Interpretable Machine Learning for Predicting Term Deposit (Classification)

LINDA KELLEN AYEBALE (2400721933)

KARANZI JOHNMARY (2400721928)

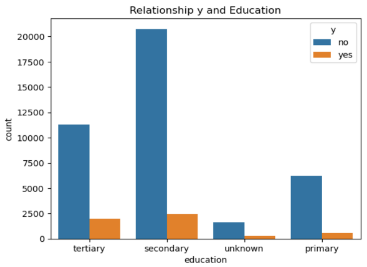
Introduction

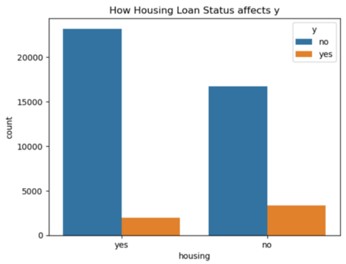
- The study uses machine learning (ML) techniques to predict client subscription to term deposits using the UCI Bank Marketing dataset. Random Forest and Logistic Regression models were employed, with Random Forest achieving 92.3% accuracy. To address class imbalance, the NearMiss-2 algorithm was applied. Explainable AI (XAI) techniques, such as SHAP and LIME, were used to enhance interpretability and provide actionable insights.

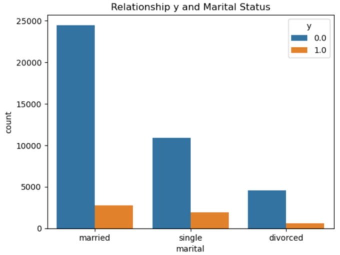
Methodology

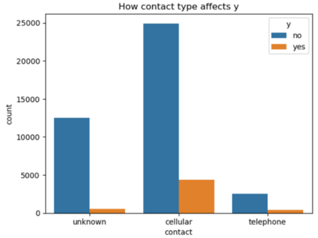
- 1. Dataset: UCI Bank Marketing dataset (45,211 rows, 17 features).
- 2. Models: Random Forest and Logistic Regression for prediction.
- 3. Class Imbalance: Managed using the NearMiss-2 algorithm.
- 4. Feature Engineering: Categorical data encoded; continuous data normalized.
- 5. Explainability: SHAP and LIME techniques applied for model interpretation.

