Machine Learning Development Using Microsoft Azure:

Project Report _Task 3

By: Jerry TSIBA

Master in Data Sciences, Deakin University, March 31, 2024

Great Learning

https://github.com/JerryTsiba

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

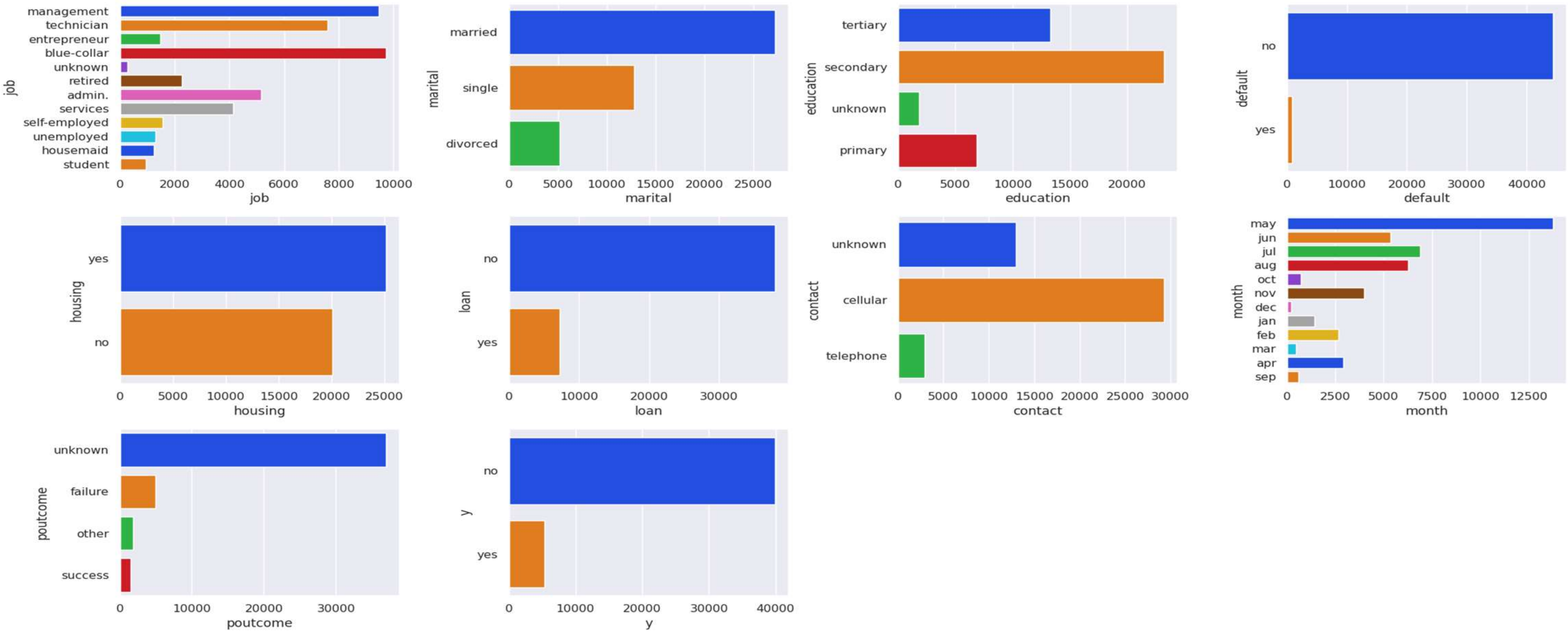
Variable	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	Age			no
job	Feature	Categorical	Occupation	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')		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

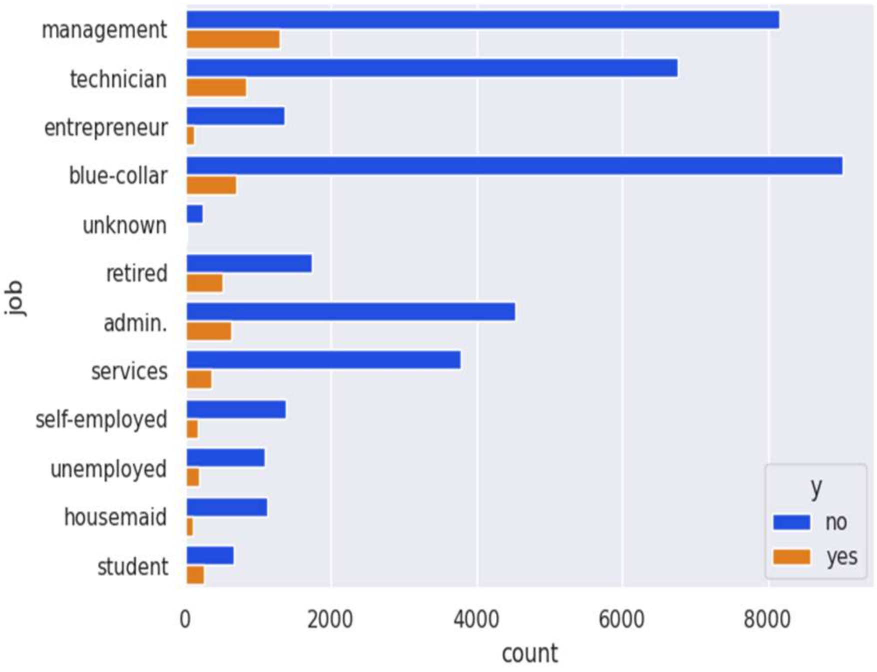
The mean age of candidate contacted is 41 years old, and the minimum is 18 years old, the maximum is 95 years old.
The average yearly balance is 1362.27 euro
The average number of contacts performed during this campaign is of 2.7 (~ 3 times)

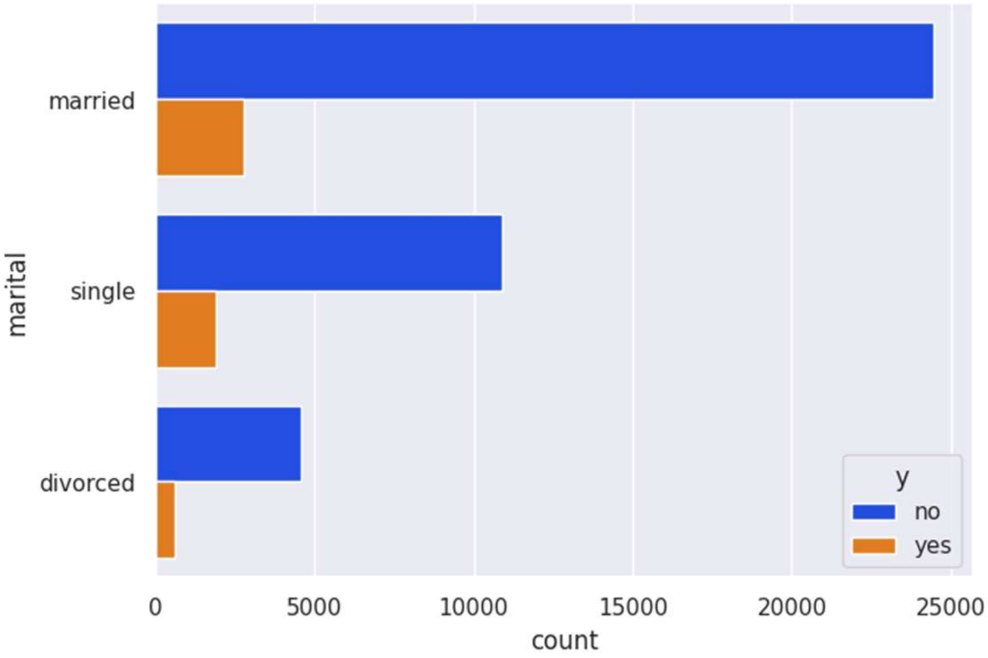
Exploratory Data Analysis



- 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

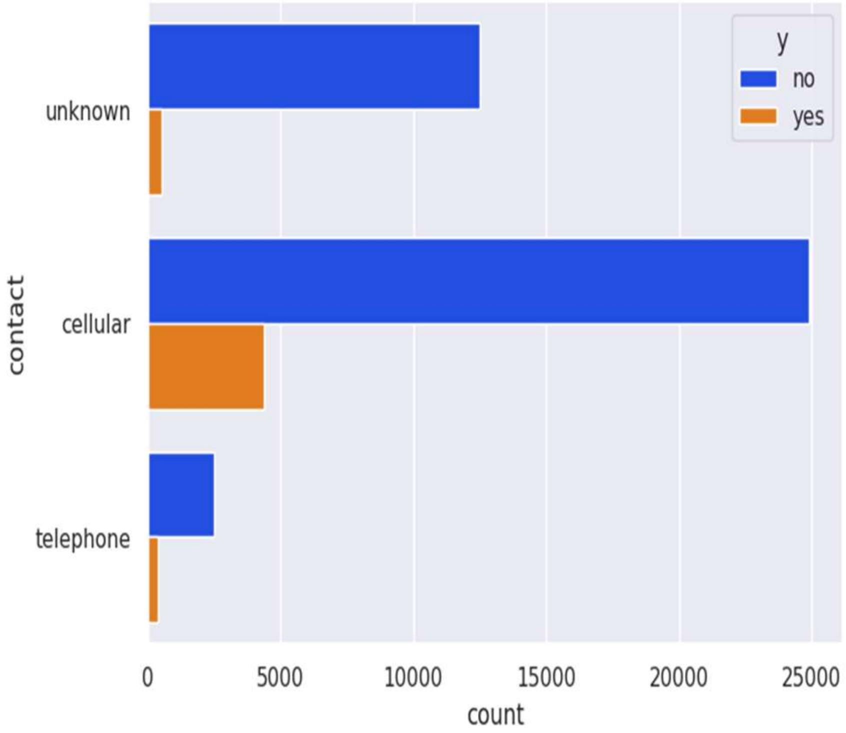


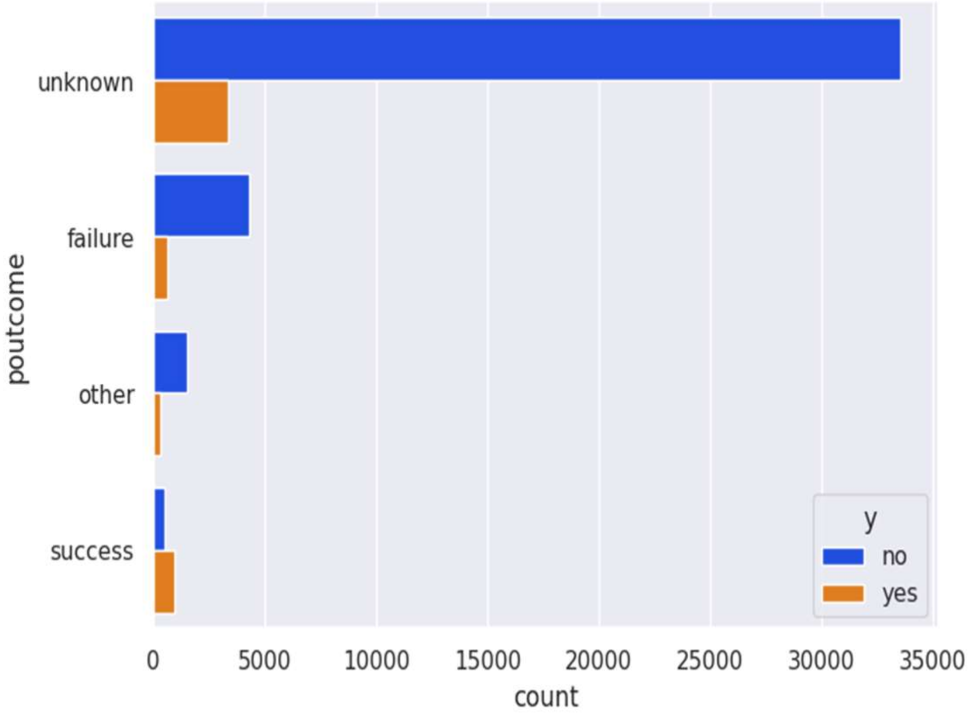


Married are predominant and the majority have not been more convinced, hence have not deposited. However, there have also more deposit compared to single and divorced

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.

