Credit Risk Analysis based on Customer Behavior

Dongyan Sun

Data Science Institute

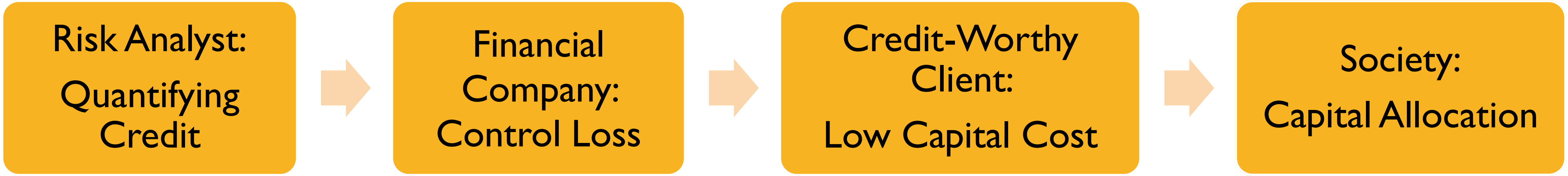
October 20, 2023

https://github.com/EmmaSun19902023/Midterm_Project_EmmaSun.git

INTRODUCTION

Predict Credit Card Repayment

according to **Clients' Background and Payment Information**



Risk Analyst: Quantifying Credit

Financial Company: Control Loss

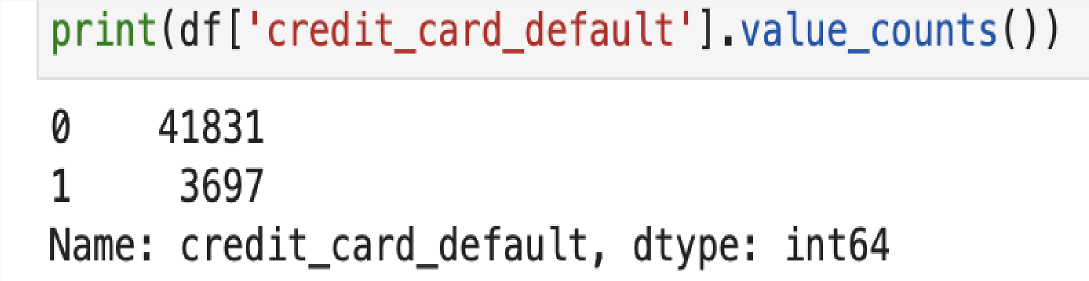
Credit-Worthy Client: Low Capital Cost

Society: Capital Allocation

Data from Kaggle: https://www.kaggle.com/datasets/pradip11/amexpert-codelab-2021/code

