Supervised Machine Learning: Classification Project: Predicting whether individual income exceeds $50K/yr

Napoleon Dewan
November 2023

Content

1. Main objective
2. Data Description
3. Exploratory Data Analysis and feature engineering
4. Classification models
5. Comparison of models
6. Key Findings and Insights
7. The next steps

Main objective

Prediction of whether an individual’s income exceeds $50K/yr based on "Census Income" dataset (Known as Adult Dataset)

Personal income often considered as one of indicators of welfare is subject to discussion in social science’ discipline. Higher income means higher opportunities for health, education, living standard and overall well being.

The ability to earn higher income depends on a number of factors ranging from education level to age. Predicting income to a certain threshold can have an important implications for the government and development organizations in designing social programs such as unemployment benefit, cash transfer, food subsidy.

Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.

Data Description

	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K

Raw dataset has 15 variables, some of which are unrecognizable without proper column name.

Data are correctly named with the following variables (2 variables are dropped due to irrelevance to this analysis

age, workclass, education, marital_status, occupation, relationship, race, sex, capital_gain , capital_loss, hours_per_week, native_country, income

Data Description contd.

Age = number of years (Integer)

Workclass = type of work (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked) (Categorical)

Education = Education Level (Categorical)

Marital-status = Married-civil-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse (Categorical)

Occupation = type of occupation (i.e, Sales, Farming-fishing) (Categorical)

Relationship (Categorical), race (Categorical), sex (Categorical), capital-gain = profit from capital (integer), Hours-per-week (Integer)

Capital-loss = loss from capital (Integer)

Native-country = origin of country (Categorical)

Income = the target variable (categorical) (<=50k or >50k)

Exploratory Data Analysis and feature engineering

' ?' category has been removed from 3 categorical variables (workclass, occupation, native_country)

There is no missing value

```
age               0
workclass         0
education         0
marital_status    0
occupation        0
relationship      0
race              0
sex               0
capital_gain      0
capital_loss      0
hours_per_week    0
native_country    0
income            0
dtype: int64
```

Shape = (30161, 13)

Categorical labels are encoded using LabelEncoder

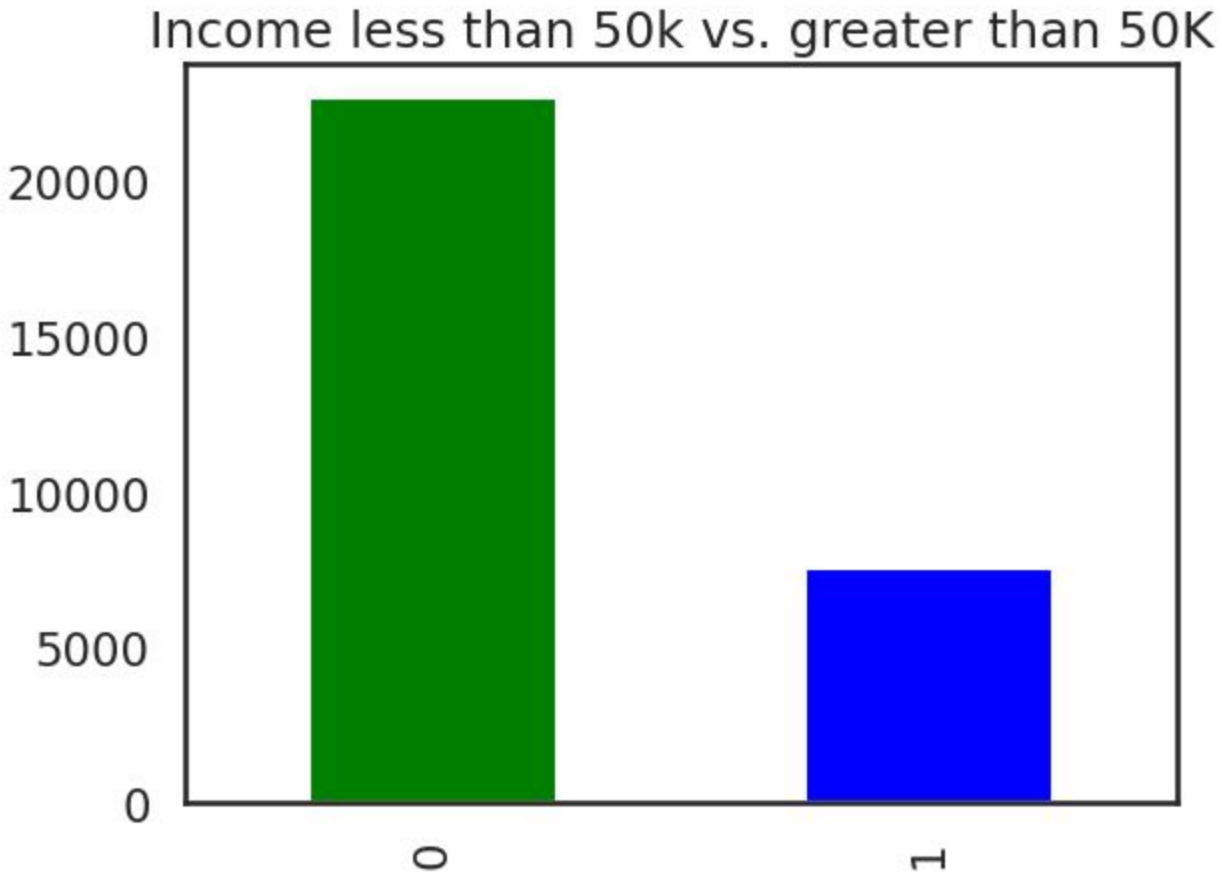
Exploratory Data Analysis and feature engineering contd.

