Machine Learning Development Using Python:
Project Report _Task 1

By: Jerry TSIBA

Master in Data Sciences, Deakin University, March 2024

Great Learning

Business Context:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Problem Statement:

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Objective:

Build machine learning models based on Decision Tree and Random Forest, compared them in terms of accuracy and some other metrics, provide justification which model is performing better and why. Furthermore, business insights and recommendations are expected.

Dataset source:

Bank Marketing - UCI Machine Learning Repository

Data description

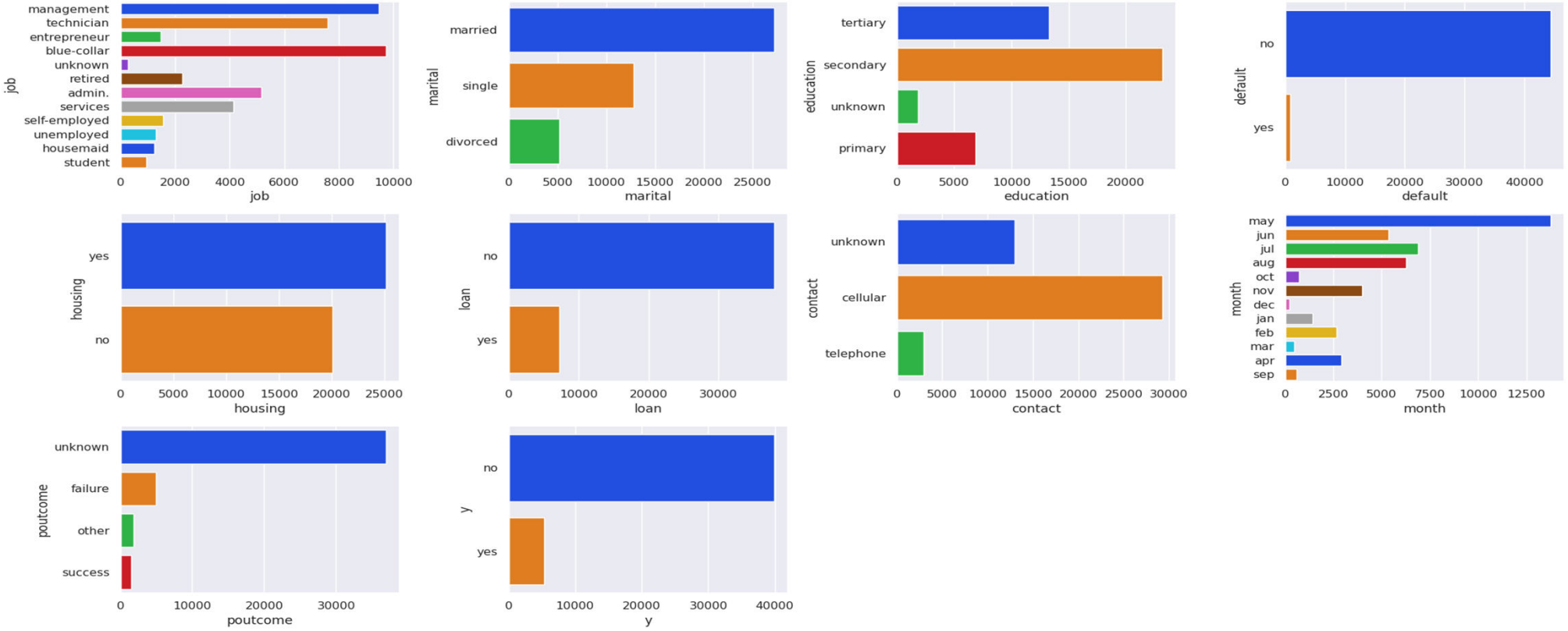
Variable	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	Age			no
job	Feature	Categorical	Occupation	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')		no
marital	Feature	Categorical	Marital Status	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)		no
education	Feature	Categorical	Education Level	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')		no
default	Feature	Binary		has credit in default?		no
balance	Feature	Integer		average yearly balance	euros	no
housing	Feature	Binary		has housing loan?		no
loan	Feature	Binary		has personal loan?		no
contact	Feature	Categorical		contact communication type (categorical: 'cellular','telephone')		yes
day_of_week	Feature	Date		last contact day of the week		

Statistics overview

	count	mean	std	min	25%	50%	75%	max
age	45211.00000	40.93621	10.61876	18.00000	33.00000	39.00000	48.00000	95.00000
balance	45211.00000	1362.27206	3044.76583	-8019.00000	72.00000	448.00000	1428.00000	102127.00000
day	45211.00000	15.80642	8.32248	1.00000	8.00000	16.00000	21.00000	31.00000
duration	45211.00000	258.16308	257.52781	0.00000	103.00000	180.00000	319.00000	4918.00000
campaign	45211.00000	2.76384	3.09802	1.00000	1.00000	2.00000	3.00000	63.00000
pdays	45211.00000	40.19783	100.12875	-1.00000	-1.00000	-1.00000	-1.00000	871.00000
previous	45211.00000	0.58032	2.30344	0.00000	0.00000	0.00000	0.00000	275.00000

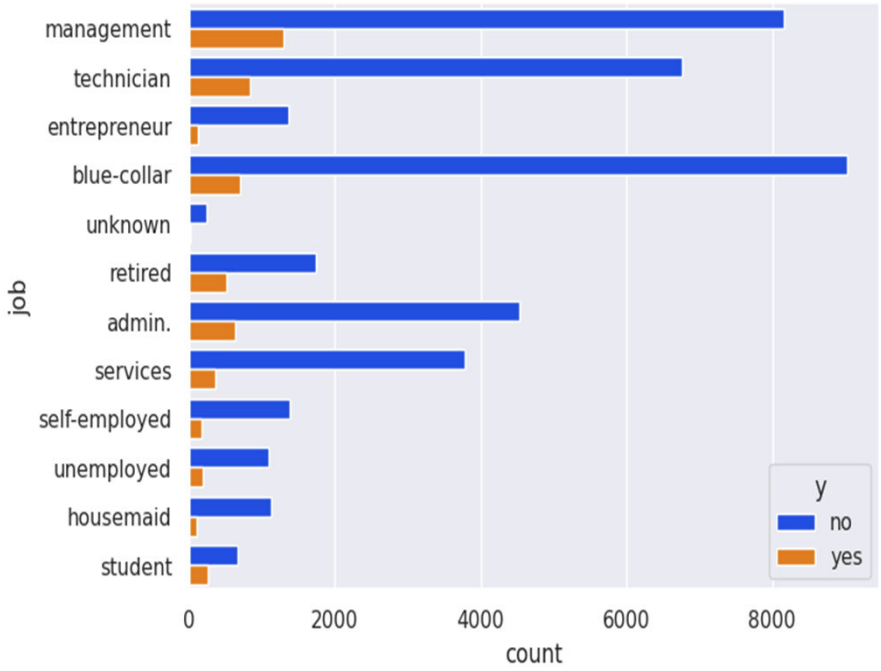
The mean age of candidate contacted is 41 years old, and the minimum is 18 years old, the maximum is 95 years old.
The average yearly balance is 1362.27 euro
The average number of contacts performed during this campaign is of 2.7 (~ 3 times)