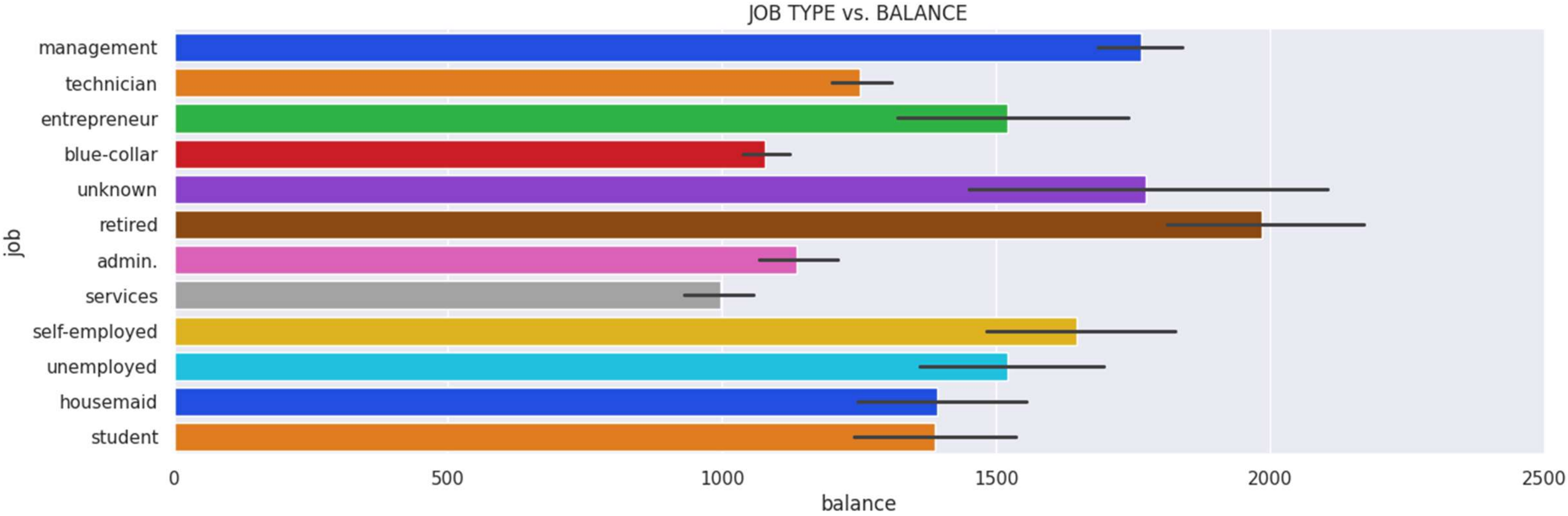
Exploratory Data Analysis





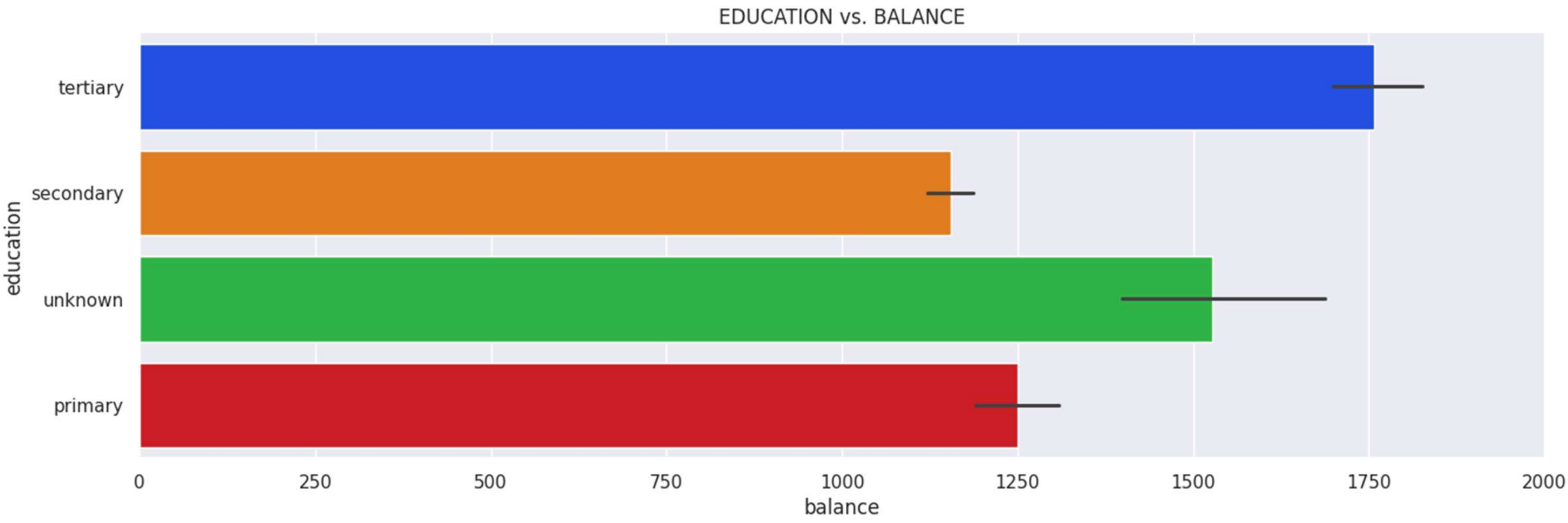
- poutcome is higher from unknown compared to those who succeeded, followed by those who failed
- Most of candidates convinced were contacted via cellular

Exploratory Data Analysis



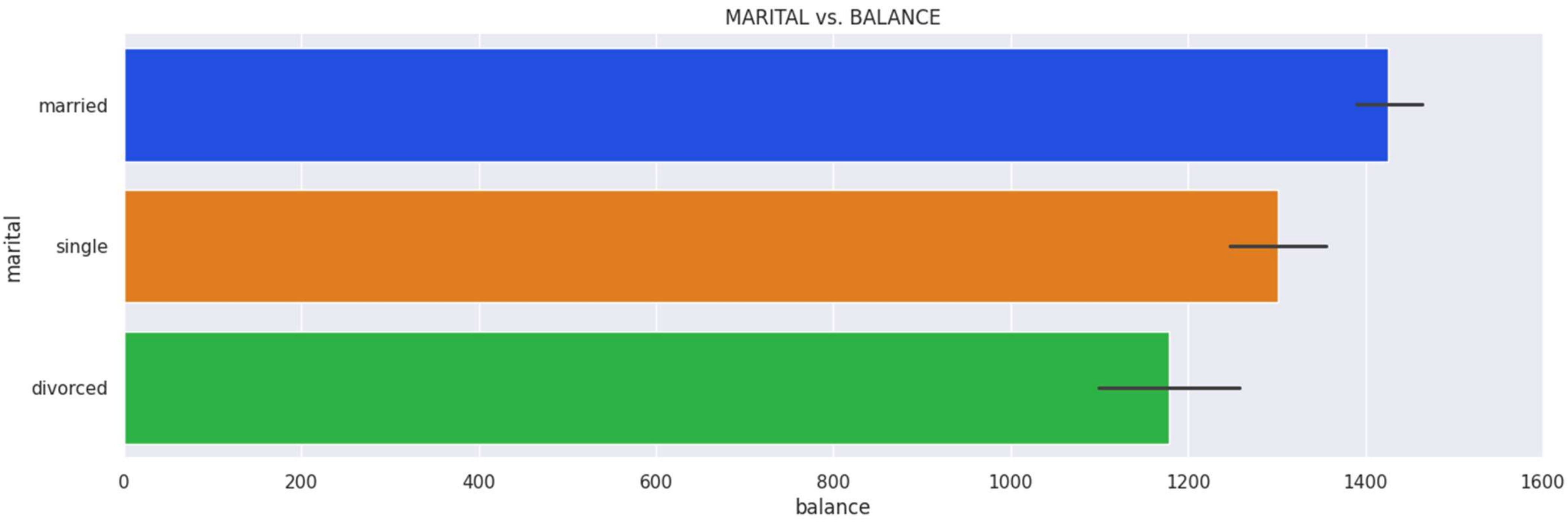
The average yearly salary of candidates in management and unknown category is equally high compared to the others, even though we observed an exception with retired people who have highest balance.

Exploratory Data Analysis



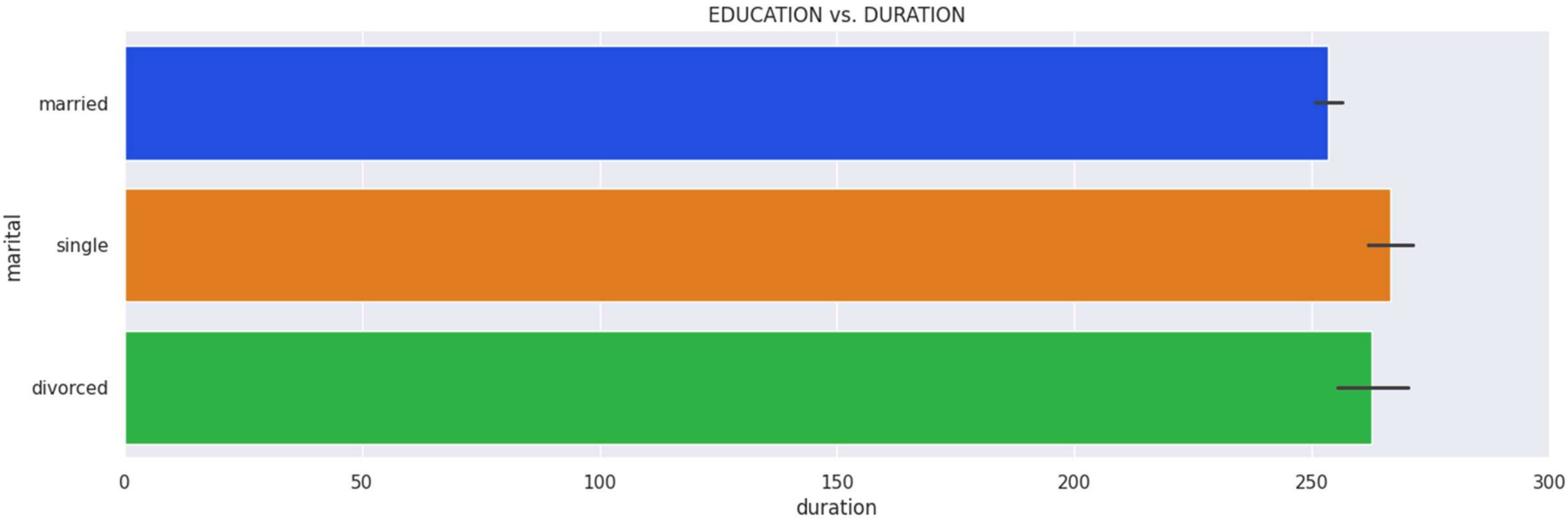
Candidates with tertiary education earn more compared to the others

