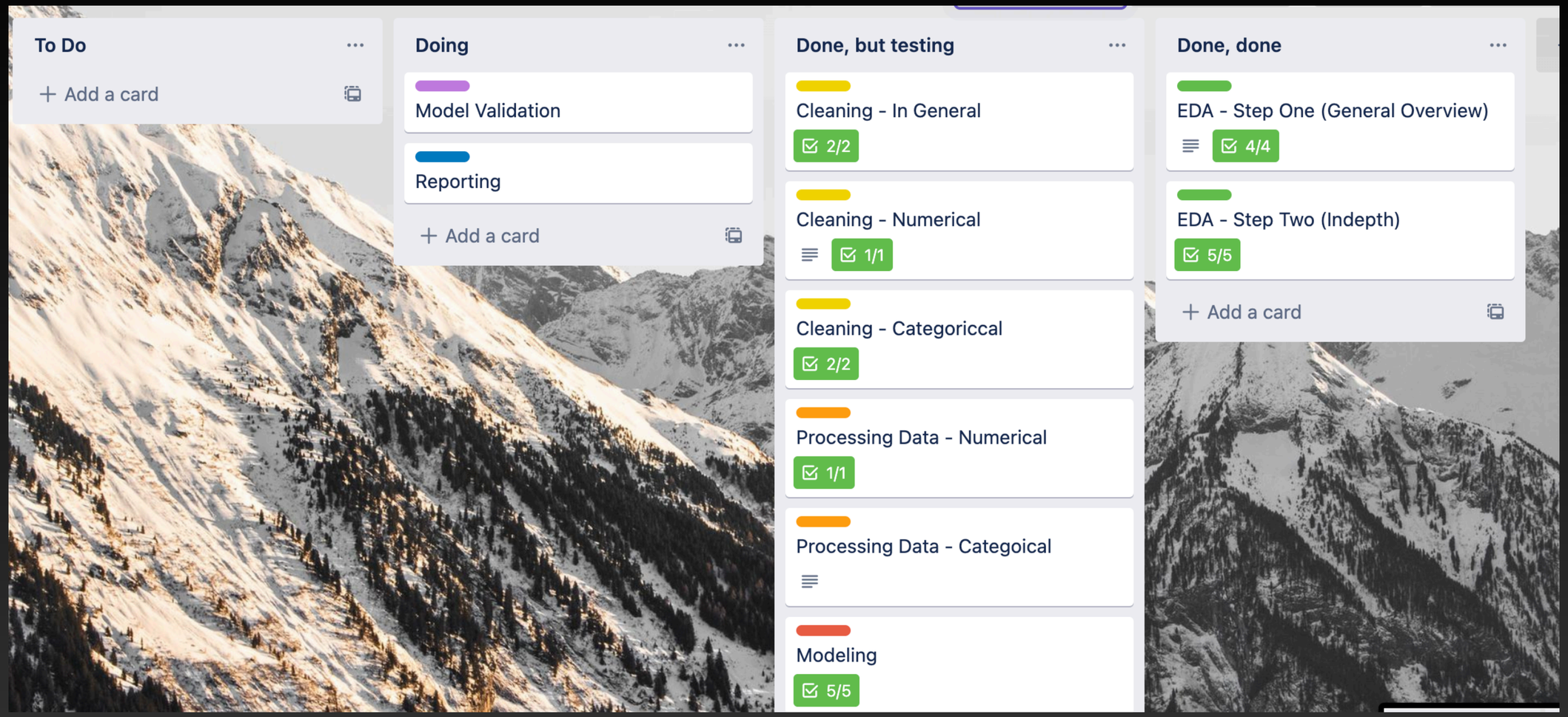
Classification Project

Analysis for marketing purpose

Johana Mar 5th 2022

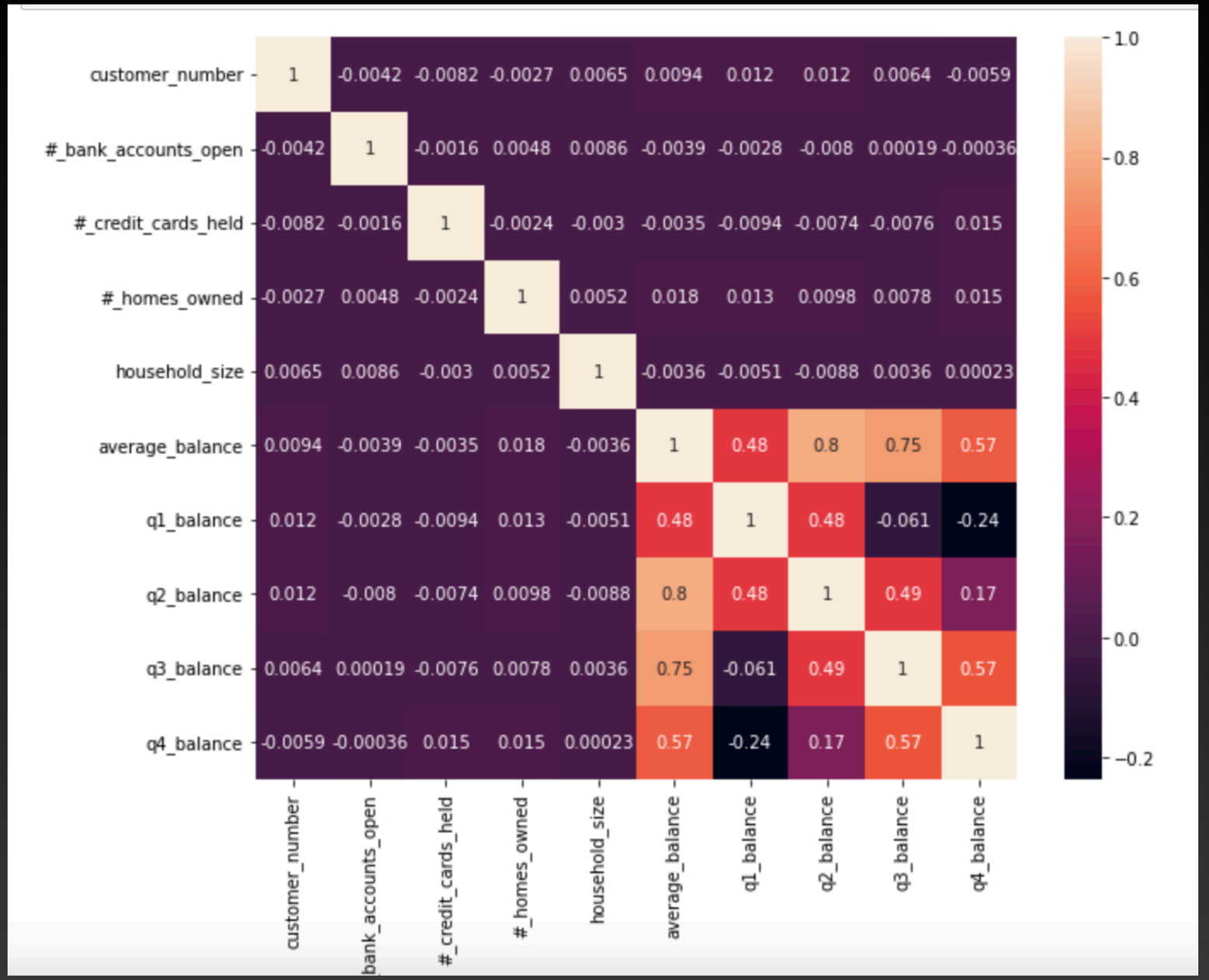
Project Roadmap

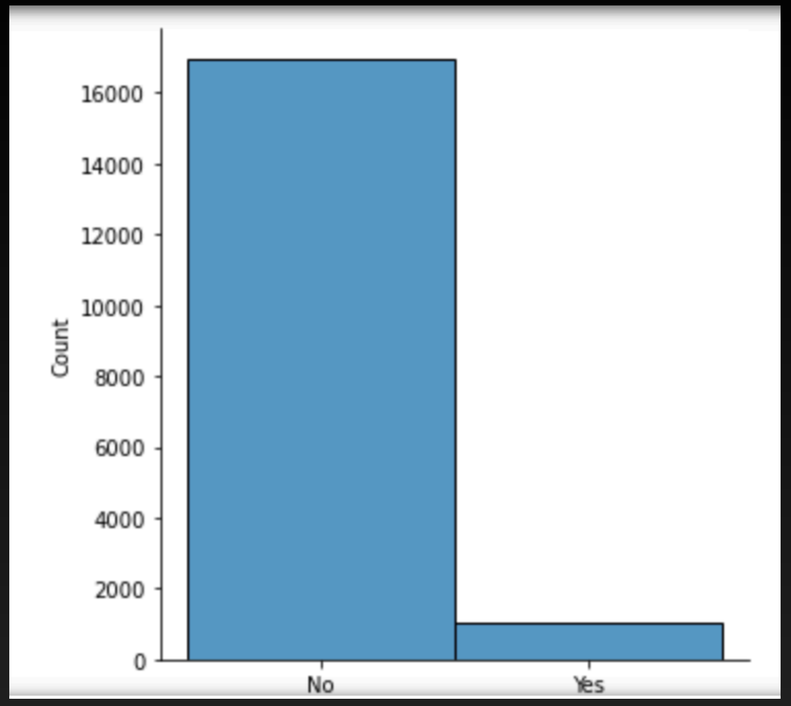