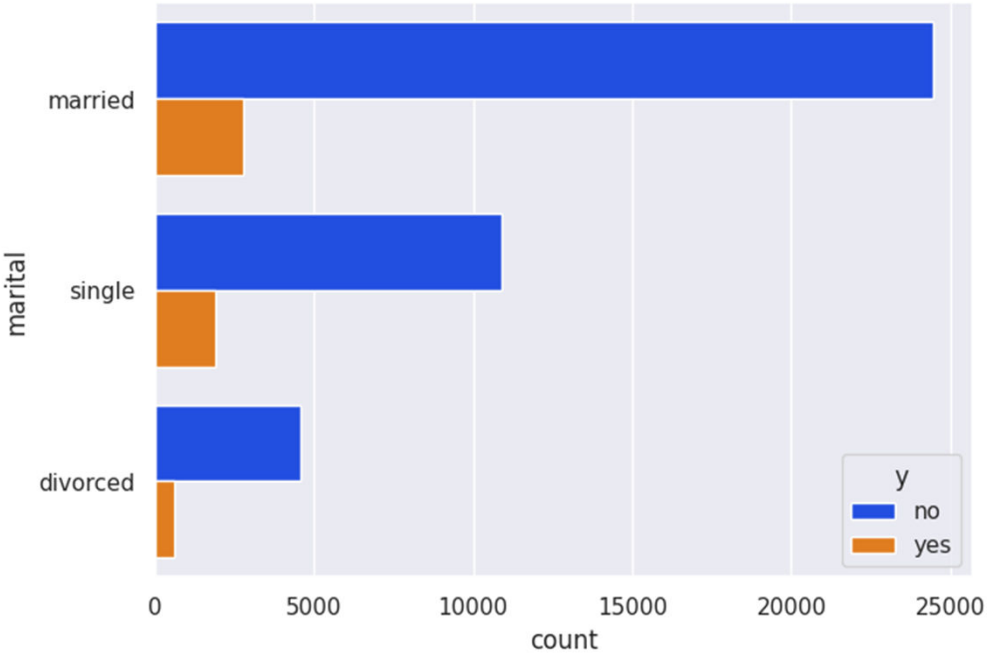
Exploratory Data Analysis



- Most of candidates have not subscribed or deposit,
- Candidates with secondary education level have more deposit, followed by candidates with tertiary then primary