Data after encoding

	age	workclass	education	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	nativ
0	50	4	9	2	3	0	4	1	0	0	13	
1	38	2	11	0	5	1	4	1	0	0	40	
2	53	2	1	2	5	0	2	1	0	0	40	
3	28	2	9	2	9	5	2	0	0	0	40	
4	37	2	12	2	3	5	4	0	0	0	40	
...	...	...	...	...	...	...	...	...	...	...	...	
32555	27	2	7	2	12	5	4	0	0	0	38	
32556	40	2	11	2	6	0	4	1	0	0	40	
32557	58	2	11	6	0	4	4	0	0	0	40	
32558	22	2	11	4	0	3	4	1	0	0	20	
32559	52	3	11	2	3	5	4	0	15024	0	40	

30161 rows × 13 columns

Exploratory Data Analysis and feature engineering contd.



Target variable

Income (binary)

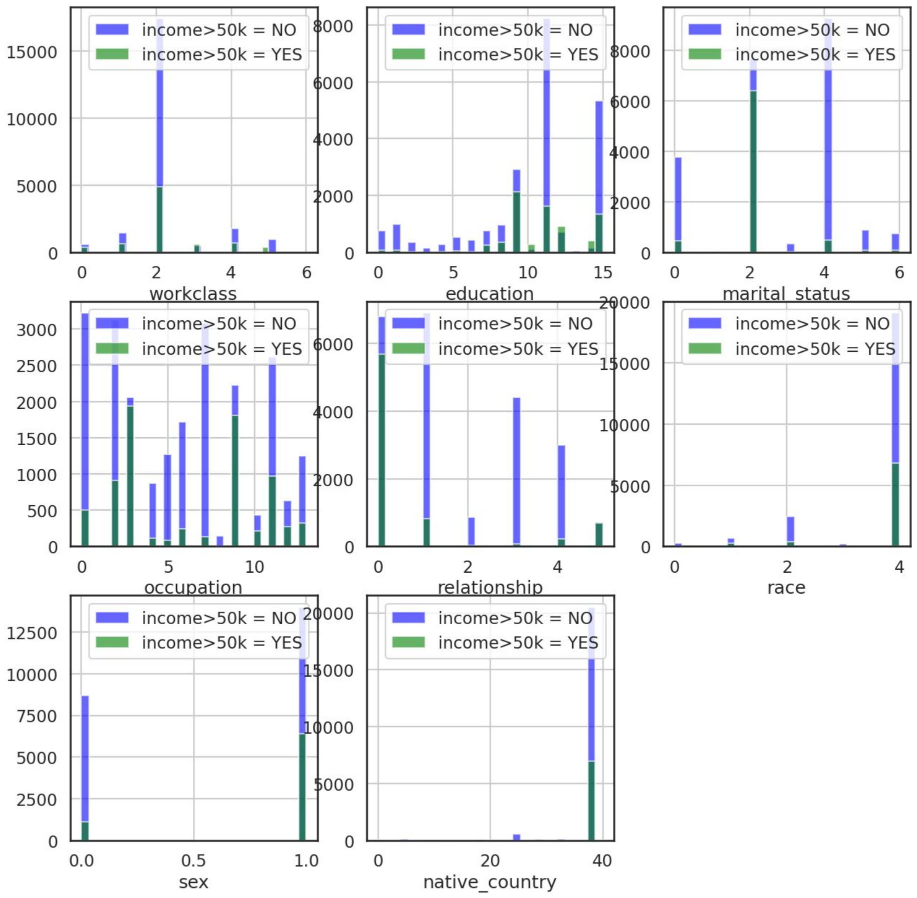
>50K = Yes = 1

<=50K = No = 0

0 = 22653 (75.11%)

1 = 7508 (24.89%)

Exploratory Data Analysis and feature engineering contd.

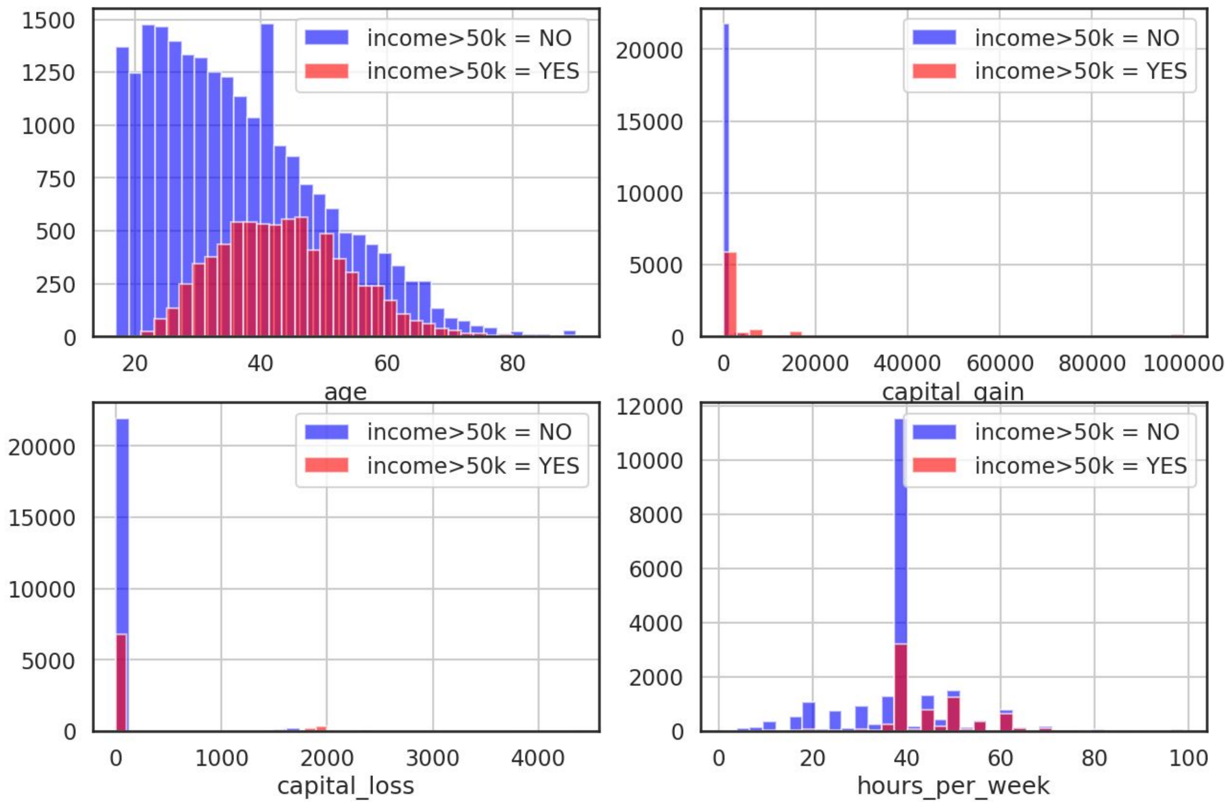


The relationship of target variable with other categorical variable is not obvious

For education, occupation and relationship, marital status, values are distributed approximately across all categories.

For other variables, values are clustered into certain categories.

