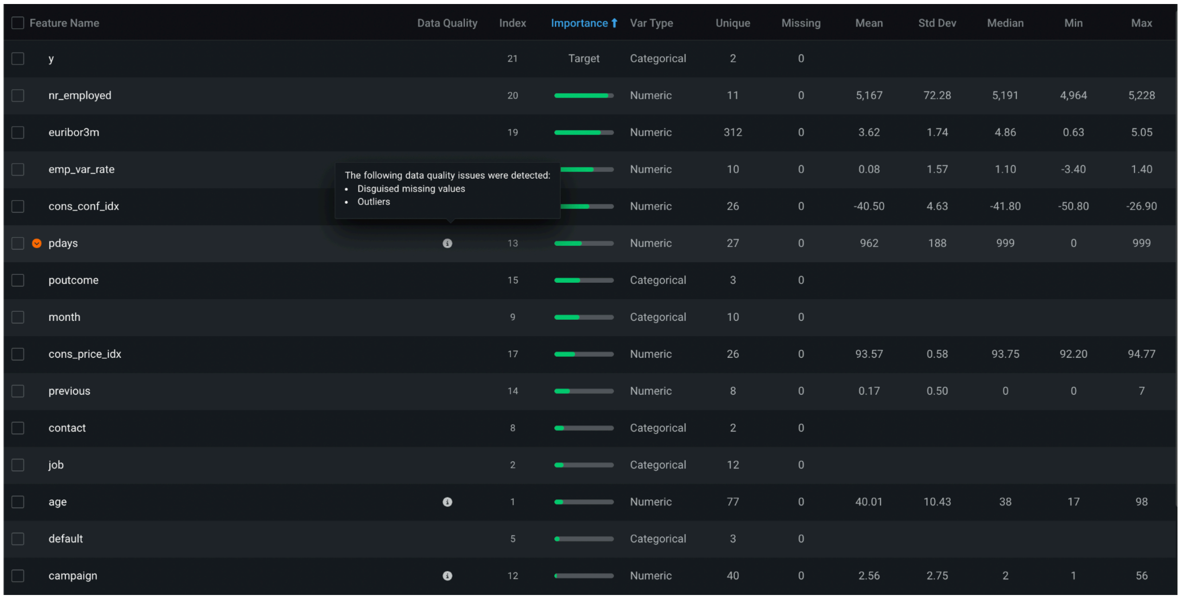
The analysis aims to predict client subscription to term deposits, a key component of the bank's strategic marketing objectives. By leveraging logistic regression and decision tree models, this initiative seeks to utilize extensive client data to forecast subscription outcomes effectively. The primary goal is to refine the bank's marketing campaigns by identifying potential subscribers with high accuracy and, as a result, optimizing resource allocation and enhancing the overall efficiency of customer engagement efforts. Evaluating model performance through metrics like Maximum Profit allows for a nuanced assessment of each campaign's financial impact. This metric is key in guiding the bank towards strategies that promise the highest returns on investment, concentrating marketing efforts on leads most likely to convert. The expectation is that a data-informed approach will not only streamline resource distribution but also significantly boost conversion rates for term deposit subscriptions.

Q1. Assess if any features are missing values

The dataset “bank-additional-full.csv” does not have any missing values across its features, indicating that it is complete with data for all observations. However, the feature “pday” was noted as having disguised missing values. A "disguised missing value" for this feature could be a specific number used to denote that the client was not previously contacted. For example, the dataset documentation might specify that a value of 999 (or another unlikely real value for this feature) indicates that the client has not been contacted before. This practice allows analysts and models to differentiate between clients who have never been contacted and those who have, even when the exact number of days since the last contact is not meaningful for the former group.

This could be treated by:

- Leaving them as-is.
- Replacing them with NaN or another indicator to explicitly mark them as missing.
- Creating a new categorical feature to indicate whether the client was previously contacted, thus preserving the information that the disguised value conveys.

