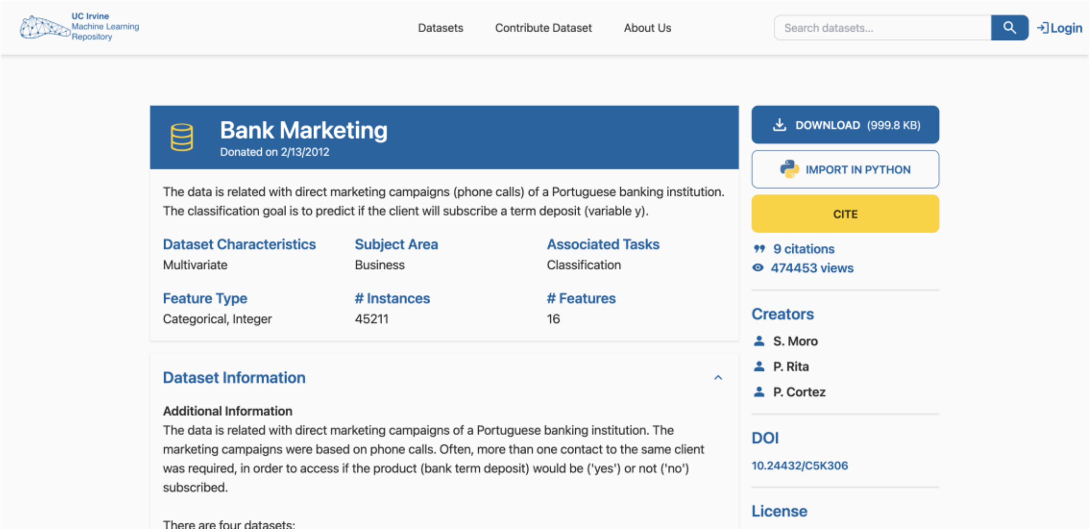
Hackathon #1 Presentation

Shaun Rimos
DI_Bootcamp_176
2025

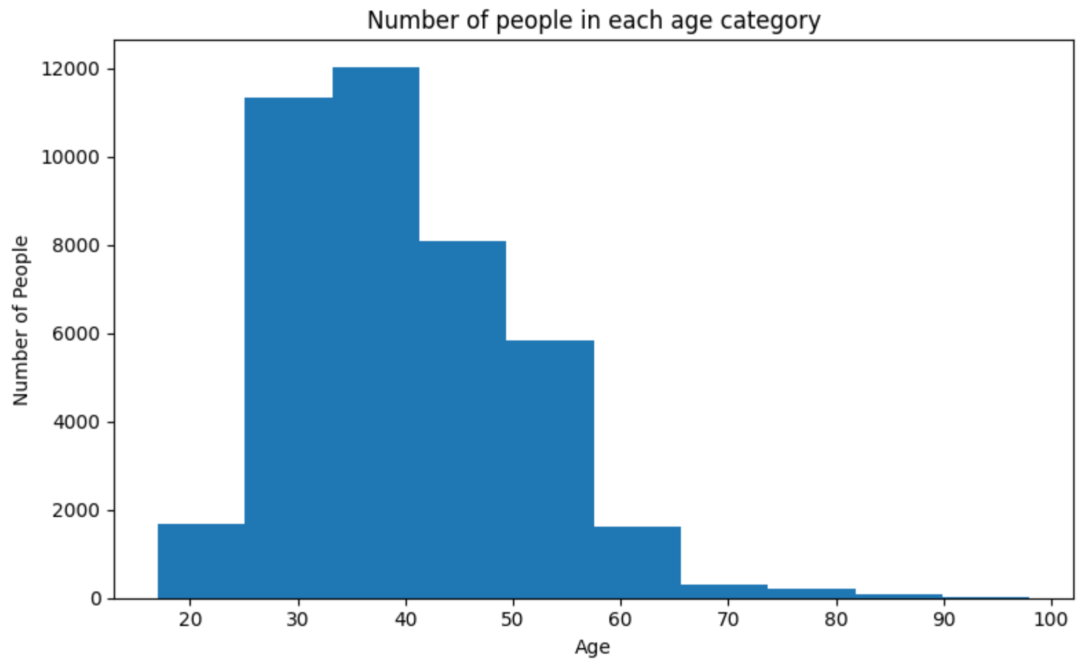
Data Description

- The data is derived from direct marketing campaigns (phone calls) of a Portugese banking institution.
- The classification goal is to predict if the client will subscribe for a term deposit (variable y).
- The newer version of the dataset 'bank-additional-full.csv' was used.



