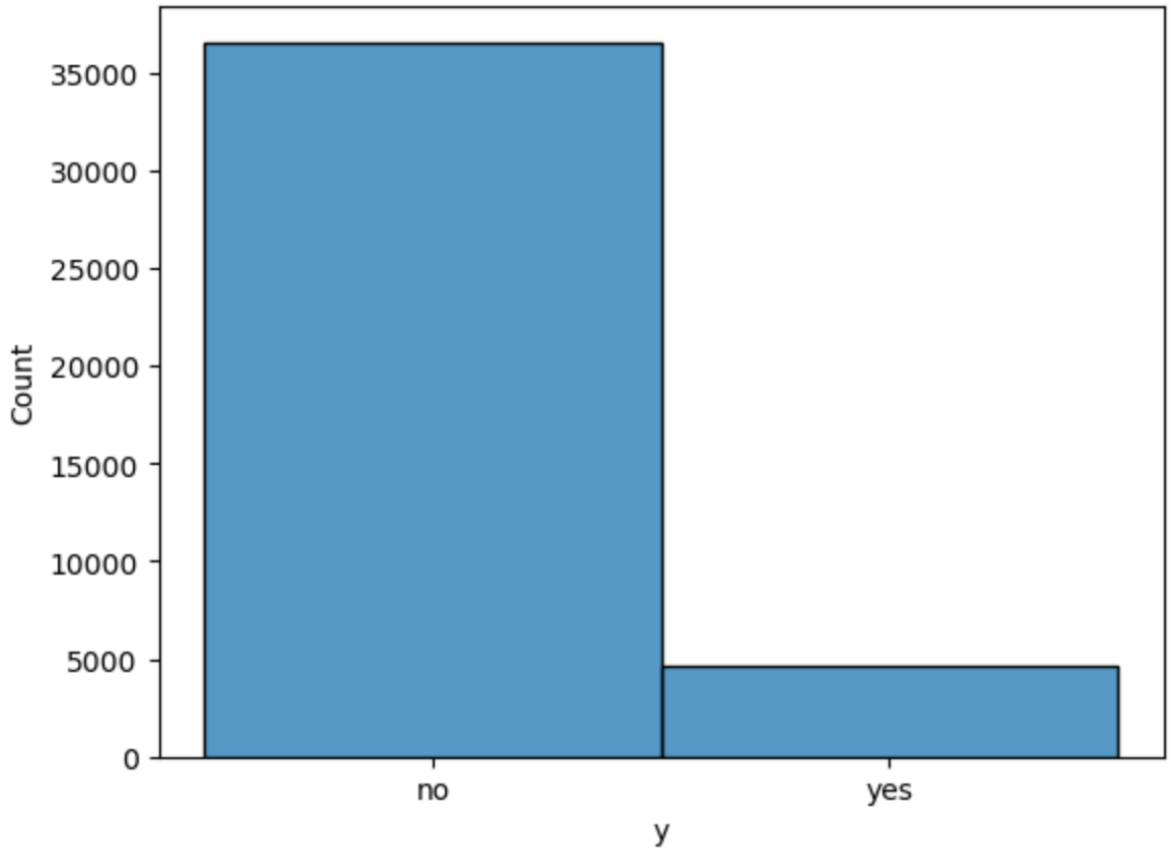
Practical-Application-3

Comparing Classifiers (Banking Data)