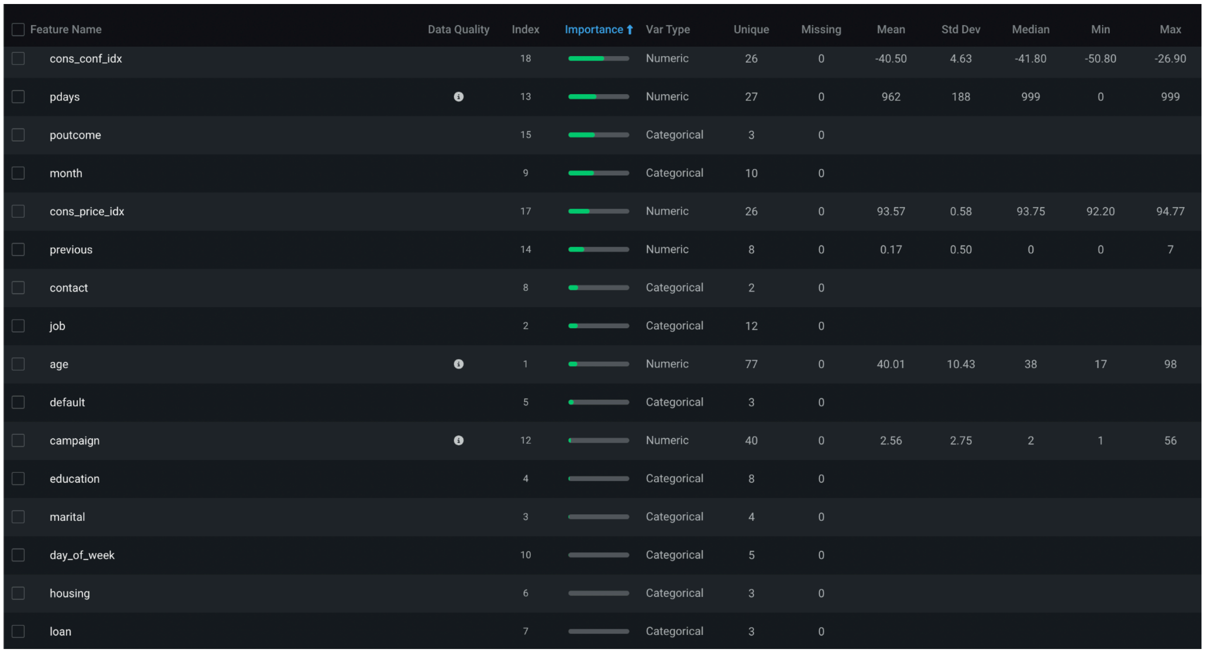


□ Feature Name	Data Quality	Index	Importance ↑	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
□ y		21	Target	Categorical	2	0					
□ nr_employed		20		Numeric	11	0	5,167	72.28	5,191	4,964	5,228
□ euribor3m		19		Numeric	312	0	3.62	1.74	4.86	0.63	5.05
□ emp_var_rate				Numeric	10	0	0.08	1.57	1.10	-3.40	1.40
□ cons_conf_idx				Numeric	26	0	-40.50	4.63	-41.80	-50.80	-26.90
□ pdays	ⓘ	13		Numeric	27	0	962	188	999	0	999
□ poutcome		15		Categorical	3	0					
□ month		9		Categorical	10	0					
□ cons_price_idx		17		Numeric	26	0	93.57	0.58	93.75	92.20	94.77
□ previous		14		Numeric	8	0	0.17	0.50	0	0	7
□ contact		8		Categorical	2	0					
□ job		2		Categorical	12	0					
□ age	ⓘ	1		Numeric	77	0	40.01	10.43	38	17	98
□ default		5		Categorical	3	0					
□ campaign	ⓘ	12		Numeric	40	0	2.56	2.75	2	1	56

The following data quality issues were detected:
- Disguised missing values
- Outliers

Q2. Assess if any features have no variance

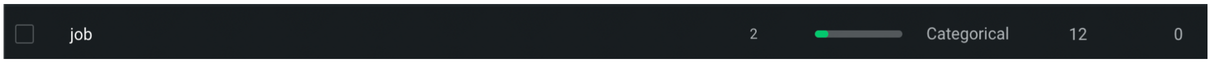
There are no features in the dataset with zero variance, meaning all numeric features have variation in their values across observations. This indicates that each numeric feature provides some level of information that could potentially contribute to the predictive modeling process, as there are no constant features that would need to be removed for lack of variability.



