Credit Risk Analysis and Modeling

Li Zou

Outline of this project:

Goal: Analyzing and modeling the Taiwan credit risk dataset to gain practical insights into real-world scenarios

Steps:
(1).Data description

(2).Data analysis

(3).Modeling

Part 1: Data description

Dataset: Default Payments of Credit Card Clients in Taiwan from 2005
Source: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/data

Dataset Information

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	ID	LIMIT_BAL	SEX	EDUCATIC	MARRIAGI	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	PAY_AMT	default.payment.next.month		
2	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1		
3	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1		
4	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0		
5	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0		
6	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0		
7	6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0		
8	7	5.00E+05	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0		
9	8	1.00E+05	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0		
10	9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0		
11	10	20000	1	3	2	35	-2	-2	-2	-2	-1	-1	0	0	0	0	13007	13912	0	0	0	13007	1122	0	0		
12	11	2.00E+05	2	3	2	34	0	0	2	0	0	-1	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66	0		
13	12	260000	2	1	2	51	-1	-1	-1	-1	-1	2	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640	0		
14	13	630000	2	2	2	41	-1	0	-1	-1	-1	-1	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0	0		
15	14	70000	1	2	2	30	1	2	2	0	0	2	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0	1		
16	15	250000	1	1	2	29	0	0	0	0	0	0	70887	67060	63561	59696	56875	55512	3000	3000	3000	3000	3000	3000	0		
17	16	50000	2	3	3	23	1	2	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0		
18	17	20000	1	1	2	24	0	0	2	2	2	2	15376	18010	17428	18338	17905	19104	3200	0	1500	0	1650	0	1		
19	18	320000	1	1	1	49	0	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0		
20	19	360000	2	1	1	49	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0		
21	20	180000	2	1	2	29	1	-2	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0		
22	21	130000	2	3	2	39	0	0	0	0	0	-1	38358	27688	24489	20616	11802	930	3000	1537	1000	2000	930	33764	0		
23	22	120000	2	2	1	39	-1	-1	-1	-1	-1	-1	316	316	316	0	632	316	316	316	0	632	316	0	1		
24	23	70000	2	2	2	26	2	0	0	2	2	2	41087	42445	45020	44006	46905	46012	2007	3582	0	3601	0	1820	1		
25	24	450000	2	1	1	40	-2	-2	-2	-2	-2	-2	5512	19420	1473	560	0	0	19428	1473	560	0	0	1128	1		
26	25	90000	1	1	2	23	0	0	0	-1	0	0	4744	7070	0	5398	6360	8292	5757	0	5398	1200	2045	2000	0		
27	26	50000	1	3	2	23	0	0	0	0	0	0	47620	41810	36023	28967	29829	30046	1973	1426	1001	1432	1062	997	0		
28	27	60000	1	1	2	27	1	-2	-1	-1	-1	-1	-109	-425	259	-57	127	-189	0	1000	0	500	0	1000	1		
29	28	50000	2	3	2	30	0	0	0	0	0	0	22541	16138	17163	17878	18931	19617	1300	1300	1000	1500	1000	1012	0		

Part 2.Data analysis

Why plot this figure:

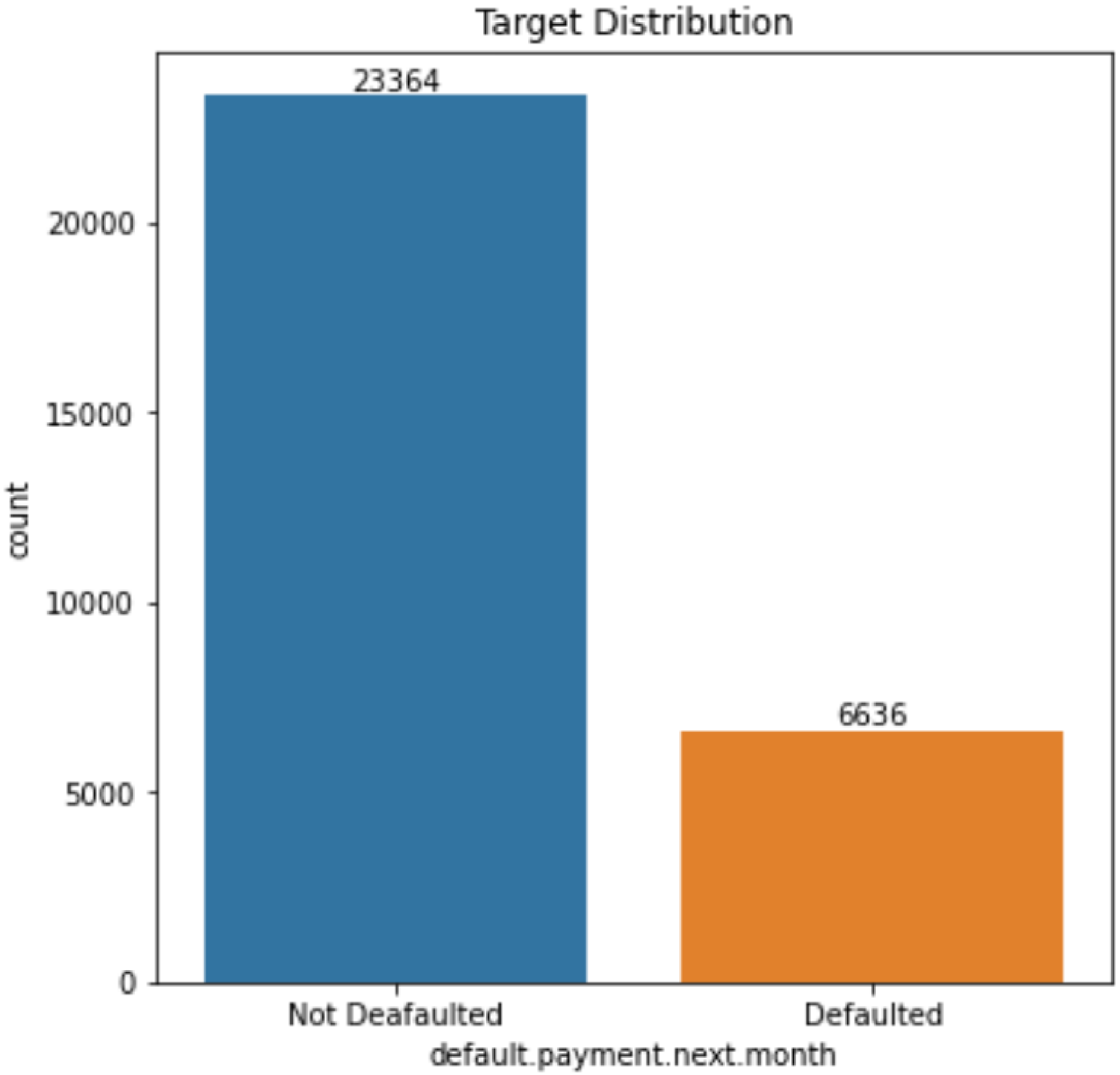
In credit analysis, the primary objective is to classify whether a client will default or not. To accomplish this, it's essential to examine the default history of our portfolio.

What is this figure:

This figure shows the distribution of default.

What is the conclusion:

77.88% of our clients are likely not to default and 22.12% are likely to default.



Why plot this figure:

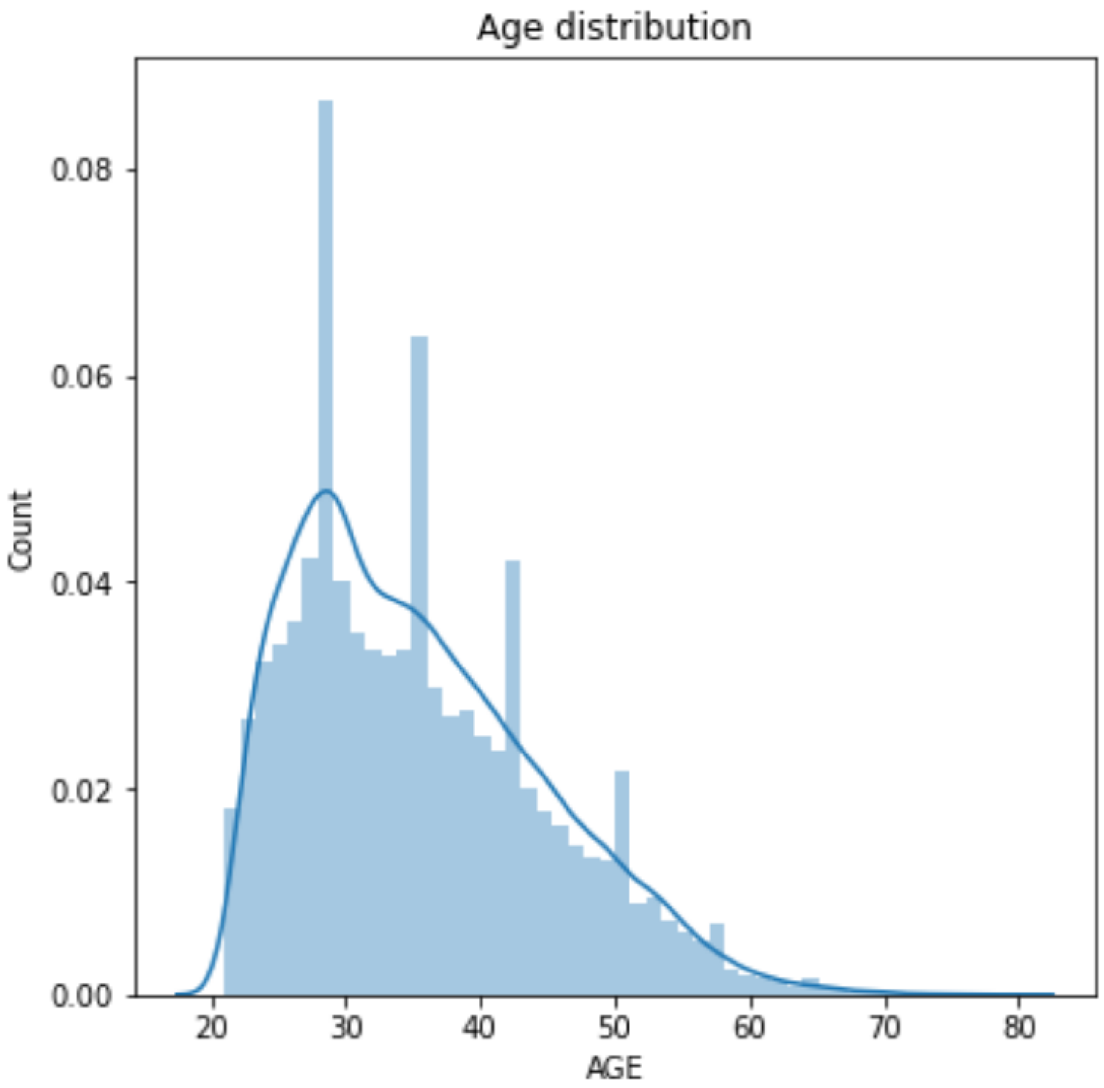
In credit analysis, the primary objective is to classify whether a client will default or not. Understanding who are our clients is important.

What is this figure:

This figure shows the distribution of the age of our clients.

What is the conclusion:

The age of most of our clients is around 30 years old.



Why plot this figure:

In credit analysis, the primary objective is to classify whether a client will default or not. Understanding which group of clients are more likely to default is important for selecting features for modeling.

What is this figure:

This figure shows the default distribution of clients under different group (gender, education and marriage).

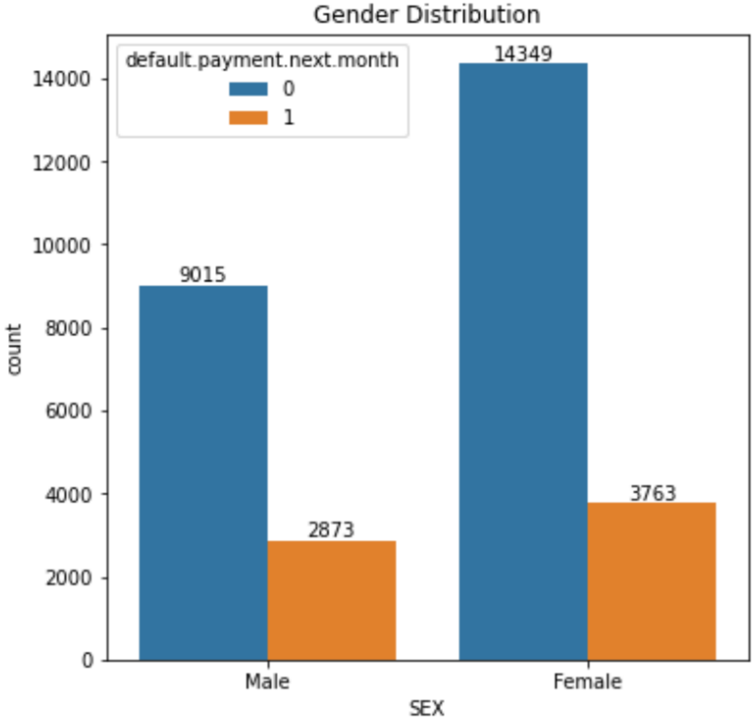
What is the conclusion:

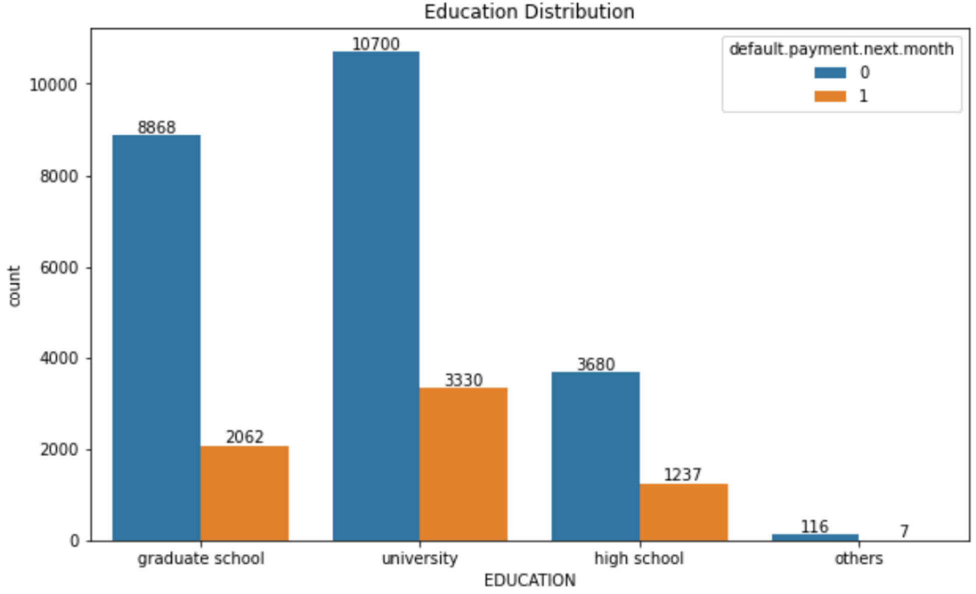
Default probability under each group:
Male: 24% Female: 20%

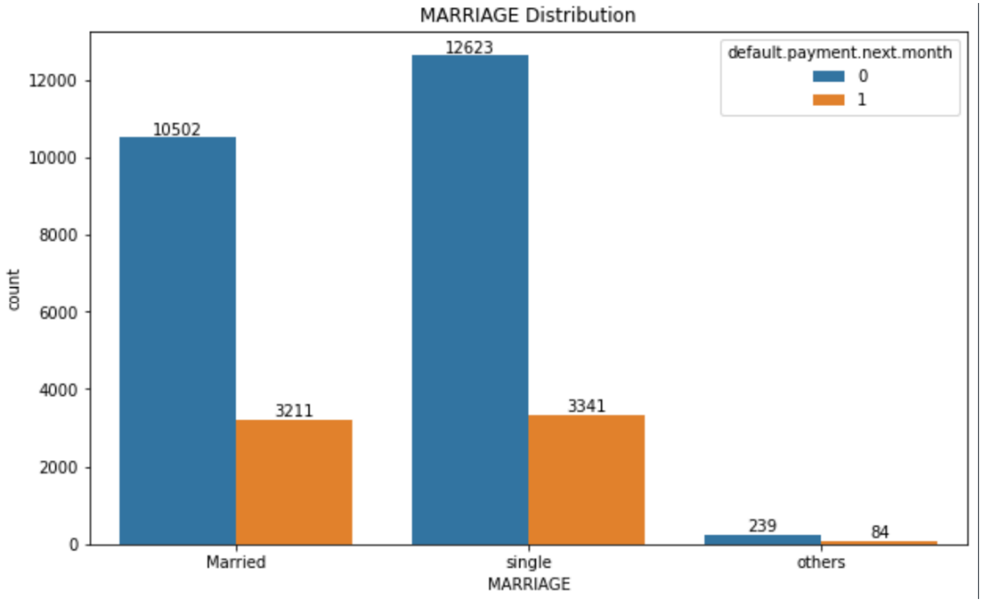
Graduate school: 19%
University: 24%
High school:25%

Married:23%
Single:21%

(1).Male have larger probability to default than female.
(2).Higher education corresponds to lower default probability.
(3).Married group has larger default probability than singer group.







Part 2.Modeling

Goal:

Select the optimal model which can be used to predict the default of a client.

Considered models: Logistic Regression, Random Forest and SVM

Why choose these models:

(1).We have historical default information of our clients. So supervised machine learning algorithms should be chosen.

(2).As we predict the outcome (default or not default), so algorithm for binary classification should be chosen.

These three models are all supervised machine learning algorithms and can be used for binary classification.

Basic information about models:

(1).Logistic Regression

Logistic regression is defined as a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.