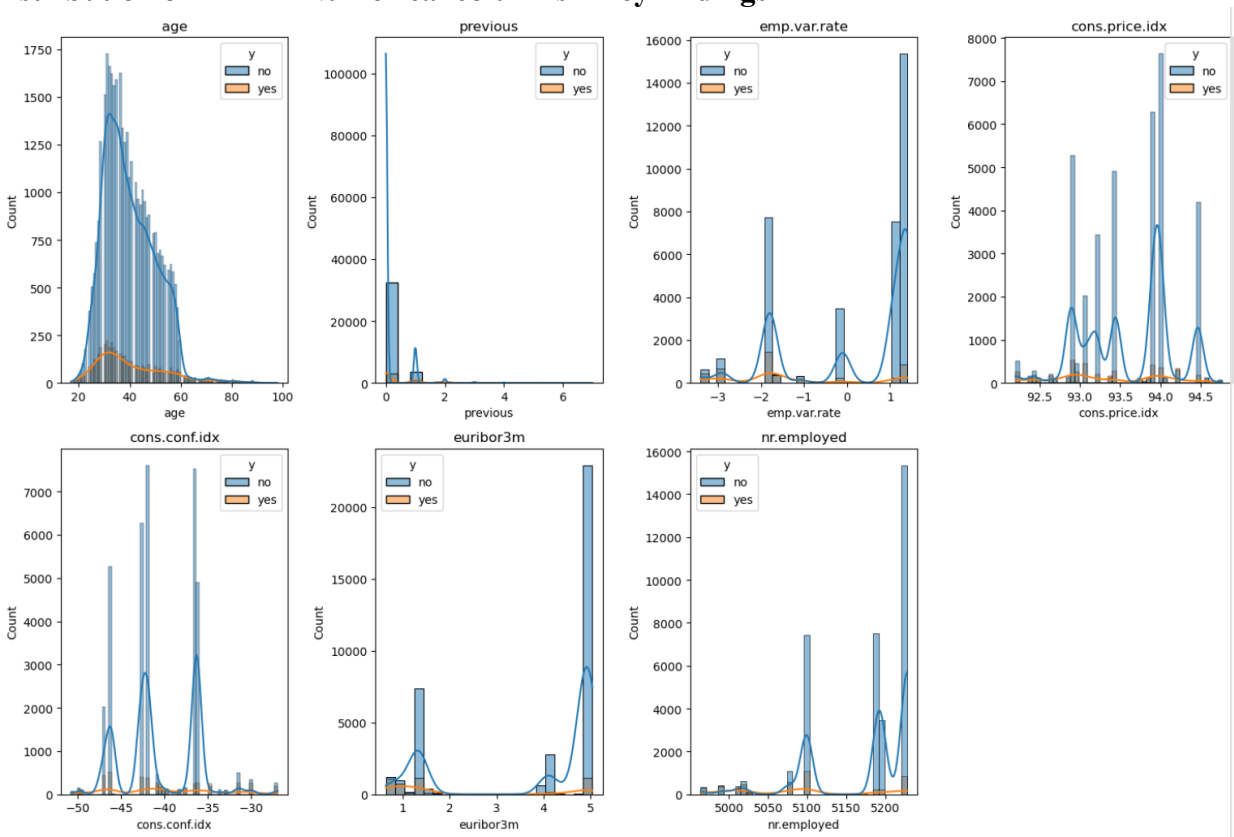
Harish Laxmi Narasimha Venugopal

Distribution of Target Variable “y” subscription - Key findings



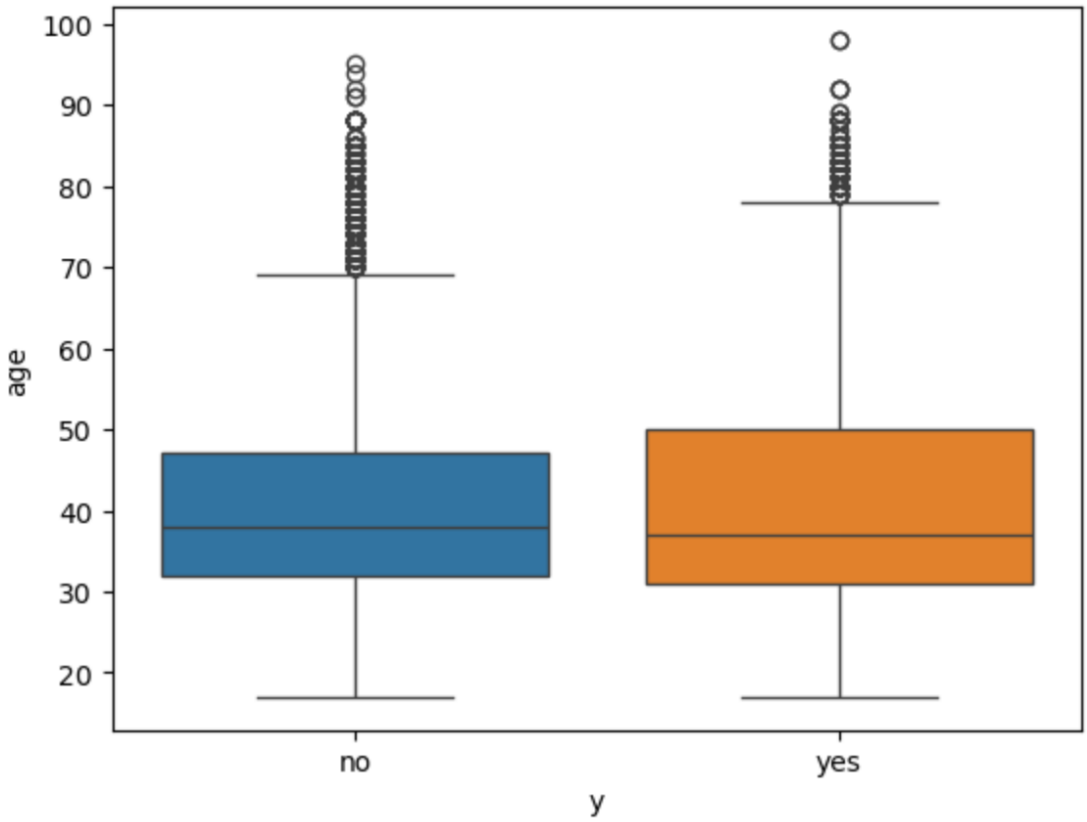
- The target variable (y) is imbalanced, meaning more customers did not subscribe to the term deposit than those who did.
- Data set is imbalanced. Majority of the clients (~88%) didnt subscribe (y = "no") to the term deposit. Using SMOTE technique can help here.

Distribution of “ALL” Numerical columns - Key findings



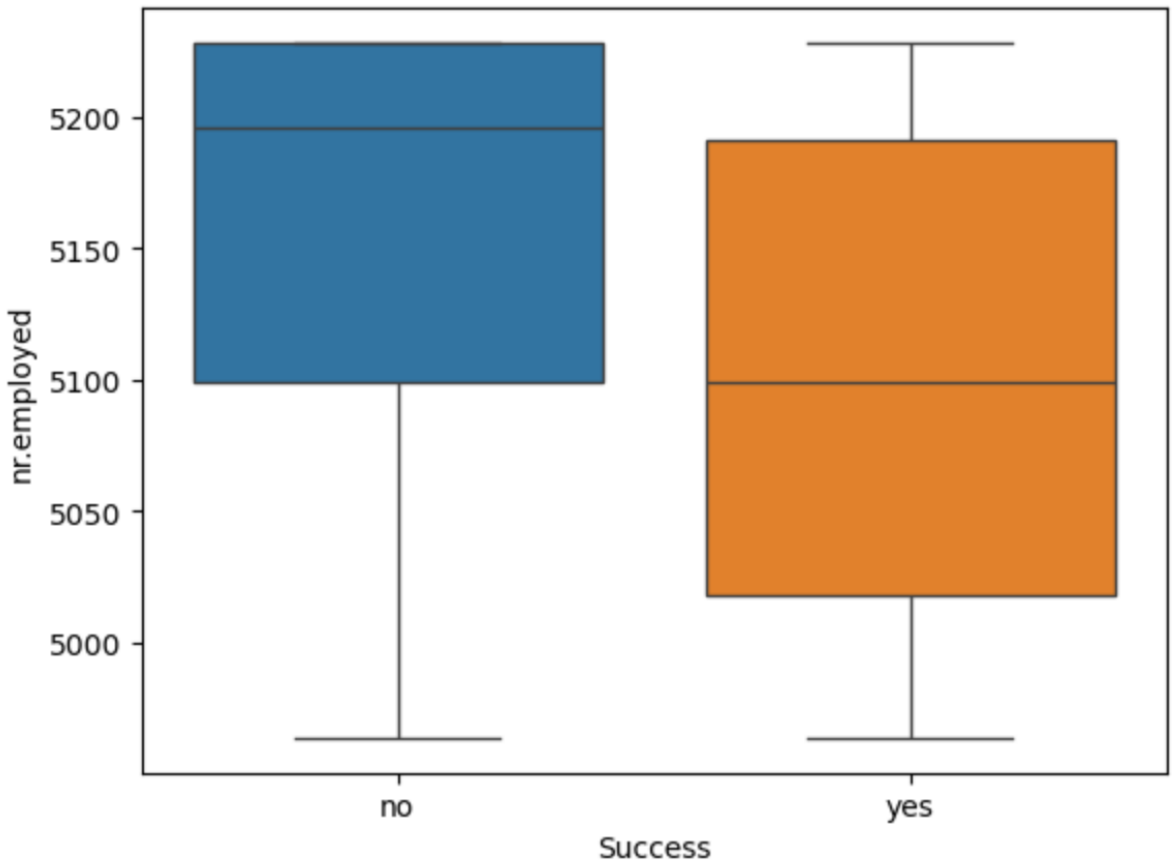
- Age: The distribution of ages is skewed to the right. A larger proportion of younger people (20-40) are not subscribed while the likelihood increases at age 40 and decreases after.
- Previous: Most clients have not been contacted in a previous campaign (0). Subscriptions appear to be slightly higher among those with "previous" campaign contact, but the vast majority are clustered at 0.
- emp.var.rate (Employment Variation Rate): The variable has modes at roughly -3, -2, 0, and +1. Subscriptions appear higher when the employment variation rate is at -3 and decreases when it is at +1.
- cons.price.idx (Consumer Price Index): This variable shows several clusters. Subscriptions seem very marginally higher when cons.price.idx is lower.
- cons.conf.idx (Consumer Confidence Index): This variable is clustered, with the highest counts at more negative values. Subscriptions appear slightly more prevalent when the consumer confidence index is less negative.
- euribor3m (Euribor 3 Month Rate): The distribution is concentrated at higher rates (around 4-5). Subscriptions are noticeably higher when the euribor3m rate is lower (around 1).
- nr.employed (Number of Employees): This has distinct clusters. Subscriptions seem to be somewhat higher when the number of employees is lower (around 5000-5100).