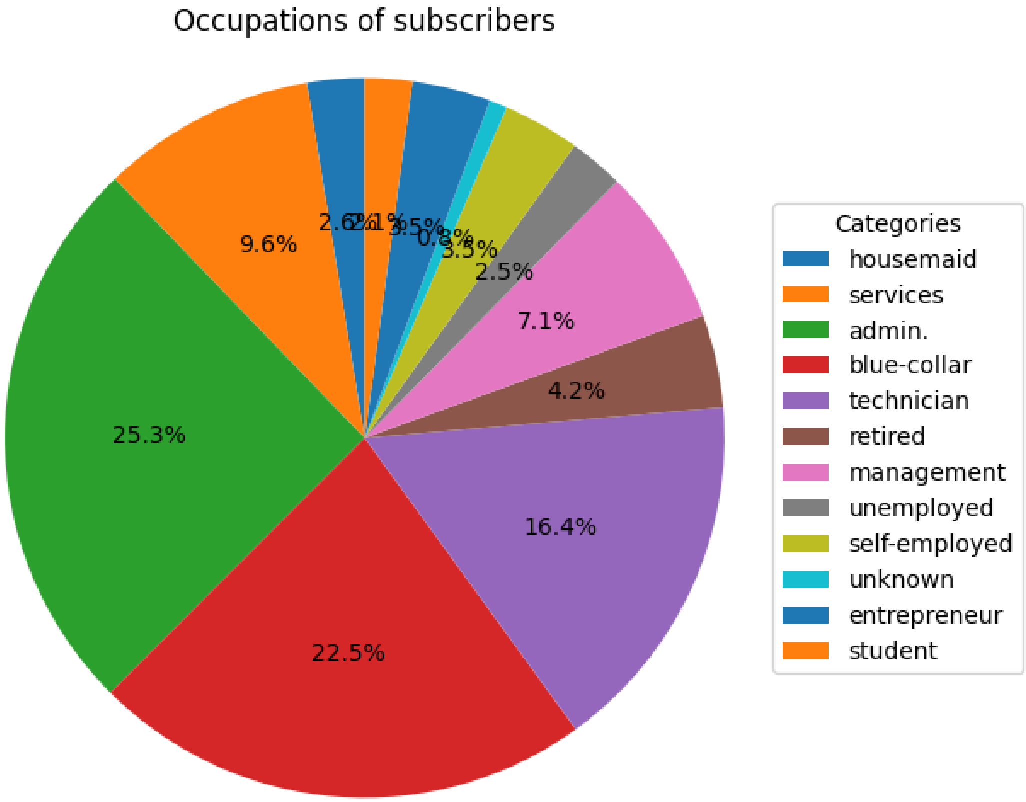
Customer Segmentation

- There are 41,188 clients in the dataset who were involved in the current campaign.
- The majority of the clients are working adults with:
  - 17.58% of clients aged 21-30
  - 39.78% of clients aged 31-40
  - 22.43% of clients aged 41-50
- Together these top 3 make up 79.79% of the whole set.