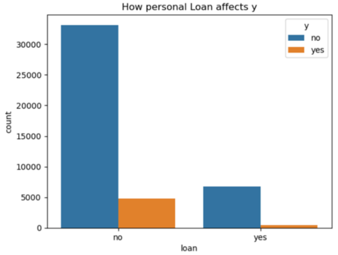
Feature Analysis

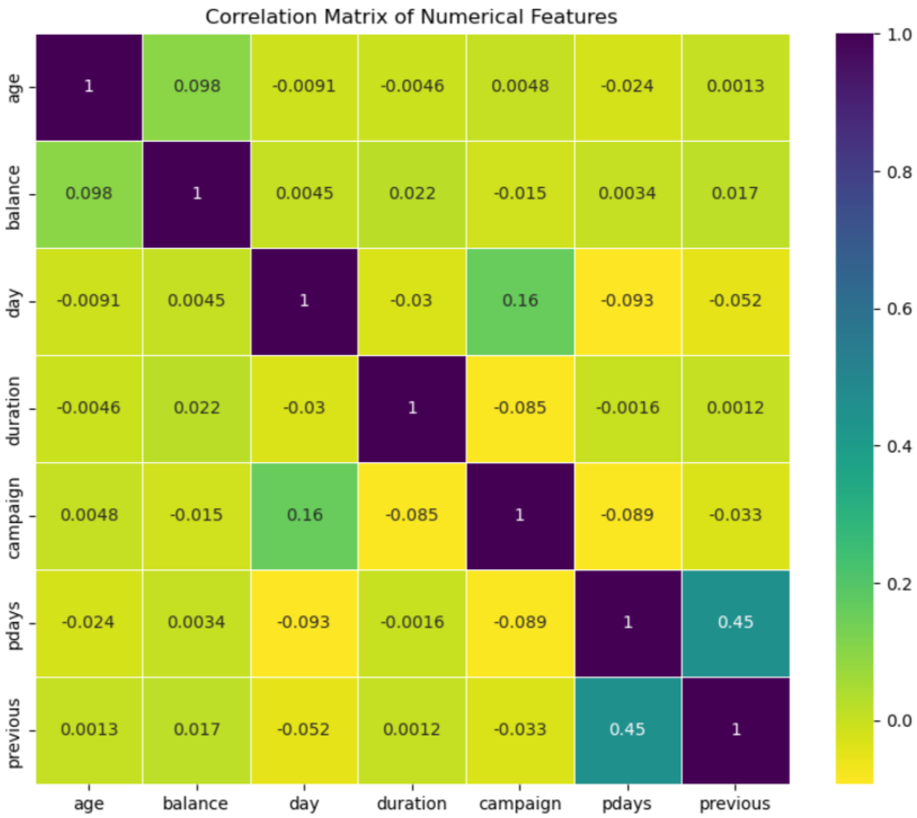








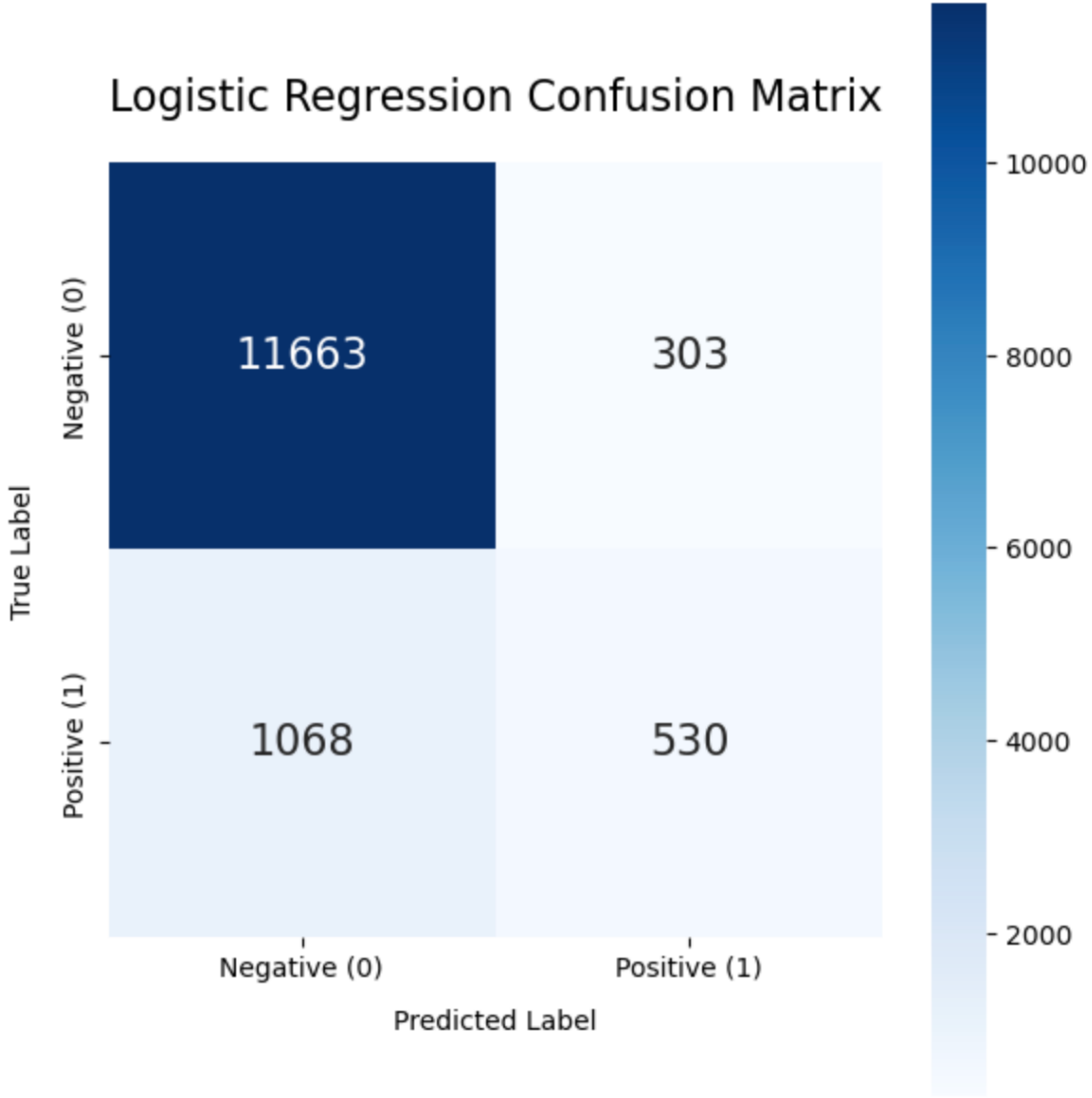




- The continuous data fields were normalized and a heat map generated in order to visualize their correlation.
- categorical feature affects the outcome. In education, clients with tertiary level have a higher likelihood of subscribing. For both housing and personal loan status, clients without these loans are more likely to subscribe to the deposit. Clients contacted via cellular phone have the highest likelihood of subscribing. For the marriage status, single clients have a higher probability of subscribing that the married or divorced.