□	Feature Name	Data Quality	Index	Importance ↑	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
□	cons_conf_idx		18		Numeric	26	0	-40.50	4.63	-41.80	-50.80	-26.90
□	pdays	ⓘ	13		Numeric	27	0	962	188	999	0	999
□	poutcome		15		Categorical	3	0					
□	month		9		Categorical	10	0					
□	cons_price_idx		17		Numeric	26	0	93.57	0.58	93.75	92.20	94.77
□	previous		14		Numeric	8	0	0.17	0.50	0	0	7
□	contact		8		Categorical	2	0					
□	job		2		Categorical	12	0					
□	age	ⓘ	1		Numeric	77	0	40.01	10.43	38	17	98
□	default		5		Categorical	3	0					
□	campaign	ⓘ	12		Numeric	40	0	2.56	2.75	2	1	56
□	education		4		Categorical	8	0					
□	marital		3		Categorical	4	0					
□	day_of_week		10		Categorical	5	0					
□	housing		6		Categorical	3	0					
□	loan		7		Categorical	3	0					

Q3. Assess if any categorical features have high cardinality

The categorical feature job has high cardinality with 12 unique values, indicating a relatively wide range of categories within this feature. High cardinality in this categorical feature can potentially pose challenges for modeling, particularly for algorithms that rely on one-hot encoding, as it may lead to a large increase in the dimensionality of the dataset.

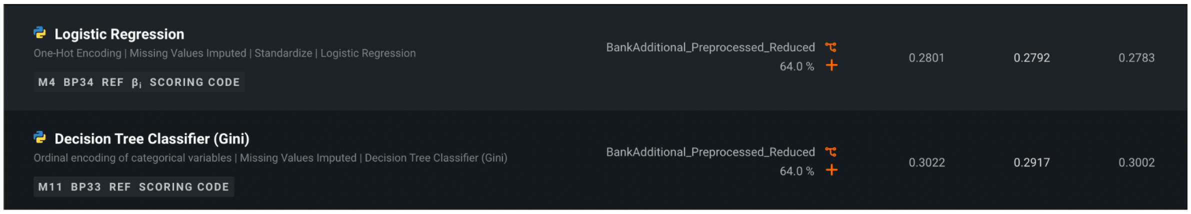


□	job	2		Categorical	12	0

Q4. Develop logistic regression and decision tree models for marketing campaign response.

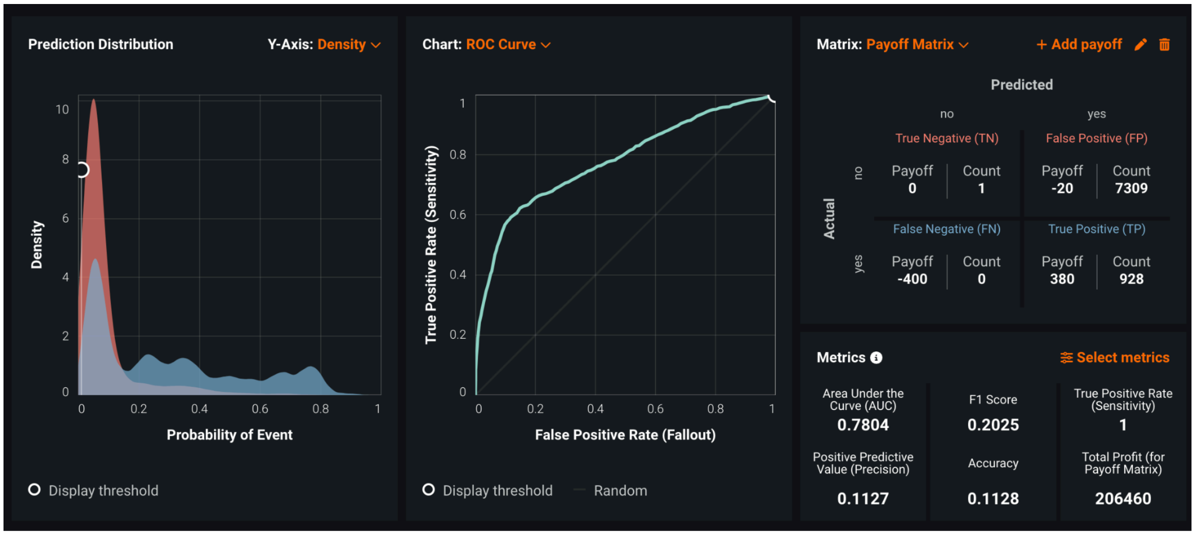
All features in the dataset were included, with the exception of duration. Excluding the "duration" variable from logistic regression and decision tree models, particularly in the context of predicting outcomes such as term deposit subscriptions in banking, is crucial primarily due to its disproportionately high impact on the prediction.

