Predicting Term Deposits in Banking

A classification model using Logistic Regression, Decision Trees,
Random Forest,
K-Nearest Neighbors and Support Vector Machines

Business Understanding

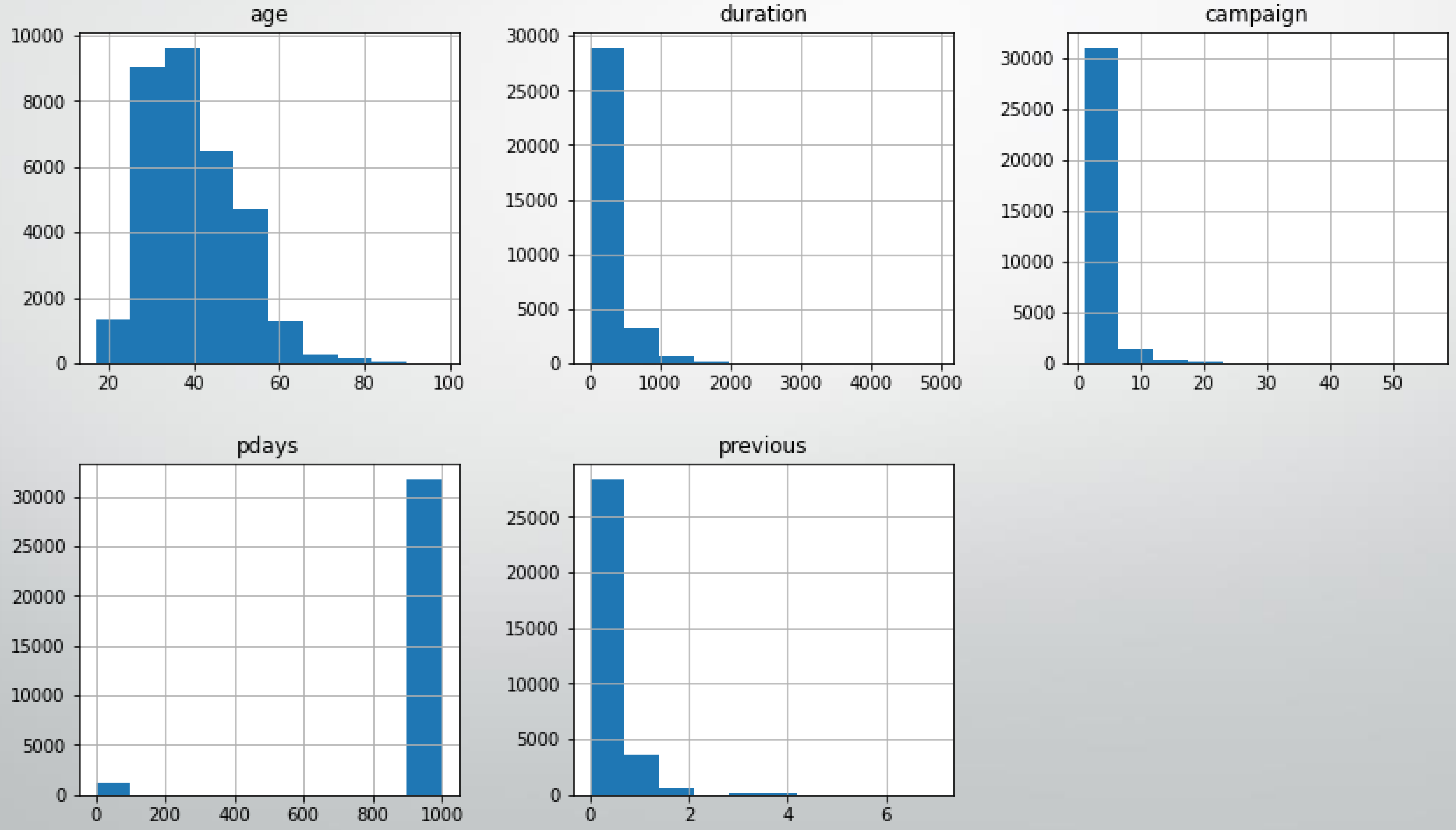
- A long term deposit is an investment product offered by banks that allows individuals to earn interest on their savings over a specified period of time.
- They're a powerful way for banks to generate substantial profits over an extended period, because they secure a stable source of funding and banks lend out the money at a higher interest rate than they offer depositors.
- Long-term deposits are less likely to be withdrawn, providing a reliable and predictable source of income.
- By identifying existing customers who are more likely to subscribe to long-term deposit plans, banks can increase their revenue while maintaining customer loyalty. With their ability to generate consistent returns, long-term deposits are a valuable tool for banks looking to build a profitable business model and achieve long-term financial stability.

Business Problem

- A bank in Portugal is experiencing a decline in revenue and seeking to boost it by increasing long term deposits.
- They intend to identify existing customers who are most likely to open long term deposits accounts.
- The bank collected data through telemarketing. The objective of the project is to develop a classification model that can predict whether a customer will open a long term deposit account or not.
- The data analysis and predictive model will help the bank in target market segmentation, thus increasing their conversion rate, and in extension, the revenue.

Data Understanding

Histograms, showing the distribution of numeric columns



Data Understanding

•Age ranges from 17 to 98 with most people around the age of 40.