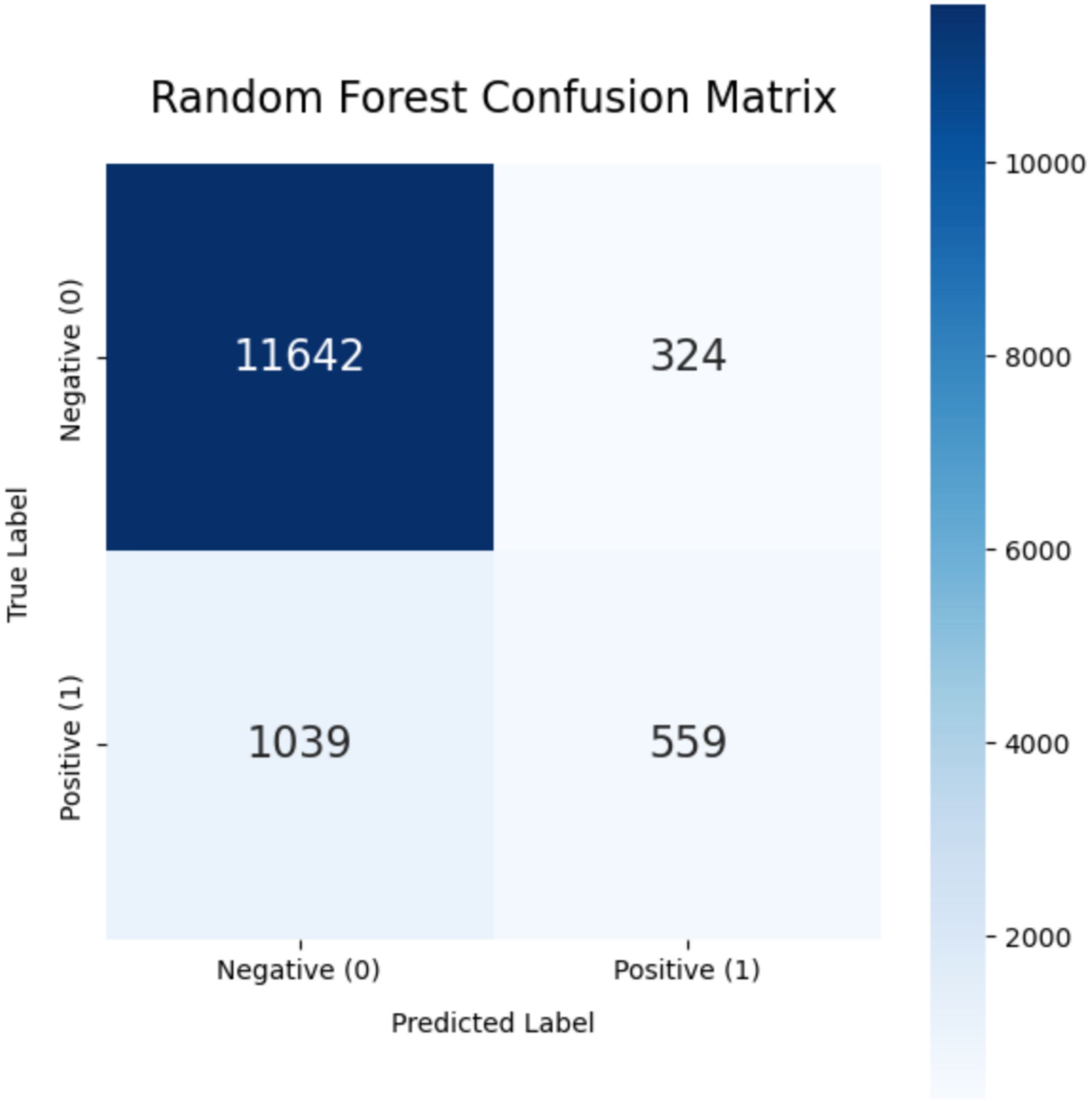
Logistic regression

Accuracy	89.89%
Precision	63.63%
Recall	33.17%
F1-Score	43.60%



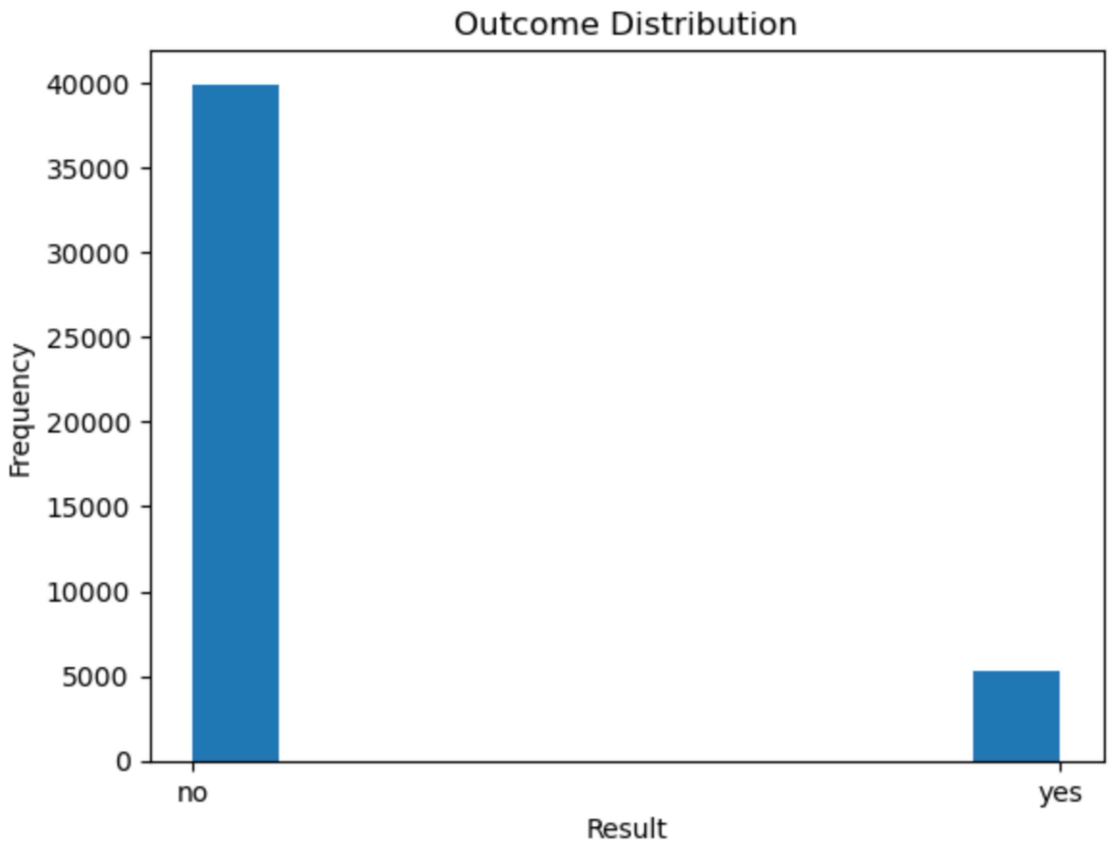
Random Forest

Accuracy	89.95%
Precision	63.31%
Recall	34.98%
F1-Score	45.06%



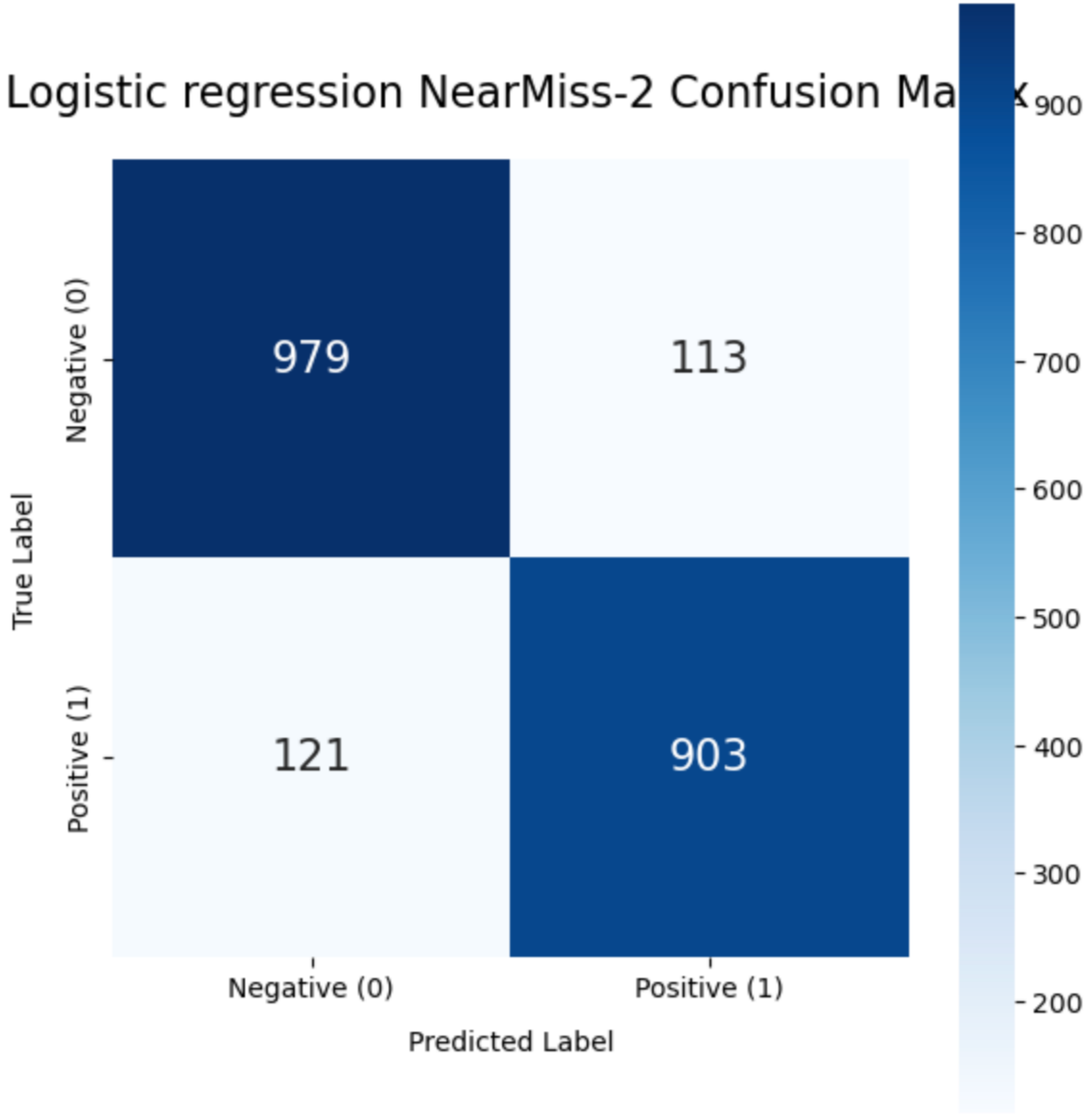
	Negative (0)	Positive (1)
Negative (0)	11642	324
Positive (1)	1039	559

- As shown in the 2 confusion matrix we realised the modals were not performing well.
- From the investigation made we found out that its due to the Outcome class y that is highly imbalanced.
- To remedy this we decided to do undersampling using NearMiss-2