Exploratory Data Analysis

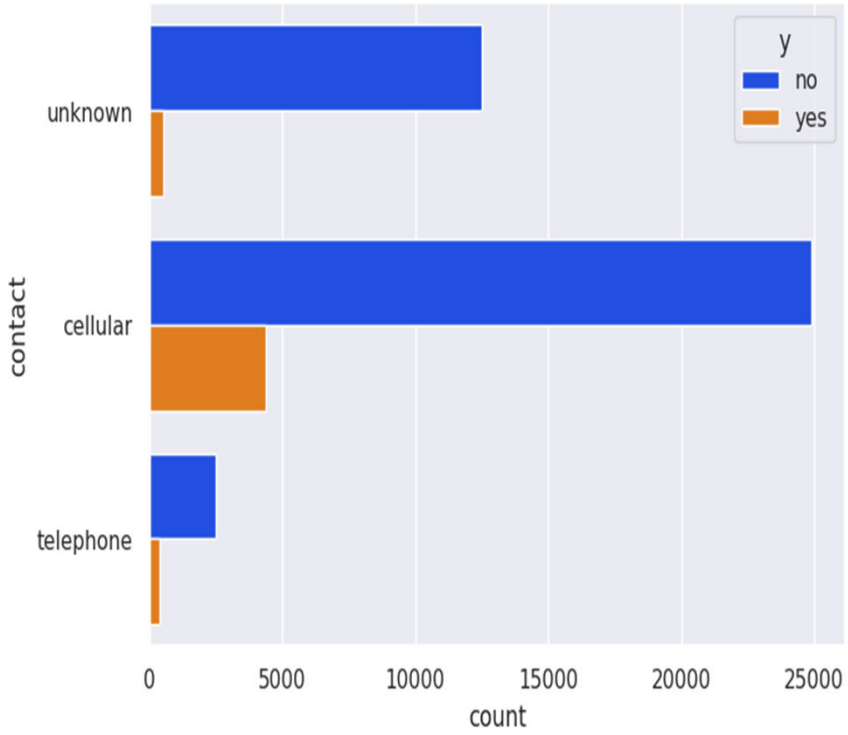


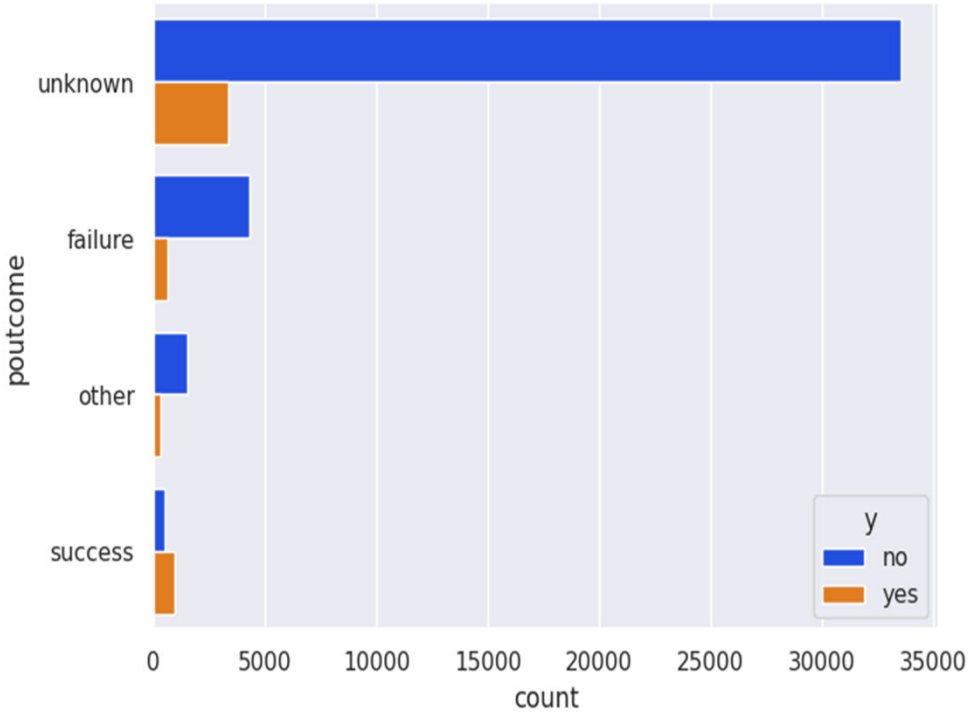


Married are predominant and the majority have not been more convinced, hence have not deposited. However, there have also more deposit compared to single and divorced

High percentage of Blue collar, managers and technicians jobs have not subscribed
- Given its number of actual managers, they also appears to be more convinced and who have mostly deposited compared to the other jobs.

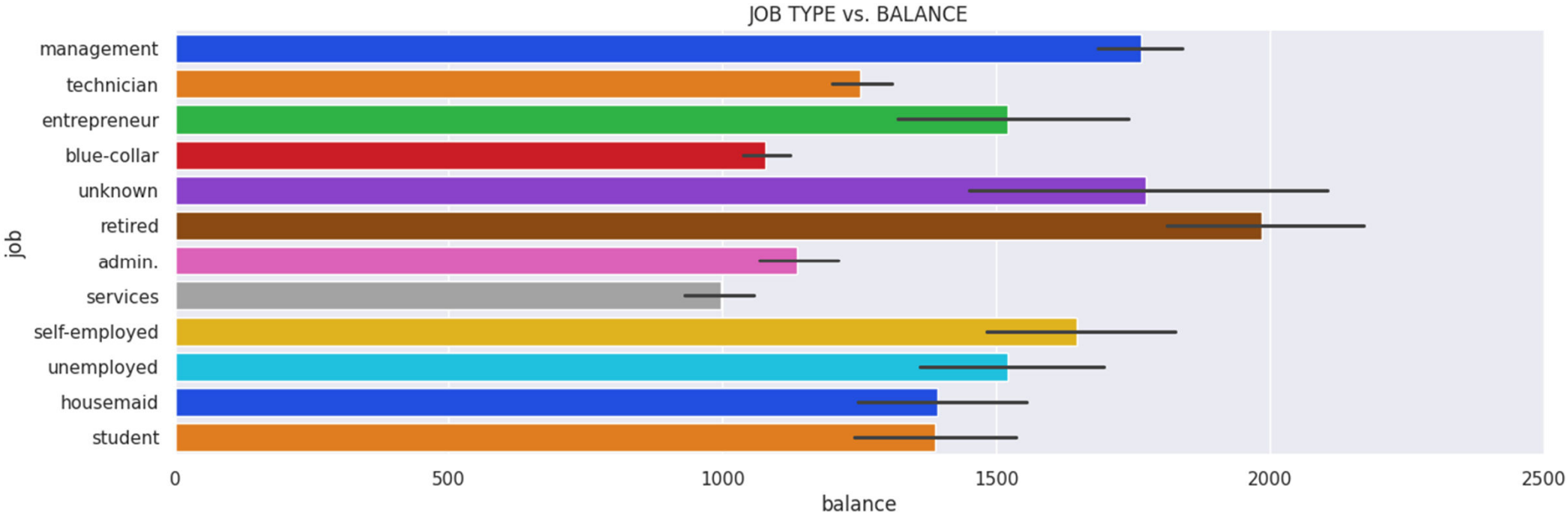
Exploratory Data Analysis





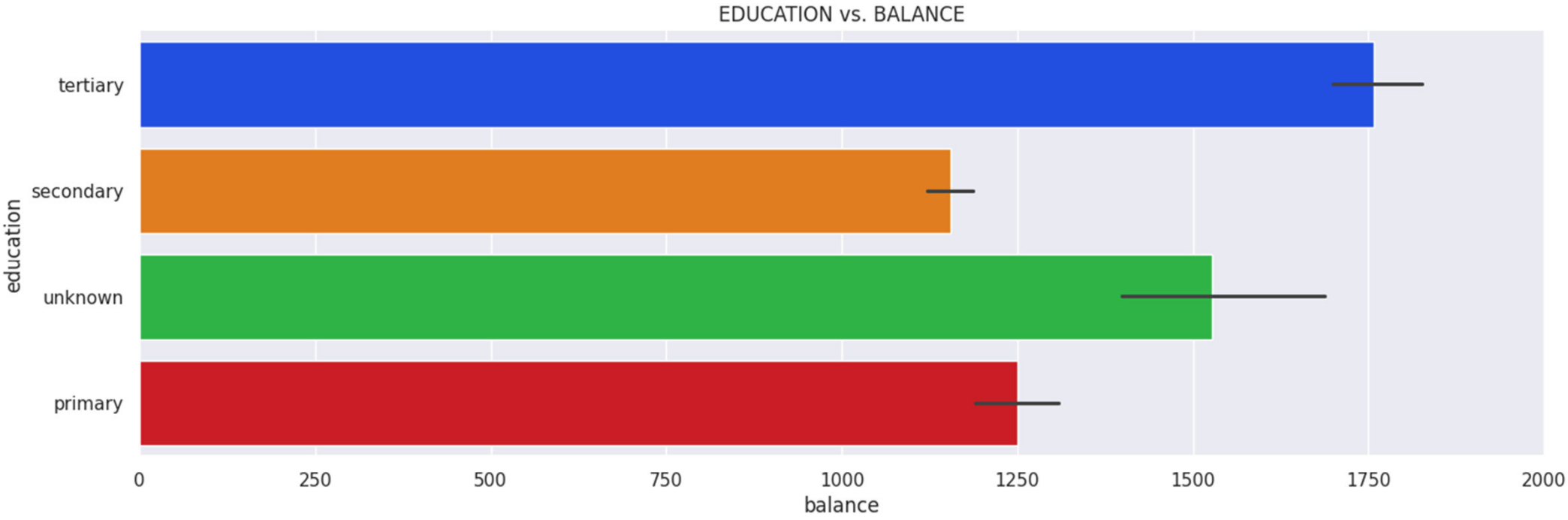
- poutcome is higher from unknown compared to those who succeeded, followed by those who failed
- Most of candidates convinced were contacted via cellular