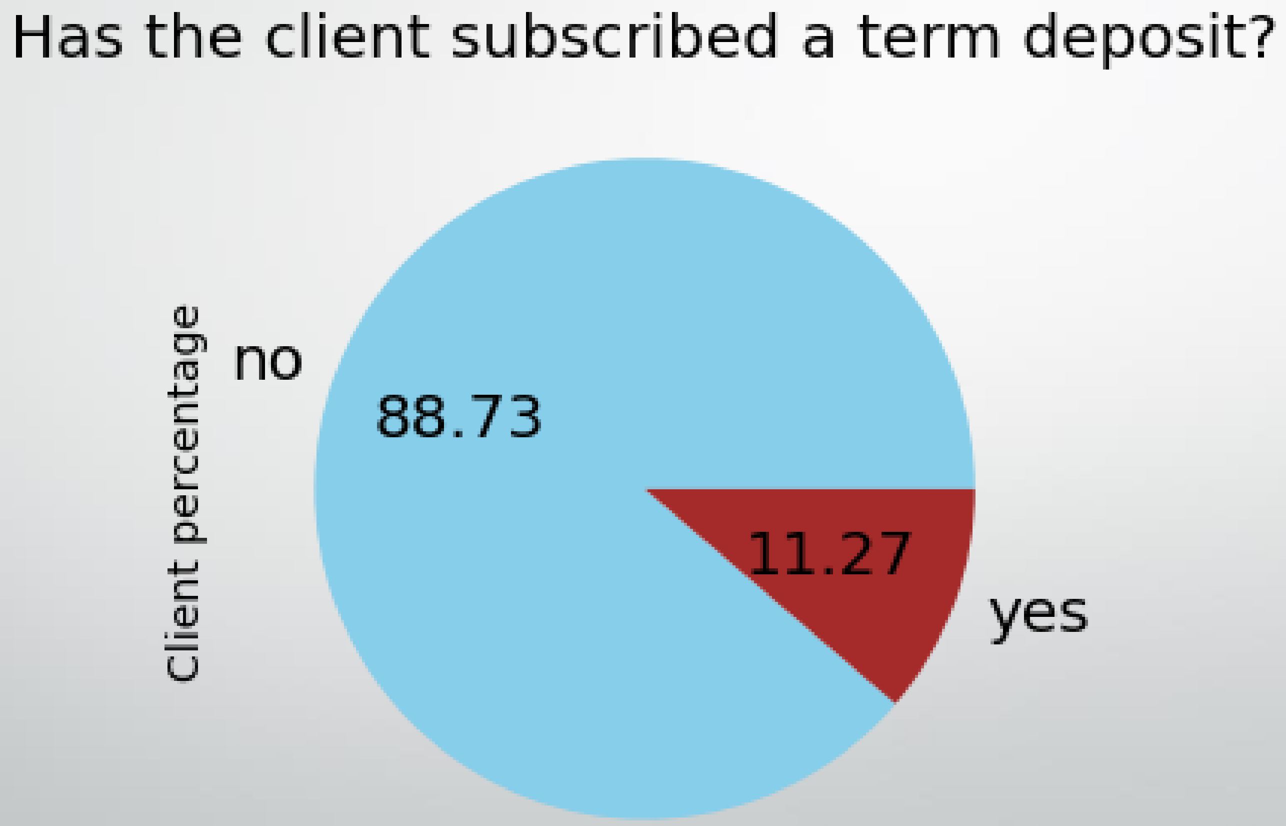
•The last contact duration is between 0 and 4918 seconds. Most people have a contact duration of less than 500 seconds.

•Most people have been contacted less than 5 times during this campaign.

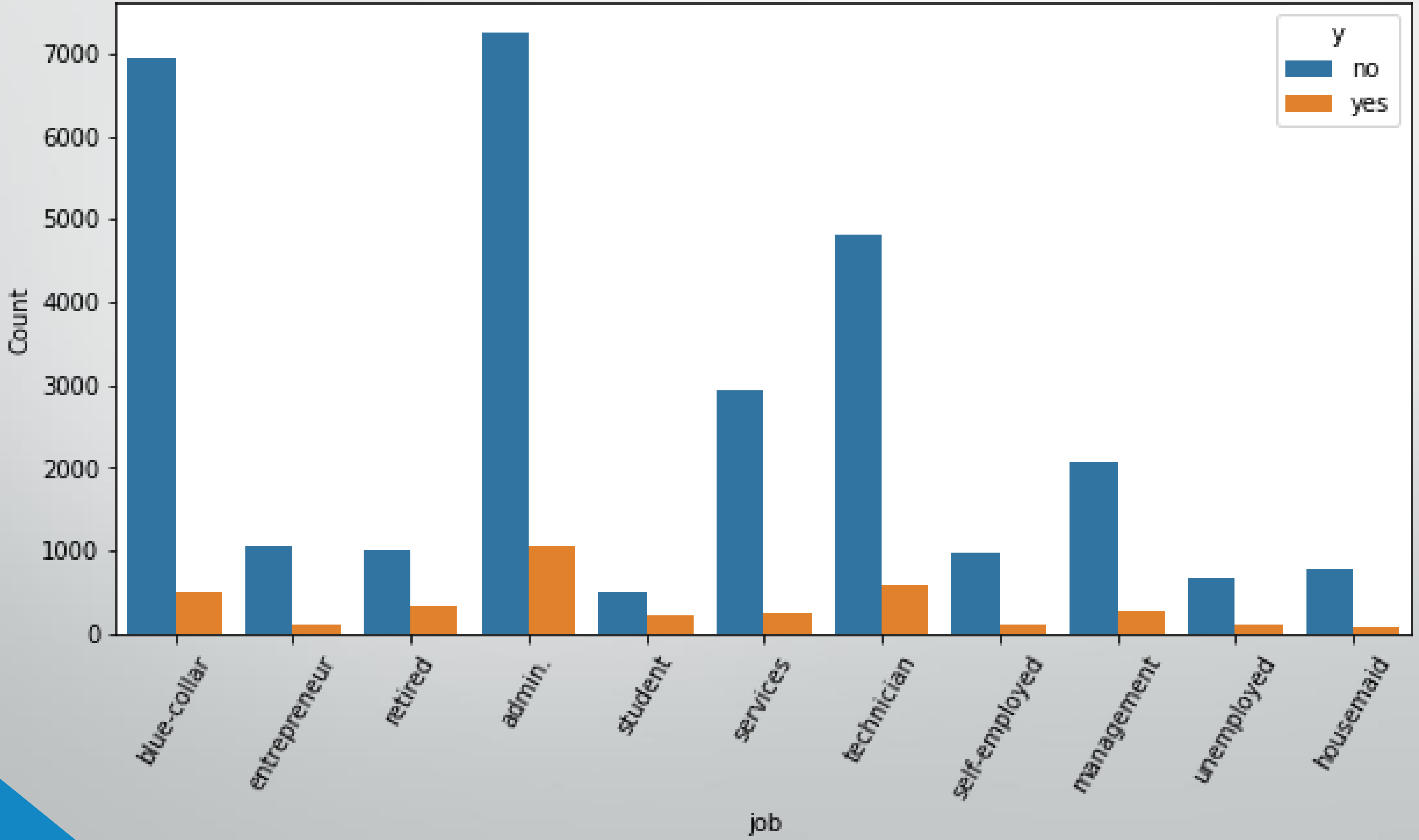
•For the majority, 999 days had passed by after the client was last contacted from a previous campaign. 999 means client was not previously contacted.