Logistic Regression

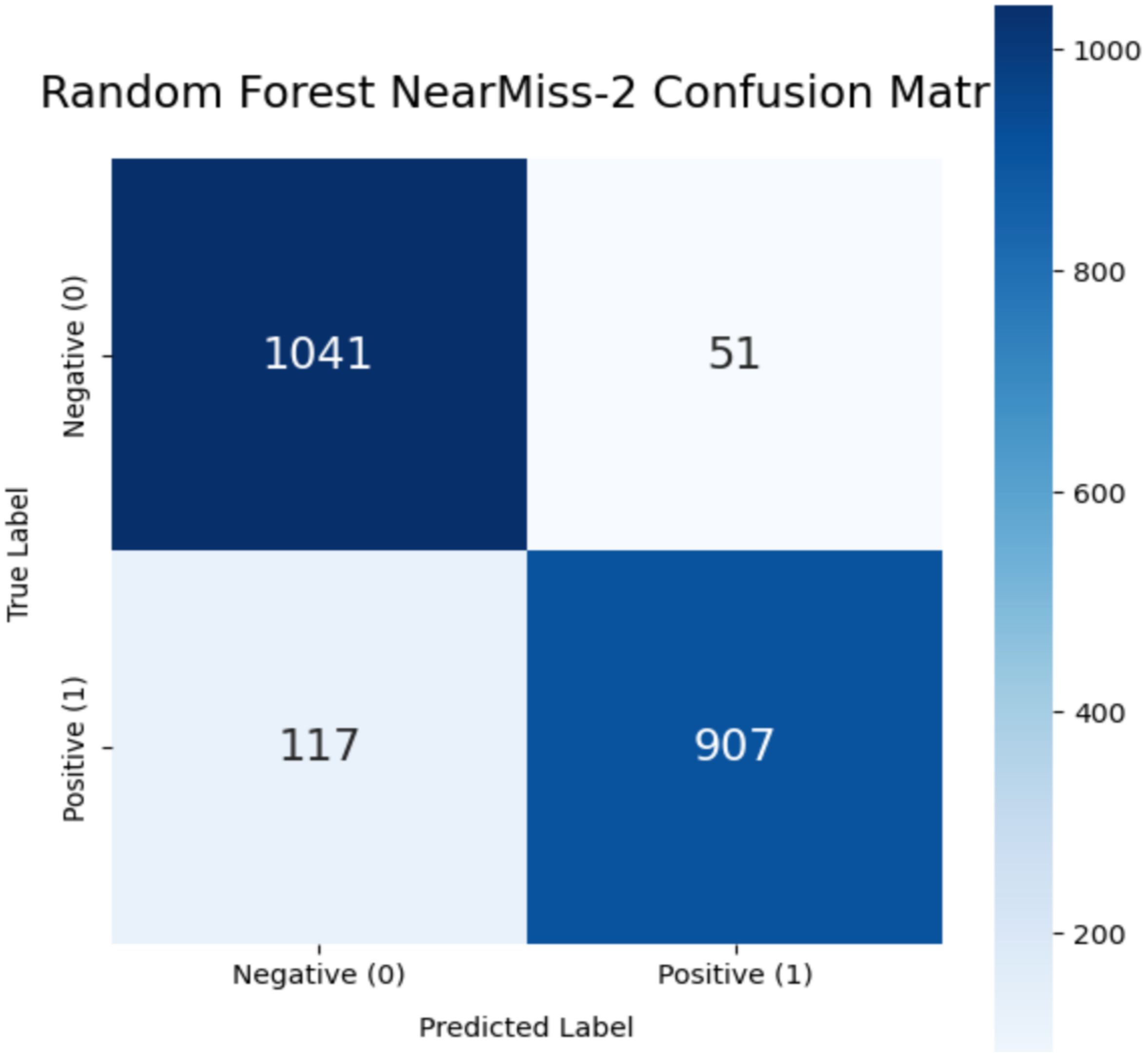
Accuracy	88.94%
Precision	88.88%
Recall	88.18%
F1-Score	88.53%