- Preparation
- EDA
- Cleaning
- Featuring
- Modeling
- Reporting

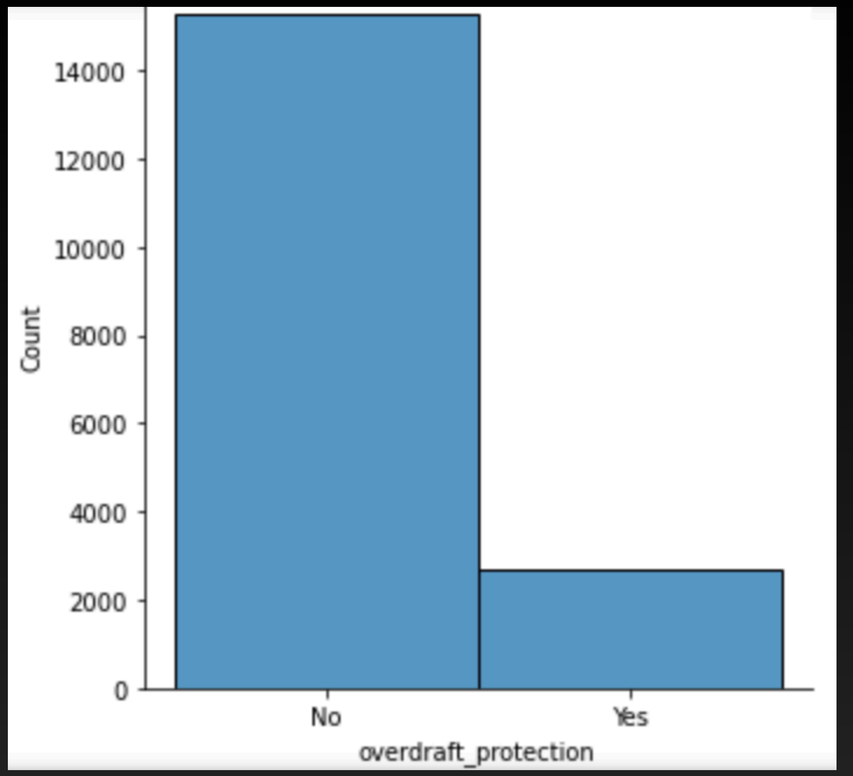


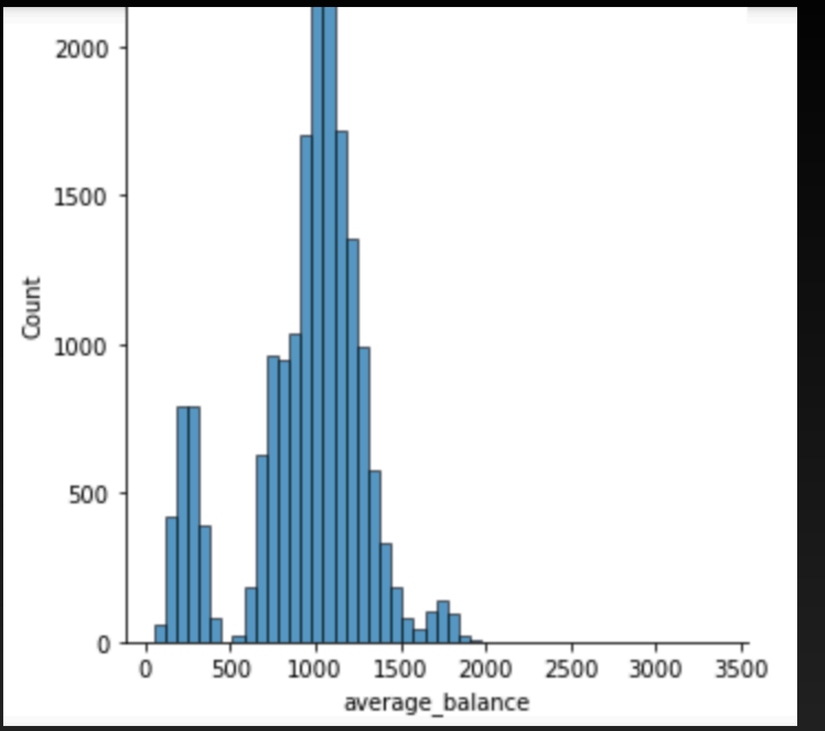
- Only 4 steps, not using kanban.