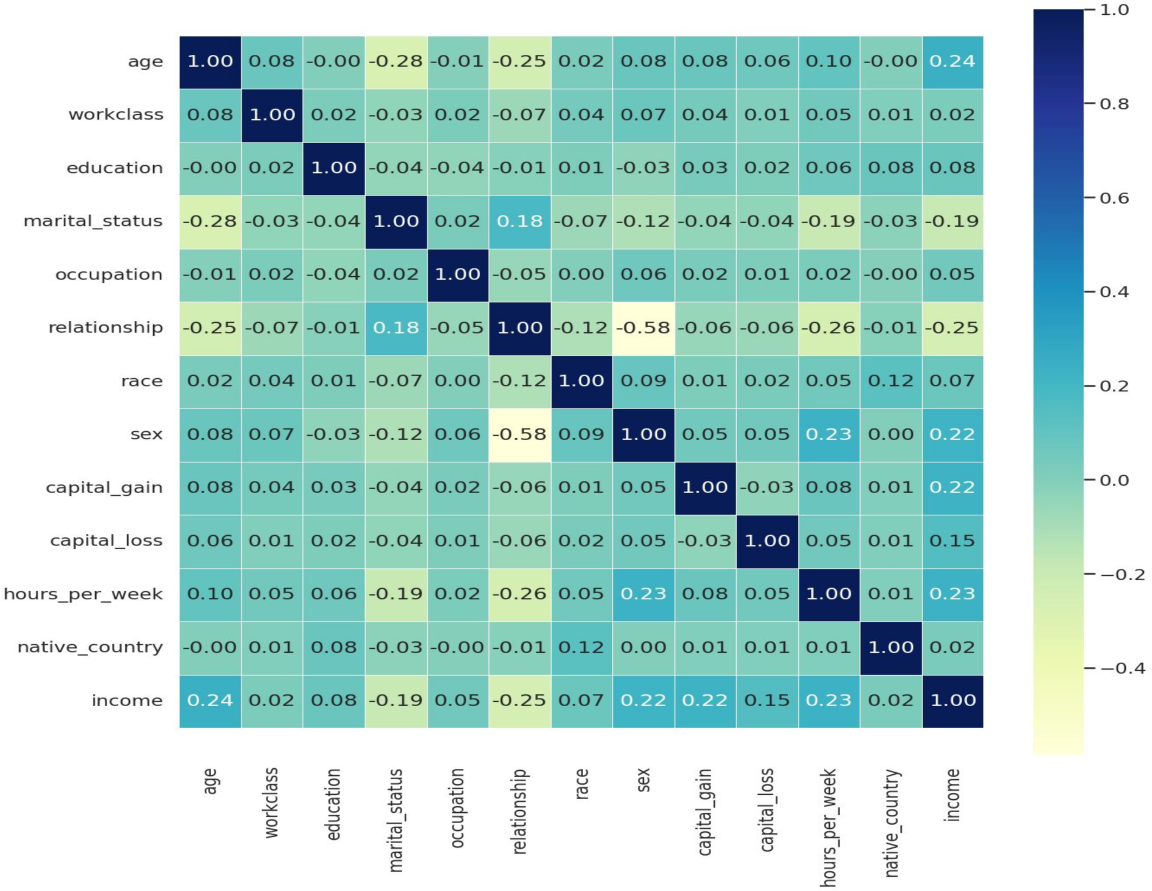
Exploratory Data Analysis and feature engineering contd.



Among the continuous variable, age with >50K and hours per week are close to symmetrical distribution.

Capital gain and capital loss are positively skewed.

Exploratory Data Analysis and feature engineering contd.



Heatmap shows the correlation among all variables.

The correlation between workclass and target variable is just 0.02.

Similar is the case with naive_country.

So they can be dropped from the analysis.

Exploratory Data Analysis and feature engineering contd.