	Negative (0)	Positive (1)
Negative (0)	979	113
Positive (1)	121	903

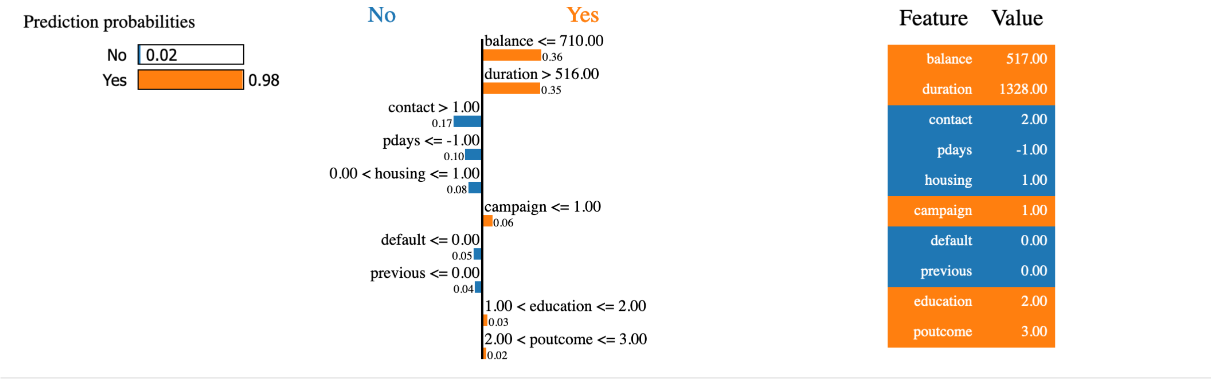
Random Forest

Accuracy	92.06%
Precision	94.68%
Recall	88.57%
F1-Score	91.52%

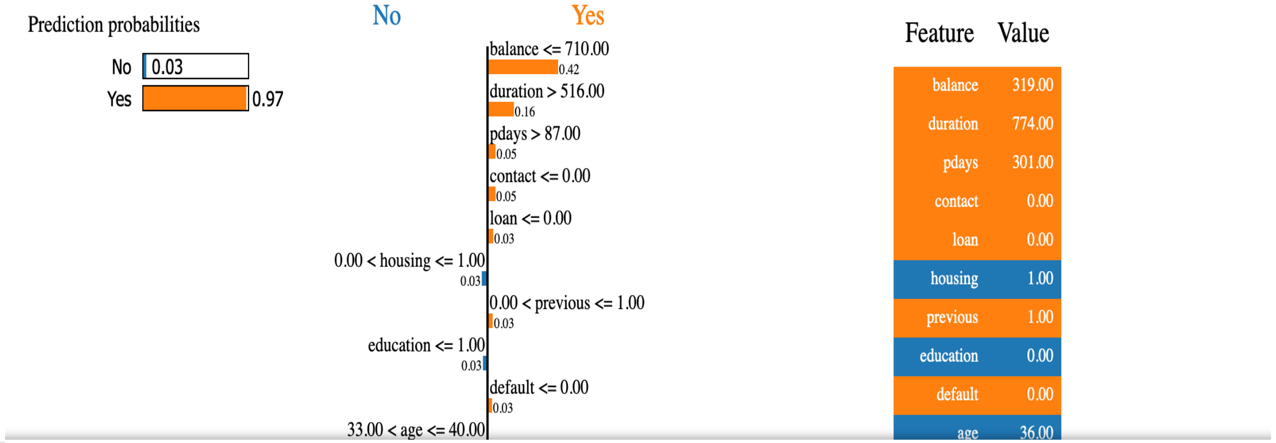


Note: From the observations of both modals, Random Forest modal gives the best results after undersampling.

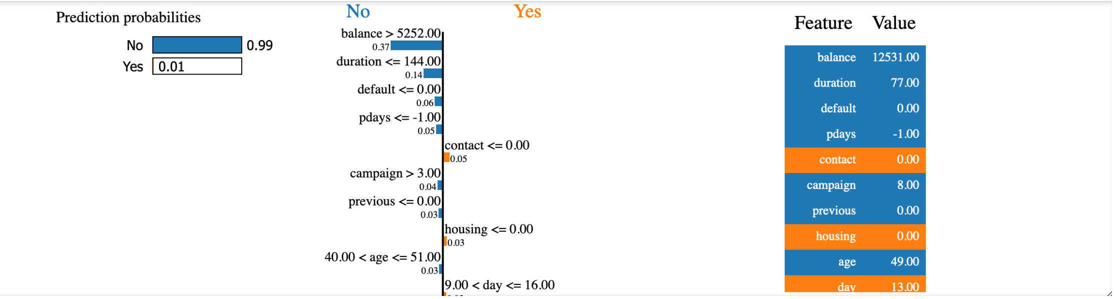
Explainability with LIME



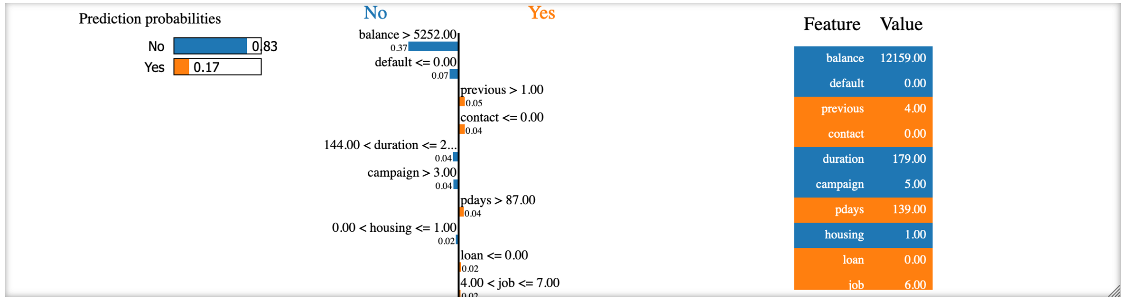
Feature	Value
balance	517.00
duration	1328.00
contact	2.00
pdays	-1.00
housing	1.00
campaign	1.00
default	0.00
previous	0.00
education	2.00
poutcome	3.00