Compare Age Column with Target Variable “y” subscription



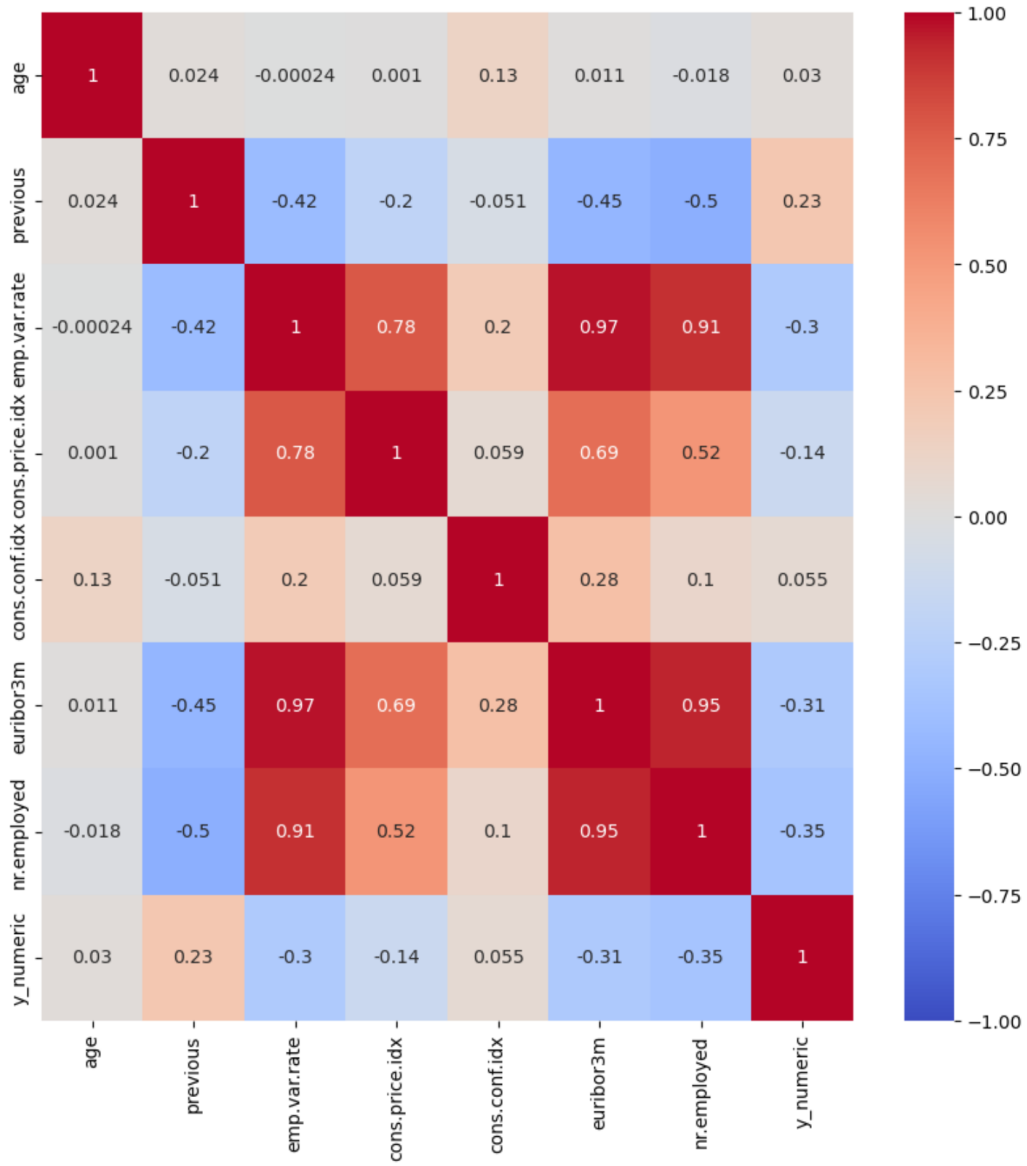
- Median Age: The median age of those who subscribed ("yes") appears to be slightly higher than those who did not subscribe ("no").
- Age Distribution: The age distribution is fairly similar between the two groups, however, there are outliers.
- Outliers: Both groups have a large number of outliers on the higher end of the age range. This indicates that there are many older individuals in the dataset, and their age doesn't necessarily preclude them from either subscribing or not subscribing.
- Range: The box representing those who did subscribe is overall slightly higher than the box for those who did not subscribe.

Compare nr.employed Column with Target Variable “y” subscription



- Median nr.employed: The median number of employees is higher for the group that did not subscribe ("no") compared to the group that did subscribe ("yes").
- Distribution: The distribution of "nr.employed" is noticeably different between the two groups.
- Quartiles: The entire boxplot for the "no" group sits higher on the y-axis than the "yes" group. This means that, in general, the number of employees tends to be higher for those who did not subscribe.
- Range: The ranges also differ somewhat.
- A lower number of employees ("nr.employed") seems to be correlated with a higher likelihood of subscription ("yes"). A higher number of employees ("nr.employed") seems to be correlated with a lower likelihood of subscription ("no").

Correlation between Numerical columns



Positive Correlation

- emp.var.rate and euribor3m: A very strong positive correlation (0.97). This makes sense, as both are indicators of the economic climate. When employment variation rates are high, interest rates tend to be high as well.
- emp.var.rate and nr.employed: Also a strong positive correlation (0.91). As employment rates increase, so does the number of people employed.
- euribor3m and nr.employed: Strong positive correlation (0.95). This is also logical, as both are related to the overall economic situation.
- emp.var.rate and cons.price.idx: Positive correlation (0.78). This suggests that as employment variation rates increase, so do consumer prices.
- euribor3m and cons.price.idx: Positive correlation (0.69). Similarly, as interest rates increase, so do consumer prices.