Numerical variables ['age', 'capital_gain', 'capital_loss', 'hours_per_week'] are scaled using StandardScaler()

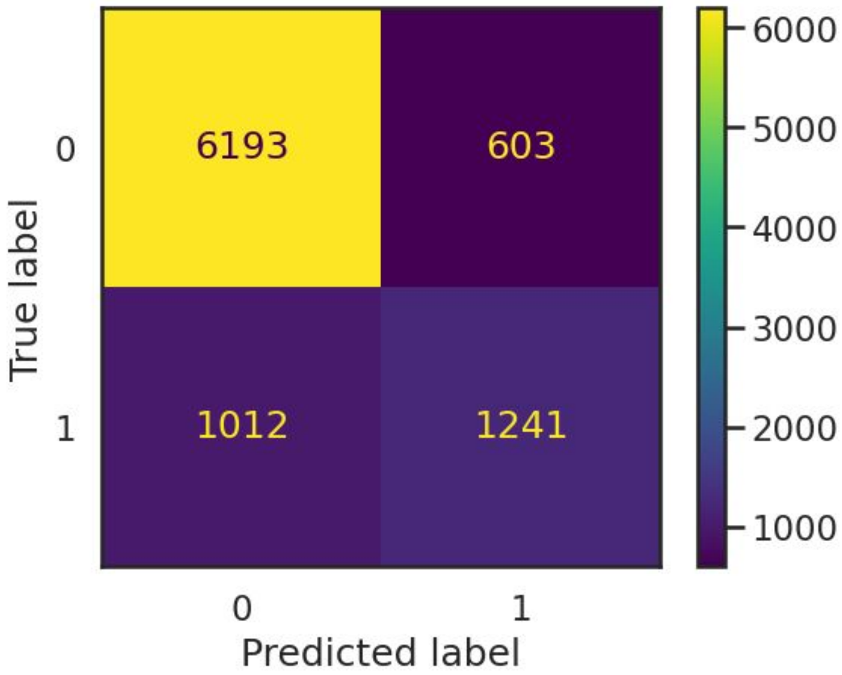
Correlation among features

	feature1	feature2	correlation	abs_correlation
41	sex	hours_per_week	0.23	0.23
18	marital_status	relationship	0.18	0.18
8	age	hours_per_week	0.10	0.10
30	relationship	race	-0.12	0.12
20	marital_status	sex	-0.12	0.12
23	marital_status	hours_per_week	-0.19	0.19
3	age	relationship	-0.25	0.25
34	relationship	hours_per_week	-0.26	0.26
1	age	marital_status	-0.28	0.28
31	relationship	sex	-0.58	0.58