•Before this campaign, most clients had not been contacted at all.
NOTE: The numeric columns are not normally distributed and have different scales.

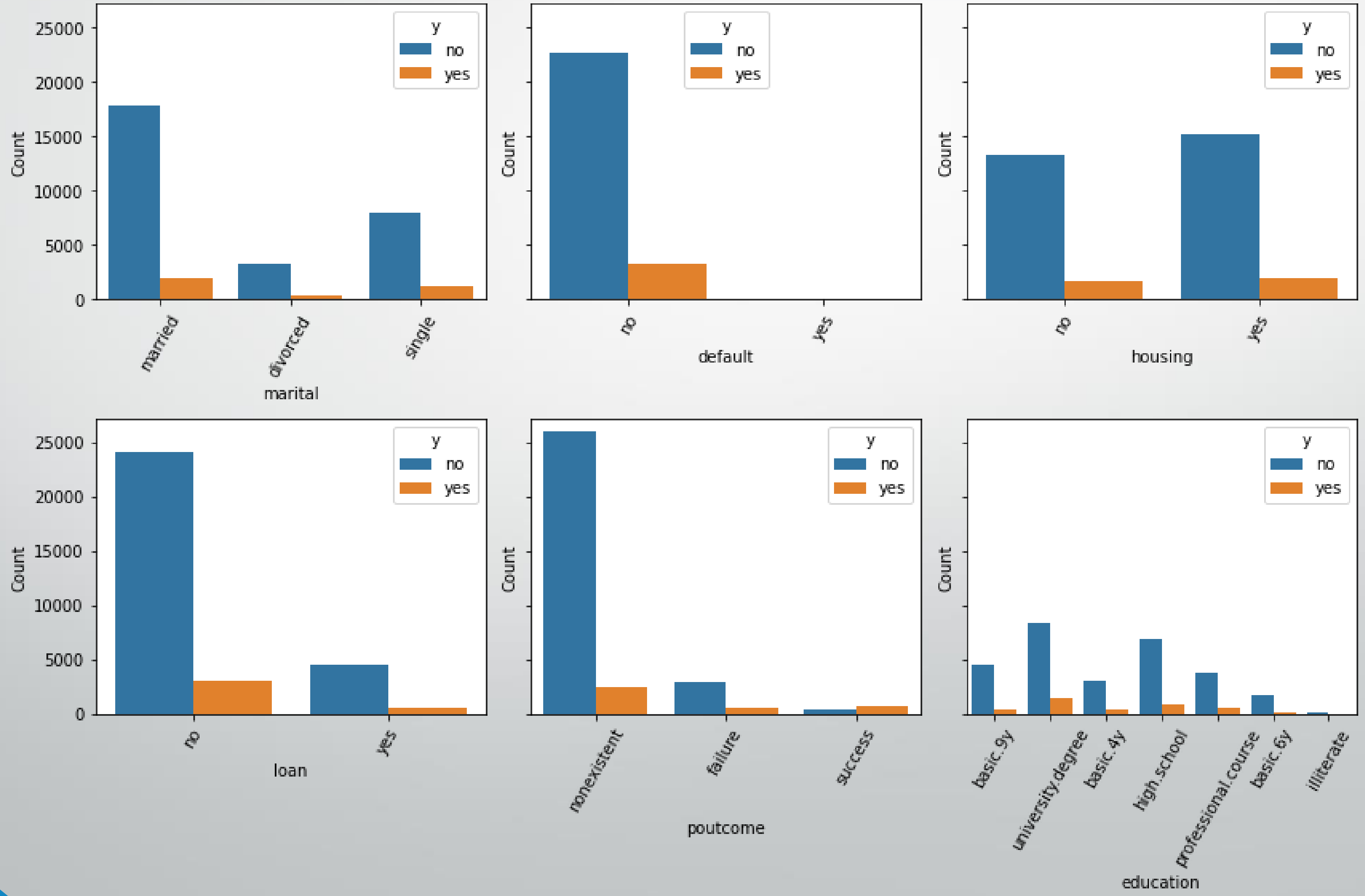
- Only 11.27% of the contacted clients, signed up for the long term deposits.