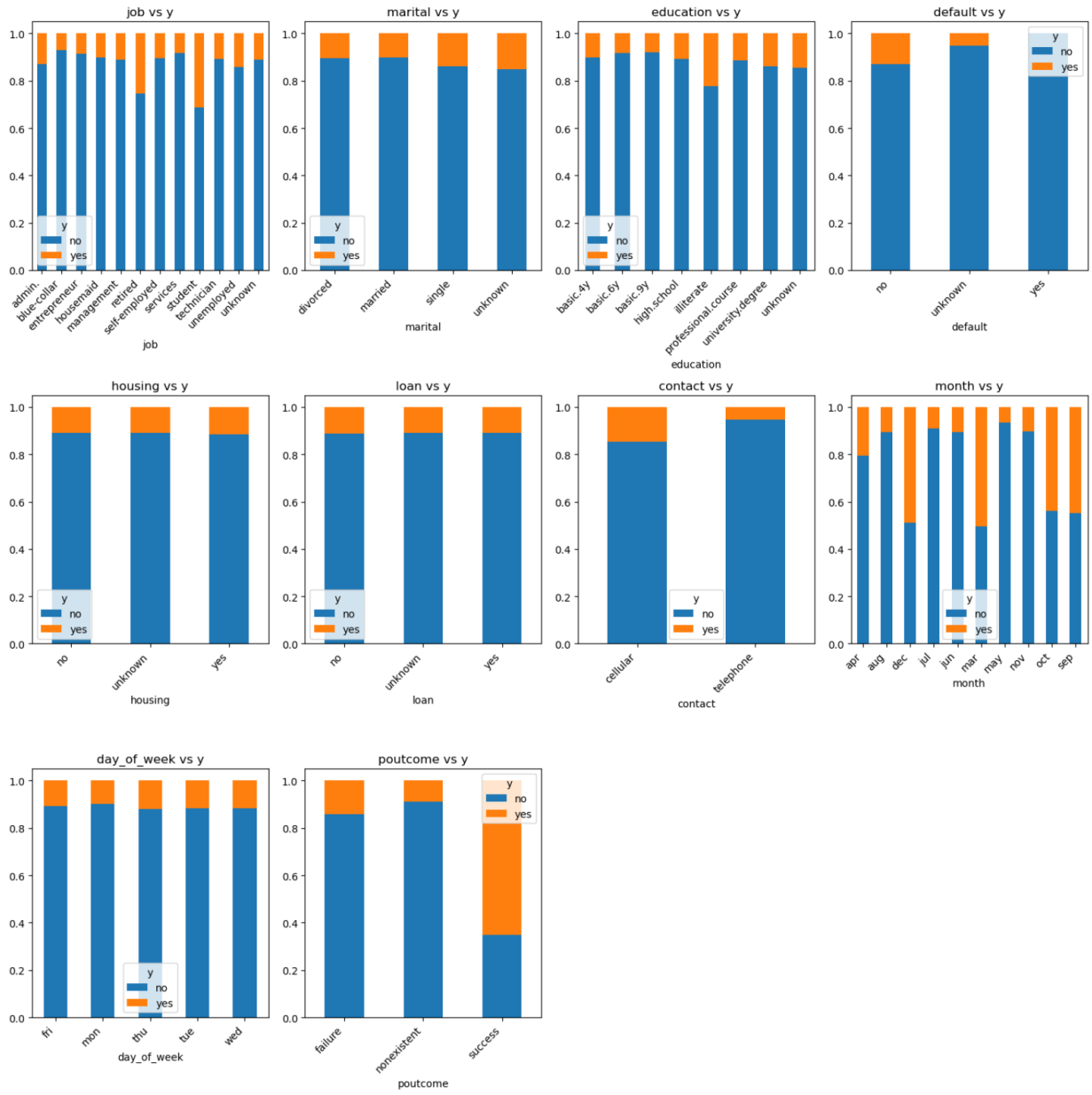
Negative Correlations

- emp.var.rate and previous: Moderate negative correlation (-0.42). This might indicate that the more contacts in previous marketing campaigns, the lower the employment variation rate. This is counterintuitive and may be more related to the timing of the campaigns and general economic trends.
- euribor3m and previous: Moderate negative correlation (-0.45). Similar to the emp.var.rate, this suggests that as interest rates increase, the effectiveness of previous campaigns decreases.
- nr.employed and previous: Moderate negative correlation (-0.50). The more people employed the less effective prior campaigns were.
- emp.var.rate and y_numeric: Weak negative correlation (-0.3). Higher employment rate correlates with less likelihood of subscribing.
- euribor3m and y_numeric: Weak negative correlation (-0.31). High interest rate correlated with less likelihood of subscribing.
- nr.employed and y_numeric: Weak negative correlation (-0.35). Higher employment correlated with less likelihood of subscribing.

Weak Correlation

- age has very weak correlations with all variables.
- cons.conf.idx has weak correlations with all variables.
- cons.price.idx has weak correlations with all variables.

Relationship between categorical features and "y" subscription



1. job vs y:
   - Students & Retired: These groups show a significantly higher proportion of "yes" subscriptions compared to other job categories.
   - Blue-collar: This group has one of the lowest subscription rates.

   - Actionable Insight: Tailor marketing messages specifically to students and retirees, highlighting the benefits relevant to their stage of life. Avoid a one-size-fits-all approach.
2. marital vs y:
   - Single: Shows a higher proportion of "yes" subscriptions compared to married or divorced individuals.
   - Married & Divorced: Have relatively similar subscription rates.