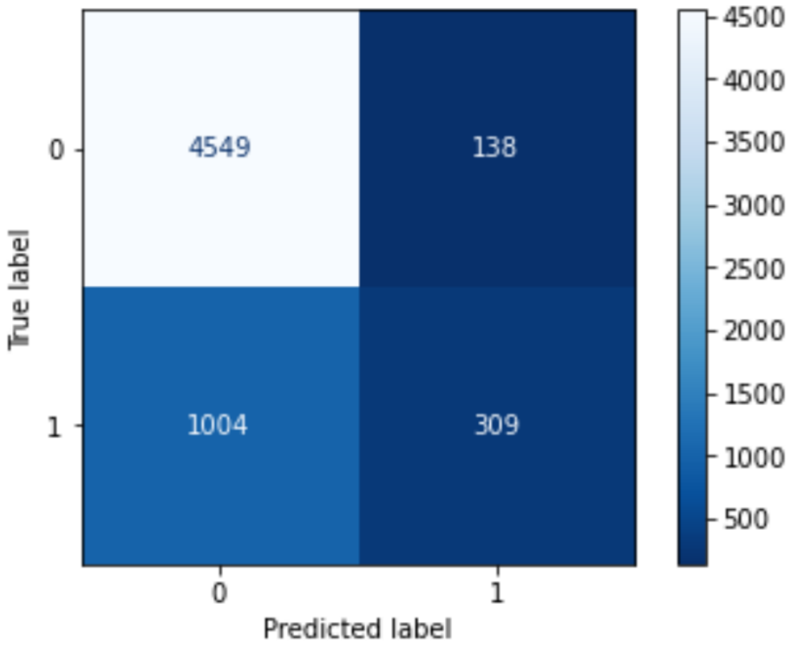
(2).Random forest

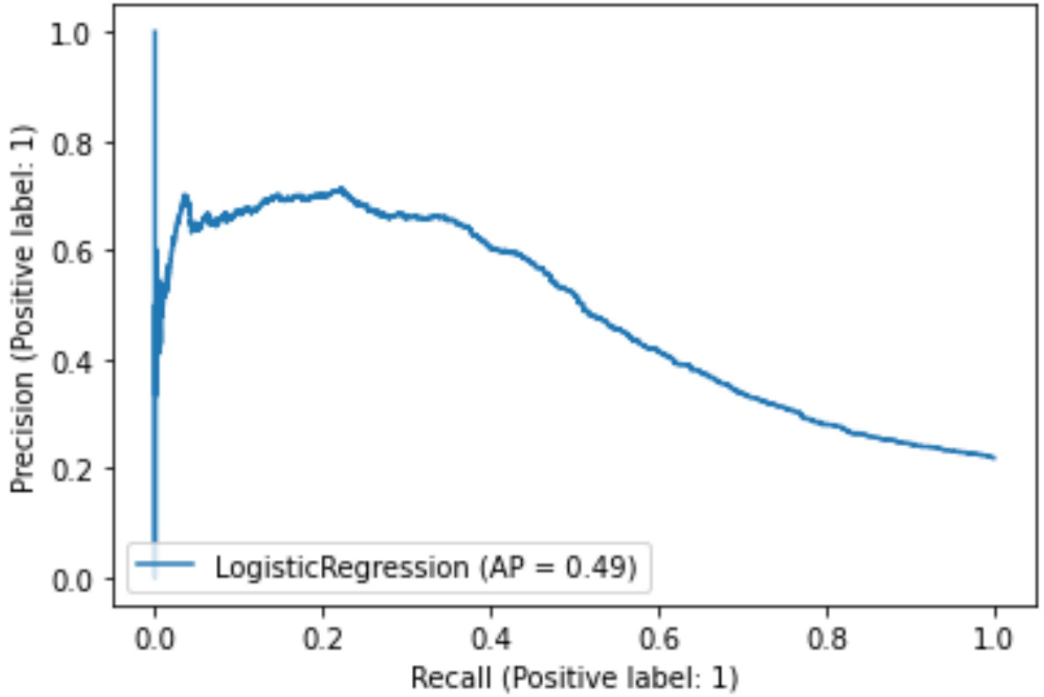
Random Forest is a non-linear ensemble learning algorithm for tasks such as classification. A large number of decision trees can be constructed from a training set. A decision tree is a flowchart-like structure in which each internal node represents a “test” on a feature, each branch represents the outcome of the test, and each leaf node represents a class label.

(3).SVM

A support vector machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space

The outcome of Logistic Regression





	precision	recall	f1-score	support
0	0.82	0.97	0.89	4687
1	0.69	0.24	0.35	1313
accuracy			0.81	6000

Result of all models

	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.82	0.97	0.89	0.81
Random Forest	0.84	0.94	0.89	0.82
SVM	0.84	0.96	0.89	0.82

Conclusion: SVM performs best among these three models.