Distribution of job type



Distribution of categorical columns

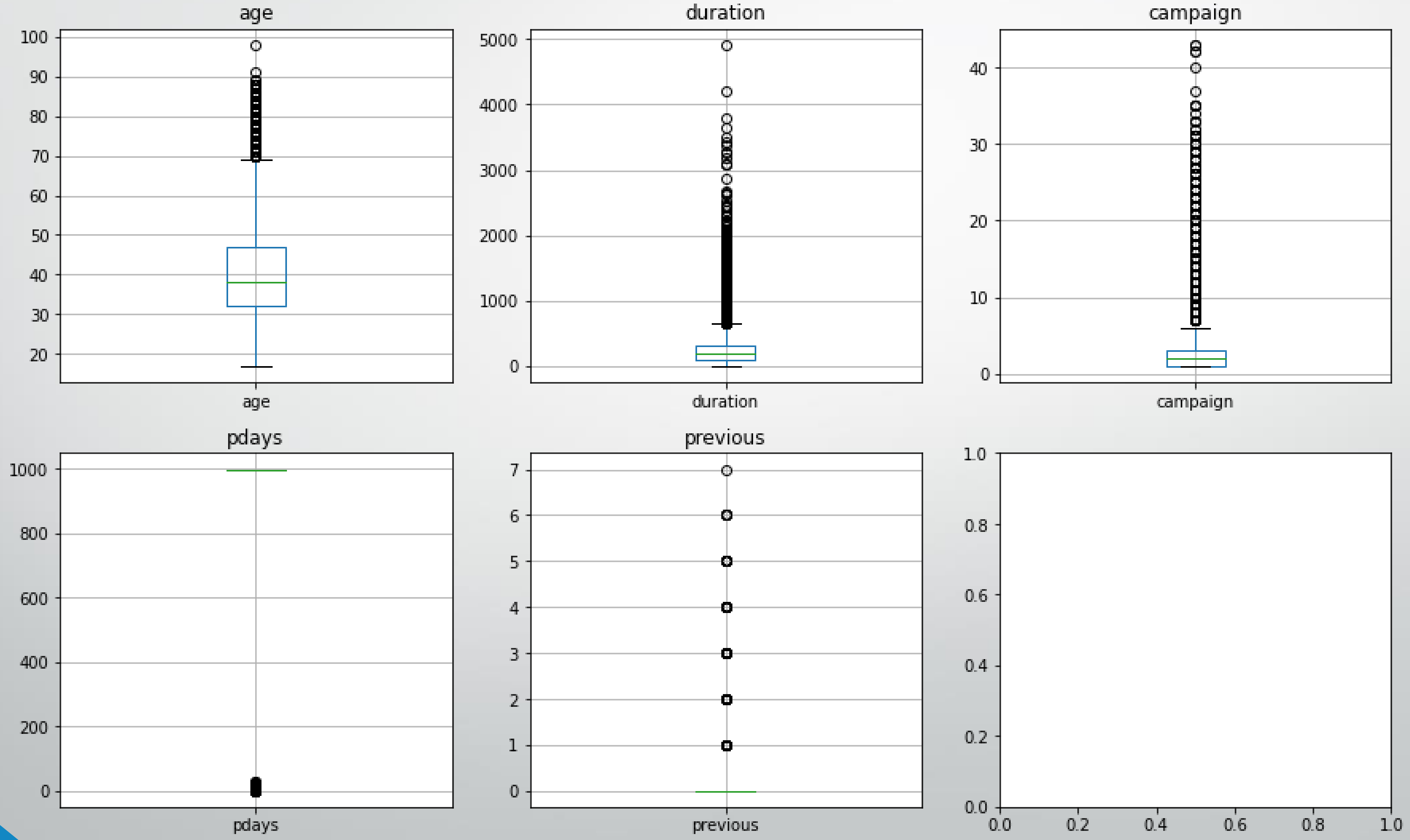


Data Preparation

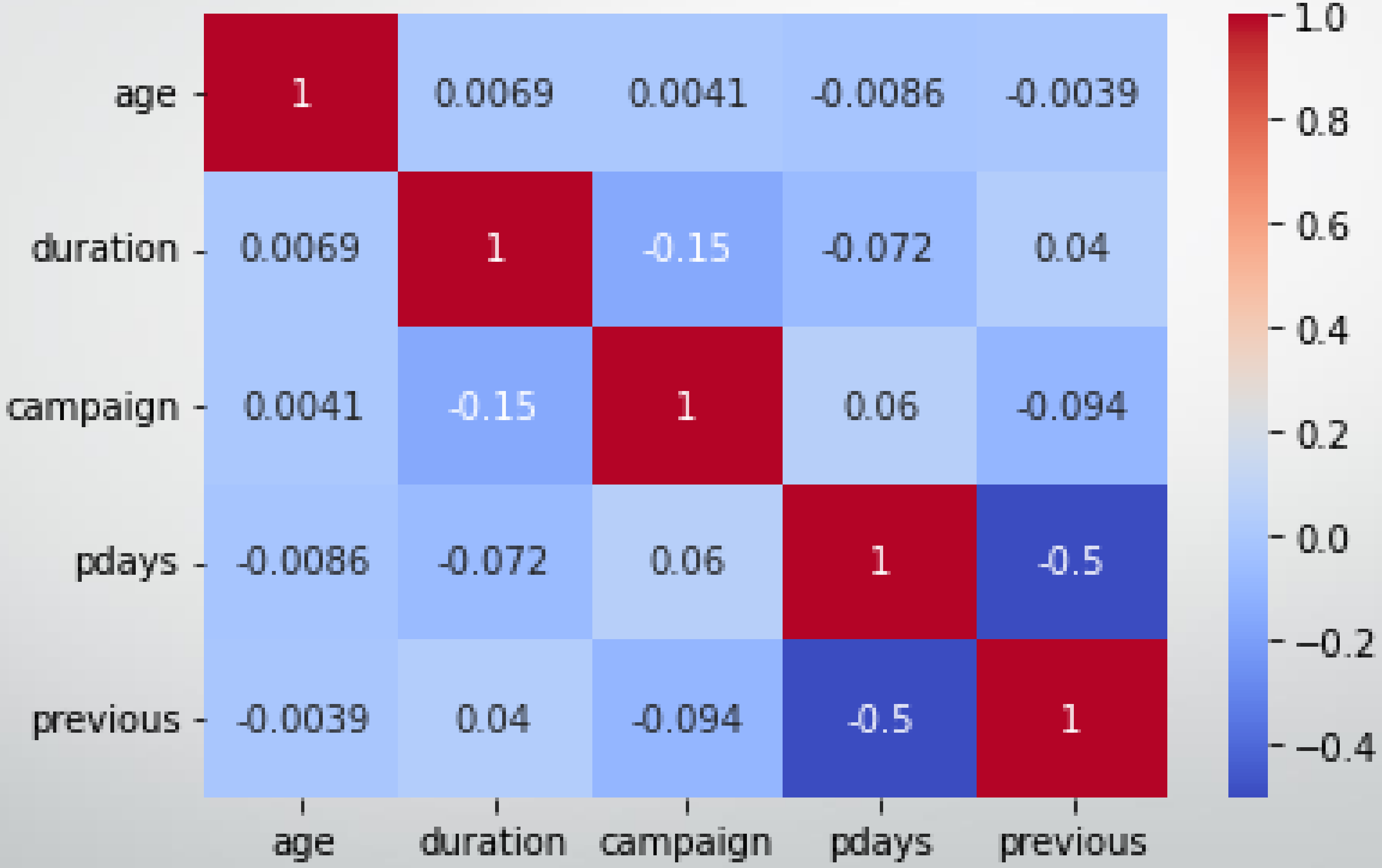
- *The team investigated and fixed;*

1. *Duplicated values – by dropping duplicated columns*
2. *Missing values – imputed with the mode and unknown values where applicable*
3. *Outliers – Log transformation*

Visual image of the data outliers



Check for multicollinearity



	age	duration	campaign	pdays	previous
age	1	0.0069	0.0041	-0.0086	-0.0039
duration	0.0069	1	-0.15	-0.072	0.04
campaign	0.0041	-0.15	1	0.06	-0.094
pdays	-0.0086	-0.072	0.06	1	-0.5
previous	-0.0039	0.04	-0.094	-0.5	1

No multicollinearity was observed between the features.