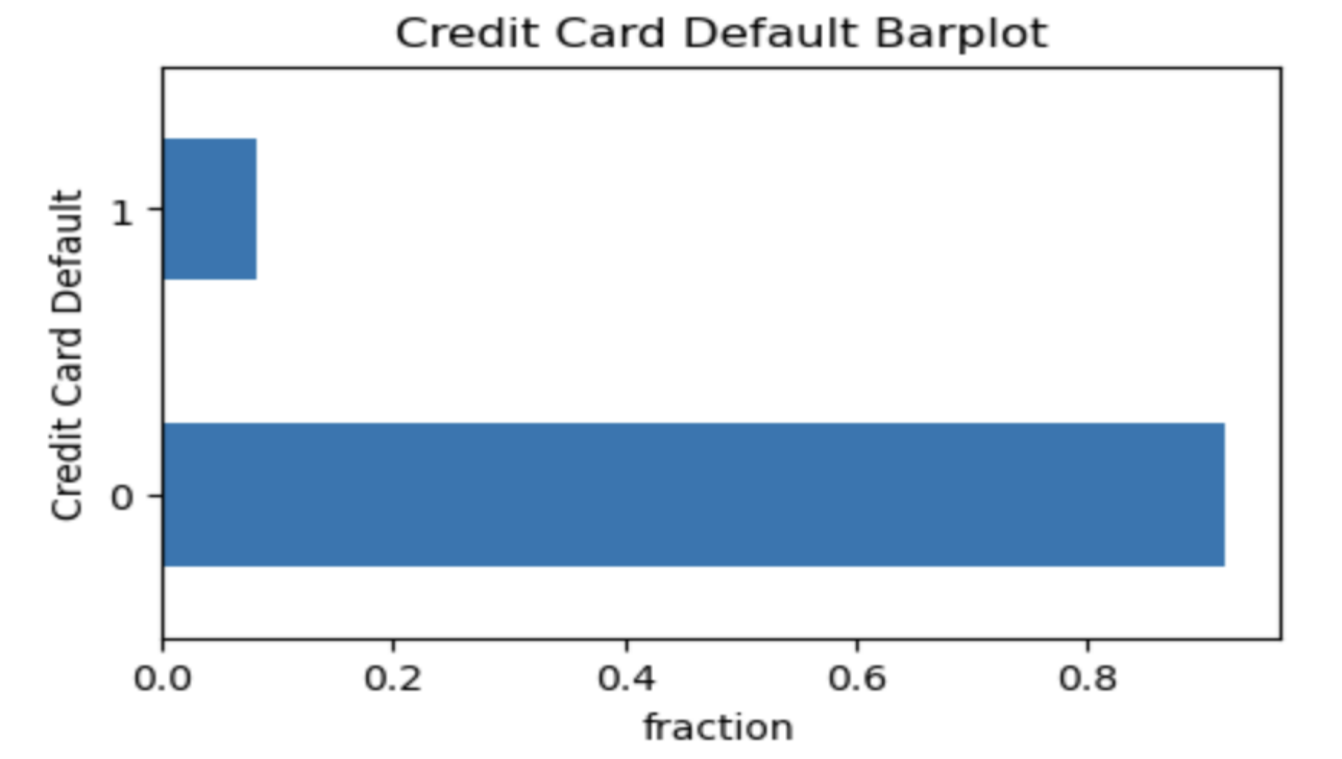
Collected by American Express Company

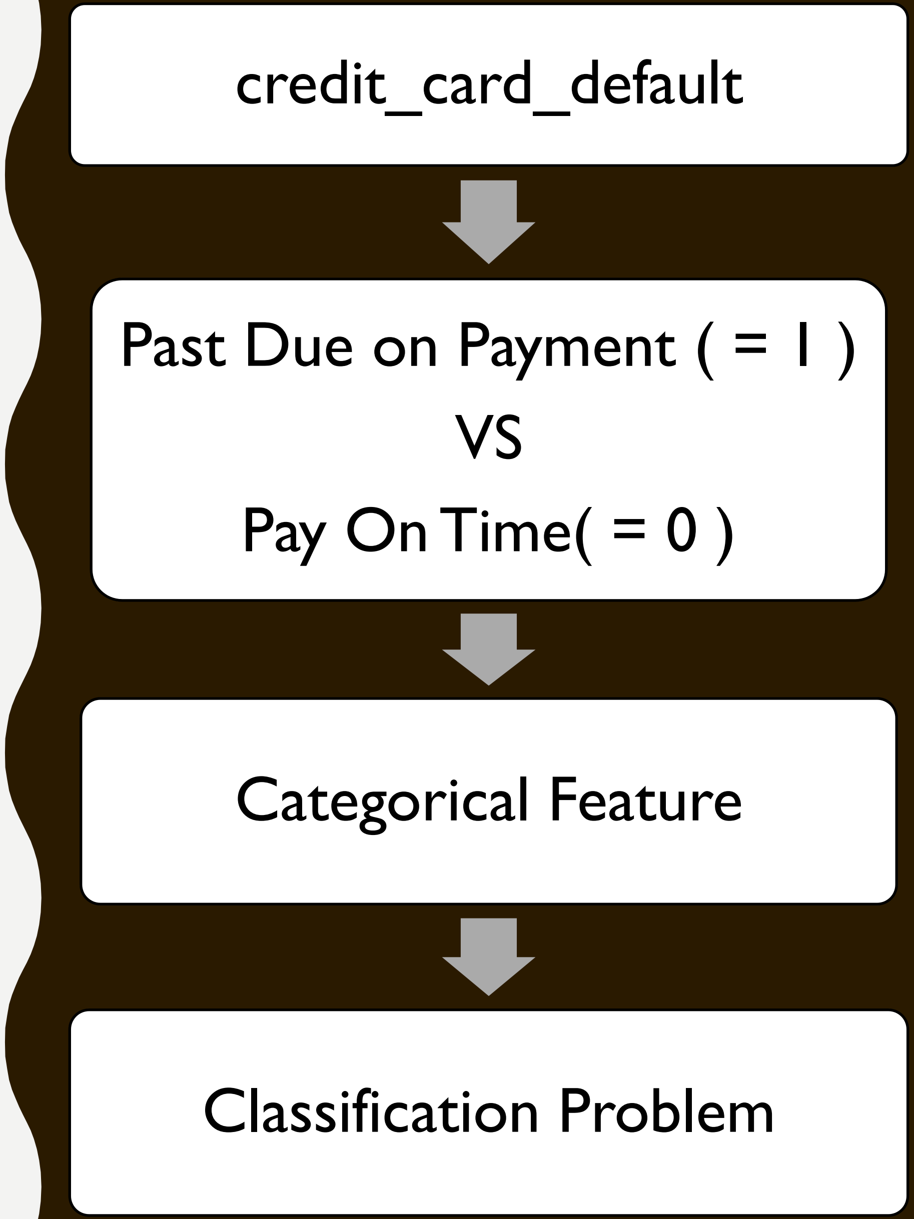
45528 ROWS *19 COLUMNS



```
print(df['credit_card_default'].value_counts())
```

```
0    41831
1     3697
Name: credit_card_default, dtype: int64
```