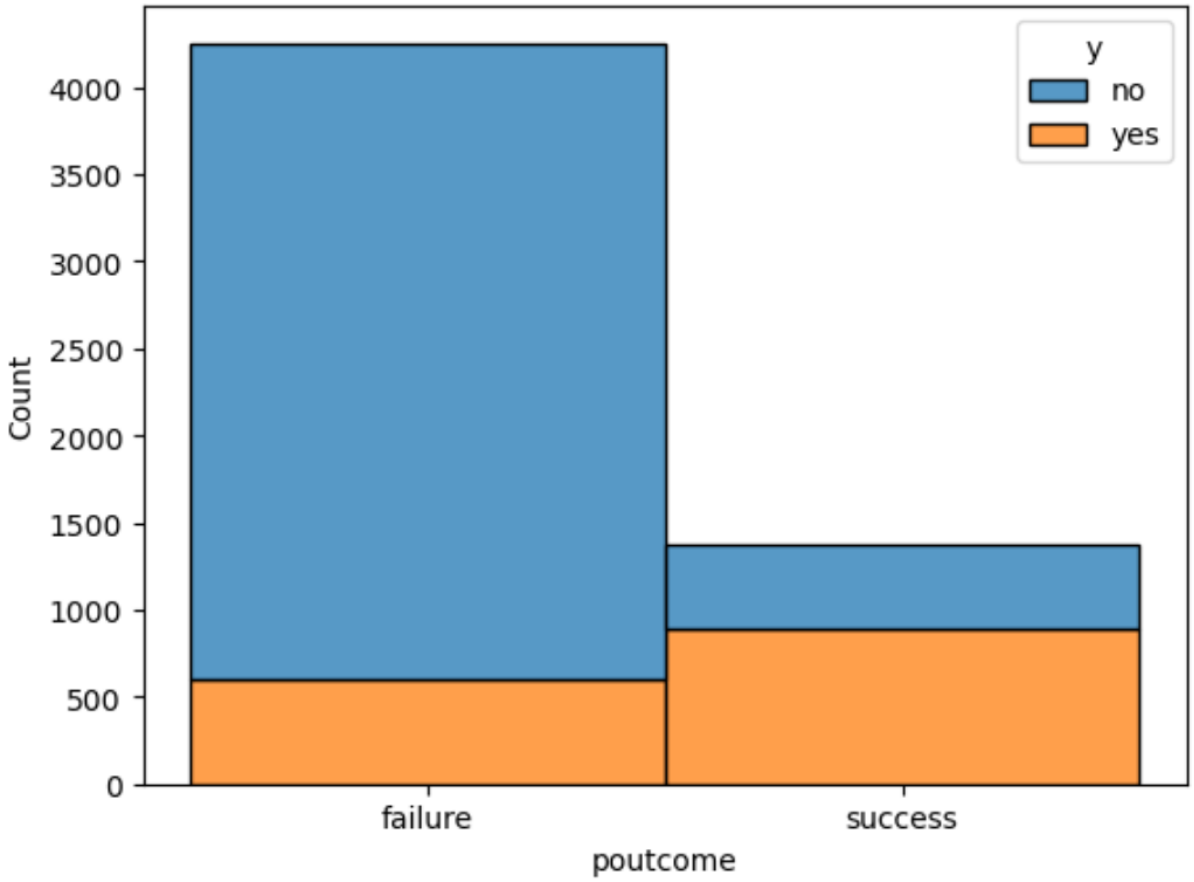
- Actionable Insight: Similar to the job category, tailor messaging. Single individuals might be more receptive to certain aspects of a term deposit (e.g., saving for a future goal), while married couples might have different priorities.

3. education vs y
   - Illiterate: This tiny group appears to have a high conversion rate (though the sample size is likely very small, so it's not reliable).
   - Basic.4y, Basic.6y, Basic.9y: These education levels have relatively lower "yes" subscription rates.
   - University.degree: Shows a higher proportion of "yes" subscriptions compared to those with basic education.
   - Actionable Insight: Explore whether more detailed education-specific messaging would be beneficial. Those with university degrees may respond to different arguments than those with only basic education
4. default vs y:
   - No (No credit in default): The vast majority fall into this category, and the "yes" subscription rate is relatively low.
   - Yes (Has credit in default): This very small group appears to have slightly higher subscription rates, however this is likely due to the small sample size.
   - Actionable Insight: Clients with no credit in default have a much lower subscription rate.
5. housing vs y
   - Yes (Has housing loan): A higher proportion of "no" subscriptions.
   - No (Does not have housing loan): The "yes" subscription is slightly higher.
   - Actionable Insight: Clients that do not have a housing loan have a higher subscription rate.

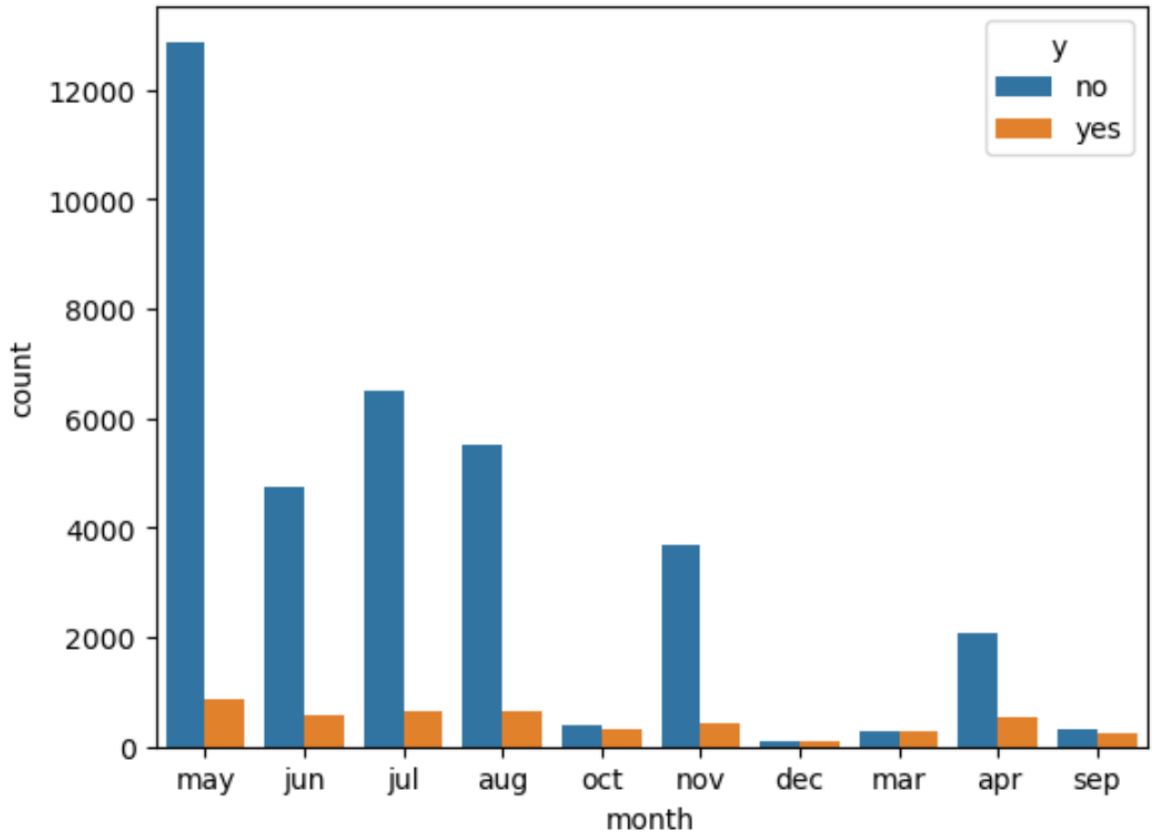
6. loan vs y:
   - Yes (Has personal loan): Smaller subscription rate.
   - No (Does not have personal loan): Slightly larger subscription rate.
   - Actionable Insight: Clients that do not have a personal loan have a higher subscription rate.
7. contact vs y:
   - Cellular: Significantly higher proportion of "yes" subscriptions compared to telephone.
   - Telephone: Significantly lower proportion of "yes" subscriptions compared to cellular.
   - Actionable Insight: Focus efforts on cellular communication channels, as they are clearly more effective.
8. month vs y:
   - Mar, Sep, Oct, Dec: These months show the highest proportion of "yes" subscriptions.
   - Aug, Jul, Jun, May: These months show a low subscription rate.
   - Actionable Insight: Concentrate marketing efforts in March, September, October, and December. Consider adjusting strategies for the less effective months. This information corresponds with the EDA from the beginning of the conversation
9. day_of_week vs y:
   - The distribution is very similar across all days of the week.
   - Actionable Insight: Day of the week likely does not play a significant role in subscription rates. This variable is not significant to the model.