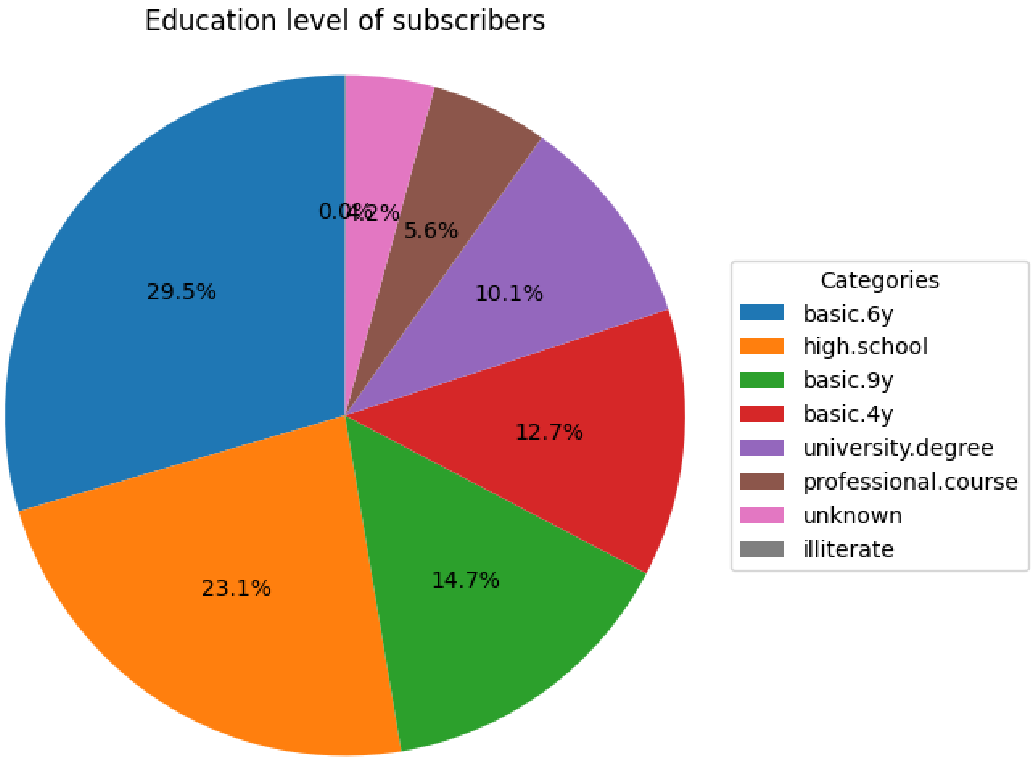
Customer Segmentation

- Out of the 41,188 clients in the dataset who were involved in the current campaign:
  - 25.3% are in admin
  - 22.5% are blue collar workers
  - 16.4% are technicians
- Together these top 3 make up 64.2% of the whole set.



Customer Segmentation

- Out of the 41,188 clients in the dataset who were involved in the current campaign:
  - 29.5% completed a 6yr basic education
  - 23.1% finished high school
  - 14.7% completed a 9yr basic education
- Together these top 3 make up 67.3% of the whole set.



Customer Segmentation

- Most clients were contacted at least 2-3 times during the campaign.
- Many clients had 999 in the 'pdays' category which may skew the data.
- Majority of clients in current campaign (86.34%) did not participate in the previous campaign.

	age	duration	campaign	pdays	previous
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963
std	10.42125	259.279249	2.770014	186.910907	0.494901
min	17.00000	0.000000	1.000000	0.000000	0.000000
25%	32.00000	102.000000	1.000000	999.000000	0.000000
50%	38.00000	180.000000	2.000000	999.000000	0.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000
max	98.00000	4918.000000	56.000000	999.000000	7.000000

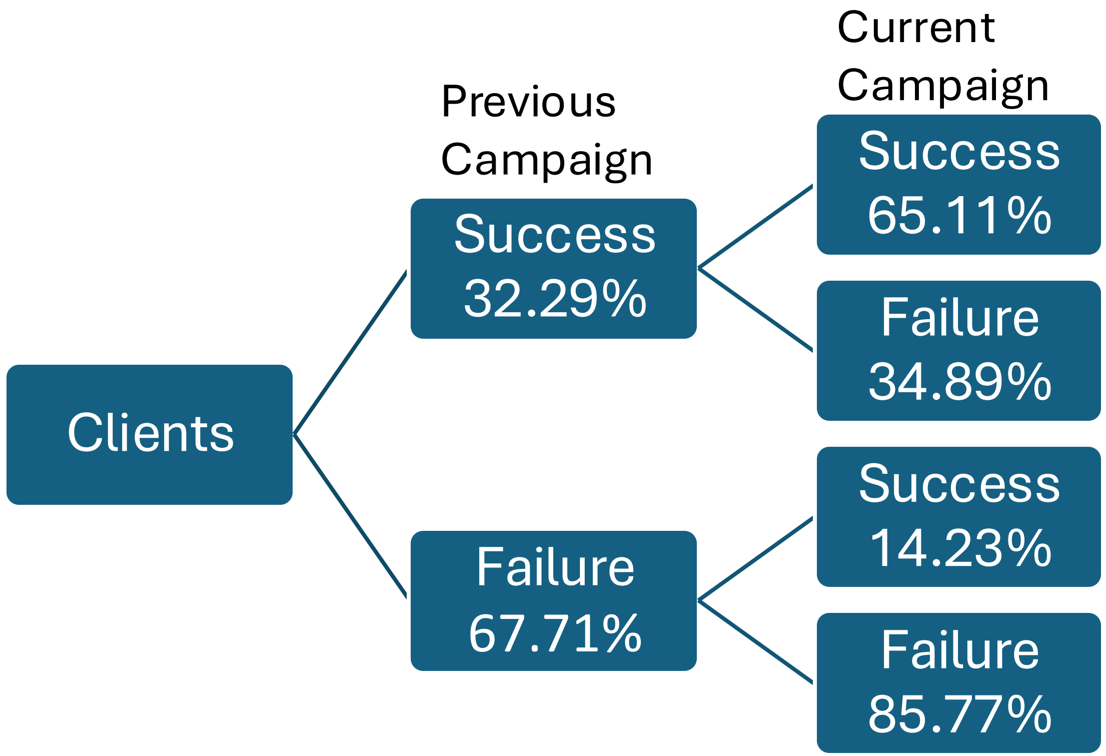
Campaign Effectiveness Analysis (Previous)