- Categorical Variables: y, outcome, month, contact, job, campaign, default, education, marital, day_of_week, housing, and loan
- Numeric Variables: nr_employed, euribor3m, emp_var_rate, cons_conf_indx, pdays, cons_price_idx, previous, age, campaign



Logistic Regression One-Hot Encoding | Missing Values Imputed | Standardize | Logistic Regression M4 BP34 REF βᵢ SCORING CODE	BankAdditional_Preprocessed_Reduced 64.0 %	0.2801	0.2792	0.2783
Decision Tree Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | Decision Tree Classifier (Gini) M11 BP33 REF SCORING CODE	BankAdditional_Preprocessed_Reduced 64.0 %	0.3022	0.2917	0.3002

Logistic Regression

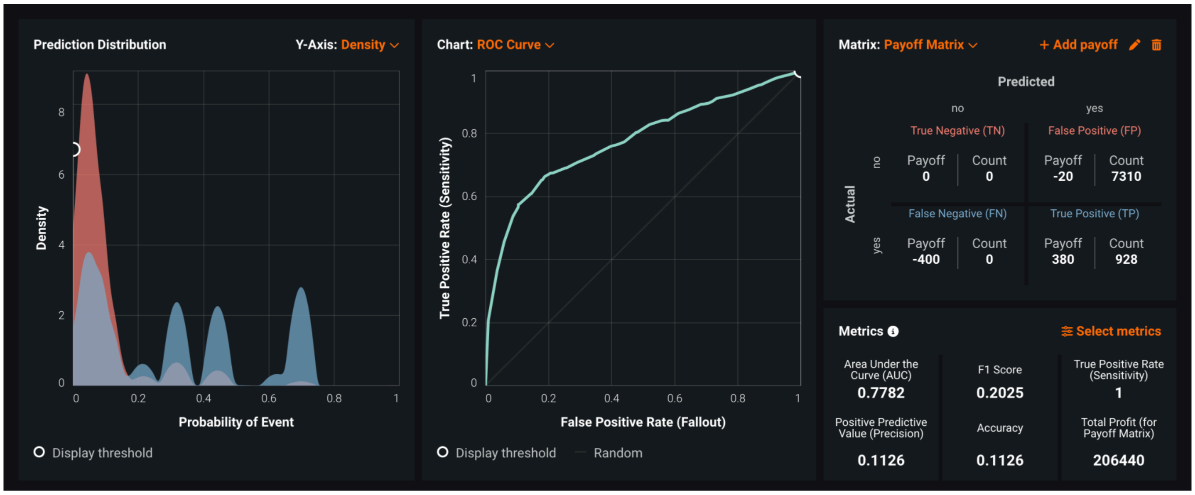


Predicted

		no	yes
Actual	no	True Negative (TN) Payoff 0 | Count 1	False Positive (FP) Payoff -20 | Count 7309
	yes	False Negative (FN) Payoff -400 | Count 0	True Positive (TP) Payoff 380 | Count 928

Area Under the Curve (AUC) 0.7804	F1 Score 0.2025	True Positive Rate (Sensitivity) 1
Positive Predictive Value (Precision) 0.1127	Accuracy 0.1128	Total Profit (for Payoff Matrix) 206460

Decision Tree



Predicted

		no	yes
Actual	no	True Negative (TN) Payoff 0 | Count 0	False Positive (FP) Payoff -20 | Count 7310
	yes	False Negative (FN) Payoff -400 | Count 0	True Positive (TP) Payoff 380 | Count 928

Area Under the Curve (AUC) 0.7782	F1 Score 0.2025	True Positive Rate (Sensitivity) 1
Positive Predictive Value (Precision) 0.1126	Accuracy 0.1126	Total Profit (for Payoff Matrix) 206440

Q5. Report recall, precision, F1, accuracy, ROC AUC, maximum payoff for each of the models. Explicate your payoff matrix and the underlying assumptions.