Exploratory Data Analysis



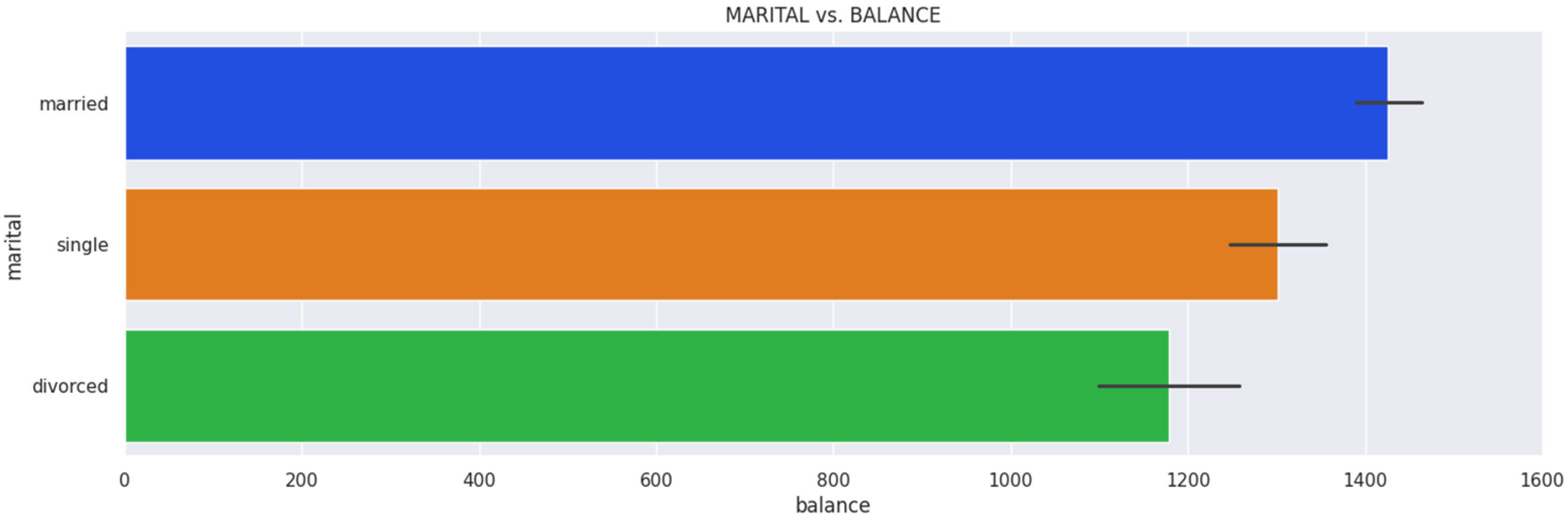
The average yearly salary of candidates in management and unknown category is equally high compared to the others, even though we observed an exception with retired people who have highest balance.

Exploratory Data Analysis



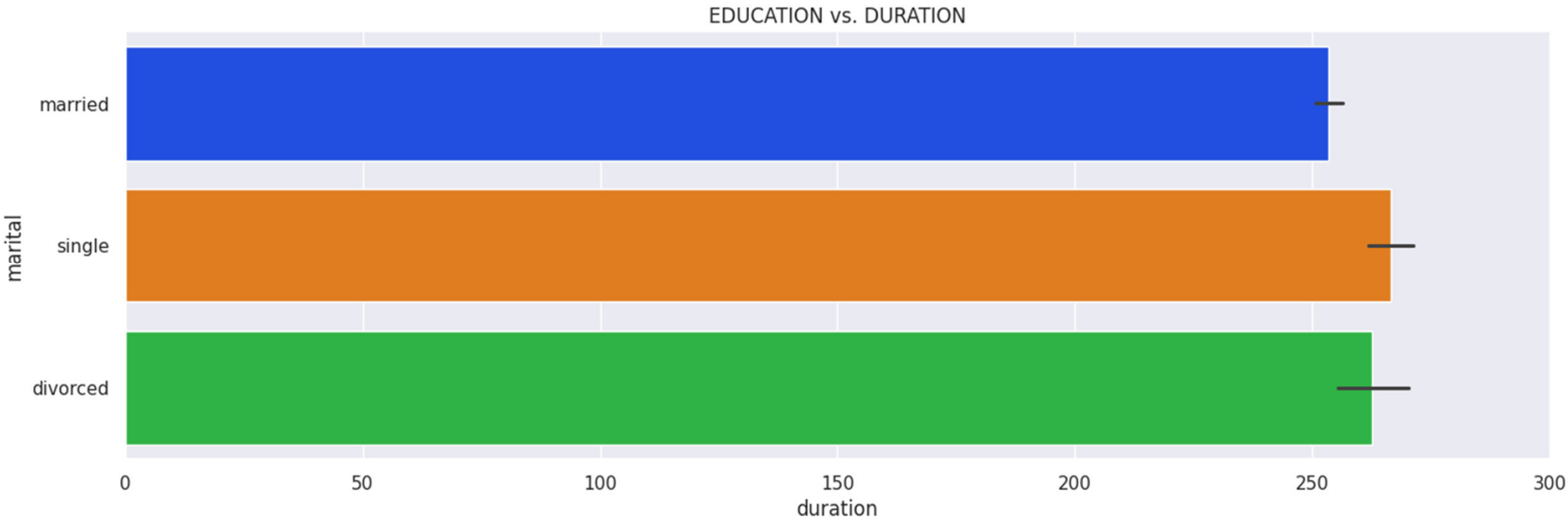
Candidates with tertiary education earn more compared to the others