- The previous campaign was held over a smaller dataset (5,625 clients).
- The success rate of the previous campaign was 32.29%.
- The current campaign had a bigger dataset (41,188 clients) but a **lower success rate of 11.3%** (4,640 subscribers)

```
poutcome
nonexistent    35563
failure         4252
success         1373
Name: count, dtype: int64
```

Campaign Effectiveness Analysis

- In the previous campaign, 5,625 clients were contacted and 1,373 subscribed.
- From these 1,373 subscribers from the previous campaign, 894 renewed their subscription.
- 605 clients who previously rejected a subscription decided to subscribe in the current campaign (conversion rate = 14.23%)
- Out of the original clients (5,625), the previous campaign produced 1,373 subscribers and the current campaign produced 1,499 subscribers.



Insight: Previous clients who had subscribed to a term deposit in the previous campaign were more likely to subscribe again in the current campaign.

Campaign Effectiveness Analysis (Previous)

- In the previous campaign, the top 3 job categories with the most subscribers are:
  - admin. (428)
  - technician (211)
  - retired (158)
- On the other hand, the top 3 job categories with the highest success rates are:
  - student (41.99%)
  - retired (40.72%)
  - unemployed (39.74%)

poutcome	failure	success	success_rate
job			
admin.	1091	428	0.281764
blue-collar	886	119	0.118408
entrepreneur	154	25	0.139665
housemaid	74	38	0.339286
management	331	95	0.223005
retired	230	158	0.407216
self-employed	145	30	0.171429
services	448	70	0.135135
student	163	118	0.419929
technician	618	211	0.254524
unemployed	94	62	0.397436
unknown	18	19	0.513514

Campaign Effectiveness Analysis (Previous)

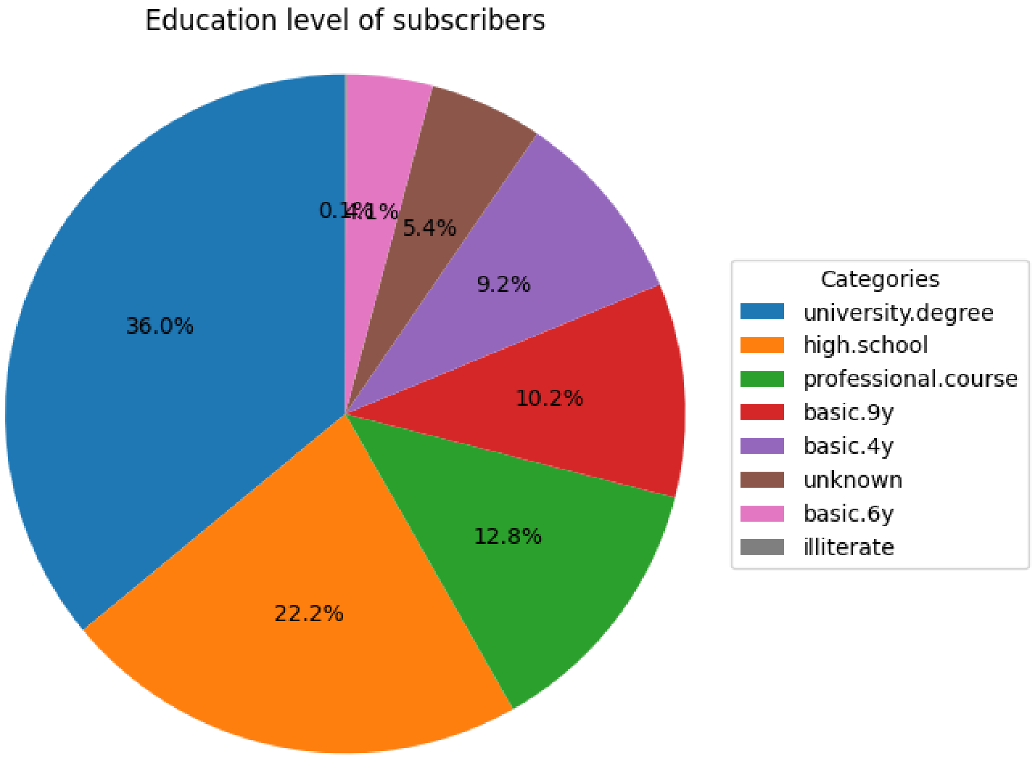
- Marital status
  - For all job categories, married clients make up the majority of the subscribers compared to divorced clients and single clients
  - The sole exception are the student clients who are mostly single