Data Modeling

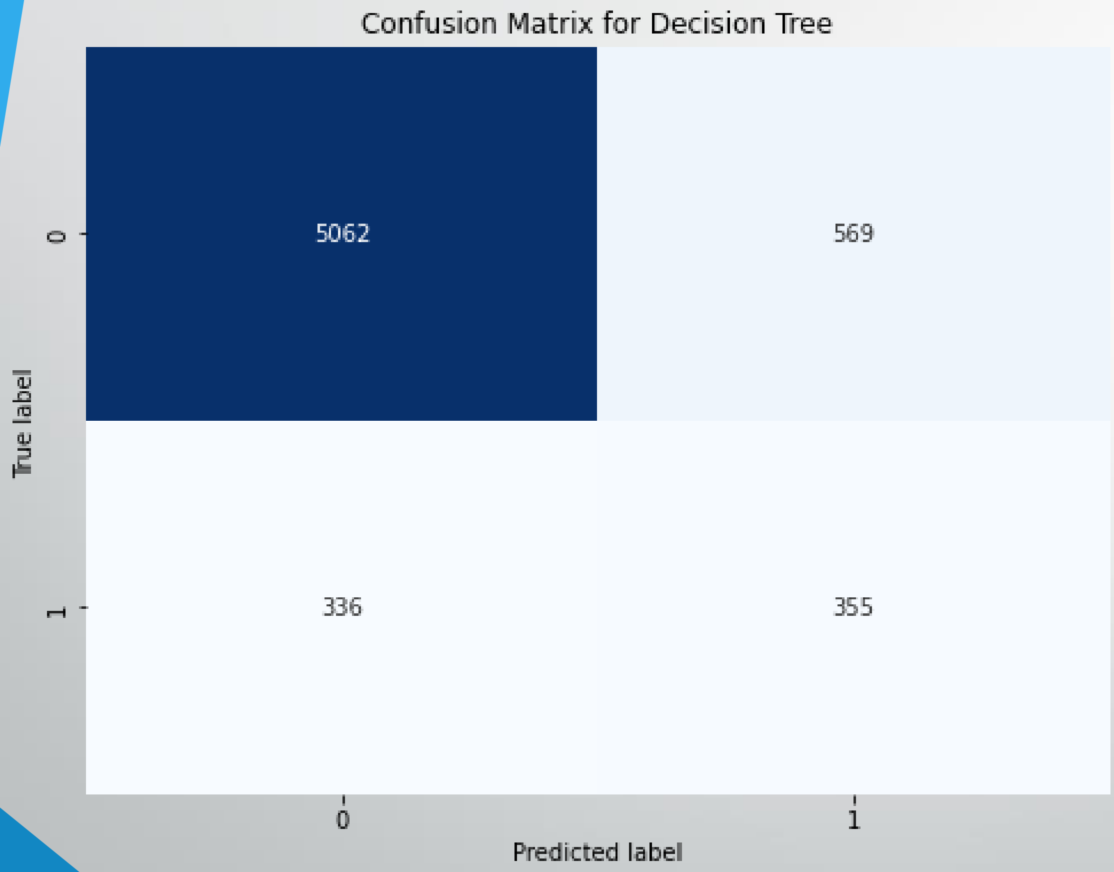
- *This is a classification problem, where we are predicting if a bank customer will open a long term deposit account or not.*
- *We used the following models:*

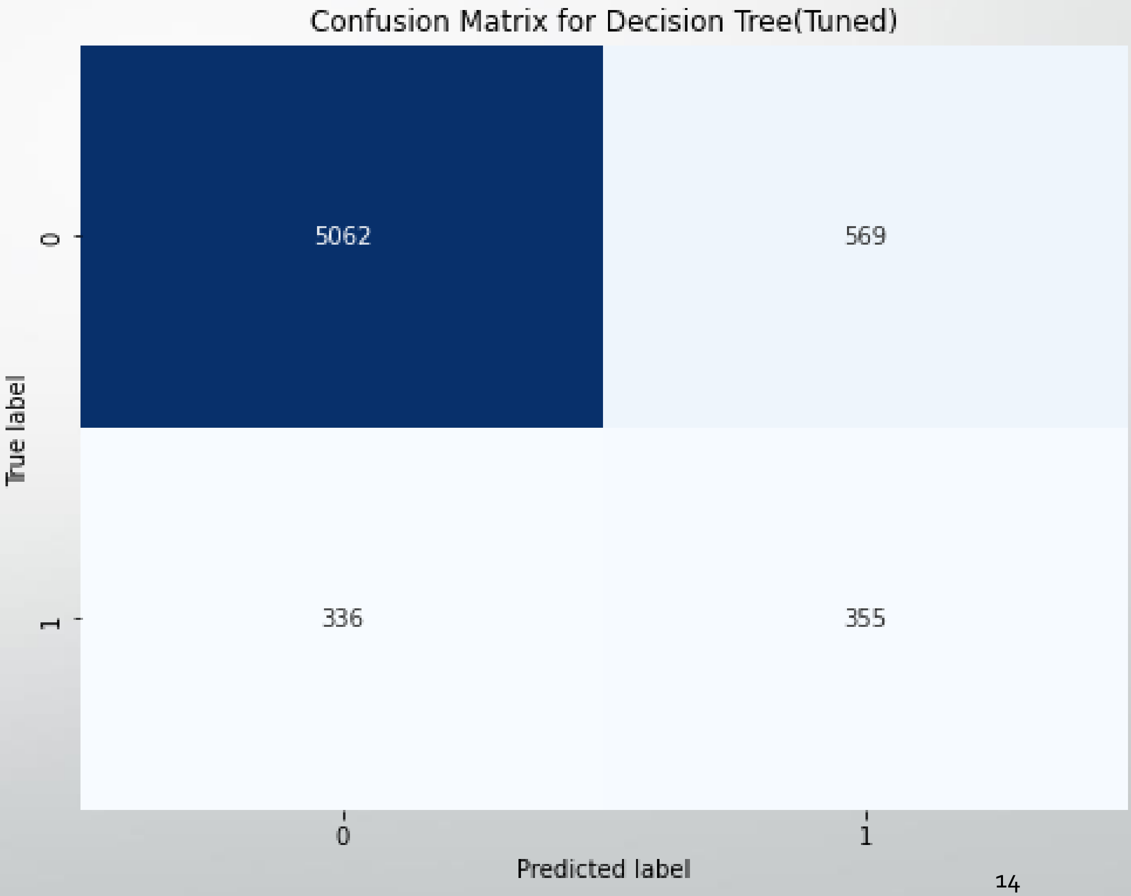
1. *Decision Tree Classifier*
2. *Logistic Regression*
3. *Random Forest Classifier*
4. *K-Nearest Neighbors*
5. *Support Vector Machine*