Exploratory Data Analysis



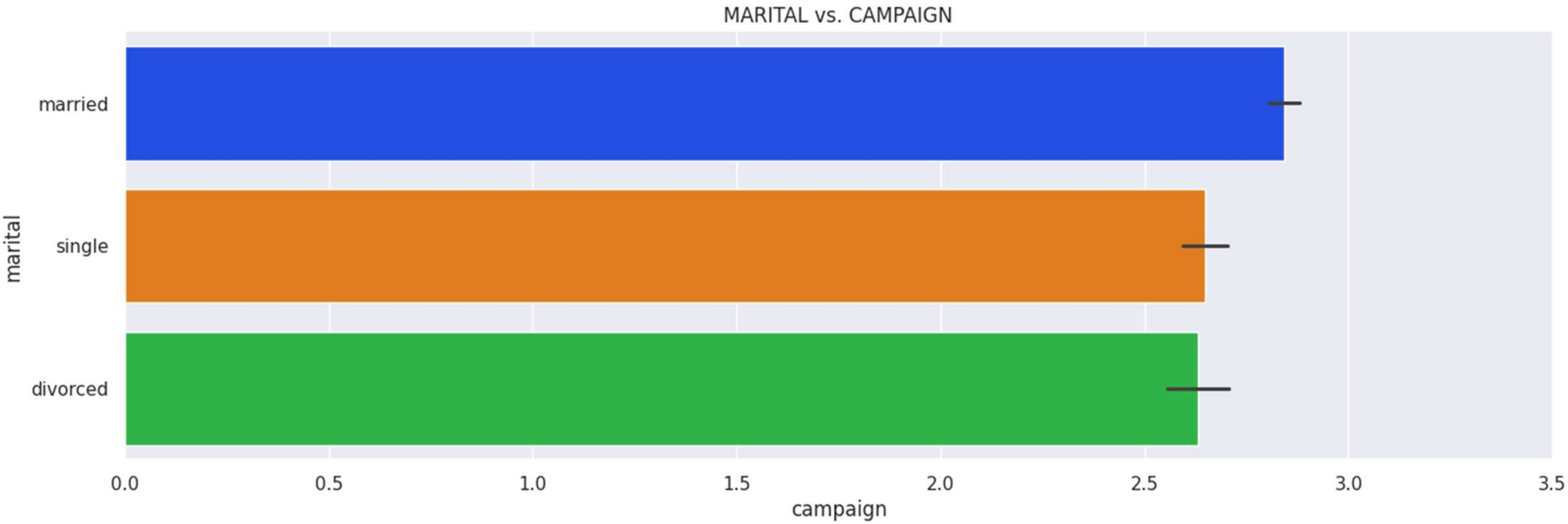
Married candidates have highest average yearly balance compared to single and divorced candidates

Exploratory Data Analysis



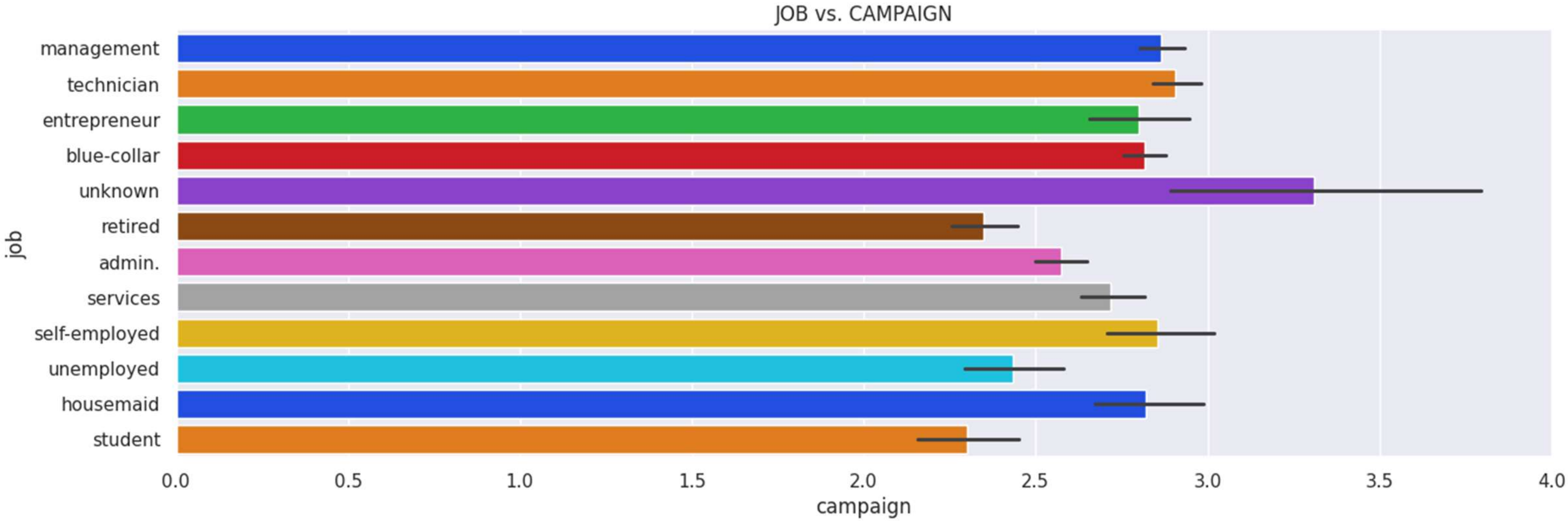
Singles followed by divorced were the most being in contact during the marketing campaign

Exploratory Data Analysis



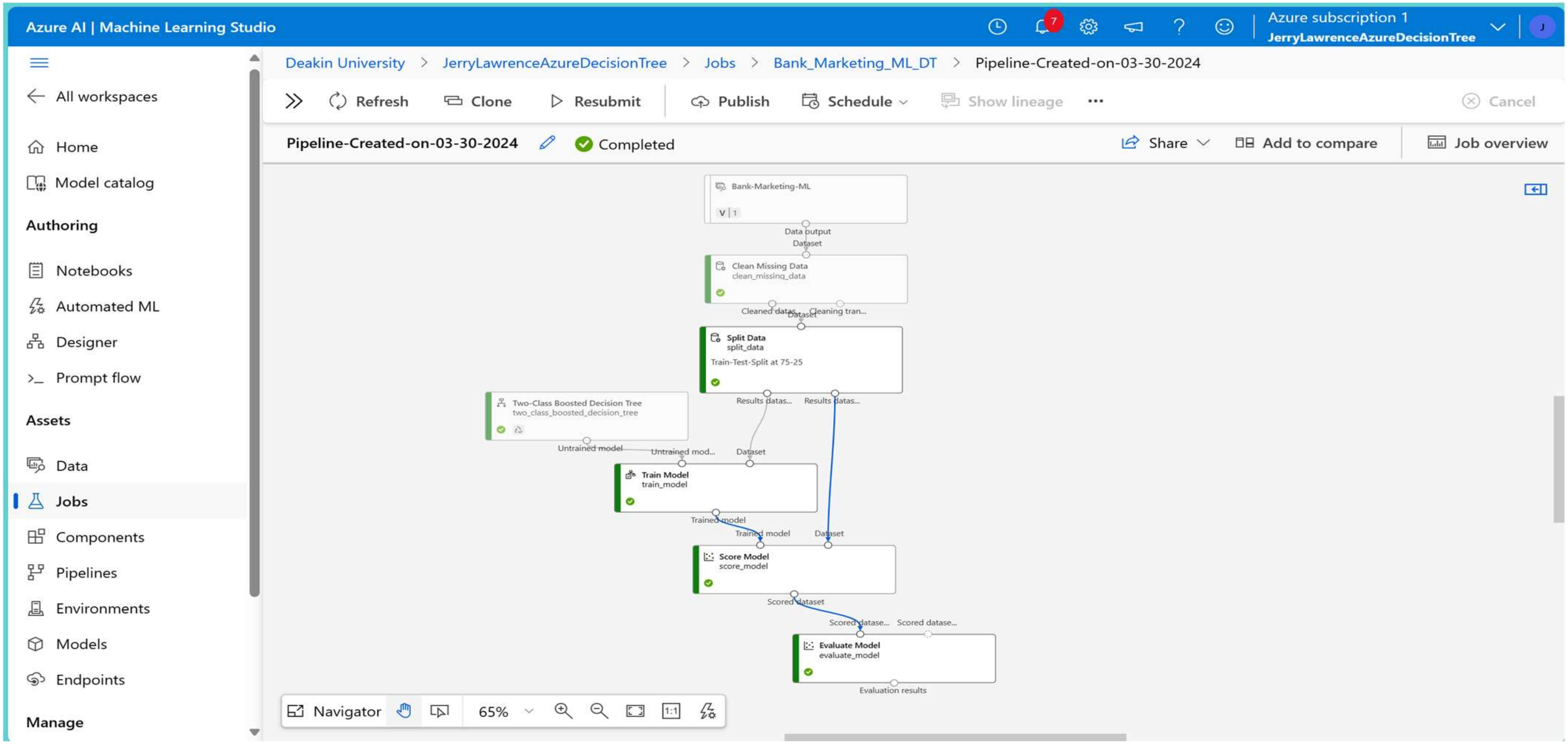
Married participated the most in the campaign, followed by single then finally divorced. This clearly illustrates that married people care most about savings, deposit etc compared to the others

Exploratory Data Analysis

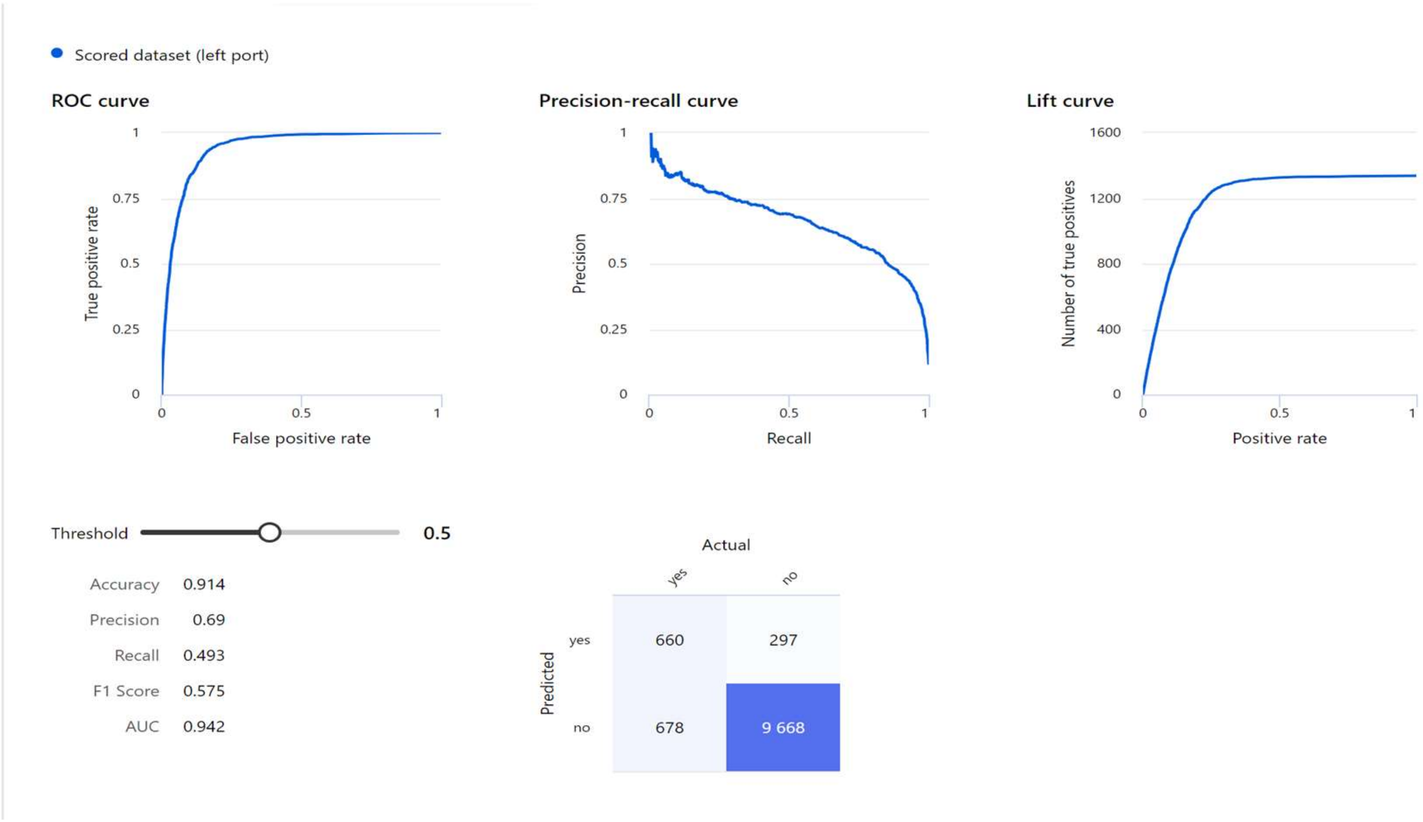


- In term of job categories, unknown, technicians, managers, self-employed, maid participated mostly in the campaign.
- Students and retired have not really participated compared to the others, may be because they likely don't have relative and income.