Exploratory Data Analysis



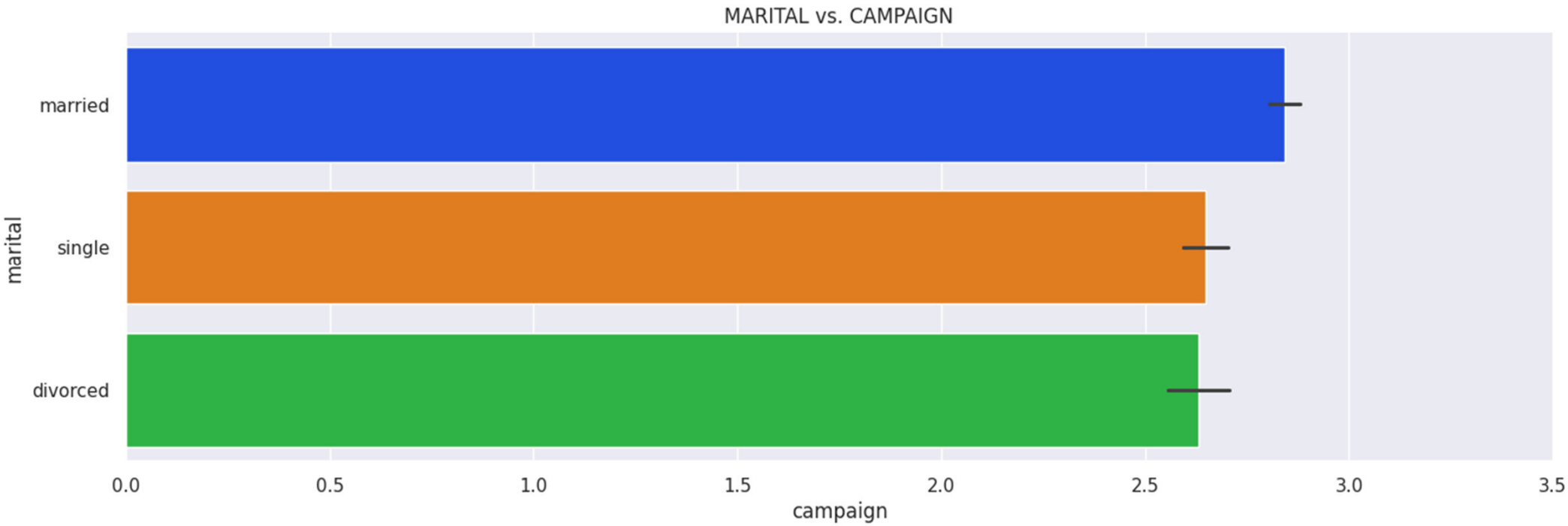
Married candidates have highest average yearly balance compared to single and divorced candidates

Exploratory Data Analysis



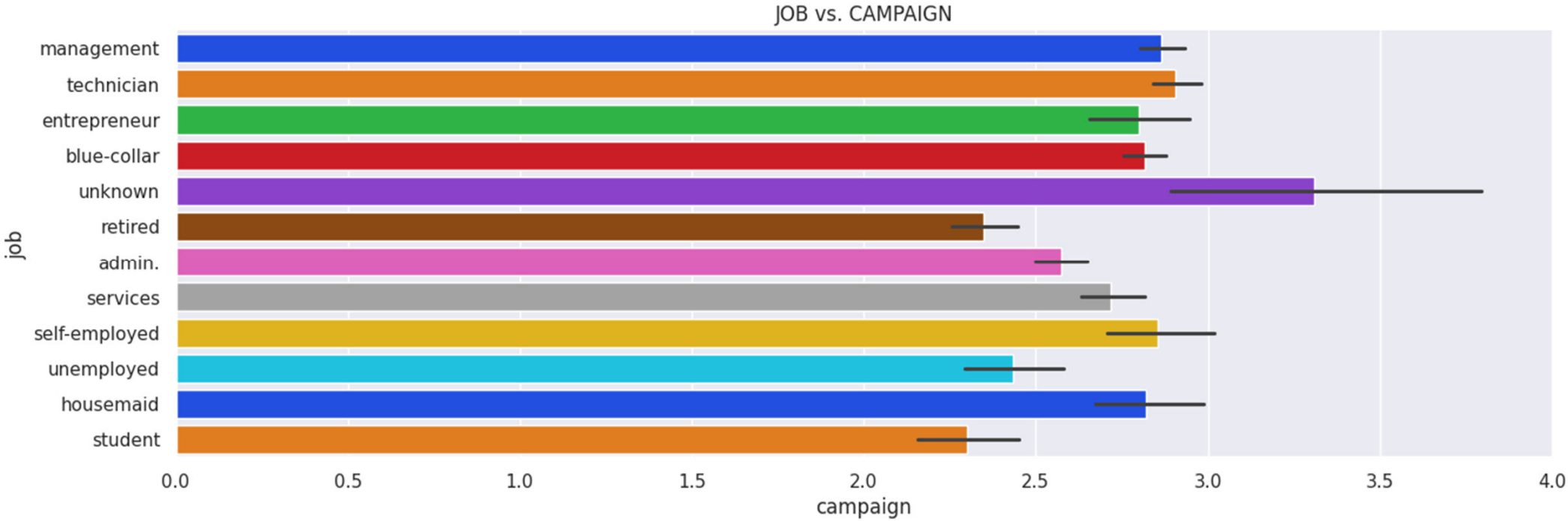
Singles followed by divorced were the most being in contact during the marketing campaign

Exploratory Data Analysis



Married participated the most in the campaign, followed by single then finally divorced. This clearly illustrates that married people care most about savings, deposit etc compared to the others