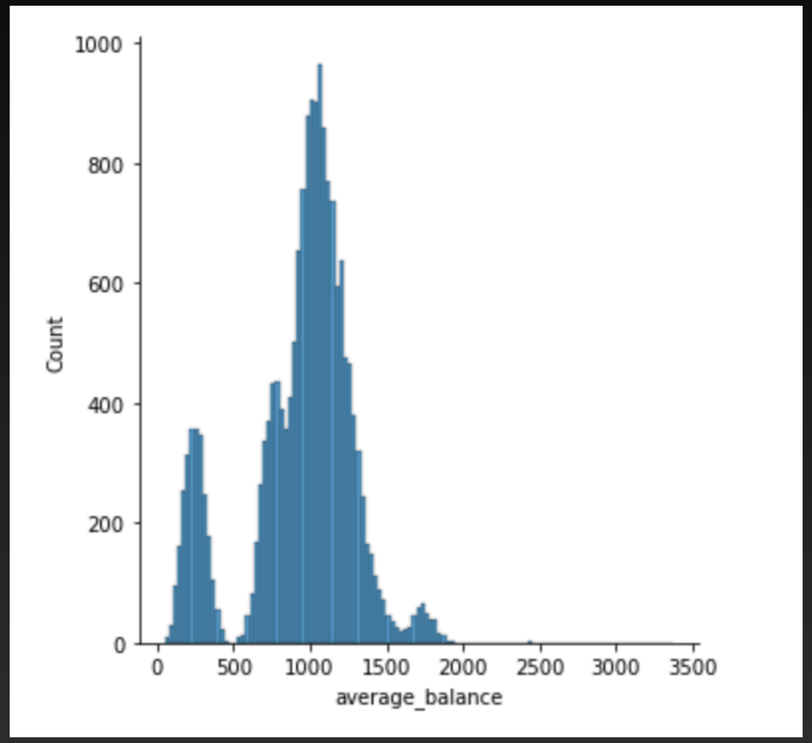
EDA

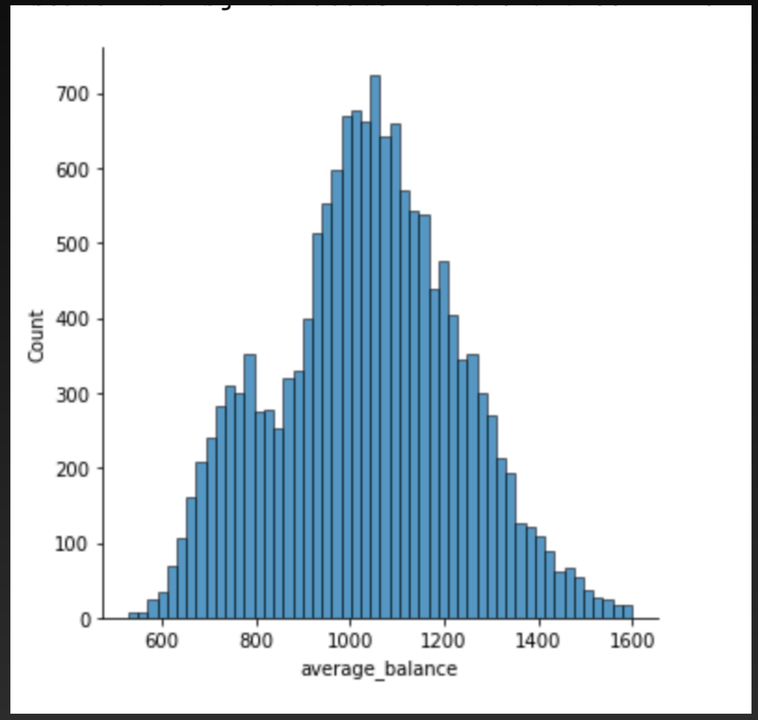

customer_number 1 -0.0042 -0.0082 -0.0027 0.0065 0.0094 0.012 0.012 0.0064 -0.0059
#_bank_accounts_open -0.0042 1 -0.0016 0.0048 0.0086 -0.0039 -0.0028 -0.008 0.00019 -0.00036
#_credit_cards_held -0.0082 -0.0016 1 -0.0024 -0.003 -0.0035 -0.0094 -0.0074 -0.0076 0.015
#_homes_owned -0.0027 0.0048 -0.0024 1 0.0052 0.018 0.013 0.0098 0.0078 0.015
household_size 0.0065 0.0086 -0.003 0.0052 1 -0.0036 -0.0051 -0.0088 0.0036 0.00023
average_balance 0.0094 -0.0039 -0.0035 0.018 -0.0036 1 0.48 0.8 0.75 0.57
q1_balance 0.012 -0.0028 -0.0094 0.013 -0.0051 0.48 1 0.48 -0.061 -0.24
q2_balance 0.012 -0.008 -0.0074 0.0098 -0.0088 0.8 0.48 1 0.49 0.17
q3_balance 0.0064 0.00019 -0.0076 0.0078 0.0036 0.75 -0.061 0.49 1 0.57
q4_balance -0.0059 -0.00036 0.015 0.015 0.00023 0.57 -0.24 0.17 0.57 1
customer_number
bank_accounts_open
#_credit_cards_held
#_homes_owned
household_size
average_balance
q1_balance
q2_balance
q3_balance
q4_balance
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

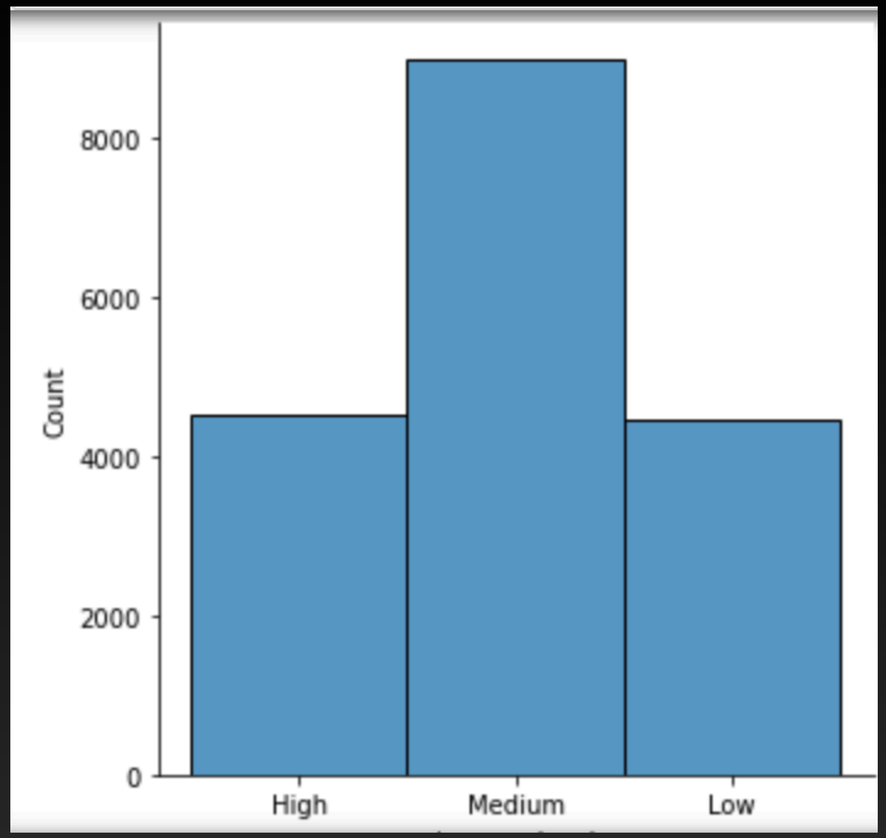

16000
14000
12000
10000
8000
6000
4000
2000
0
Count
No
Yes

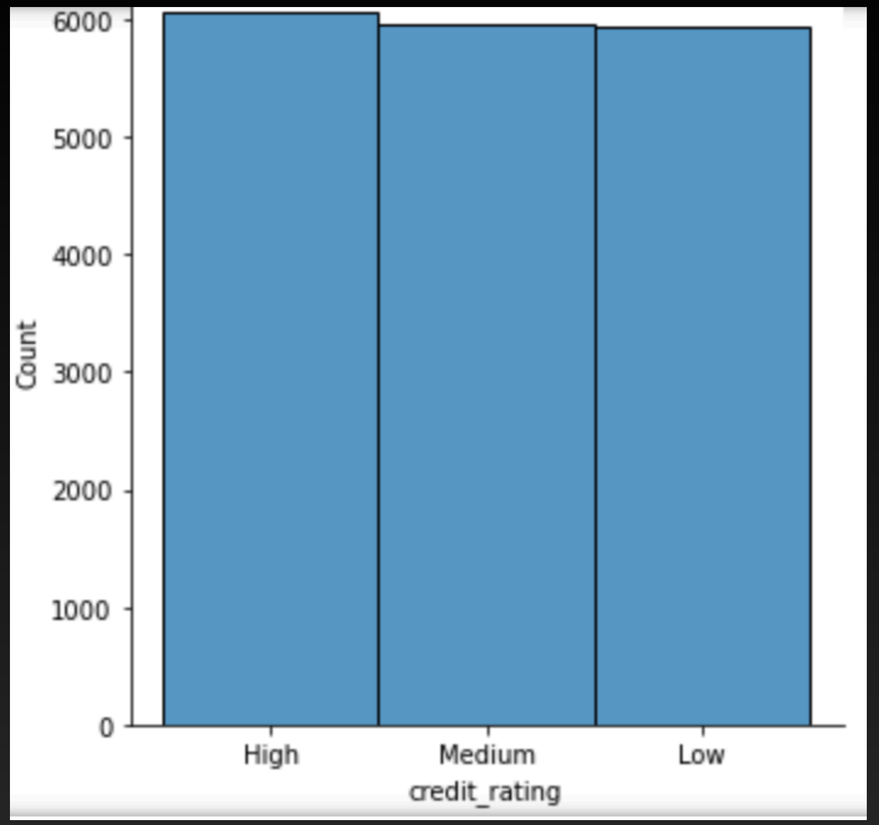

14000
12000
10000
8000
6000
4000
2000
0
Count
No
Yes
overdraft_protection


2000
1500
1000
500
0
Count
0
500
1000
1500
2000
2500
3000
3500
average_balance


1000
800
600
400
200
0
Count
0
500
1000
1500
2000
2500
3000
3500
average_balance


700
600
500
400
300
200
100
0
Count
600
800
1000
1200
1400
1600
average_balance






```
##income_level, credit_rating convert to numerical features, as there is a level increasing in these features

data_avg1['income_level'].replace({"High":"3", "Medium" : "2", "Low" : "1"}, inplace=True)
data_avg1['credit_rating'].replace({"High":"3", "Medium" : "2", "Low" : "1"}, inplace=True)
```