Two-Class Boosted Decision Tree



Two-Class Boosted Decision Tree

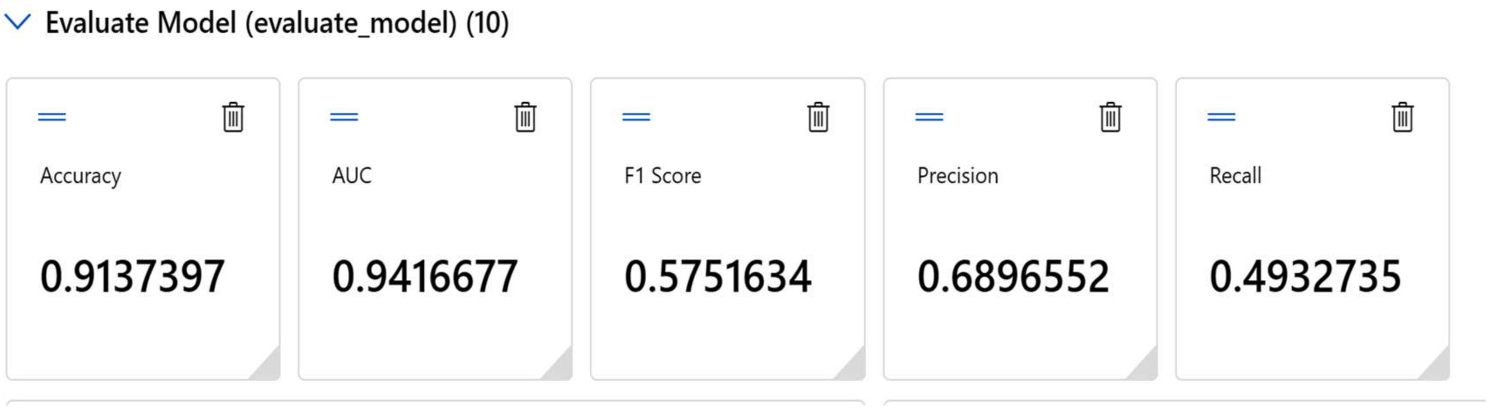


Threshold 0.5

Accuracy	0.914
Precision	0.69
Recall	0.493
F1 Score	0.575
AUC	0.942

		Actual	
		yes	no
Predicted	yes	660	297
	no	678	9 668

Two-Class Boosted Decision Tree

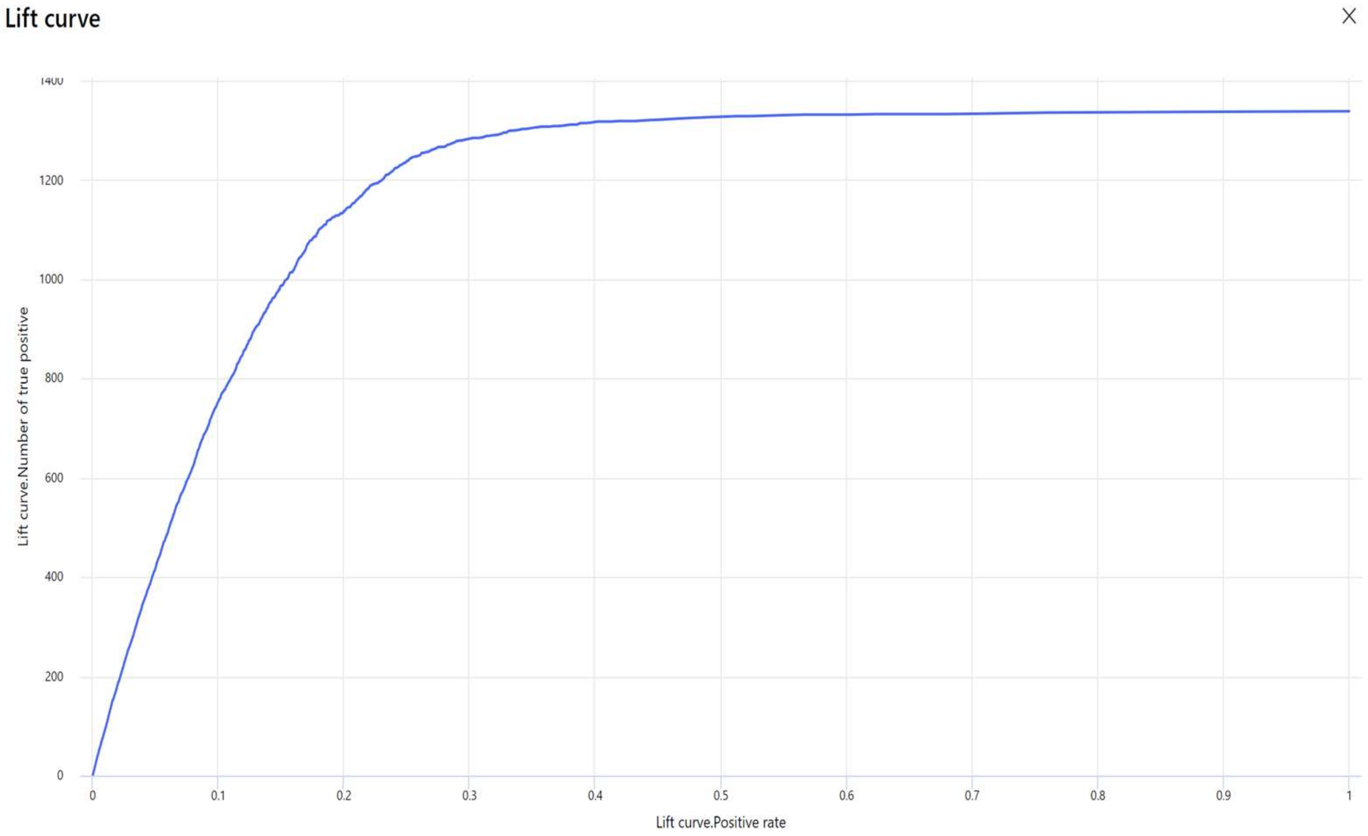


Evaluate Model (evaluate_model) (10)

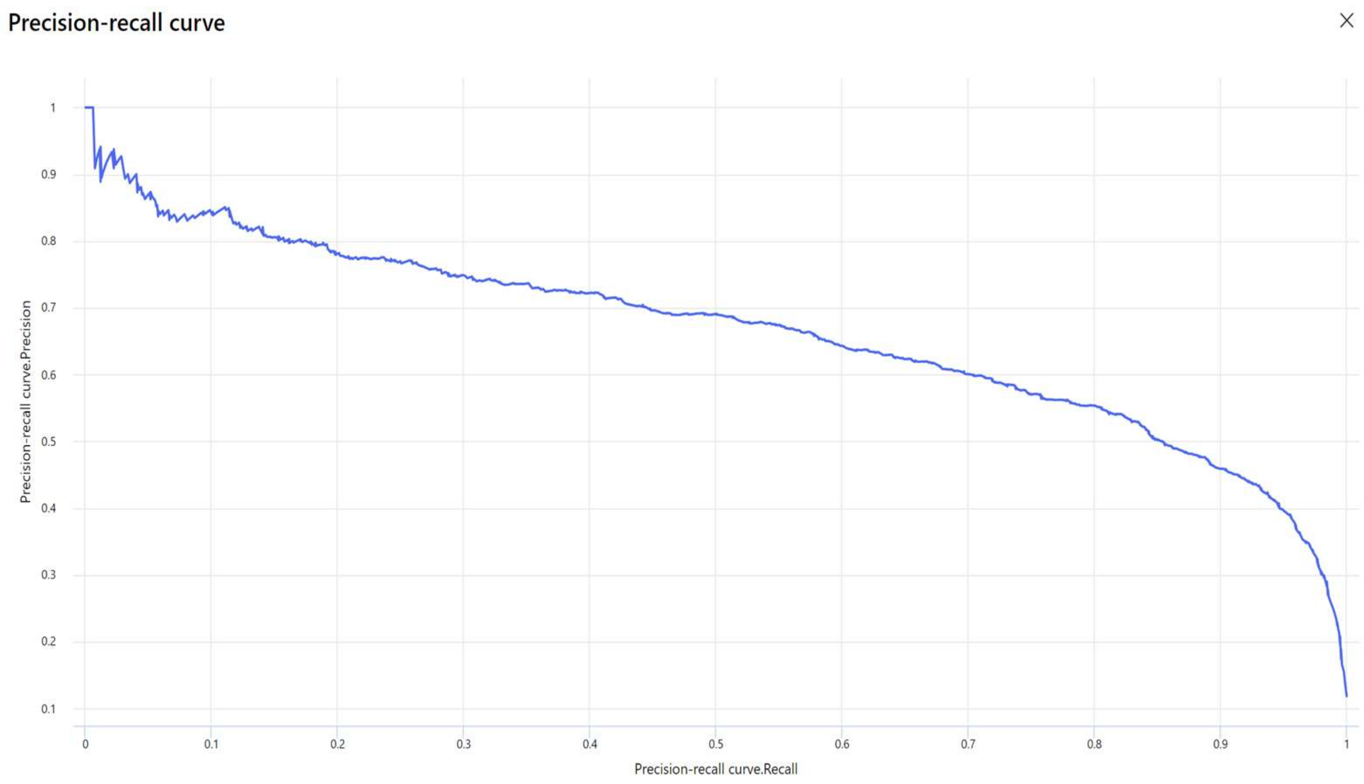
Accuracy	AUC	F1 Score	Precision	Recall
0.9137397	0.9416677	0.5751634	0.6896552	0.4932735

- Two-Class Boosted Decision Tree gives us high AUC of 94.16% and Accuracy 91.37% but very low Recall 49%
- This model performs is not bad