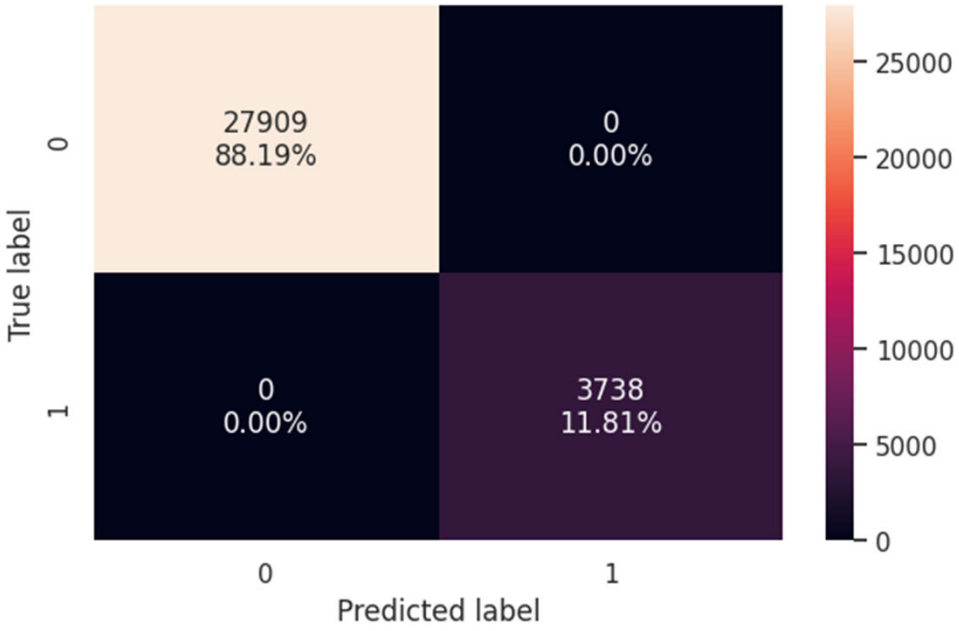
Exploratory Data Analysis



- In term of job categories, unknown, technicians, managers, self-employed, maid participated mostly in the campaign.
- Students and retired have not really participated compared to the others, may be because they likely don't have relative and income.

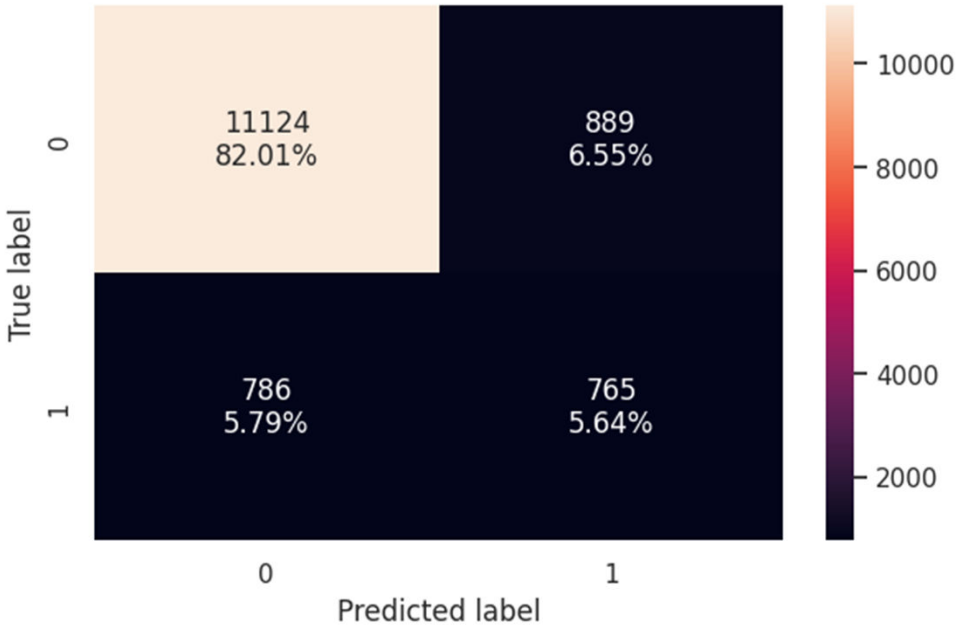
Decision Tree: Sklearn

Train



	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

Test

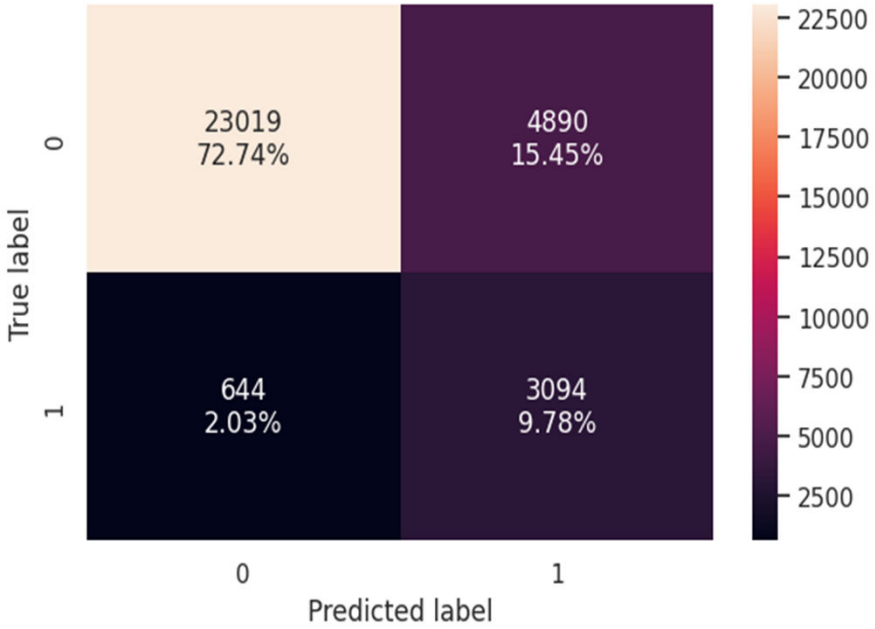


	Accuracy	Recall	Precision	F1
0	0.87651	0.49323	0.46252	0.47738

-Accuracy is high when the model is trained on test dataset
-duration and balance are the most important features