10. poutcome vs y:
    - Success: Those who had a "successful" outcome in the previous campaign have a very high subscription rate this time around.
    - Failure: Much lower conversion rate.
    - Nonexistent: Most people fall in this category and the subscription rate is relatively low.
    - Actionable Insight: Prior campaign success is a strong predictor of future success. Focus efforts on those who were previously successful.

Returning (Repeat) customers subscribed more to the term deposit



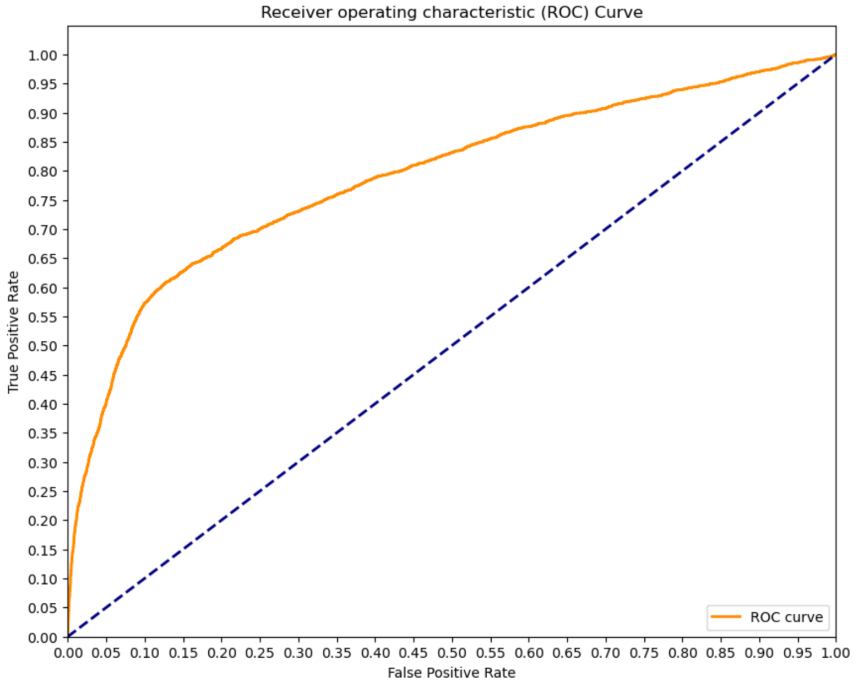
Relationship between "month" and the target variable 'y' (subscription)



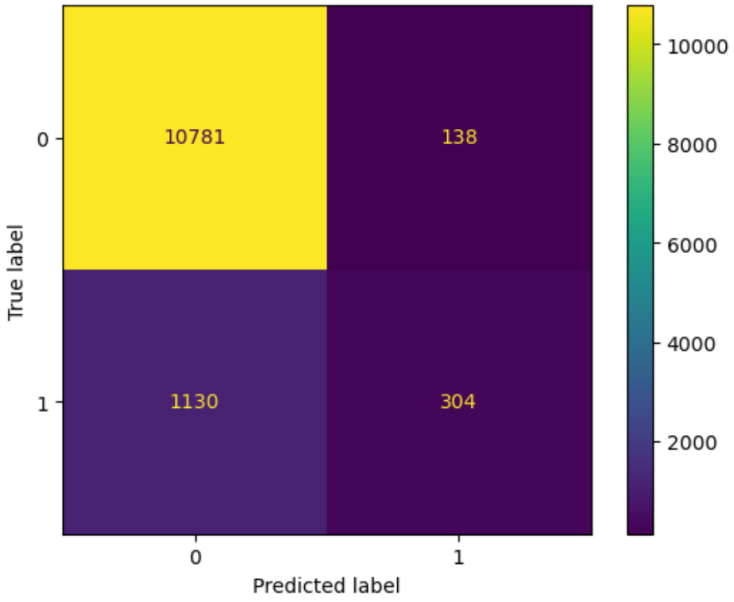
- May, June, July, August: Low conversion rates These months exhibit the highest number of contacts ("no" subscriptions) and some of the lowest rates of "yes" subscriptions.
- March, September, October, December: High Conversion Rates. Note: These months have much fewer contacts overall than the summer months.