marital	divorced	married	single	unknown
job				
admin.	132.0	652.0	566.0	2.0
blue-collar	53.0	421.0	161.0	3.0
entrepreneur	14.0	88.0	21.0	1.0
housemaid	16.0	74.0	16.0	NaN
management	39.0	226.0	63.0	NaN
retired	92.0	329.0	12.0	1.0
self-employed	16.0	82.0	51.0	NaN
services	33.0	166.0	124.0	NaN
student	3.0	8.0	264.0	NaN
technician	65.0	384.0	279.0	2.0
unemployed	10.0	86.0	48.0	NaN
unknown	3.0	16.0	15.0	3.0

Insight: Except for students, married clients have a higher likelihood to subscribe for a term deposit.

Campaign Effectiveness Analysis (Previous)

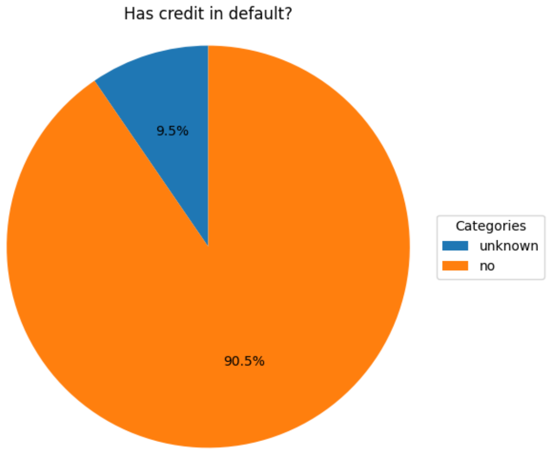
- Education level
  - 36% of subscribers have at least a university degree. Many of them work in admin., management, as technicians or entrepreneurs, or are self-employed.
  - 22.2% of subscribers have finished high school. These include students and those working in services.
  - 12.8% of subscribers completed a professional course and the majority are technicians.
  - The majority of clients who completed 9yr and 6yr basic education are blue collar workers.
  - 9.2% completed only 4yr basic schooling and the majority are retired and are blue collar workers.
  - Very few illiterate clients subscribed for a term deposit

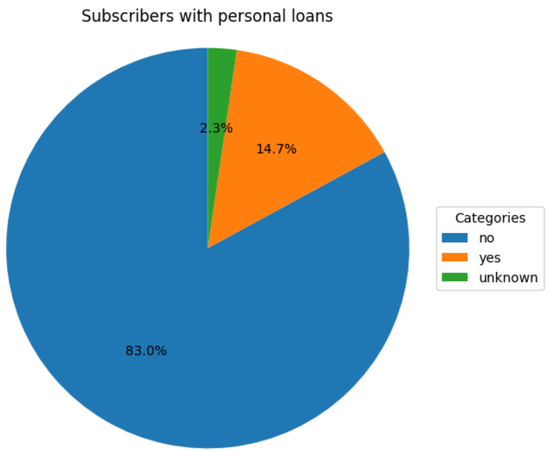


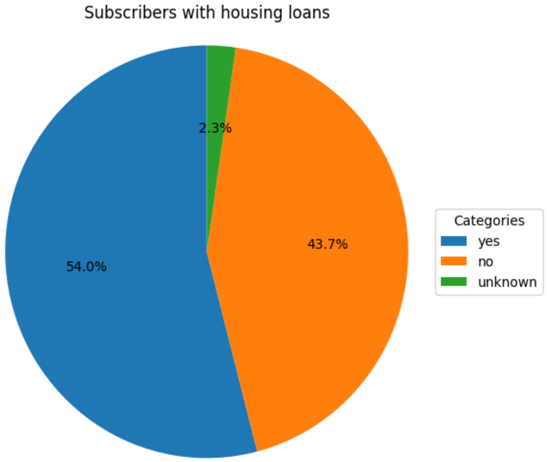
Insight: In general, the higher the education level, the more likely the client will subscribe to a term deposit.

Campaign Effectiveness Analysis (Previous)

- Housing loan
  - 54% of subscribers have a housing loan
- Personal loan
  - 83% of subscribers have a personal loan
- Default credit
  - 90.5% of subscribers have no credit in default, the rest is unknown







Insight: In general, all subscribers have a loan of some kind, but almost all of them have no credit in default, which is a determining factor whether they subscribed for a term deposit or not.