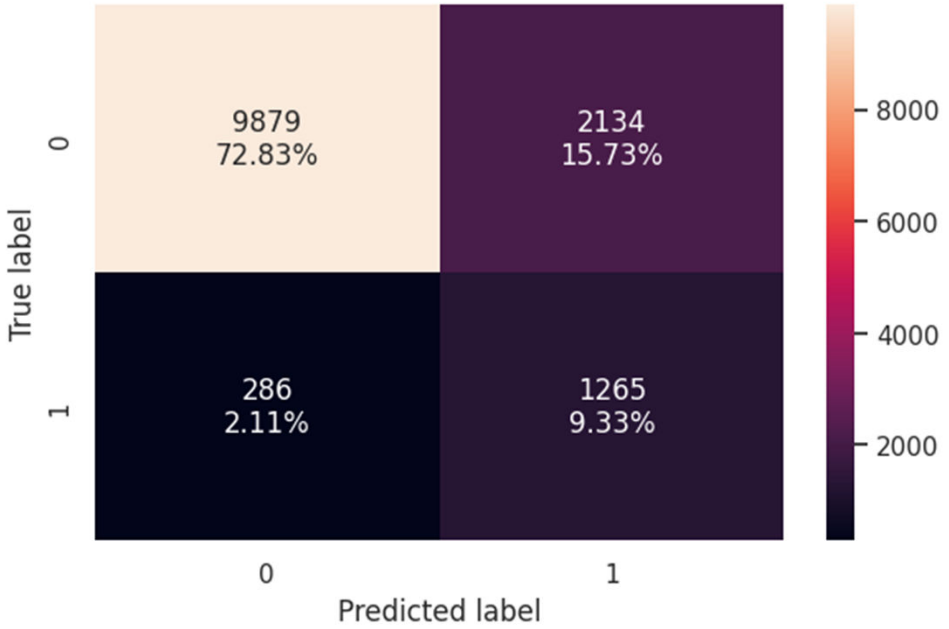
Decision Tree: Pre-pruning

Train



	Accuracy	Recall	Precision	F1
0	0.82513	0.82772	0.38753	0.52790

Test

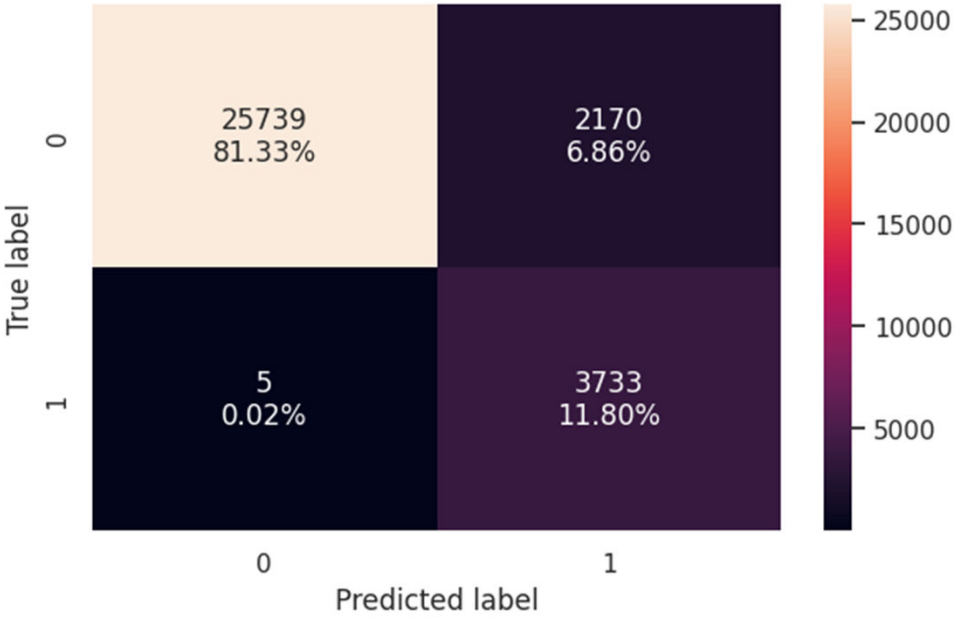


	Accuracy	Recall	Precision	F1
0	0.82159	0.81560	0.37217	0.51111

-Accuracy remains almost the same when the model sees the test dataset
-duration and poutcome are the most important features