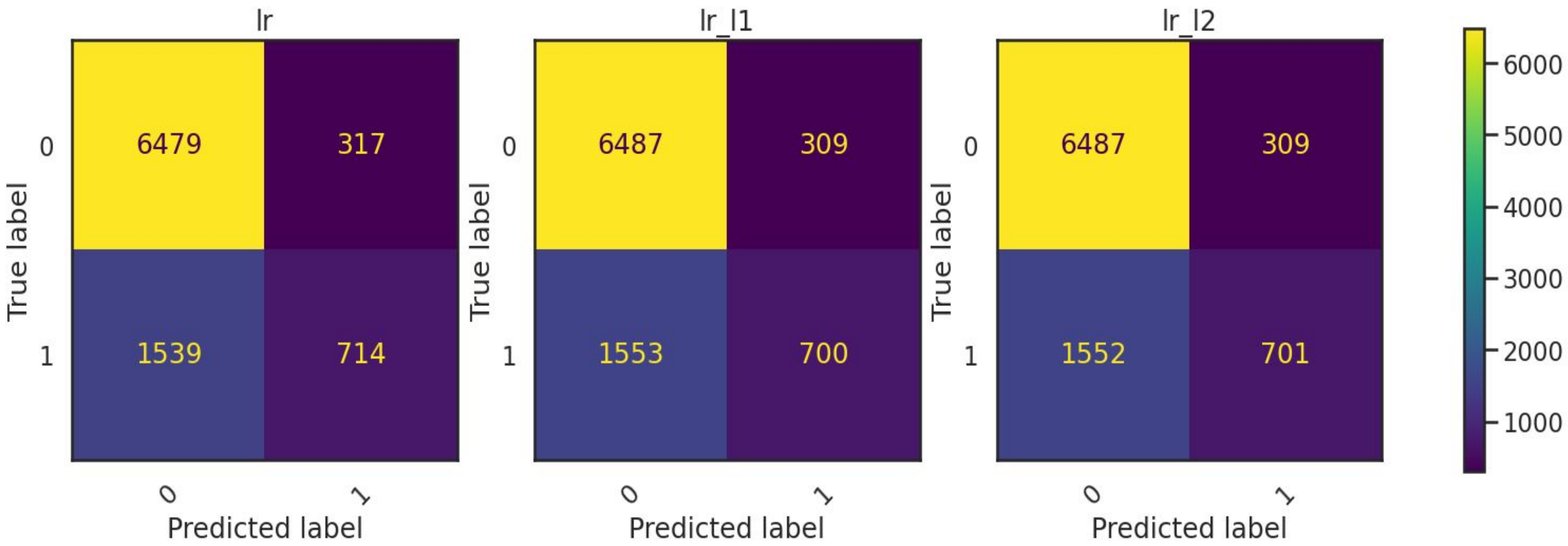
Classification models

Logistic Regression

	0	1	accuracy	macro avg	weighted avg
precision	0.81	0.69	0.79	0.75	0.78
recall	0.95	0.32	0.79	0.64	0.79
f1-score	0.87	0.43	0.79	0.65	0.77
support	6796.00	2253.00	0.79	9049.00	9049.00



Classification models contd.