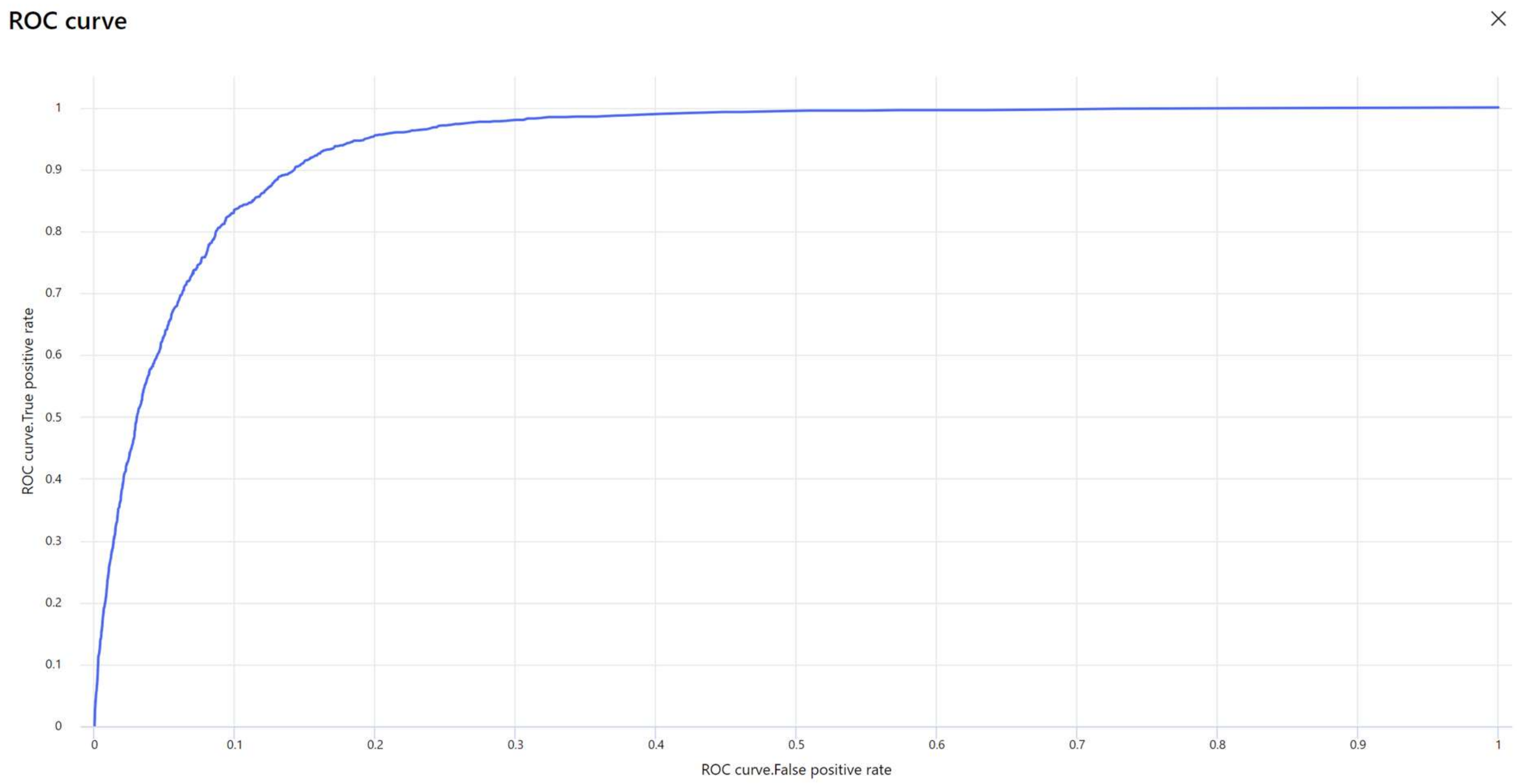
Two-Class Boosted Decision Tree



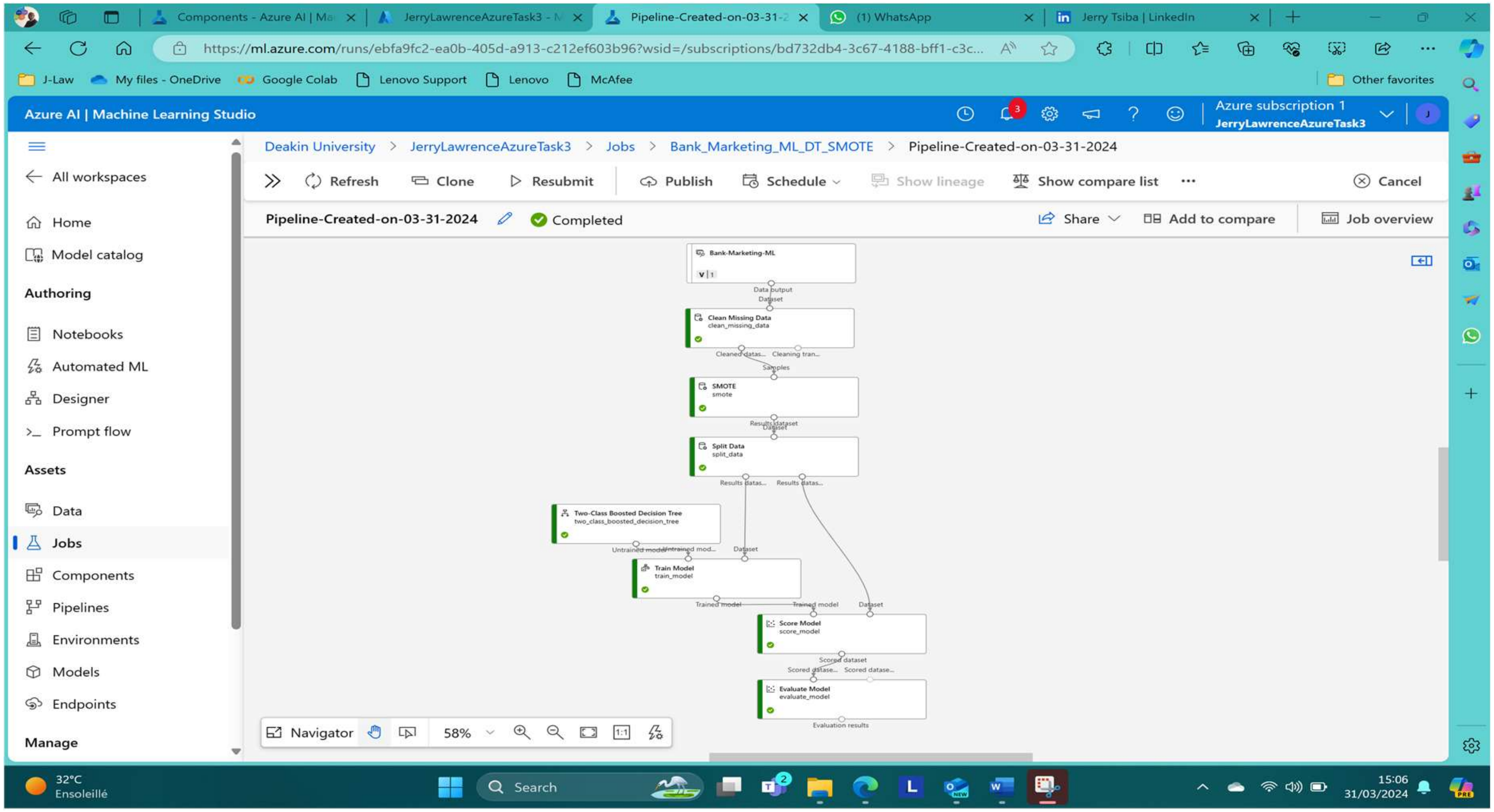
Two-Class Boosted Decision Tree


Precision-recall curve

Two-Class Boosted Decision Tree

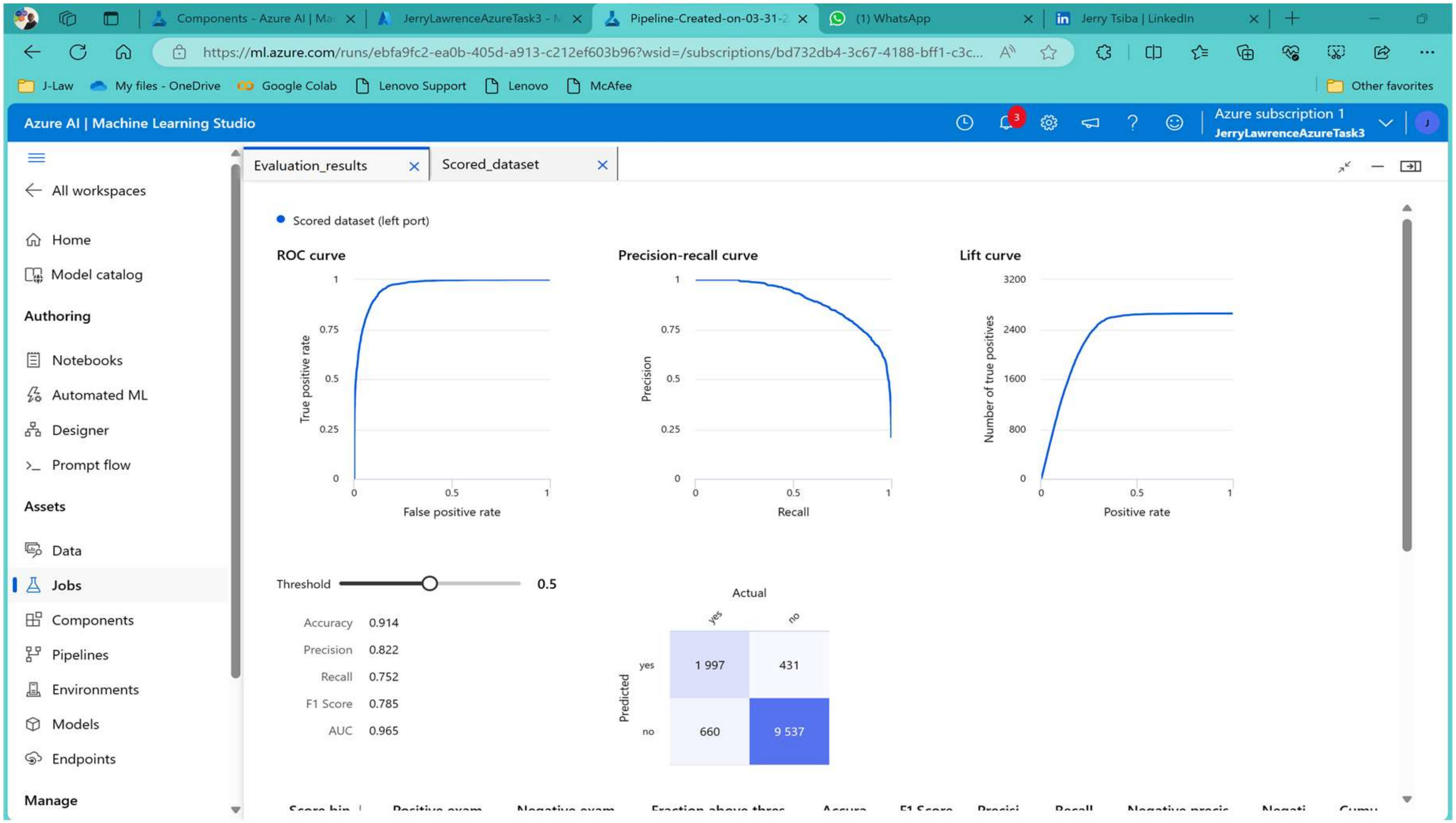


Two-Class Boosted Decision Tree-Using SMOTE



- Let us try to use SMOTE to see if the model performance will be enhanced

Two-Class Boosted Decision Tree-Using SMOTE

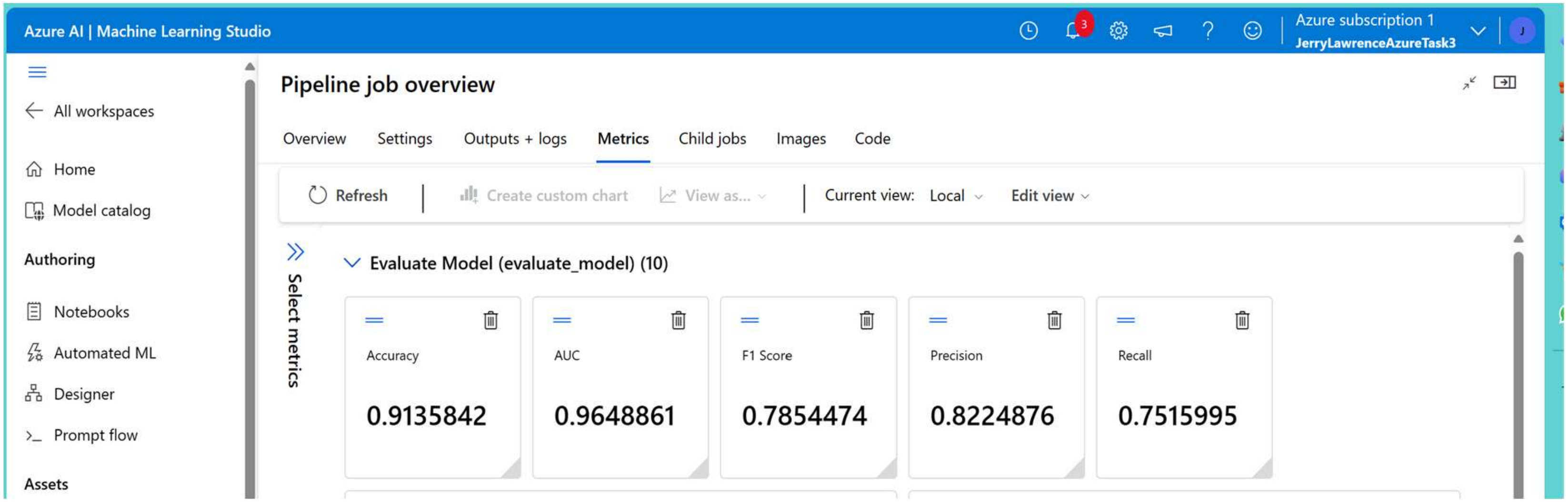


Accuracy	0.914
Precision	0.822
Recall	0.752
F1 Score	0.785
AUC	0.965

Predicted \ Actual	yes	no
yes	1 997	431
no	660	9 537

Score bin | Positive exam... | Negative exam... | Fraction above thres... | Accura... | F1 Score | Precisi... | Recall | Negative precis... | Negati... | Cumu...