Campaign Effectiveness Analysis (Conclusion)

- The current campaign has a lower success rate of 11.3% compared to the previous campaign (32.3%).
- The current campaign was directed to:
  - a large number of clients with low education levels (below university degree)
  - a large number of blue-collar workers.
- The previous campaign showed that these clients were less likely to subscribe for a term deposit.
- The current campaign also failed to capitalise on students and retired clients who have higher success rates.

Predictive Modelling

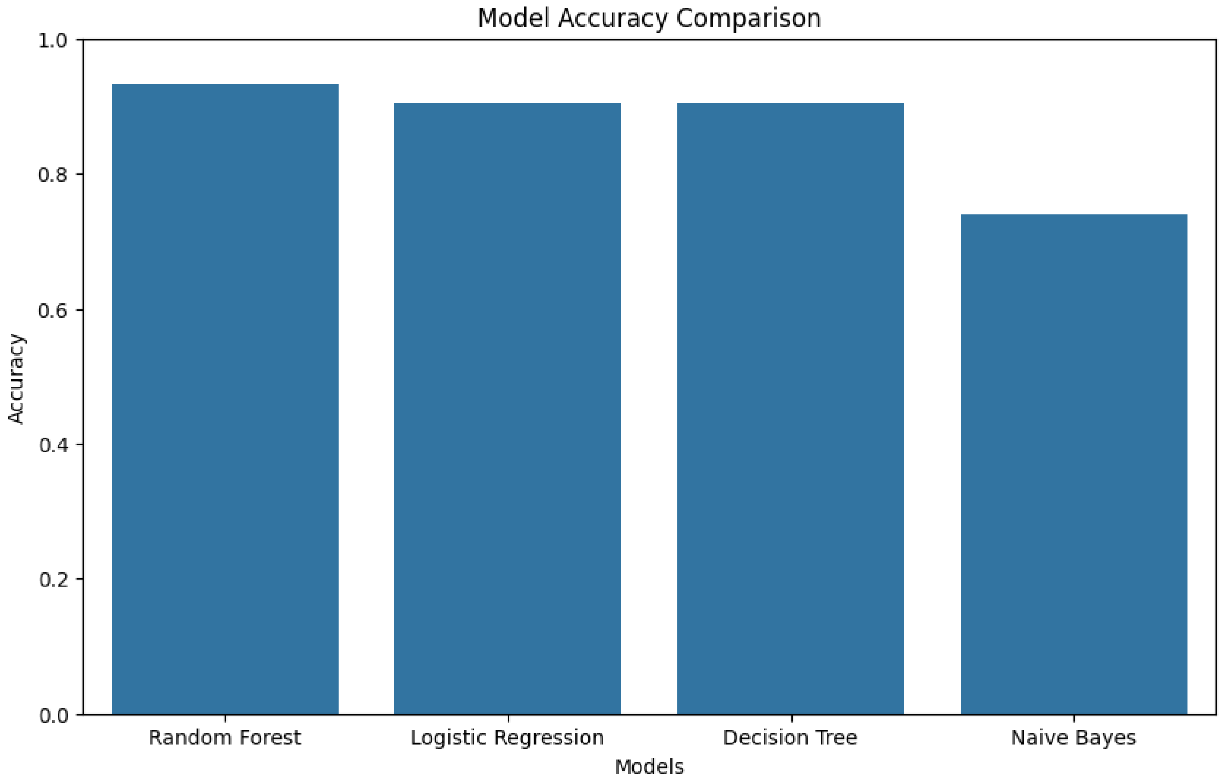
- 4 classification models are tested using the data:
    - Logistic Regression
    - Random Forest Classifier
    - Naïve Bayes (MultinomialNB)
    - Decision Tree Classifier

Predictive Modelling

- Preprocessing the data
  - The 'age' category was categorised into bins of '1-20', '21-40',...
  - One hot encoding was applied to categorical columns, e.g. 'job'
  - Columns which have no impact were dropped, e.g. 'unknown' values
  - 999 in 'pdays' was replaced with '0'
  - Numerical columns were standardised using Standard Scaler, e.g. 'emp.var.rate'
  - Oversampling was used to address the issue of imbalance as the number of 'no' responses far outweighed the number of 'yes' responses.