Model Evaluation

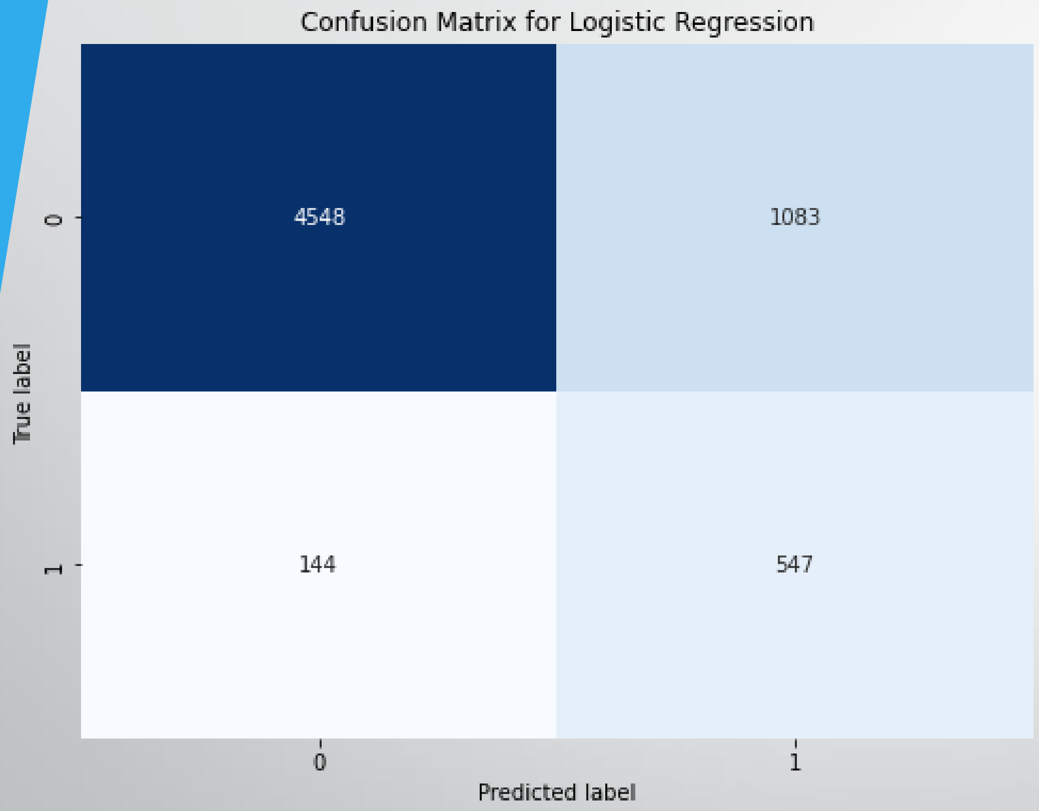
- The evaluation metrics we focused on were:
  1. Accuracy
  2. Precision

Decision Tree Confusion Matrix


Confusion Matrix for Decision Tree

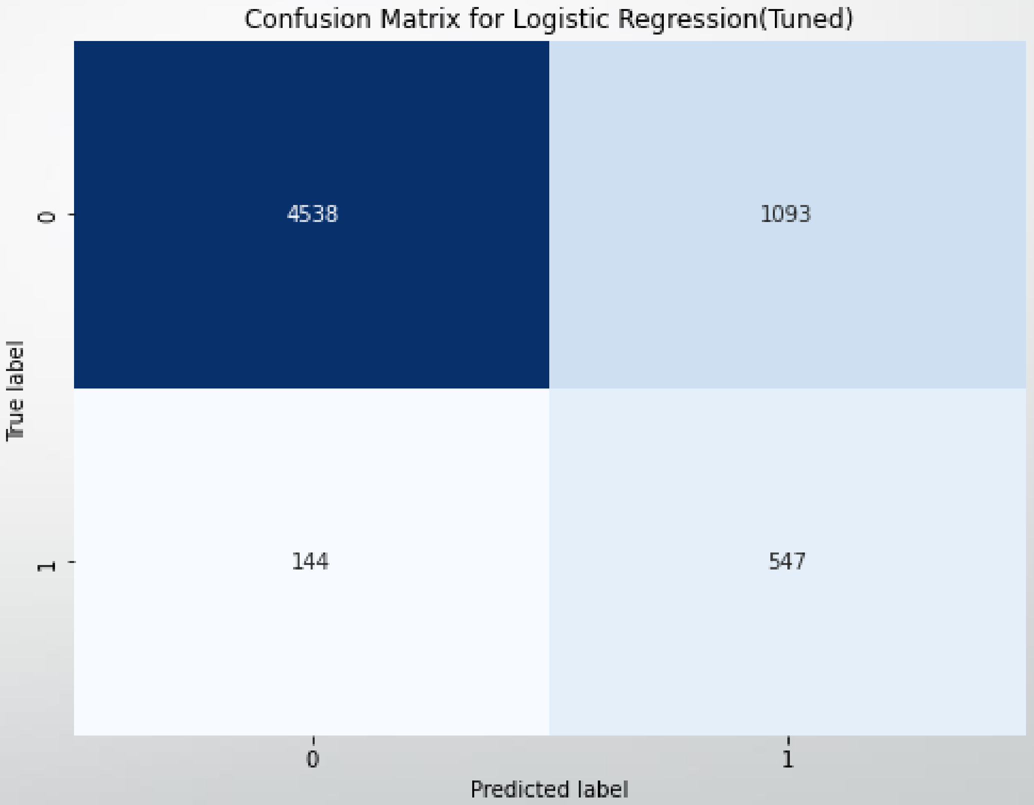

Confusion Matrix for Decision Tree(Tuned)