Two-Class Boosted Decision Tree-Using SMOTE

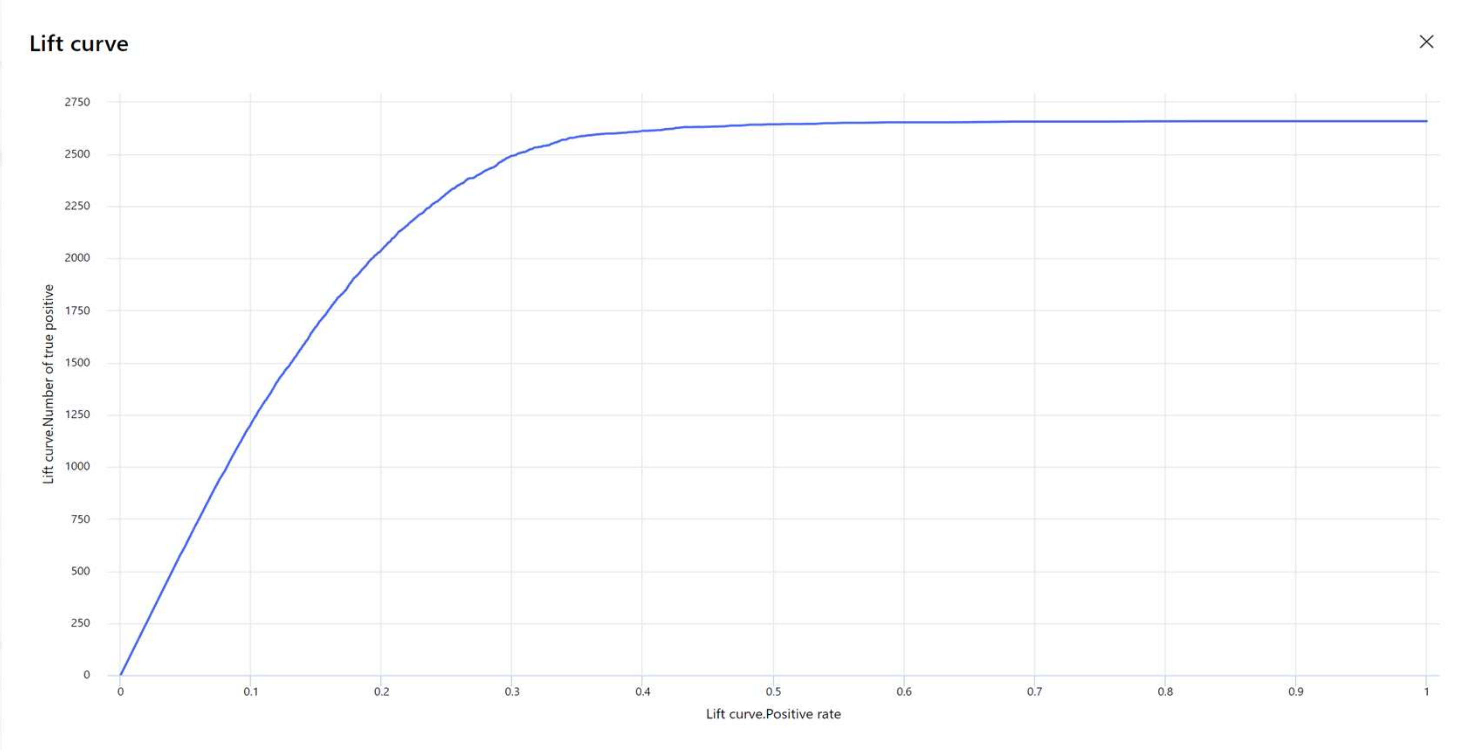


Evaluate Model (evaluate_model) (10)

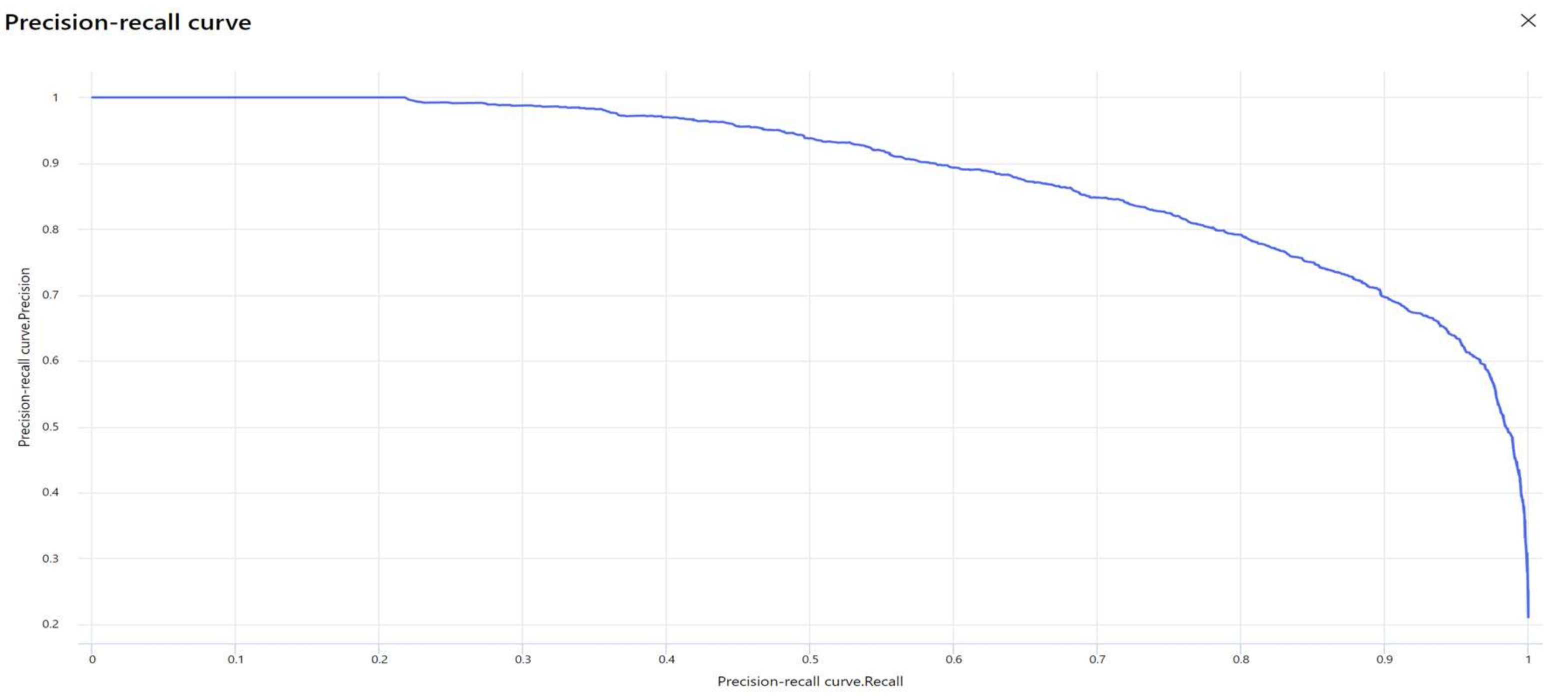
Accuracy	AUC	F1 Score	Precision	Recall
0.9135842	0.9648861	0.7854474	0.8224876	0.7515995

- Compared to simple Two-Class Boosted Decision Tree, Two-Class Boosted Decision Tree using SMOTE gives us much better AUC of 96.48% compared to 94.16% on the previous model and an Accuracy of 91.35% compared to 91.37% on the previous model. However, the F1 Score, precision and Recall have all been improved
- This model performs better. This clearly explains that data are really imbalanced, SMOTE has helped.

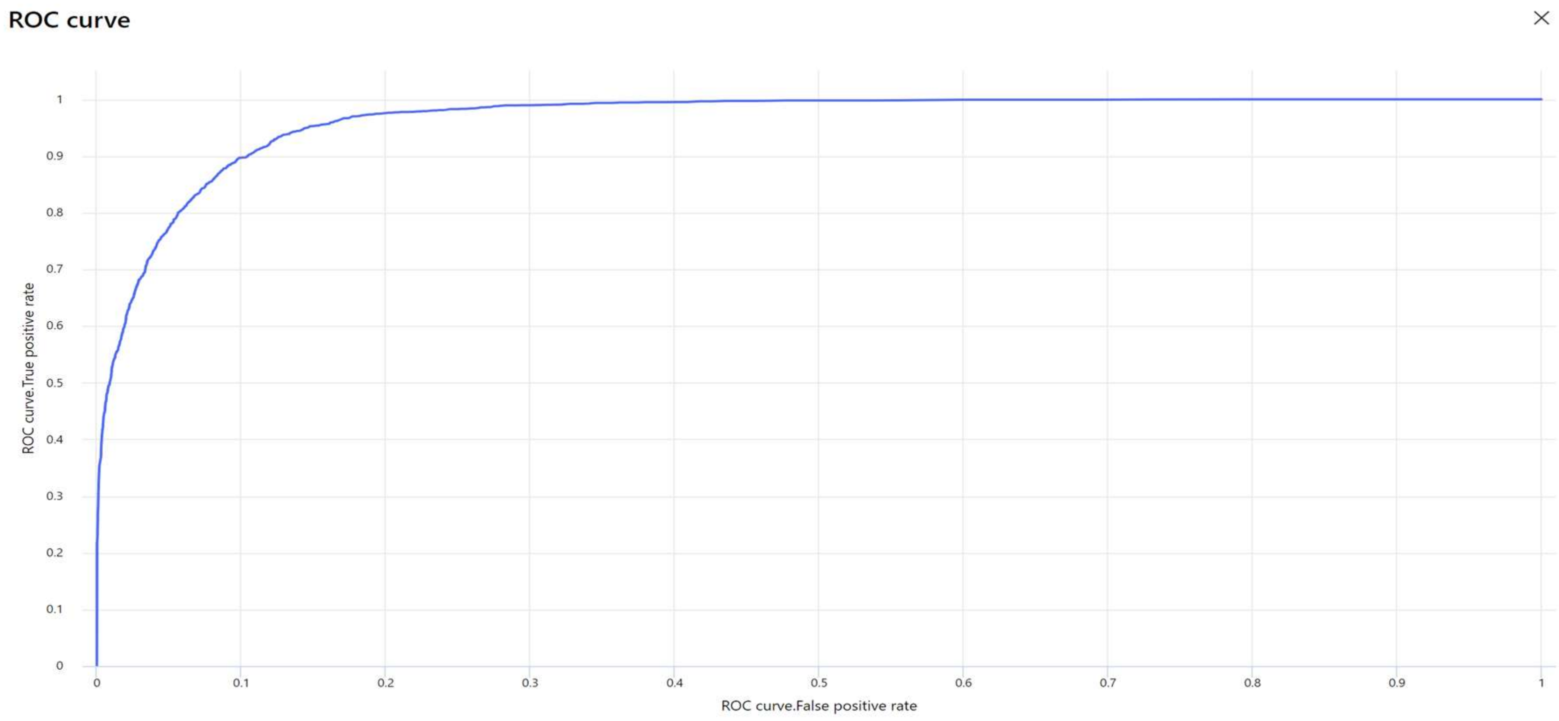
Two-Class Boosted Decision Tree-Using SMOTE