Predictive Modelling

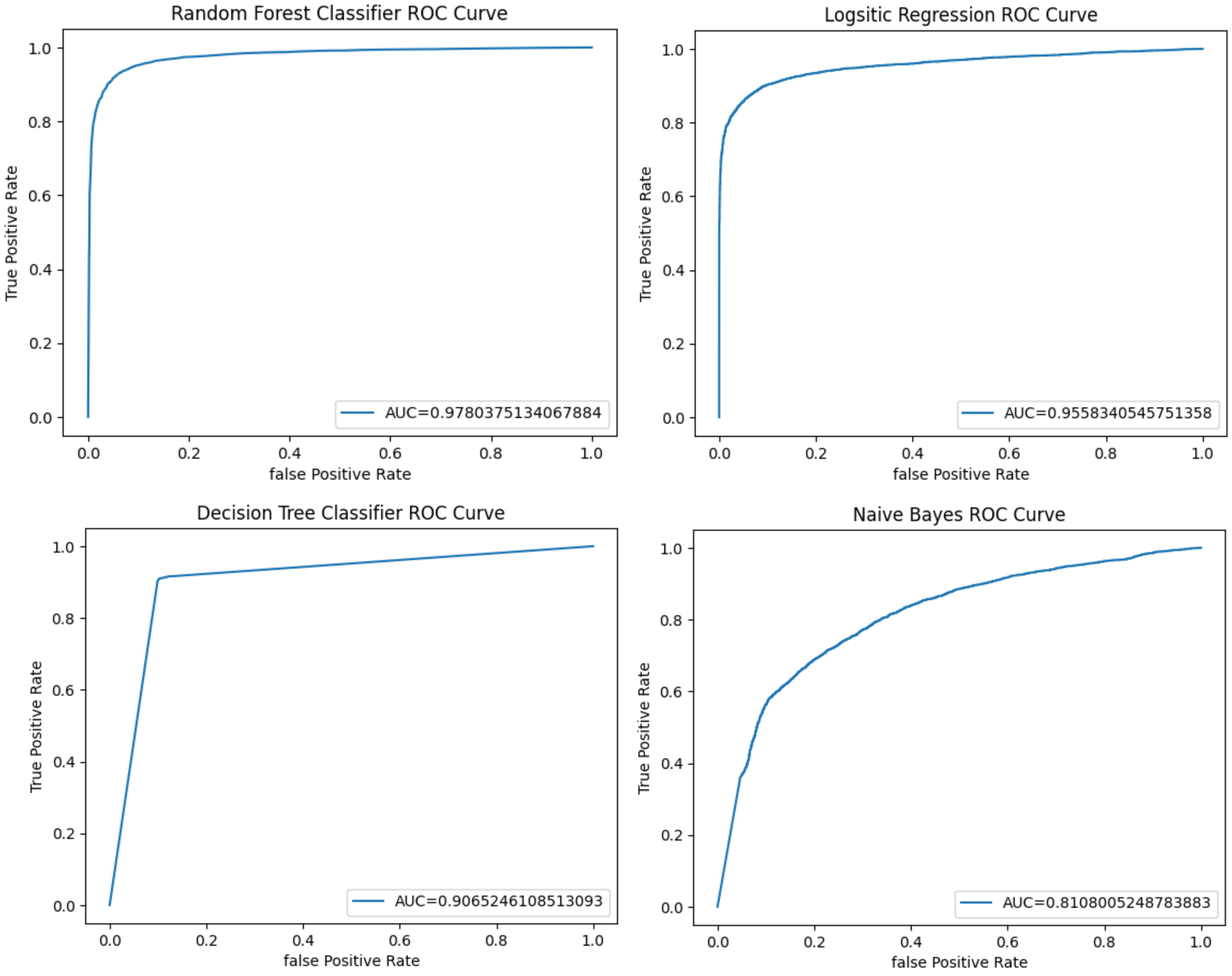
- The Random Forest Classifier model had the highest accuracy and f1-score (93%) out of the 4 models.
- Both the Logistic Regression model and the Decision Tree Classifier model have accuracy and f1-scores of 90%
- The Naïve Bayes model scored the lowest with 0.74 for accuracy and f1-score.



	Accuracy	Precision	Recall	f1-score
Random Forest	0.93	0.93	0.94	0.93
Logistic Regression	0.90	0.94	0.86	0.90
Decision Tree	0.90	0.90	0.91	0.90
Naïve Bayes	0.74	0.74	0.74	0.74