Logistic Regression Confusion Matrix



Confusion Matrix for Logistic Regression

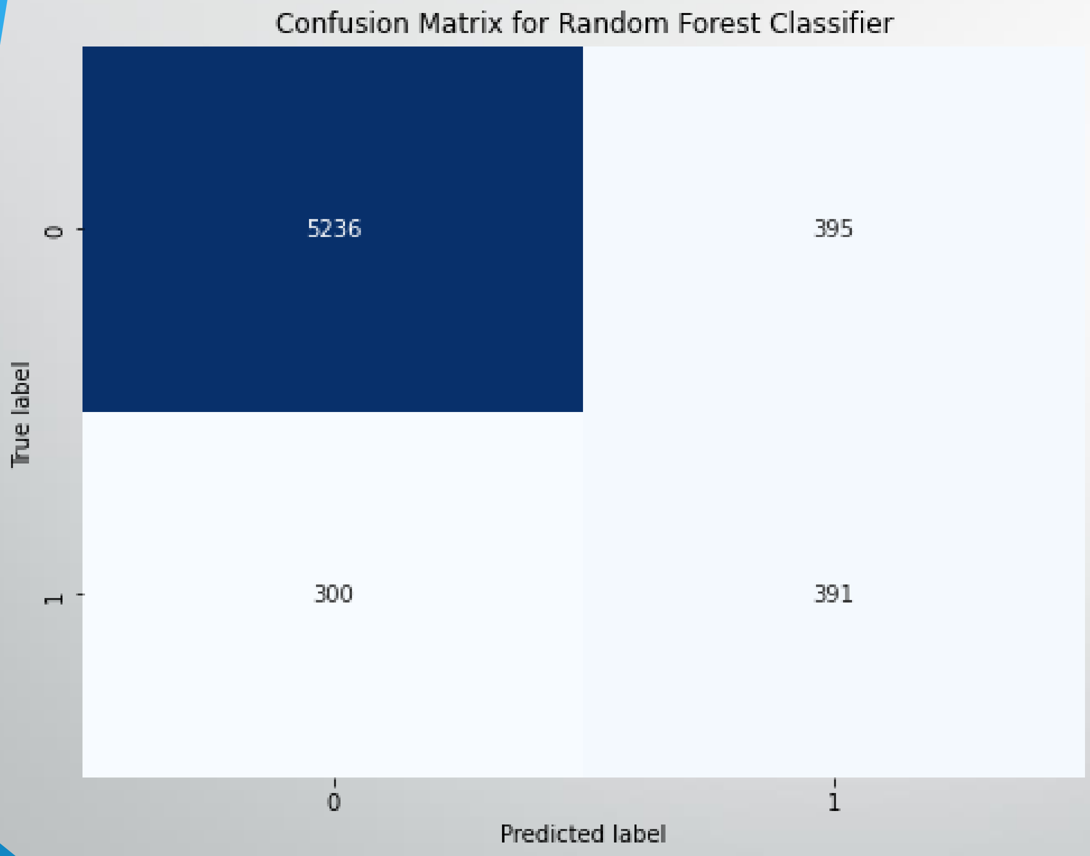
	0	1
0	4548	1083
1	144	547



Confusion Matrix for Logistic Regression(Tuned)