Logistic Regression Output

AUC: 0.7890830708565997

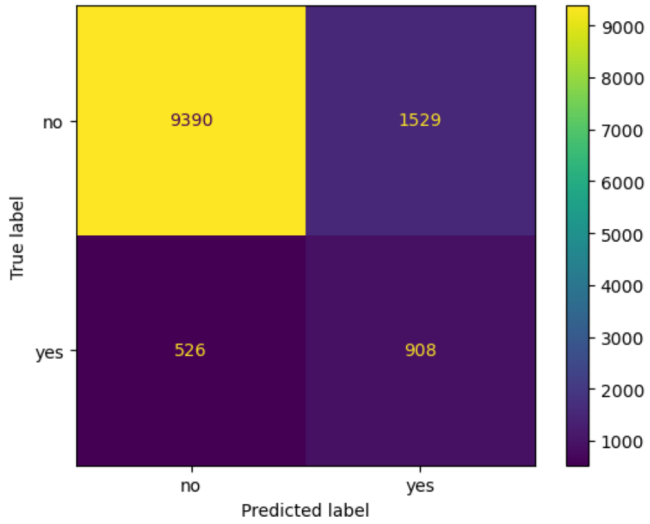


Confusion Matrix



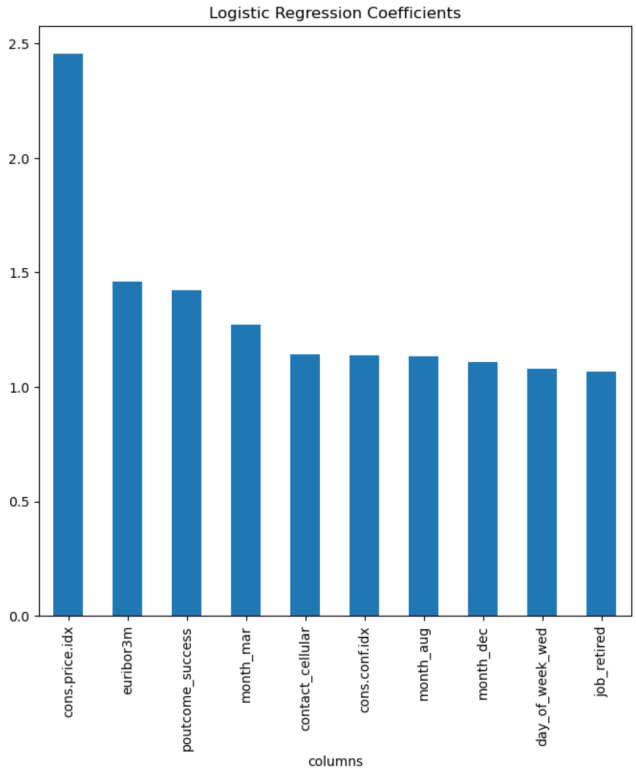
Logistic Regression Grid Search

```
Fitting 5 folds for each of 8 candidates, totalling 40 fits
<Figure size 300x300 with 0 Axes>
```

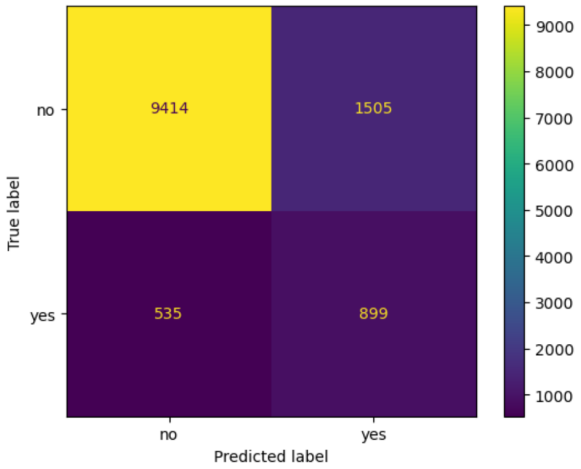


	precision	recall	f1-score	support
no	0.95	0.86	0.90	10919
yes	0.37	0.63	0.47	1434
accuracy			0.83	12353
macro avg	0.66	0.75	0.69	12353
weighted avg	0.88	0.83	0.85	12353

Logistic Regression Coefficients



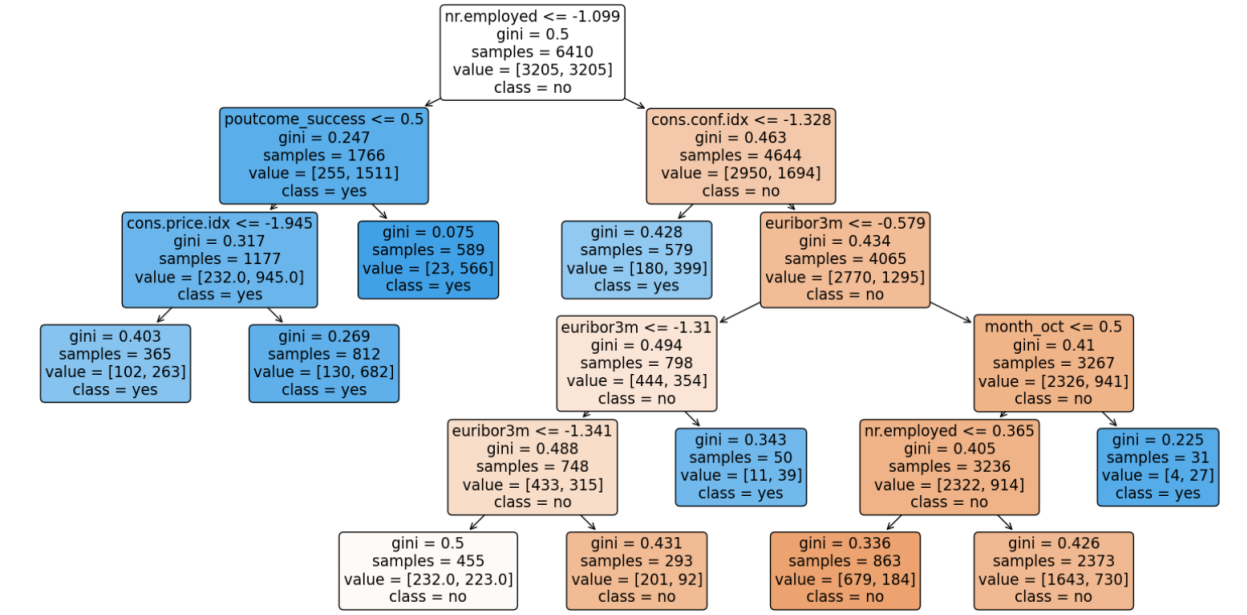
Model Evaluation



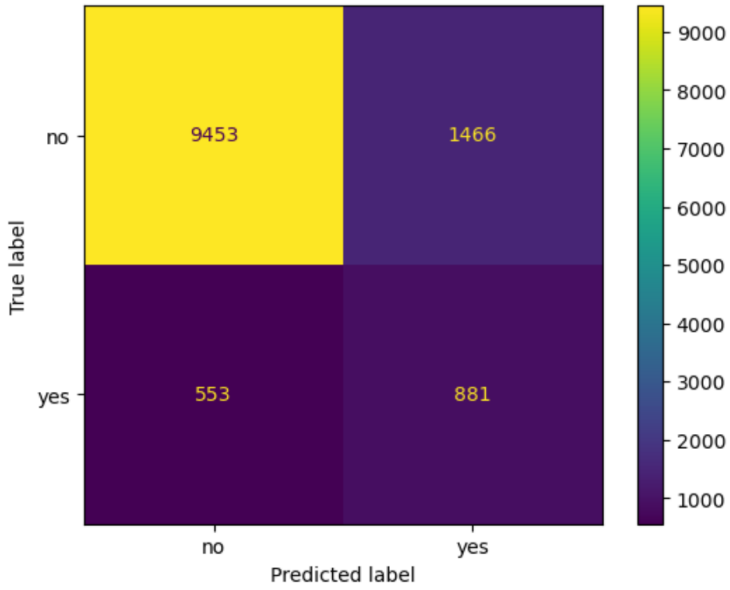
	precision	recall	f1-score	support
no	0.95	0.86	0.90	10919
yes	0.37	0.63	0.47	1434
accuracy			0.83	12353
macro avg	0.66	0.74	0.69	12353
weighted avg	0.88	0.83	0.85	12353

Decision Tree

Decision tree highlights the importance of the number of employees, consumer confidence index, consumer price index, outcome of previous campaigns, and the euribor 3 month rate in predicting term deposit subscriptions. Lower number of employees and euribor 3 month rate seem to generally correlated to a higher likelihood of subscription.