Financial Assumptions:

- CLV Calculation: The CLV for a deposit would include the net interest margin the bank earns over a certain period. Assuming an average deposit amount of $20,000 and a net interest rate of 4%, we can calculate the annual profit from one customer's deposit over the duration of a year. Annual Profit = $20,000 × 4% = $800

- Banks profit margin on interest can be assumed to be 50%. This means the actual profit the bank earns from the interest income, after covering its costs related to the deposit (e.g., interest paid to customers, operational costs), is 50% of the interest income. Bank's Profit from Interest per Customer: $800 × 50% = $400
- Cost of outreach per phone call is assumed to be $20 per phone call, which encompasses the marketing costs.

Payoff Matrix:

- True Positive (TP): Represents acquiring a new customer who makes a deposit as a result of the marketing campaign. Profit from a TP: $400 (bank's profit from interest per customer after assuming a 50% profit margin on the $800 interest income) - $20 (cost of outreach per phone call).
    - Hypothetical Value: +$380
- True Negative (TN): Represents correctly identifying a customer who would not have made a deposit, hence saving the cost of outreach.
    - Hypothetical Value: $0 (since there's no direct profit from a TN, but there's a cost saving from not making an unnecessary outreach call).
- False Positive (FP): Represents the cost of outreach to customers who do not make a deposit.
    - Hypothetical Value: -$20 (the cost of the outreach call that did not result in a deposit).
- False Negative (FN): Represents missing out on a customer who would have made a deposit if they had been contacted. Since the FN does not incur the outreach cost but represents a lost profit opportunity, the lost profit is equivalent to the bank's profit from interest per customer.
    - Hypothetical Value for FN: -$400 (reflecting the lost profit opportunity from not acquiring the deposit).

Explaining Metrics:

- Maximum Payoff: Appropriate if a business case can be reduced to evaluation of financial outcomes.
- ROC AUC: Appropriate for comparing model performance because it does not depend on the choice threshold.
- Recall/Precision/Specificity/Accuracy: Once you decide what’s the best model, assess these metrics and think about implications of model performance in relation to the business case you’re trying to solve.

	Logistic Regression (Maximizing Profit)	**Decision Tree (Maximizing Profit)**
Recall	1	1

Precision	0.1127	0.1126
F1	0.2025	0.2025
Accuracy	0.1128	0.1126
Error	0.8872	0.8874
ROC AUC	0.7804	0.7782
Maximum Payoff	206,460	206,440
Threshold	0.0104	0

Q6. What is the best metric to evaluate model performance and why? Which is the better model?

Maximum Payoff is the best metric to evaluate the model because it directly correlates with the financial outcome of the marketing campaign. It translates model performance into tangible business value, reflecting the net financial gain or loss resulting from the model's predictions. In scenarios where financial impact is more important than anything else, such as in marketing campaigns aiming to maximize profit or minimize loss, Maximum Payoff provides a clear measure of success. It incorporates the cost-benefit analysis of different outcomes (TP, FP, TN, FN), making it a comprehensive metric that accounts for both the effectiveness of identifying potential customers (reflected in Recall, Precision) and the cost implications of the model's decisions.

Model Comparison Based on Maximum Payoff:

- Logistic Regression: Maximum Payoff of $206,460
- Decision Tree: Maximum Payoff of $206,440

While both models exhibit identical Recall, very similar Precision, F1, and Accuracy scores, and comparable ROC AUC scores, the Logistic Regression model achieves a slightly higher Maximum Payoff than the Decision Tree model. The difference in Maximum Payoff is minimal ($20), yet in a scenario where maximizing financial outcomes is the priority, even small differences can be significant.

So which model is better?

- The Logistic Regression model is the better model in this context, given its marginally higher Maximum Payoff. This suggests that, for this specific marketing campaign and under the provided conditions, Logistic Regression is slightly more efficient at generating profit than the Decision Tree model.
- The superiority of the Logistic Regression model here is based purely on the financial metric of Maximum Payoff. However, it's important to note that the choice of "best" model can vary depending on other factors, such as interpretability, scalability, or specific business objectives beyond the immediate financial outcome.

- The ROC AUC score, which is slightly higher for the Logistic Regression model, supports its selection by indicating a marginally better ability to discriminate between positive and negative classes across different thresholds.
- Despite the similarity in other metrics, when the primary goal is to maximize profit, the model with the highest Maximum Payoff should be chosen, reinforcing the decision to favor the Logistic Regression model in this scenario.