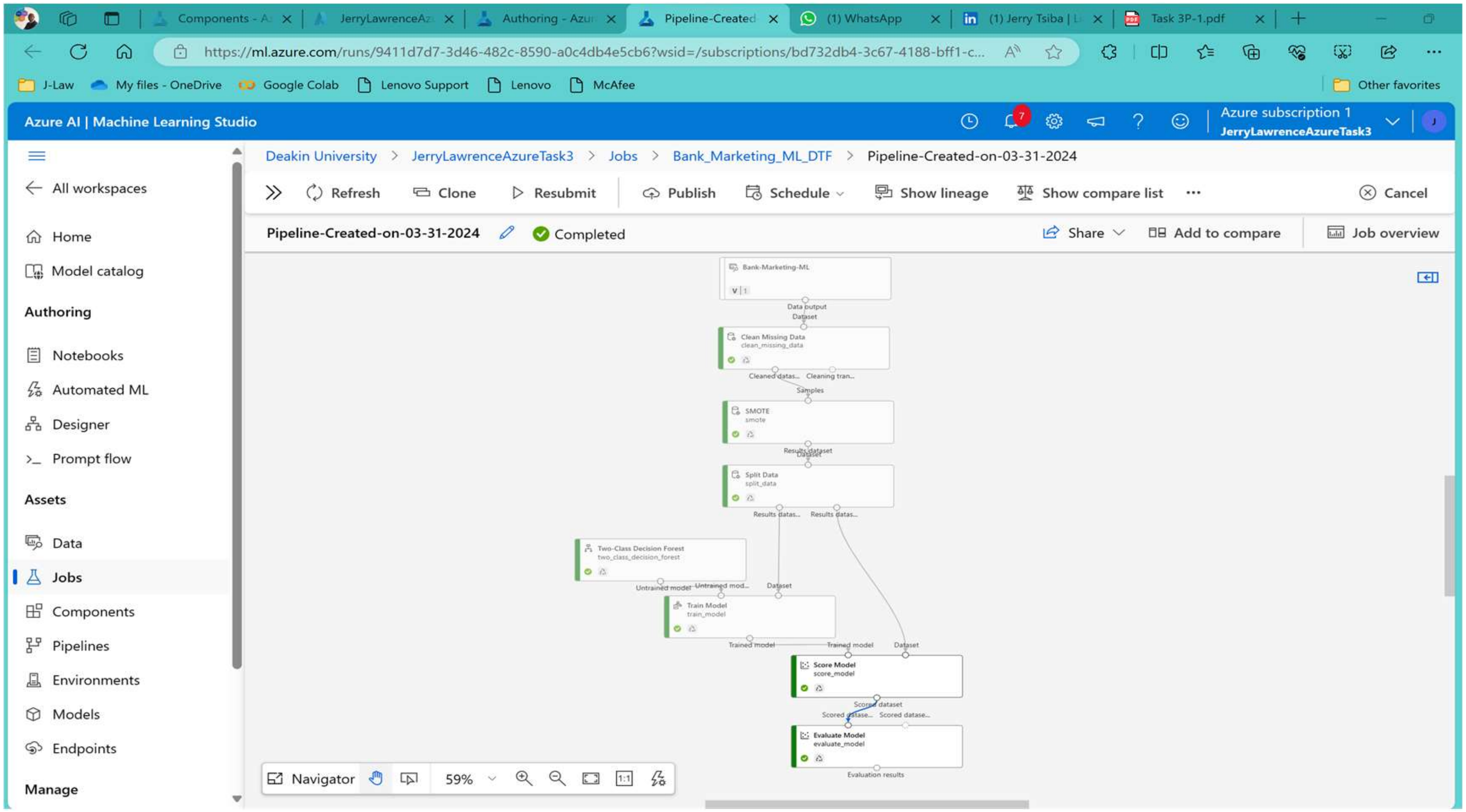
Two-Class Boosted Decision Tree-Using SMOTE



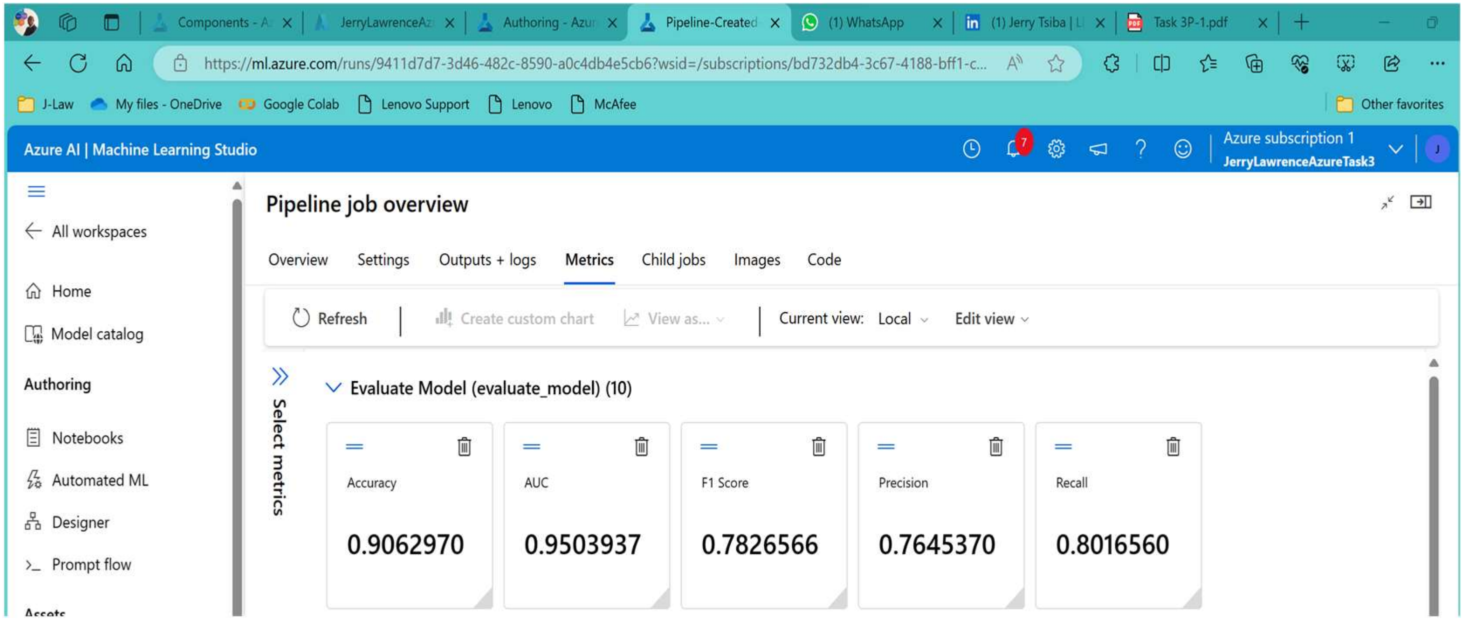
Two-Class Boosted Decision Tree-Using SMOTE



Two-Class Decision Forest-Using SMOTE



Two-Class Decision Forest-Using SMOTE



Pipeline job overview

Evaluate Model (evaluate_model) (10)

Accuracy 0.9062970

AUC 0.9503937

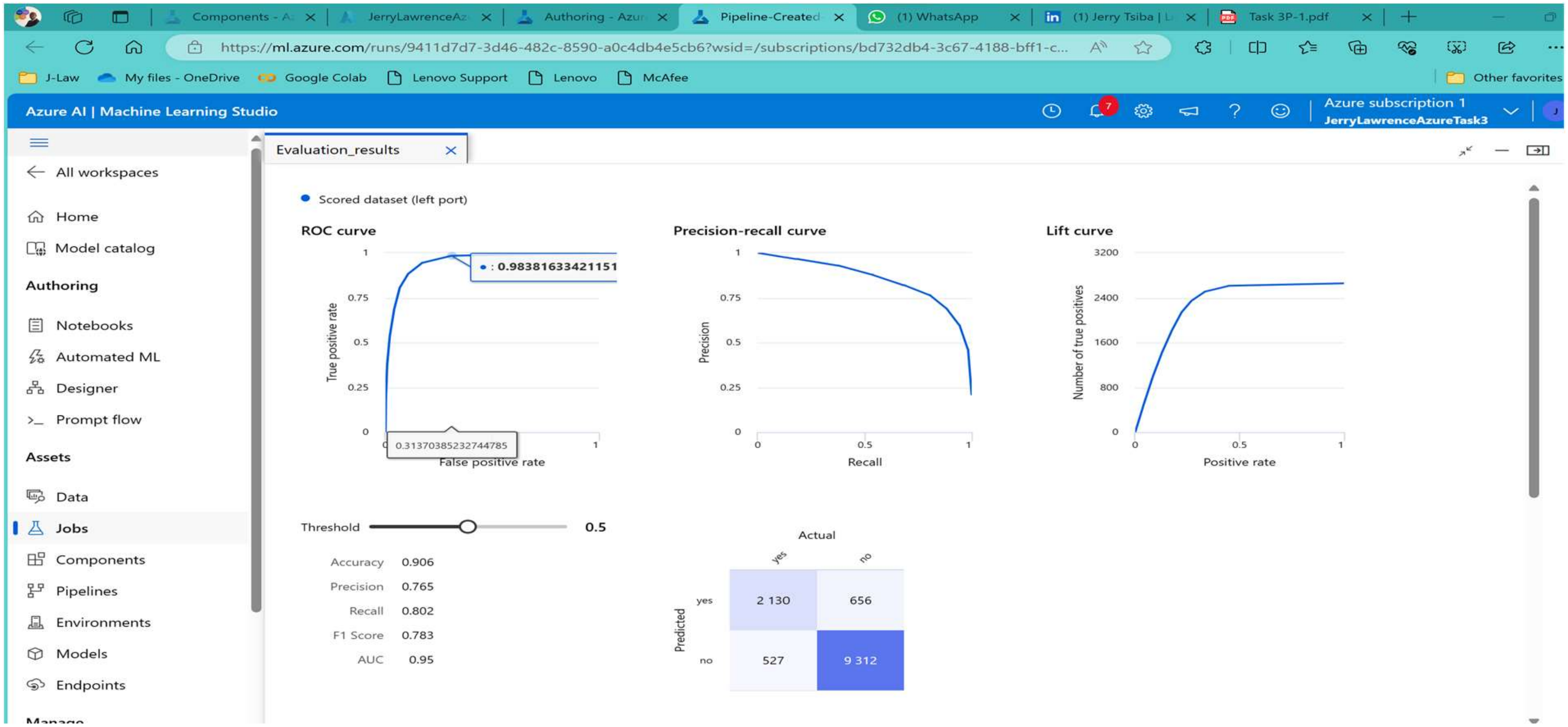
F1 Score 0.7826566

Precision 0.7645370

Recall 0.8016560

- Compared to simple Two-Class Boosted Decision Tree and Two-Class Boosted Decision Tree using SMOTE, this Two-Class Decision Forest performs enough good and better than simple Two-Class Boosted Decision Tree