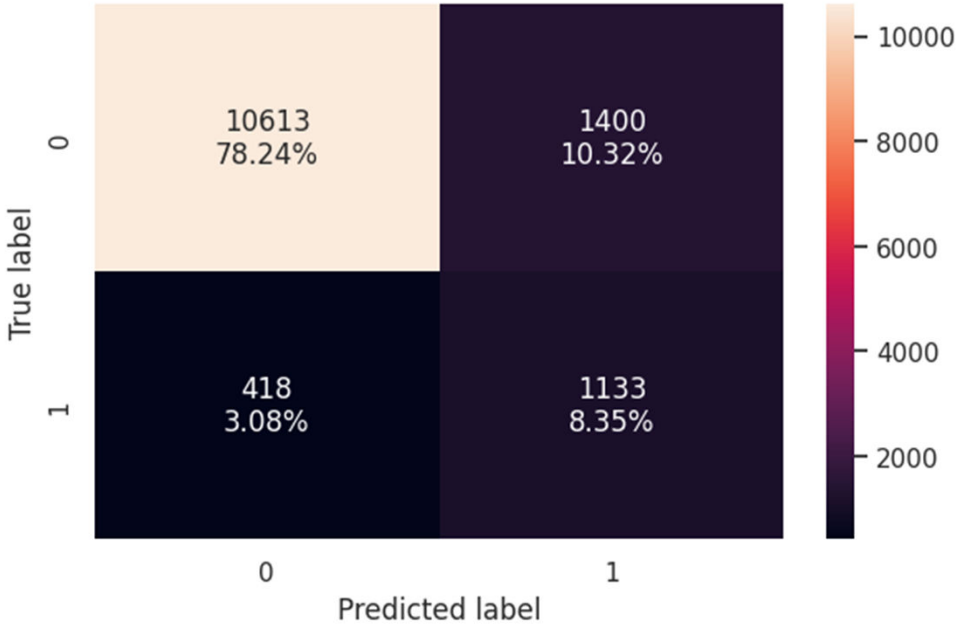
Decision Tree: Post-pruning

Train



	Accuracy	Recall	Precision	F1
0	0.93127	0.99866	0.63239	0.77440

Test

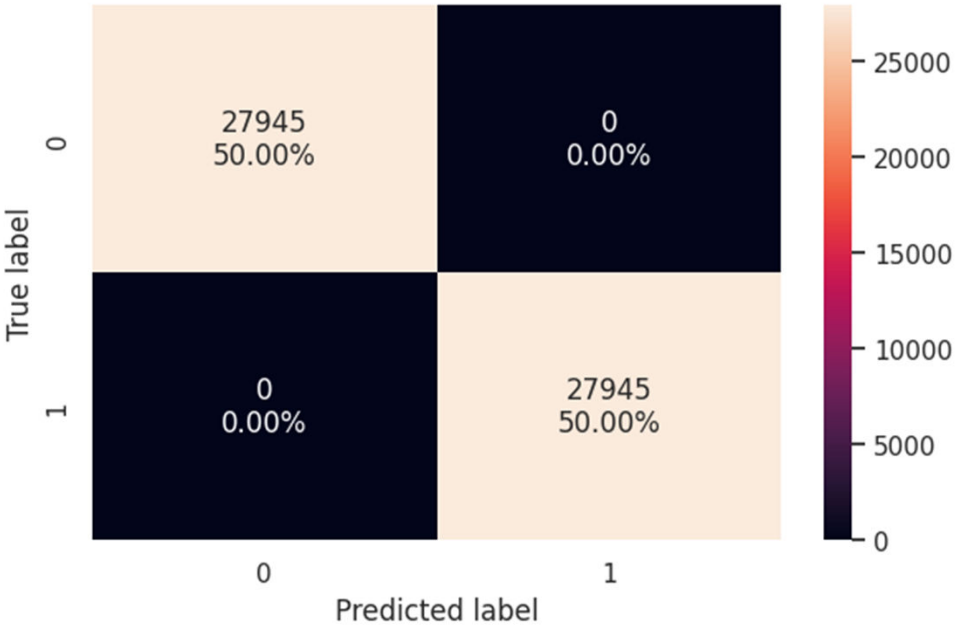


	Accuracy	Recall	Precision	F1
0	0.86597	0.73050	0.44730	0.55485

-After post pruning the decision tree the performance has generalized on training and test set. We are getting reasonable accuracy value compared to the other metrics that highly fluctuate.
-duration and poutcome are the most important features

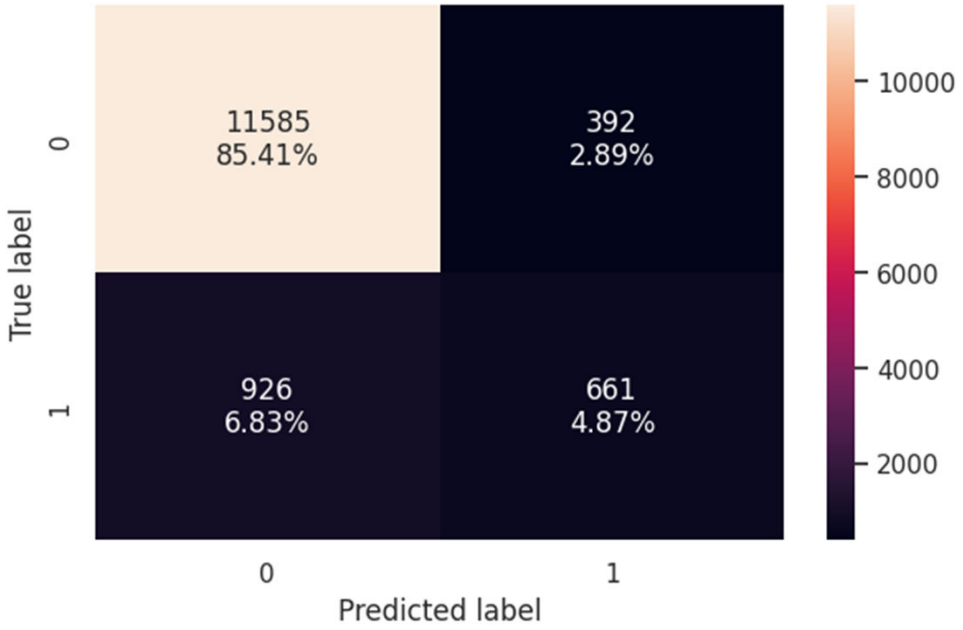
Random Forest: Sklearn

Train



	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

Test



	Accuracy	Recall	Precision	F1
0	0.90283	0.41651	0.62773	0.50076

-After post pruning the decision tree the performance has generalized on training and test set. We are getting reasonable accuracy value compared to the other metrics that highly fluctuate.
-duration and poutcome are the most important features

Models performance comparison

Training				
	Accuracy	Recall	Precision	F1
Decision Tree sklearn	1.00000	1.00000	1.00000	1.00000
Decision Tree (Pre-Pruning)	0.82513	0.82772	0.38753	0.52790
Decision Tree (Post-Pruning)	0.93127	0.99866	0.63239	0.77440
Random Forest (resampled)	1.00000	1.00000	1.00000	1.00000

Testing				
	Accuracy	Recall	Precision	F1
Decision Tree sklearn	0.87651	0.49323	0.46252	0.47738
Decision Tree (Pre-Pruning)	0.82159	0.81560	0.37217	0.51111
Decision Tree (Post-Pruning)	0.86597	0.73050	0.44730	0.55485
Random Forest (resampled)	0.90283	0.41651	0.62773	0.50076

Random forest has a high accuracy compared to Decision Tree sklearn model, using the test dataset

MODEL EVALUATION AND CONCLUSION

Looking back at the Exploration Data Analysis, Data are highly imbalanced
Duration remains the most important feature.
We also think that candidates that was contacted during the previous campaign, were most likely to subscribe or deposit after being contacted during the forthcoming campaigns.

Married participated the most in the campaign, followed by single then finally divorced. This clearly illustrates that married people care most about savings, deposit etc compared to the others

In term of job categories, unknown, technicians, managers, self-employed, maid participated mostly in the campaign

Students and retired have not really participated compared to the others, may be because they likely don't have relative and income.

Random forest (resampled) seems to be the better model with high accuracy on test dataset. We think its better performance may be due to the fact that data were balanced using SMOTE (Synthetic Minority Oversampling TEchnique)