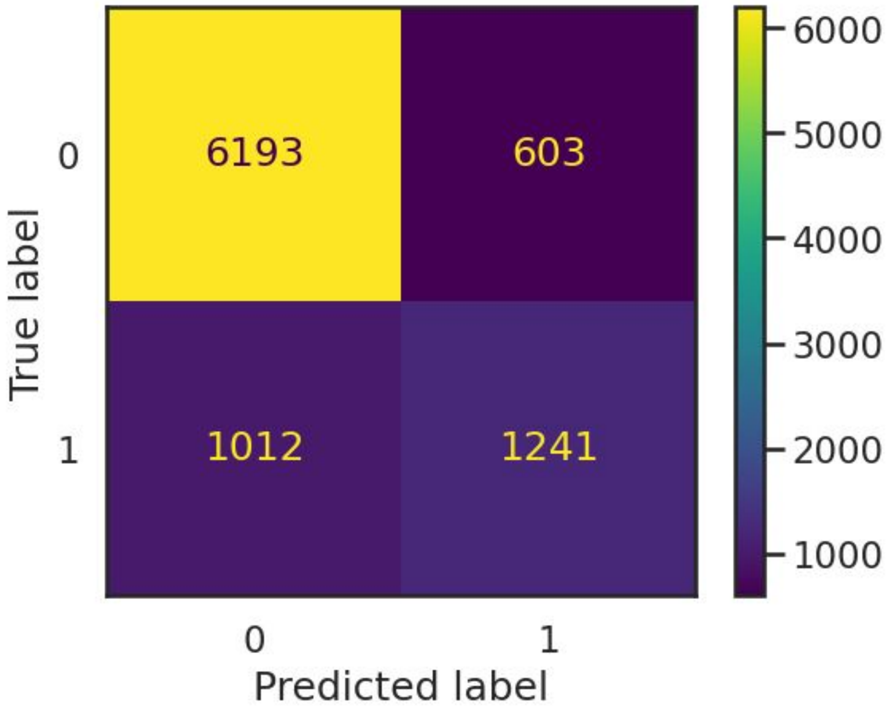


Regularization with Lasso and Ridge. It makes no difference.

Classification models contd.

K-Nearest Neighbors

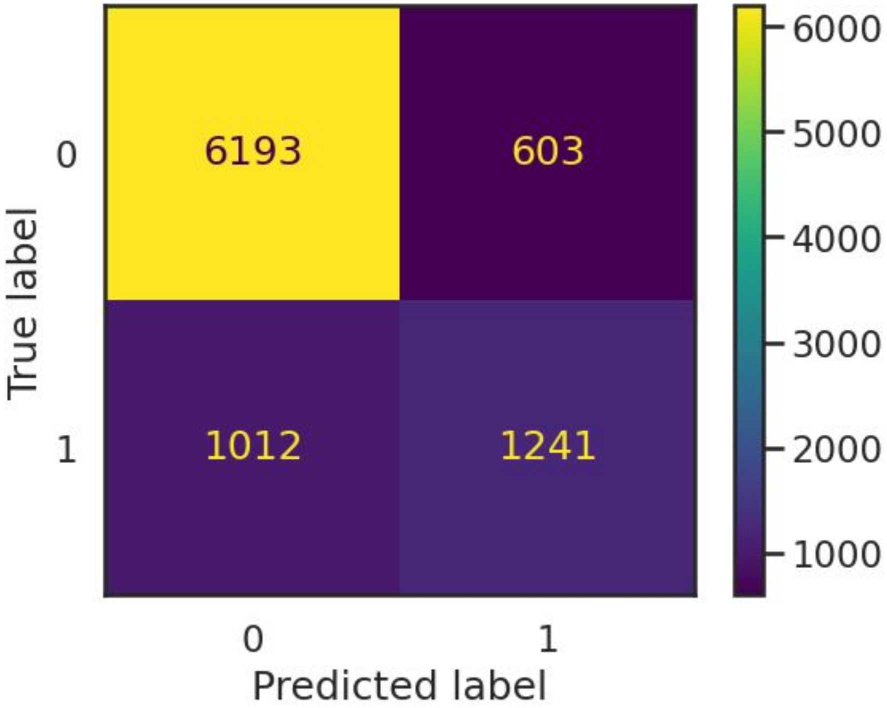
	0	1
precision	0.86	0.64
recall	0.89	0.57
f1-score	0.88	0.60
support	6796.00	2253.00