Two-Class Decision Forest-Using SMOTE

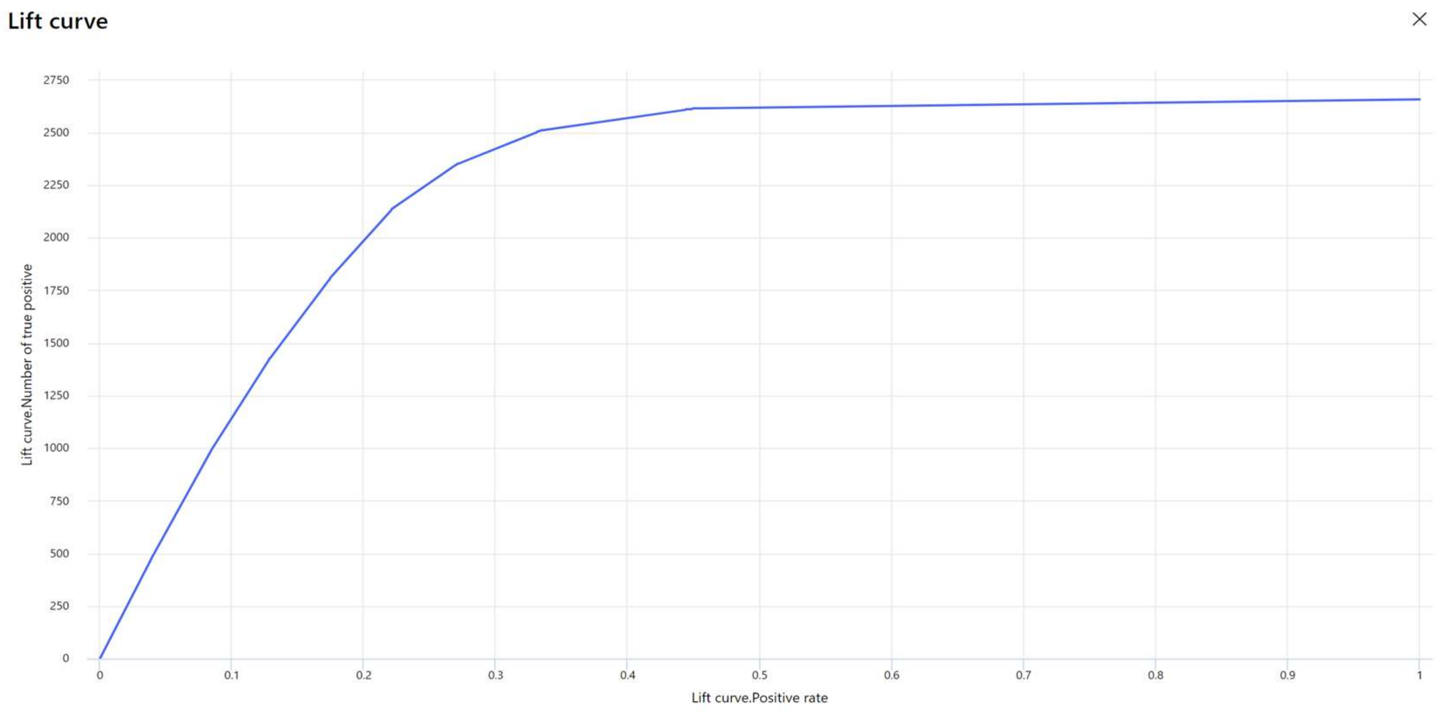


Threshold 0.5

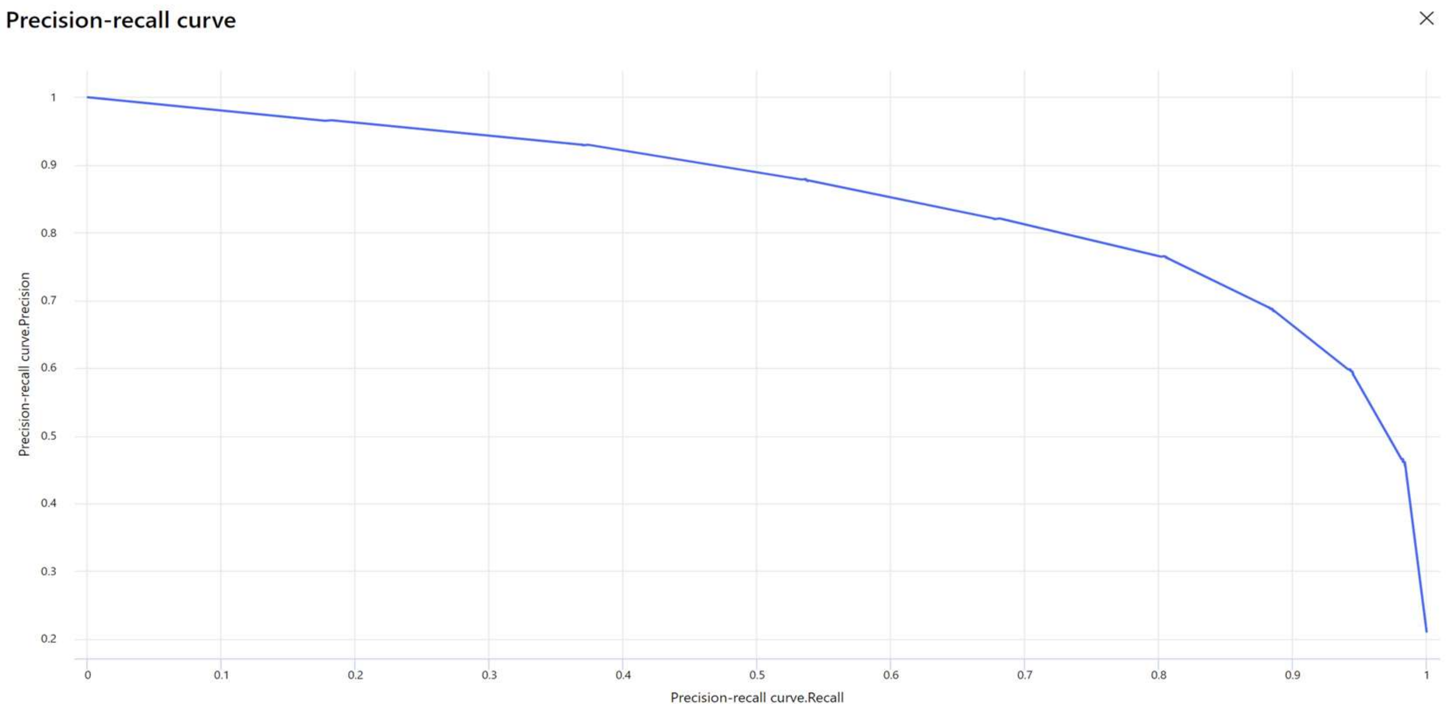
Accuracy	0.906
Precision	0.765
Recall	0.802
F1 Score	0.783
AUC	0.95

		Actual	
		yes	no
Predicted	yes	2 130	656
	no	527	9 312

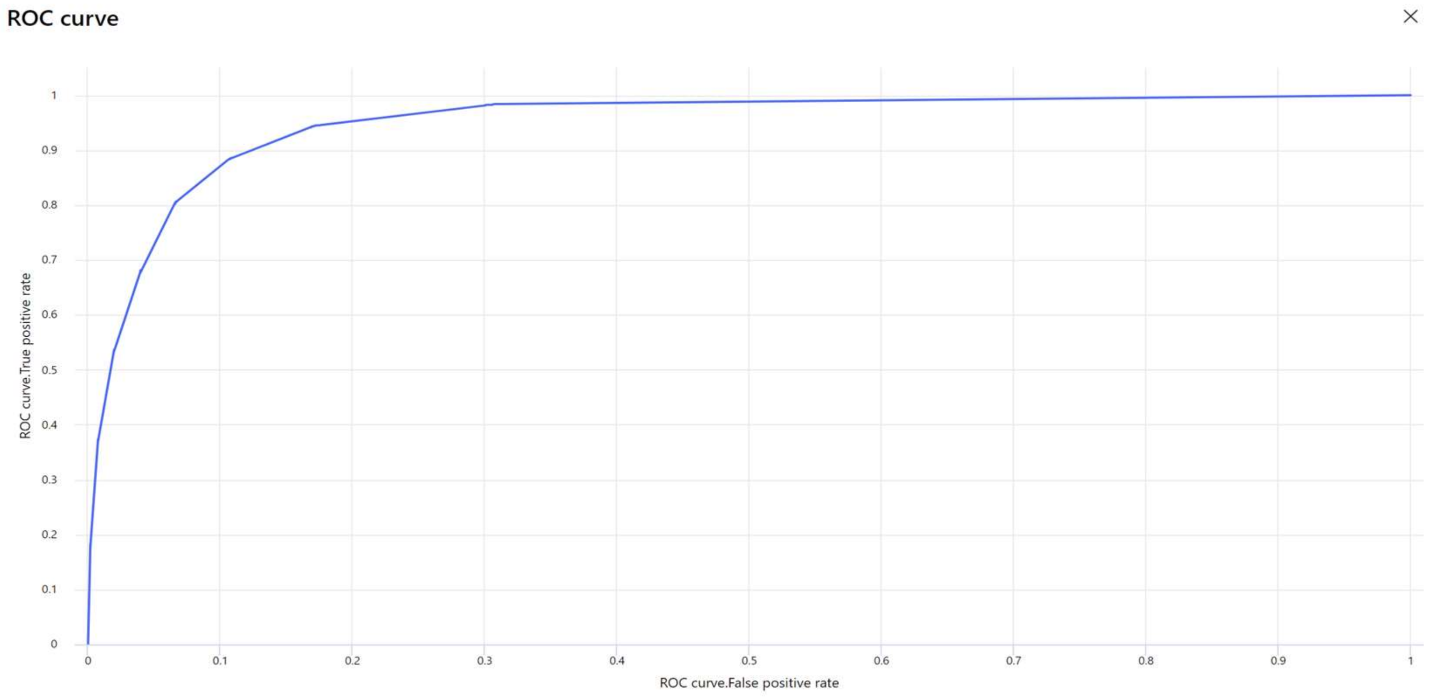
Two-Class Decision Forest-Using SMOTE



Two-Class Decision Forest-Using SMOTE



Two-Class Decision Forest-Using SMOTE



MODEL EVALUATION AND CONCLUSION

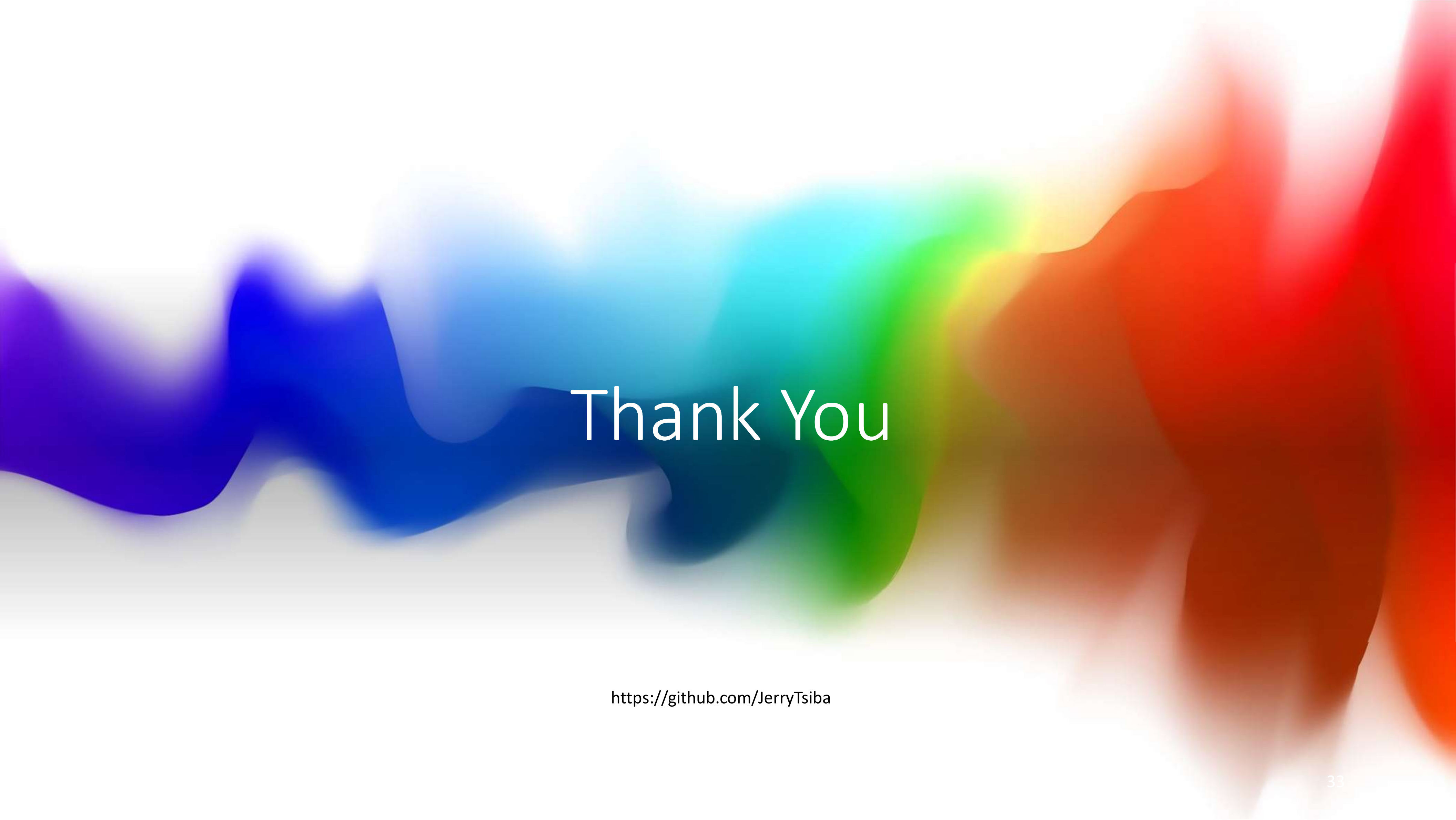
Looking back at the Exploration Data Analysis, Data are highly imbalanced. The performance of the models either Decision Tree or Random Forest is improved only when using SMOTE.
Two–Class Decision Tree using SMOTE performs much better.

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.



Thank You

https://github.com/JerryTsiba