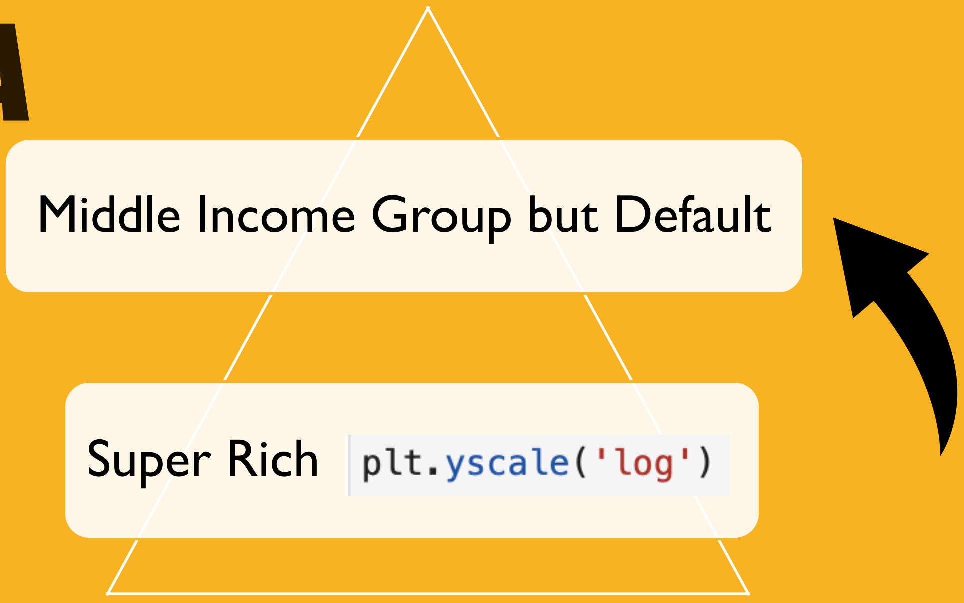
credit_card_default

Past Due on Payment (= 1)
VS
Pay On Time(= 0)