Modeling

Modeling - different featuring

- Avg_income outliers
- Numerical Feature Scaling
- Under Sampling
- Over Sampling - RandomOverSampler, SMOTE
- Coefficient score
- Selective features
- Categorical feature to discrete numerical data
- Etc

Modeling Validation

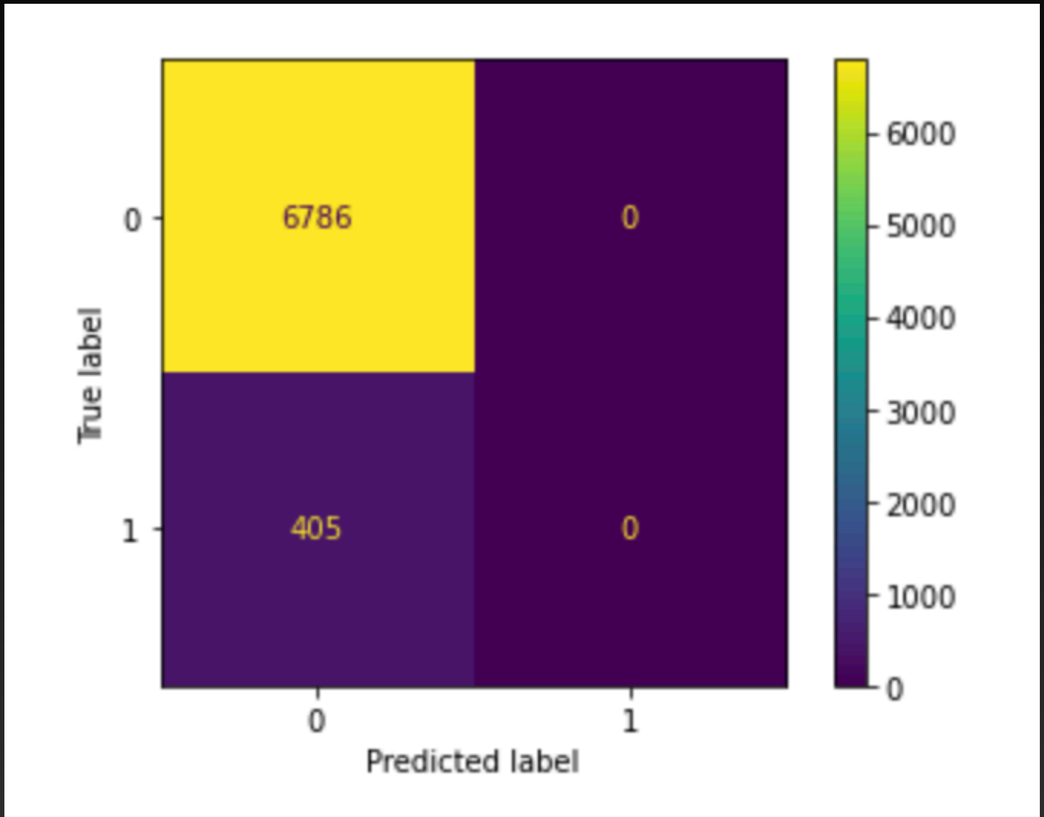
Model 1:

- Transform categorical columns
- It doesn’t have any feature engineering

94 accuracy.

That’s all! Thank you!

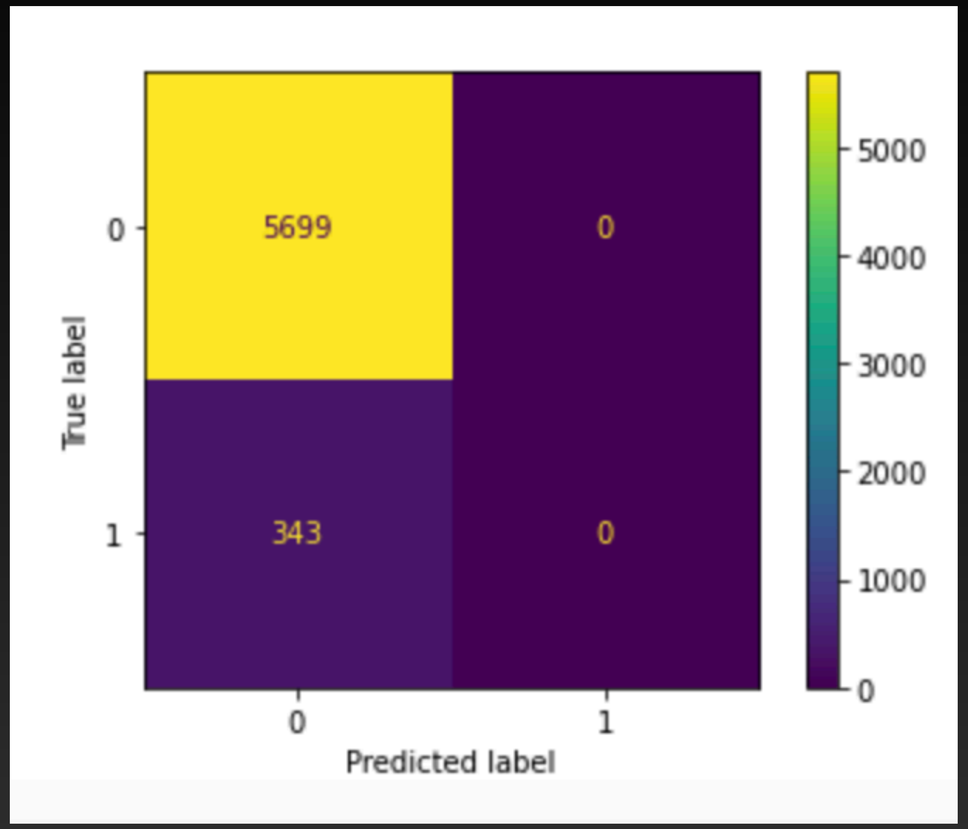
That’s all! Thank you!