Classification models contd.

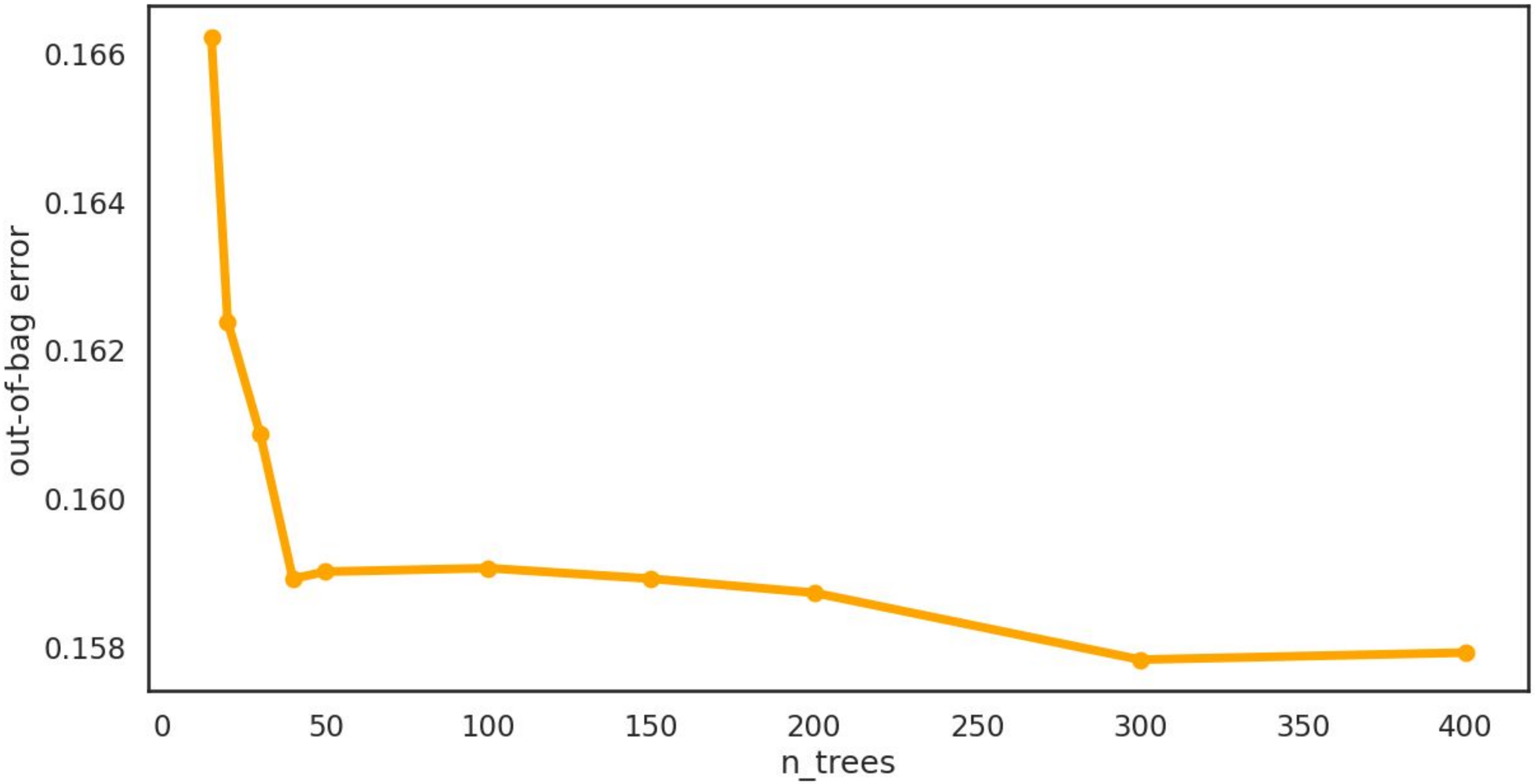
Decision Tree

	0	1
precision	0.87	0.65
recall	0.89	0.59
f1-score	0.88	0.62
support	6796.00	2253.00



Classification models contd.

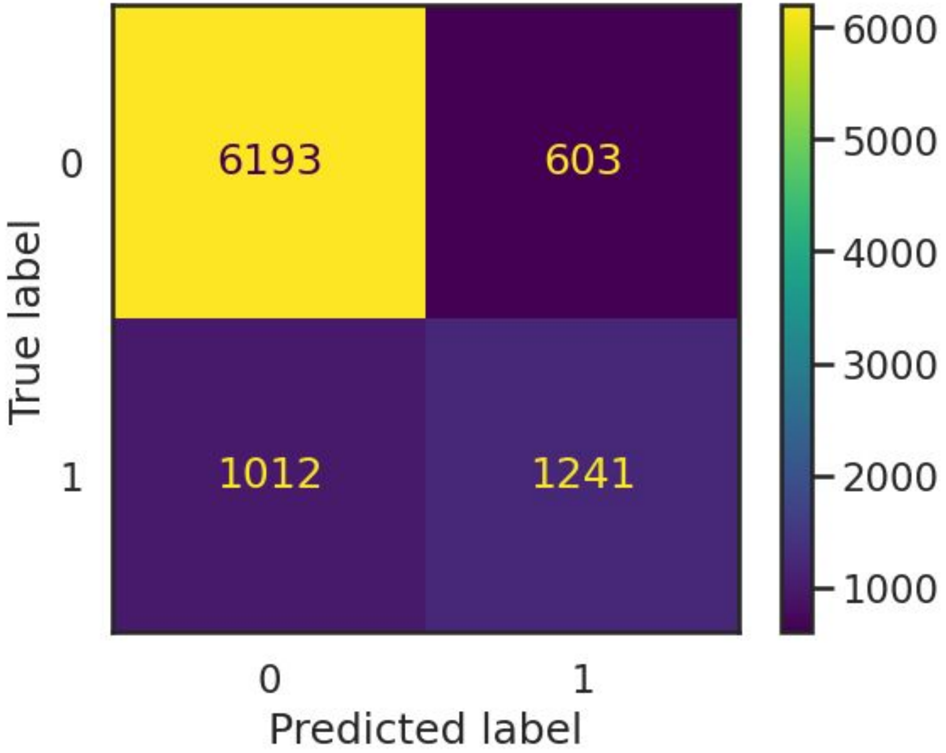
Random Forest



Classification models contd.

Random Forest

	0	1
precision	0.88	0.70
recall	0.91	0.62
f1-score	0.90	0.66
support	6796.00	2253.00



Comparison of models

Combined metrics

	precision	recall	accuracy	f1score	auc
Logistic Regression	0.78	0.79	0.79	0.43	0.64
KNN	0.81	0.81	0.81	0.60	0.73
Decision Tree	0.81	0.82	0.82	0.62	0.74
Random Forest	0.83	0.84	0.84	0.66	0.77

Key Findings and Insights

Among the four models, precision, recall and accuracy are mostly similar.

However, F1score and AUC are comparatively lower for Logistic regression model, suggesting a trade off between precision and recall.

Based on the metrics derived, the following ranking can be made in terms of preference.

1. Random Forest
2. Decision Tree
3. KNN
4. Logistic Regression

For overall performance, Random Forest may be chosen.

The next steps

- The target variable is unbalanced (75% vs. 25%)
- Other method such as Boosting, SVM, and Bagging may be tried.
- Further, oversampling and downsampling method may be employed to address unbalanced dataset.

THANK YOU