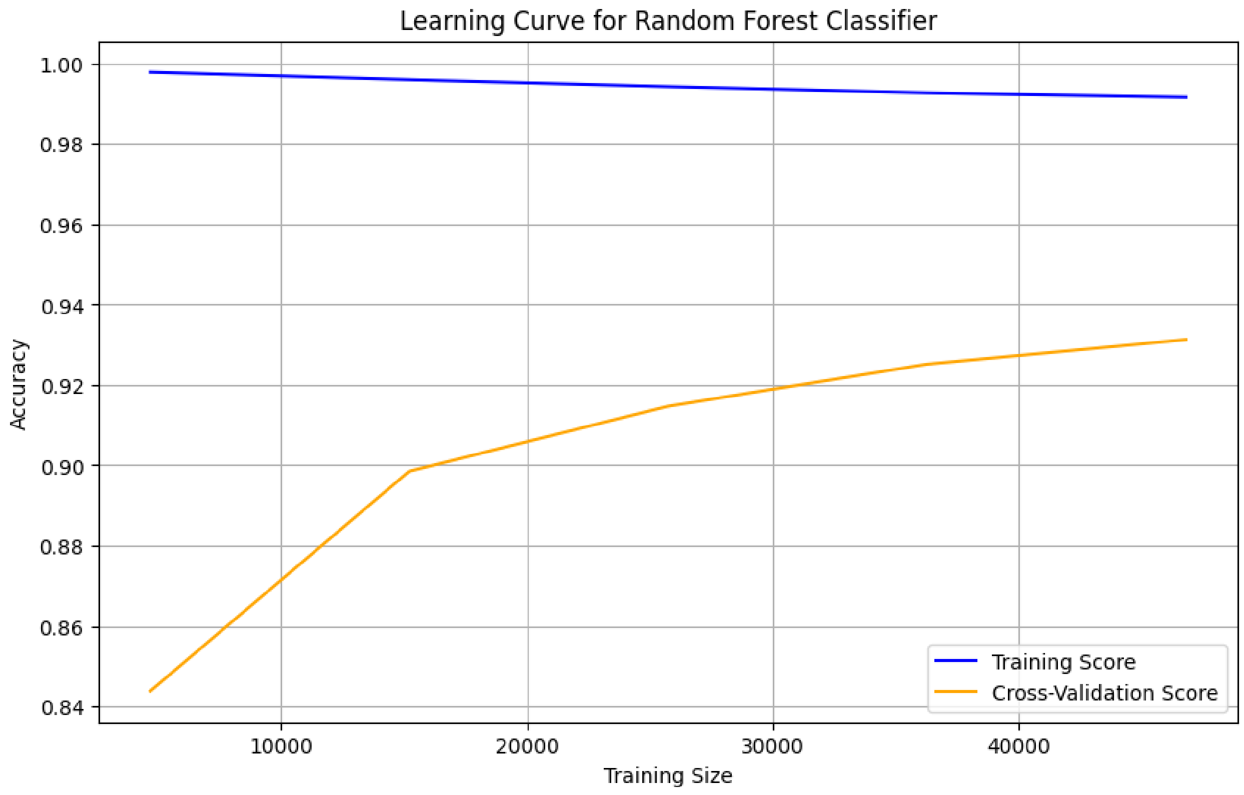
Predictive Modelling

- ROC Curves were plotted for all 4 models.
- The Random Forest Classifier model had the highest ROC AUC score



Predictive Modelling

- The Random Forest Classifier model was chosen as the best model to be used to predict customer response to future marketing campaigns
- A visualisation of the learning curve for the Random ForestClassifier model was plotted.



Predictive Modelling (Conclusions)

- The Random Forest Classifier model scored the highest (93%) in accuracy, precision and recall out of the 4 models tested.
- After preprocessing, the dataset contained a total of 61 columns. This was due to the large number of categories produced after one hot encoding of the categorical columns, e.g. job, education…
- The model initially faced imbalance issues ('no' outnumbered 'yes') which had to be rectified using oversampling.
- The model cannot be applied directly to real-life data. The data will need to undergo the same preprocessing as the dataset used for training and testing before the model can be used.
- The model can be further refined using boosting and bagging methods.

Feature Importance and Interpretability

- Using SelectKBest for feature selection, the top 3 features were:
  - 'pdays' - number of days that passed by after the client was last contacted from a previous campaign
  - 'previous' - number of contacts performed before this campaign and for this client
  - 'p_outcome_success' - outcome of the previous marketing campaign, success only
- This further confirms that the results from the previous campaign have a big impact on the success of future campaigns.

	Features	Score
51	month_sep	645.541017
50	month_oct	763.644573
47	month_mar	842.916583
11	age_bin_61-80	894.106258
7	nr.employed	973.556589
3	emp.var.rate	1096.769360
6	euribor3m	1115.715645
59	poutcome_success	3982.548056
2	previous	4543.394485
1	pdays	9835.989418

Recommendations

- The bank should learn from previous campaigns as they provide valuable data regarding factors that influence customer choices.
- The bank should perform personalised campaigns targeted towards students, graduates, retired and married clients.
  - Examples include higher interest rates, bonuses and special offers.
- The bank should be aware that having a credit in default can be a determining factor whether the client will subscribe to a term deposit (no credit =more likely to subscribe)