Model 1:

- Transform categorical columns
- It doesn’t have any feature engineering

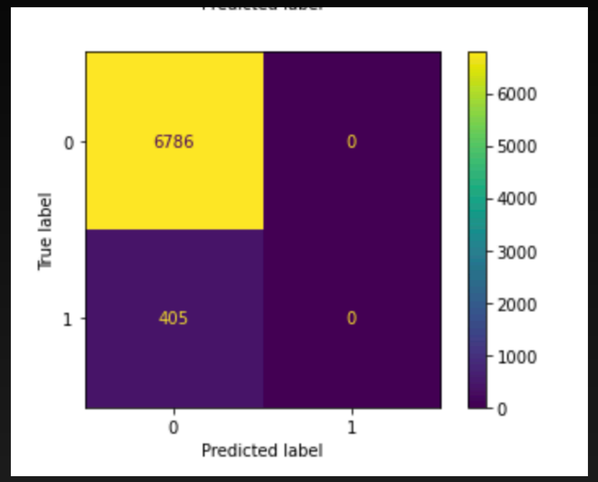
	precision	recall	f1-score	support
0	0.94	1.00	0.97	6786
1	0.00	0.00	0.00	405
accuracy			0.94	7191
macro avg	0.47	0.50	0.49	7191
weighted avg	0.89	0.94	0.92	7191



Model 2:

- Transform categorical columns
- Clean outliers
- Scaling numerical data

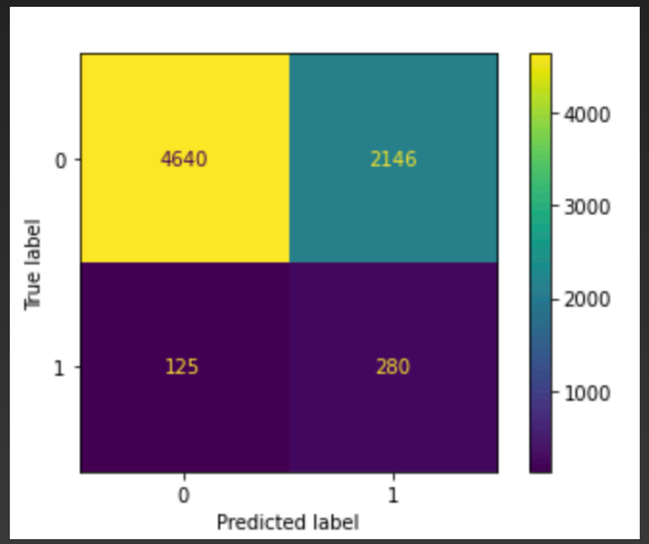
	precision	recall	f1-score	support
0	0.94	1.00	0.97	5699
1	0.00	0.00	0.00	343
accuracy			0.94	6042
macro avg	0.47	0.50	0.49	6042
weighted avg	0.89	0.94	0.92	6042



Model 3:

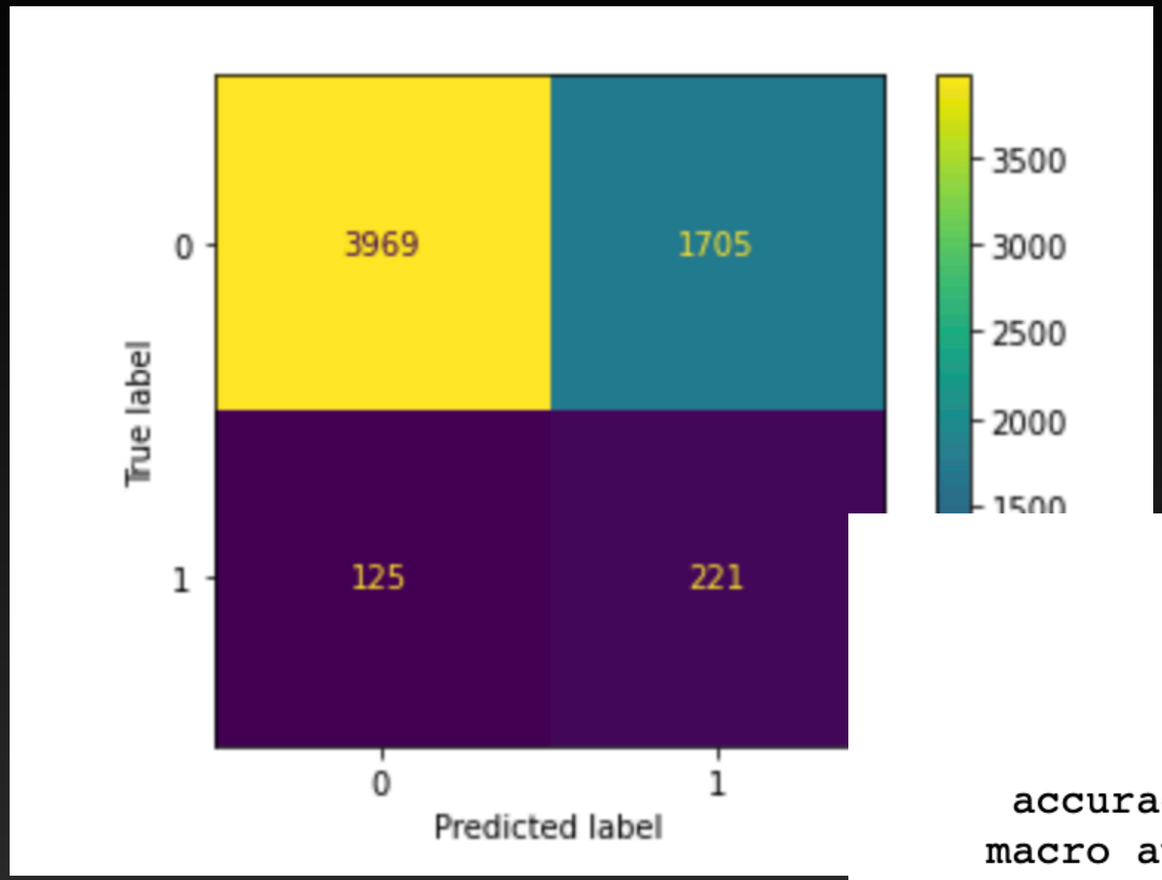
- Transform categorical columns
- Over Sampling and Under Sampling

	precision	recall	f1-score	support
0	0.97	0.66	0.79	6786
1	0.11	0.70	0.19	405
accuracy			0.67	7191
macro avg	0.54	0.68	0.49	7191
weighted avg	0.93	0.67	0.76	7191



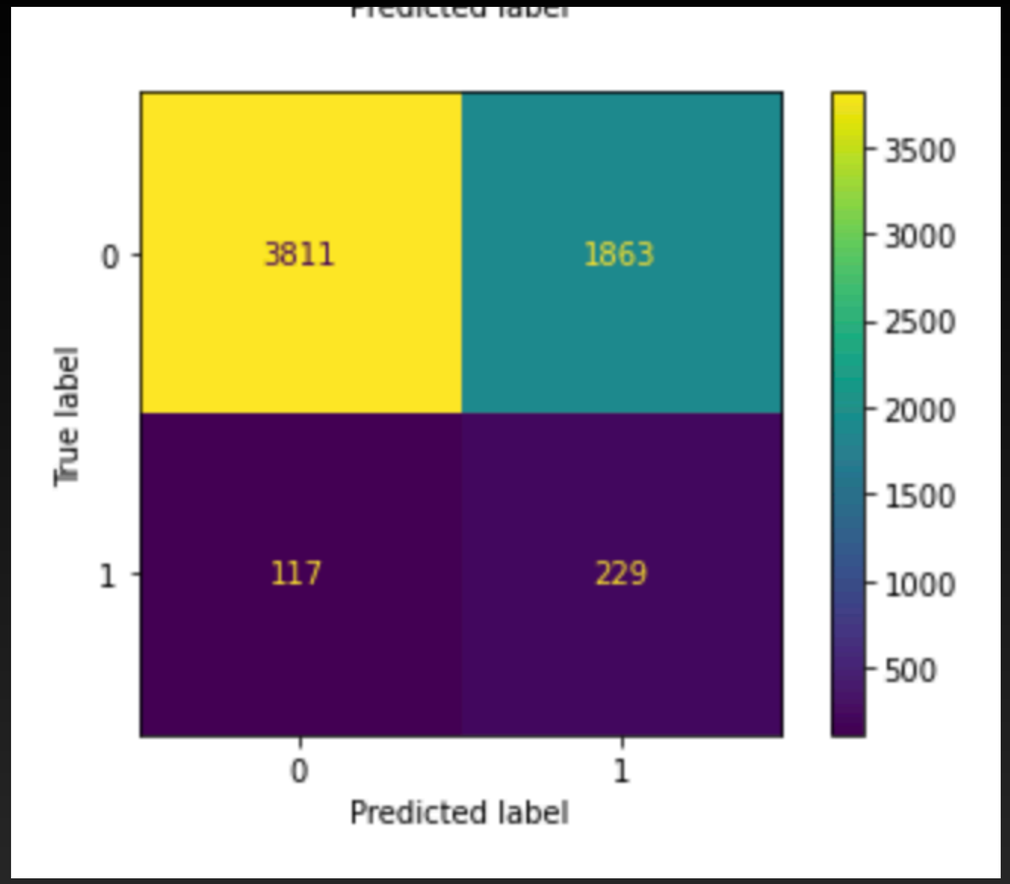
Model 3:

- Transform categorical columns
- Clean outliers
- Scaling numerical data
- Over Sampling - SMOTE



	precision	recall	f1-score	support
0	0.75	0.71	0.73	8521
1	0.72	0.76	0.74	8521
accuracy			0.74	17042
macro avg	0.74	0.74	0.74	17042
weighted avg	0.74	0.74	0.74	17042

	precision	recall	f1-score	support
0	0.97	0.70	0.81	5674
1	0.11	0.64	0.19	346
accuracy			0.70	6020
macro avg	0.54	0.67	0.50	6020
weighted avg	0.92	0.70	0.78	6020

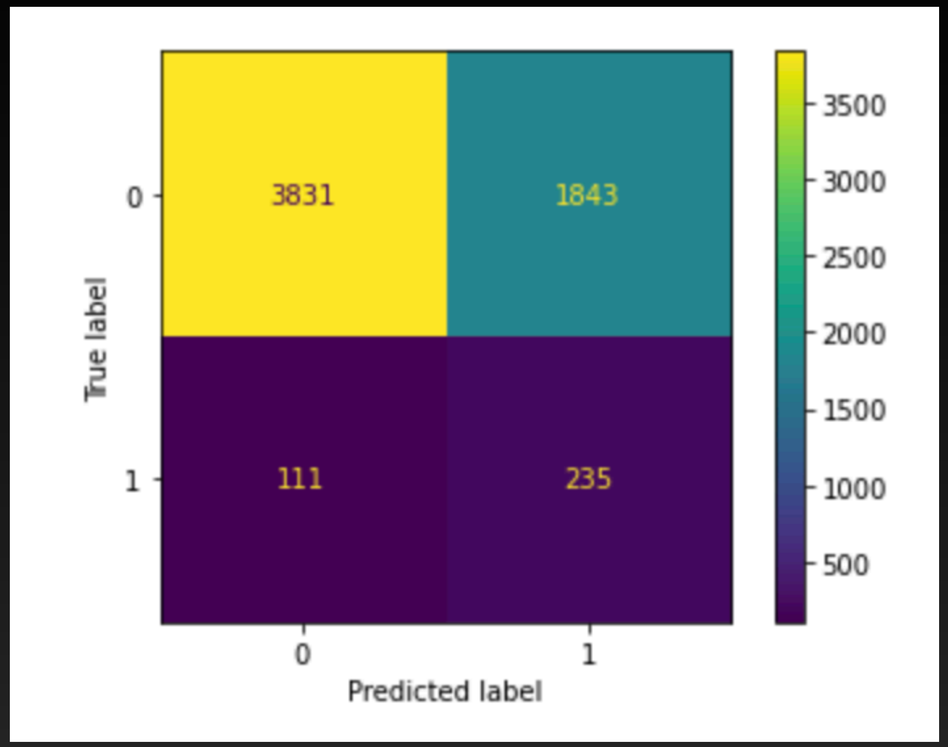


Model 3 _2:

- Transform categorical columns
- Clean outliers
- Scaling numerical data
- Over Sampling - RandomOverSample

	precision	recall	f1-score	support
0	0.72	0.68	0.70	8521
1	0.70	0.73	0.71	8521
accuracy			0.71	17042
macro avg	0.71	0.71	0.71	17042
weighted avg	0.71	0.71	0.71	17042

	precision	recall	f1-score	support
0	0.97	0.67	0.79	5674
1	0.11	0.66	0.19	346
accuracy			0.67	6020
macro avg	0.54	0.67	0.49	6020
weighted avg	0.92	0.67	0.76	6020

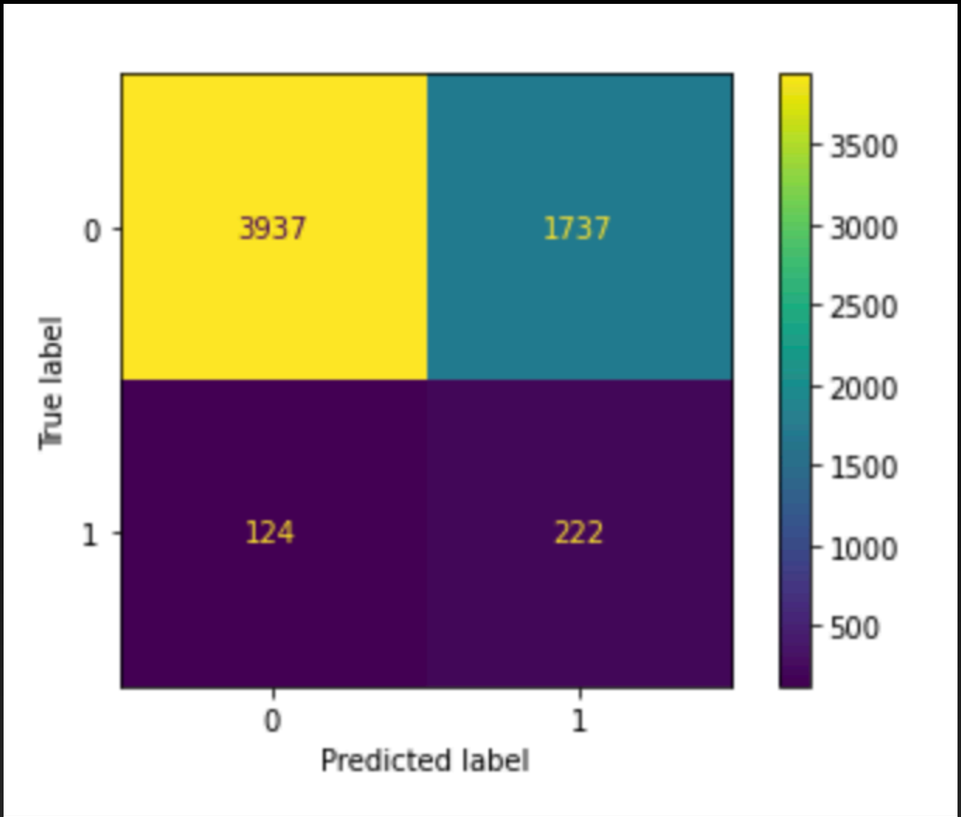


Model 4:

- Transform categorical columns
- Clean outliers
- Scaling numerical data
- Over Sampling - SMOTE
- Only include features with high coef
- [‘income_level','#_bank_accounts_open',
- ‘credit_rating','household_size','reward_Cash

	precision	recall	f1-score	support
0	0.67	0.68	0.68	8521
1	0.68	0.67	0.67	8521
accuracy			0.68	17042
macro avg	0.68	0.68	0.68	17042
weighted avg	0.68	0.68	0.68	17042

	precision	recall	f1-score	support
0	0.97	0.68	0.80	5674
1	0.11	0.68	0.19	346
accuracy			0.68	6020
macro avg	0.54	0.68	0.50	6020
weighted avg	0.92	0.68	0.76	6020



Model 5:

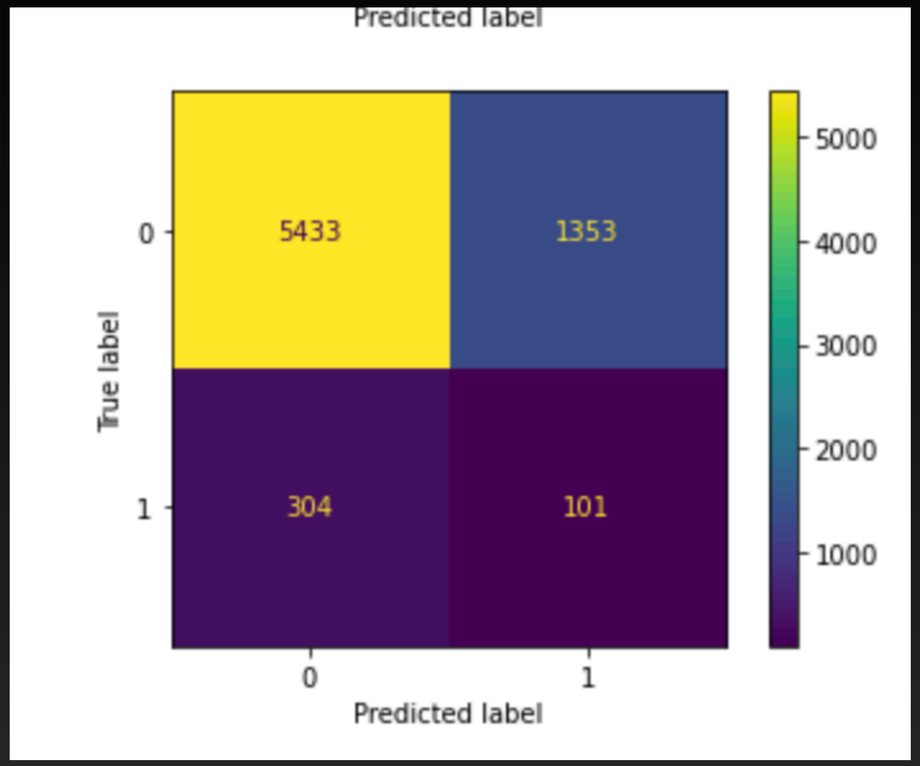
- Transform categorical columns
- Clean outliers
- Scaling numerical data
- Over Sampling - SMOTE
- Only include features that I think it's important

	precision	recall	f1-score	support
0	0.74	0.70	0.72	8521
1	0.72	0.76	0.74	8521
accuracy			0.73	17042
macro avg	0.73	0.73	0.73	17042
weighted avg	0.73	0.73	0.73	17042

	precision	recall	f1-score	support
0	0.97	0.69	0.81	5674
1	0.11	0.64	0.19	346
accuracy			0.69	6020
macro avg	0.54	0.67	0.50	6020
weighted avg	0.92	0.69	0.77	6020

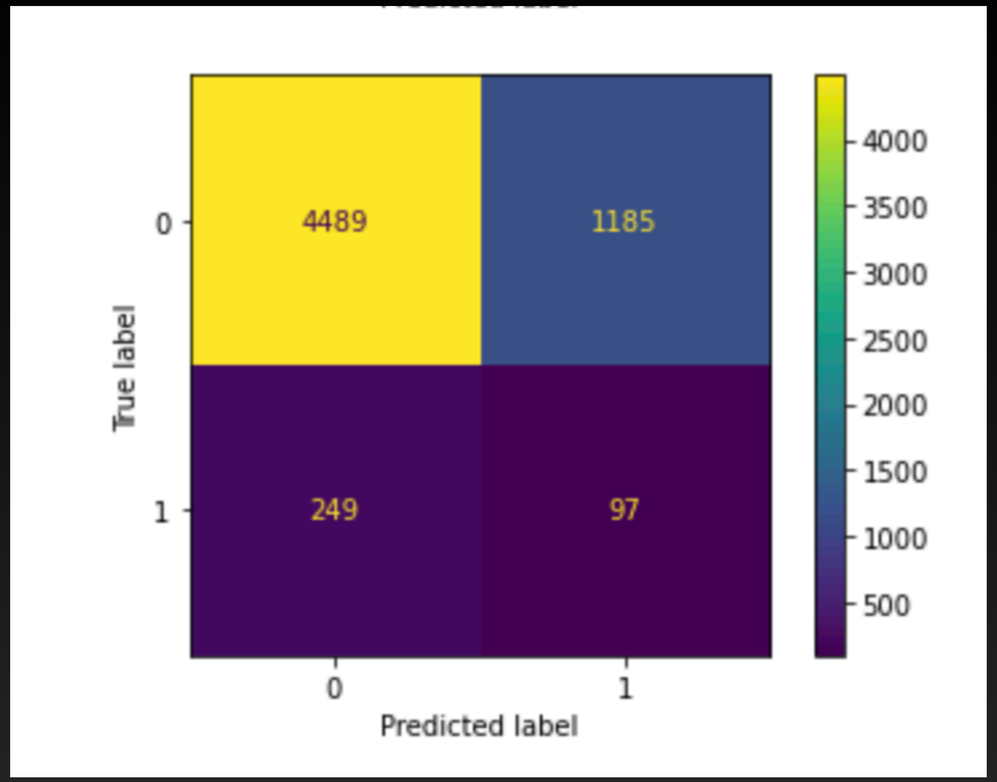
Model 7:

- Encoding
- Clean outliers
- Over Sampling - SMOTE



	precision	recall	f1-score	support
0	0.83	0.80	0.81	10169
1	0.81	0.83	0.82	10169
accuracy			0.82	20338
macro avg	0.82	0.82	0.82	20338
weighted avg	0.82	0.82	0.82	20338

	precision	recall	f1-score	support
0	0.95	0.80	0.87	6786
1	0.07	0.25	0.11	405
accuracy			0.77	7191
macro avg	0.51	0.53	0.49	7191
weighted avg	0.90	0.77	0.82	7191



Model 8:

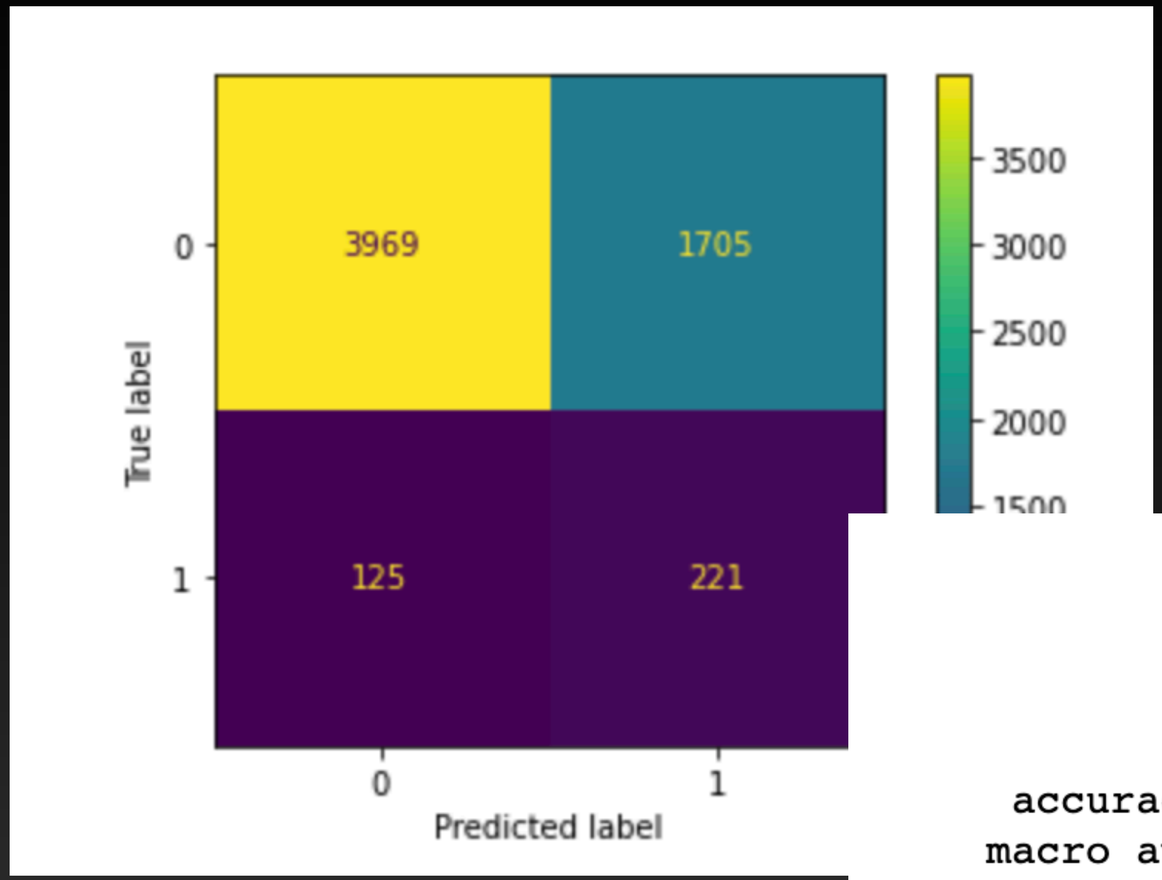
- Encoding categorical feature (credit rate, incon
- Cleaning outliers
- Scaling numerical data
- Over Sampling - SMOTE