Feature	Value
balance	319.00
duration	774.00
pdays	301.00
contact	0.00
loan	0.00
housing	1.00
previous	1.00
education	0.00
default	0.00
age	36.00

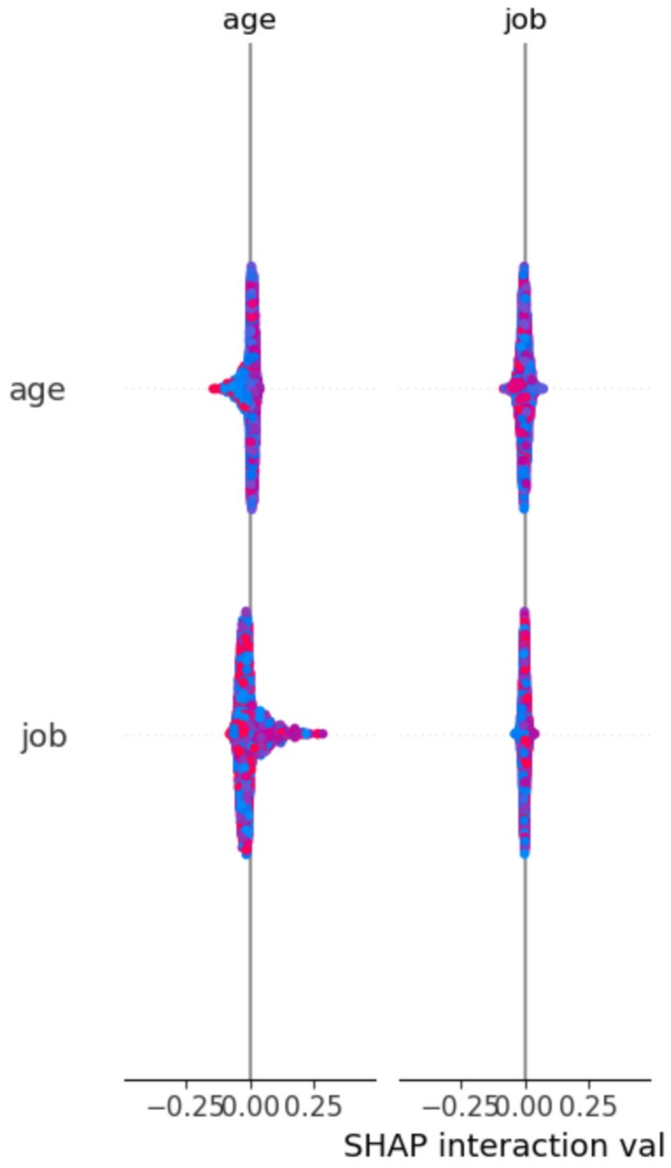


Feature	Value
balance	12531.00
duration	77.00
default	0.00
pdays	-1.00
contact	0.00
campaign	8.00
previous	0.00
housing	0.00
age	49.00
day	13.00



Feature	Value
balance	12159.00
default	0.00
previous	4.00
contact	0.00
duration	179.00
campaign	5.00
pdays	139.00
housing	1.00
loan	0.00
job	6.00

Explainability with SHAP



Contributions & Future Work

- Contributions:
- - Implemented ensemble learning for class imbalance.
- - Applied XAI techniques for model transparency.
- - Enhanced marketing strategies via actionable insights.

- Future Work:
- - Develop advanced ensemble methods and cost-sensitive learning.
- - Implement real-time prediction systems.
- - Explore deep learning for complex patterns.

Conclusion

- Random Forest performed best with high accuracy, precision, recall and F1 score
- Lime shows that a combination of features affect whether a customer will make a fixed deposit or not for example duration and balance that is to say if a customer has a high balance and had a long call with marketing there were high chances of them using the product while if one of them was low, greatly reduced chances of the customer making a deposit.