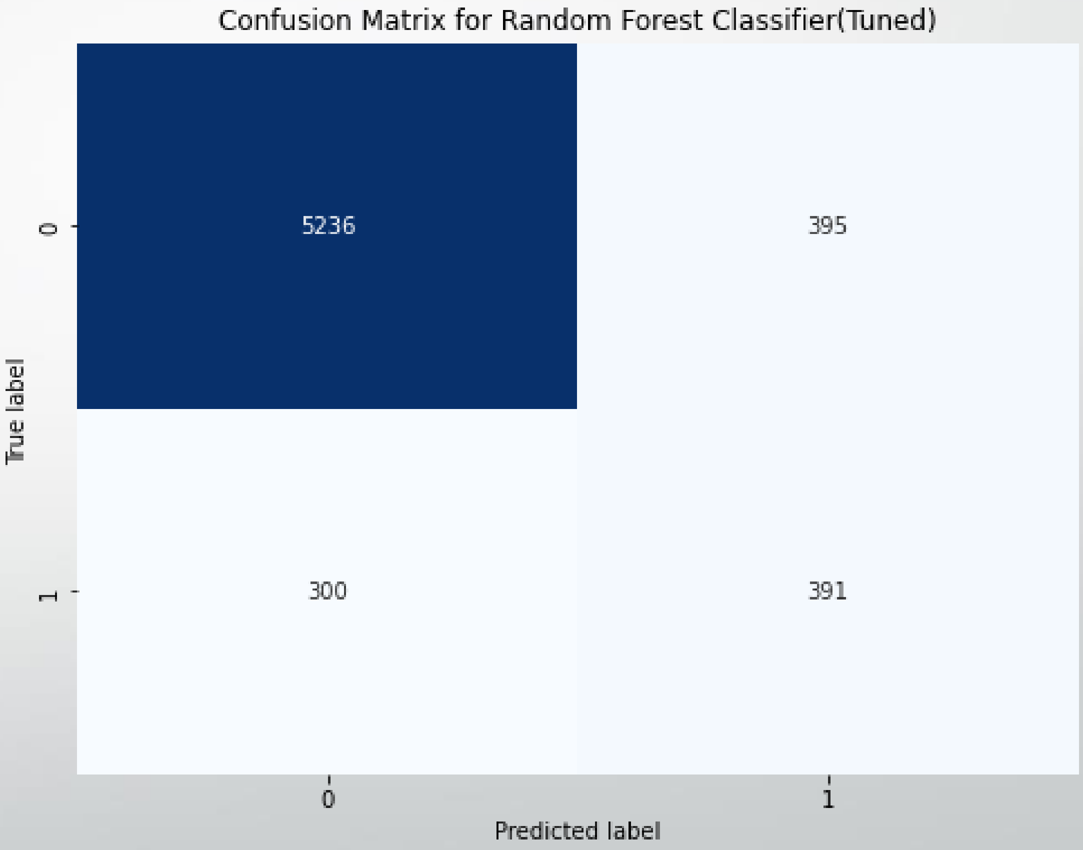
	0	1
0	4538	1093
1	144	547

Random Forest Confusion Matrix



Confusion Matrix for Random Forest Classifier

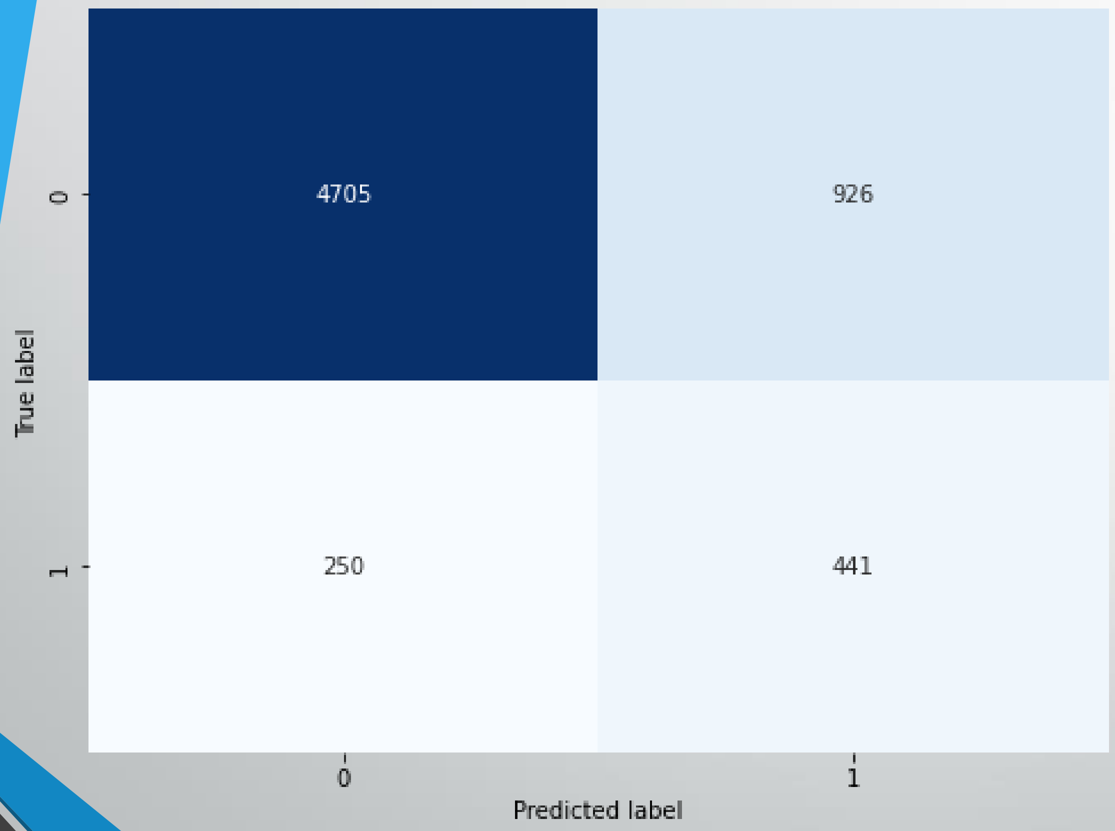
	0	1
0	5236	395
1	300	391

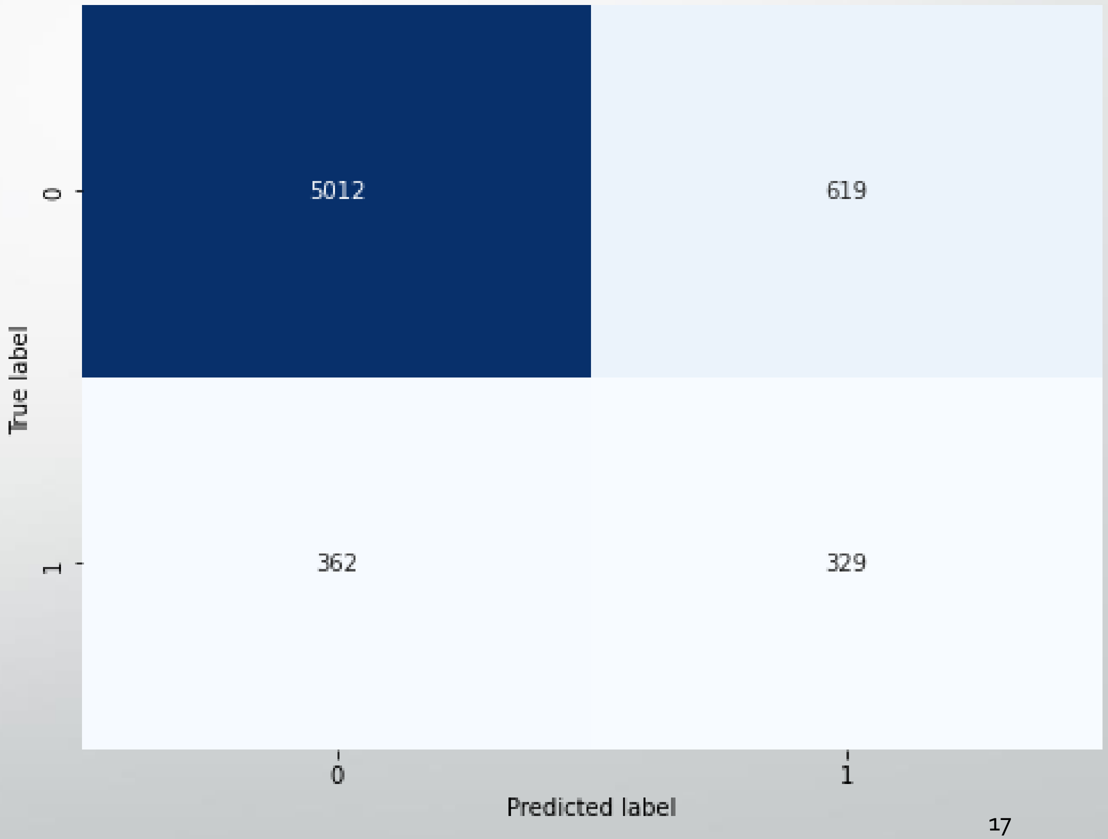


Confusion Matrix for Random Forest Classifier(Tuned)

	0	1
0	5236	395
1	300	391

KNN Confusion Matrix


Confusion Matrix for KNN Classifier


Confusion Matrix for KNN Classifier(Tuned)

Conclusion

- We trained and evaluated five different models, as mentioned in previous slides.
- Random Forest Classifier performed the best on both the training and test sets, with an accuracy of 1.00 on the training set and 0.86 on the test set.
- It also had the highest precision, recall, and F1 scores for the positive class ('yes') on the test set.
- Therefore, the Random Forest Classifier is the best model for this classification problem.
- The machine learning pipeline developed and evaluated in this project could be a useful tool for predicting whether or not a customer will subscribe to a term deposit based on their demographics, previous marketing interactions, and economic indicators.