	precision	recall	f1-score	support
0	0.81	0.79	0.80	8521
1	0.80	0.81	0.80	8521
accuracy			0.80	17042
macro avg	0.80	0.80	0.80	17042
weighted avg	0.80	0.80	0.80	17042

	precision	recall	f1-score	support
0	0.95	0.79	0.86	5674
1	0.08	0.28	0.12	346
accuracy			0.76	6020
macro avg	0.51	0.54	0.49	6020
weighted avg	0.90	0.76	0.82	6020

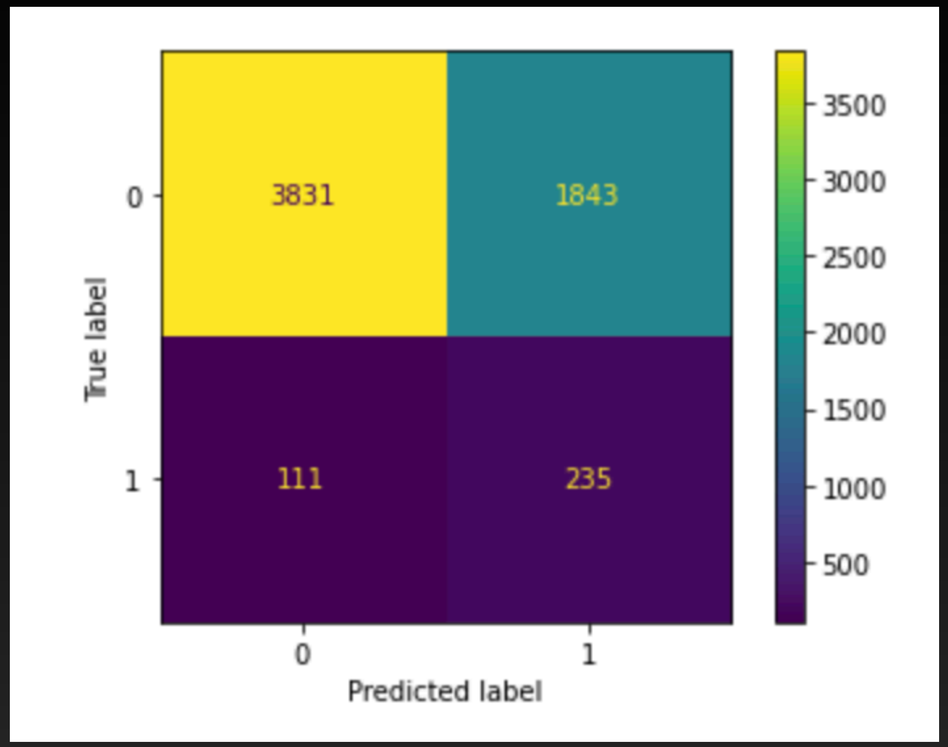
Model 3:

- Transform categorical columns
- Clean outliers
- Scaling numerical data
- Over Sampling - SMOTE



	precision	recall	f1-score	support
0	0.75	0.71	0.73	8521
1	0.72	0.76	0.74	8521
accuracy			0.74	17042
macro avg	0.74	0.74	0.74	17042
weighted avg	0.74	0.74	0.74	17042

	precision	recall	f1-score	support
0	0.97	0.70	0.81	5674
1	0.11	0.64	0.19	346
accuracy			0.70	6020
macro avg	0.54	0.67	0.50	6020
weighted avg	0.92	0.70	0.78	6020



Model 4:

- Transform categorical columns
- Clean outliers
- Scaling numerical data
- Over Sampling - SMOTE
- Only include features with high coef
- [‘income_level','#_bank_accounts_open',
- ‘credit_rating','household_size','reward_Cash

	precision	recall	f1-score	support
0	0.67	0.68	0.68	8521
1	0.68	0.67	0.67	8521
accuracy			0.68	17042
macro avg	0.68	0.68	0.68	17042
weighted avg	0.68	0.68	0.68	17042

	precision	recall	f1-score	support
0	0.97	0.68	0.80	5674
1	0.11	0.68	0.19	346
accuracy			0.68	6020
macro avg	0.54	0.68	0.50	6020
weighted avg	0.92	0.68	0.76	6020

	features	coef
0	income_level	-0.806725
1	#_bank_accounts_open	-0.397877
2	credit_rating	-1.382343
3	#_credit_cards_held	-0.139592
4	#_homes_owned	-0.151236
5	household_size	-0.037276
6	average_balance	0.000647
7	reward_Cash Back	-3.781316
8	reward_Points	-1.832542
9	mailer_type_Postcard	-0.150016
10	overdraft_protection_Yes	-1.514271
11	own_your_home_Yes	-0.381570

9]:

	features	coef
0	income_level	-0.635868
1	#_bank_accounts_open	-0.286355
2	credit_rating	0.724695
3	#_credit_cards_held	-0.058078
4	#_homes_owned	0.053293
5	household_size	-0.198624
6	average_balance	-0.037979
7	reward_Cash Back	-0.488377
8	reward_Points	-0.029235
9	mailer_type_Postcard	0.007080
10	overdraft_protection_Yes	-0.016451
11	own_your_home_Yes	0.024470

	features	coef
0	income_level	-0.635868
1	#_bank_accounts_open	-0.286355
2	credit_rating	0.724695
3	household_size	-0.058078
4	reward_Cash Back	0.053293

	features	coef
0	income_level	-1.218828
1	#_bank_accounts_open	-0.545751
2	credit_rating	1.382404
3	#_homes_owned	-0.401533
4	household_size	-0.072609
5	reward_Cash Back	-0.963573
6	reward_Points	0.013584
7	mailer_type_Postcard	-0.036347

Important Features

- Credit_rating
- Income_level
- Reward_cash Back
- #_of_bank_accounts_open
- household_size

Findings

- Imbalanced data:
- Over Sampling returns a much better results than Under Sampling.
- Random Over Sampling out performs SMOTE in this model.
- Without oversampling, the model is useless.

- Outliers cleaning has positive impact on the model
- Numerical scaling has positive impact

- Coefficient score is more reliable than my understanding
- Categorical data -> Numerical discrete data performs better encoding categorical data on “yes” results

Things to improve

- Deep in each feature to see the weight of the features
- Use different models to see the differences.
- Build function to automate the process
- Visualization
- The documentation