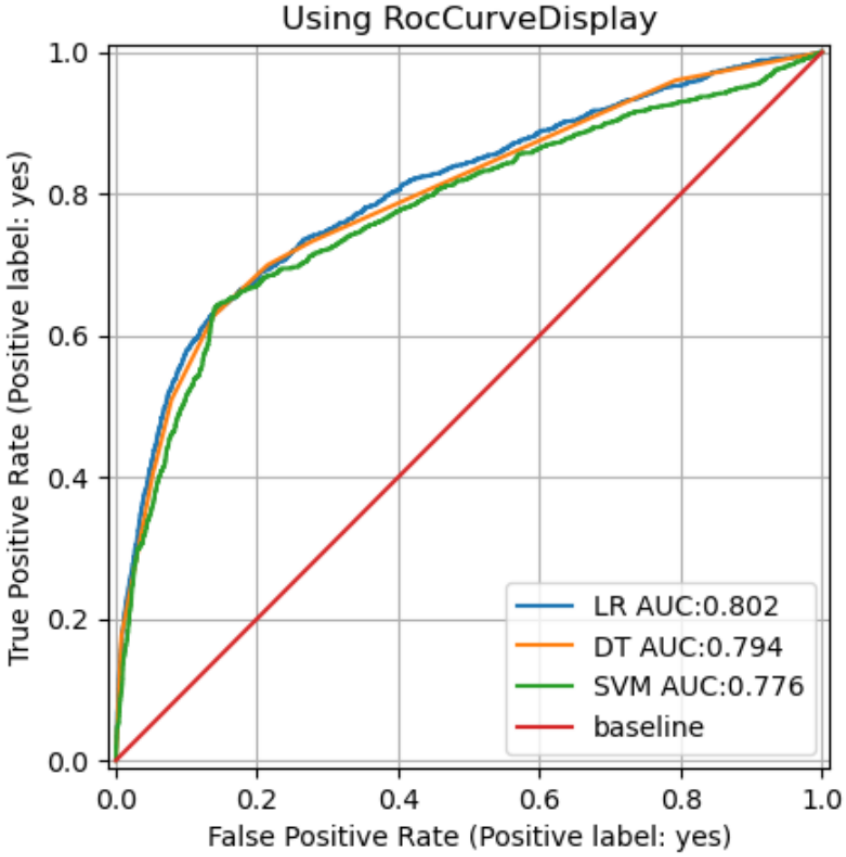
Fit-Predict-Evaluate Model



	precision	recall	f1-score	support
no	0.94	0.87	0.90	10919
yes	0.38	0.61	0.47	1434
accuracy			0.84	12353
macro avg	0.66	0.74	0.68	12353
weighted avg	0.88	0.84	0.85	12353

	model	test accuracy	precision	recall	specificity	f1	AUC	grid total fit time	best model mean fit time	grid predict time	best params
0	lr	0.834	0.373	0.633	0.860	0.469	0.802	5.62	0.355	0.03	{'lr__C': 1.0, 'lr__solver': 'liblinear'}
1	dt	0.835	0.374	0.627	0.862	0.468	0.794	1.04	0.055	0.01	{'dt__max_depth': 6, 'dt__min_impurity_decreas...
2	svm	0.837	0.375	0.614	0.866	0.466	0.776	390.60	169.323	2.85	{'svm__kernel': 'linear'}

Scoring the model



Based on the analysis of ROC curves and AUC scores, Model Performance Assessment is performed based on the analysis of ROC curves and AUC scores. The ROC (Receiver Operating Characteristic) curve visualizes the performance of a binary classifier by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC (Area Under the Curve) score quantifies the overall ability of the model to distinguish between positive and negative instances. A higher AUC indicates better performance.

- Logistic Regression (LR): AUC = 0.802 - Performs the best among the three models.
- Decision Tree (DT): AUC = 0.794 - Performs slightly worse than Logistic Regression.
- Support Vector Machine (SVM): AUC = 0.776 - Performs the worst among the three models.
- Baseline: The baseline is the diagonal line, representing random chance (AUC = 0.5). All the models perform significantly better than random chance.

- In our chosen model Logistic Regression - Undersampling helped improve the F1 scores, yet at 47% they remain quite low for the positive class. The precision was high for the negative class at 90% and F1 score of 90%
- This would imply that the chosen model will predict a lot of customers to purchase a certificate product, resulting in many false positives and impacting limited resources for the bank to do their outreach
- This model requires further refinement before it may be deployed by the bank

Model Evaluation Metrics:

- Accuracy, Precision, Recall, F1-score
- ROC-AUC curve
- Confusion Matrix

Conclusion

Based on the analysis of ROC curves and AUC scores, Model Performance Assessment is performed based on the analysis of ROC curves and AUC scores. The ROC (Receiver Operating Characteristic) curve visualizes the performance of a binary classifier by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC (Area Under the Curve) score quantifies the overall ability of the model to distinguish between positive and negative instances. A higher AUC indicates better performance.

- Logistic Regression (LR): AUC = 0.802 - Performs the best among the three models.
- Decision Tree (DT): AUC = 0.794 - Performs slightly worse than Logistic Regression.
- Support Vector Machine (SVM): AUC = 0.776 - Performs the worst among the three models.
- Baseline: The baseline is the diagonal line, representing random chance (AUC = 0.5). All the models perform significantly better than random chance

Next Steps & Recommendations

- Hyperparameter tuning - Further experimentation with hyperparameter tuning for all three models, particularly the Logistic Regression and Decision Tree, to see if you can squeeze out better performance
- Interaction Terms: Based on EDA (especially the correlations and bar charts), create interaction terms between features that seem to have a combined effect on the target variable.
- Polynomial Features: Consider adding polynomial features to capture non-linear relationships.
- Feature Selection/Dimensionality Reduction:
- Regularization: For Logistic Regression, experiment with L1 (Lasso) or L2 (Ridge) regularization to reduce overfitting and potentially improve generalization.
- PCA/Feature Importance: Use PCA or feature importance from a tree-based model (e.g., Random Forest) to select the most relevant features and reduce dimensionality.

- Subscriptions are noticeably higher when the euribor3m rate is lower. Requires further discovery and understanding.
- Recommend incorporating external data sources (e.g., demographic data, economic indicators) to enrich banking feature set
- A/B Testing: Once you have a refined model, perform A/B testing on a small segment of your customer base to compare the performance of ML(LR) model-driven marketing strategy against your current approach.
- Monitoring: Continuously monitor the performance of the model in a production environment and retrain it periodically with new data to maintain its accuracy and relevance.