Categorical Feature

Classification Problem

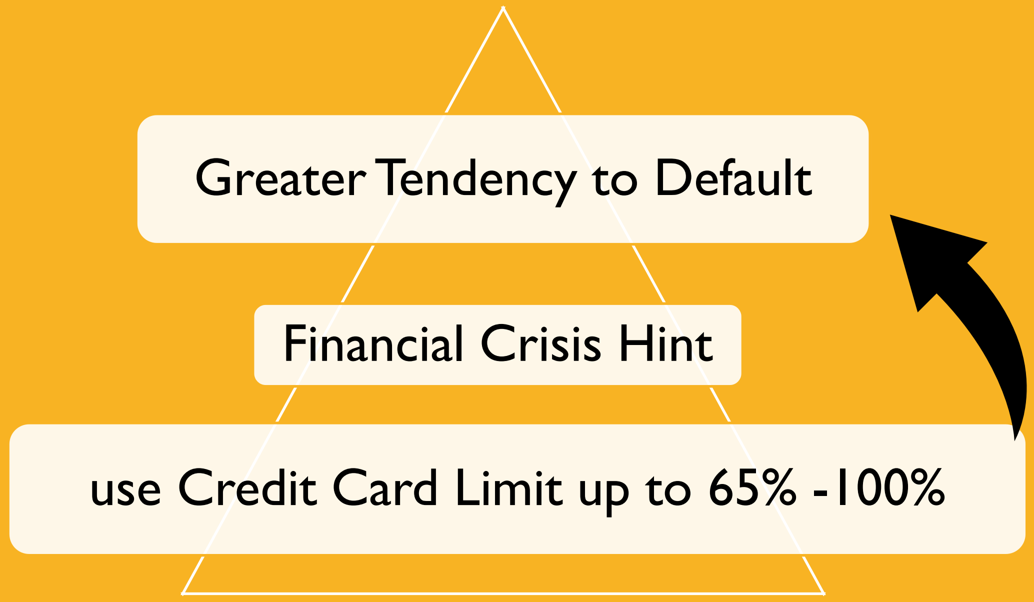
EDA



Middle Income Group but Default

Super Rich

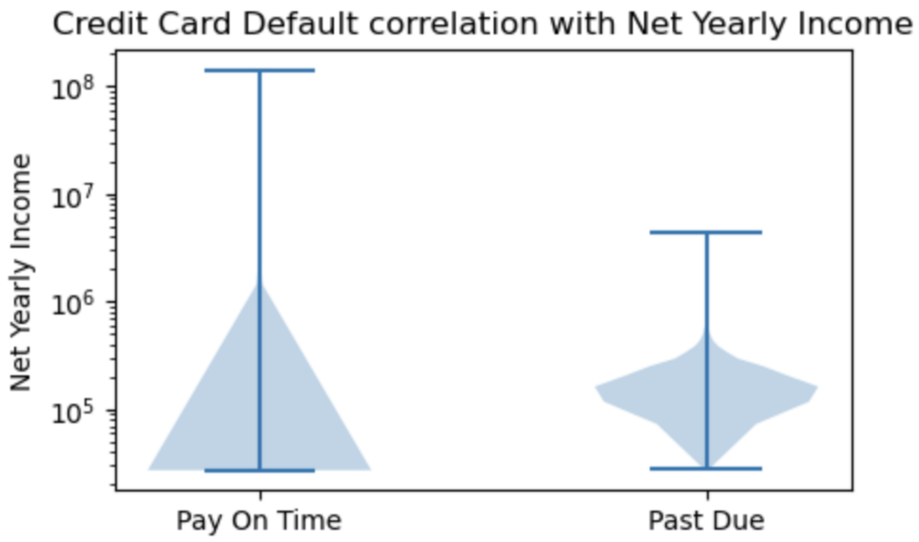
```
plt.yscale('log')
```



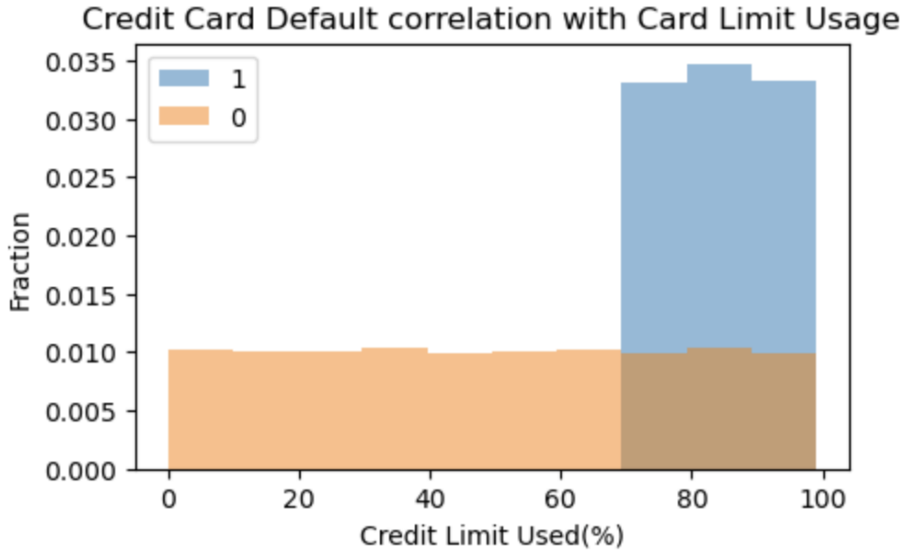
Greater Tendency to Default

Financial Crisis Hint

use Credit Card Limit up to 65% -100%



Credit Card Default correlation with Net Yearly Income



Credit Card Default correlation with Card Limit Usage

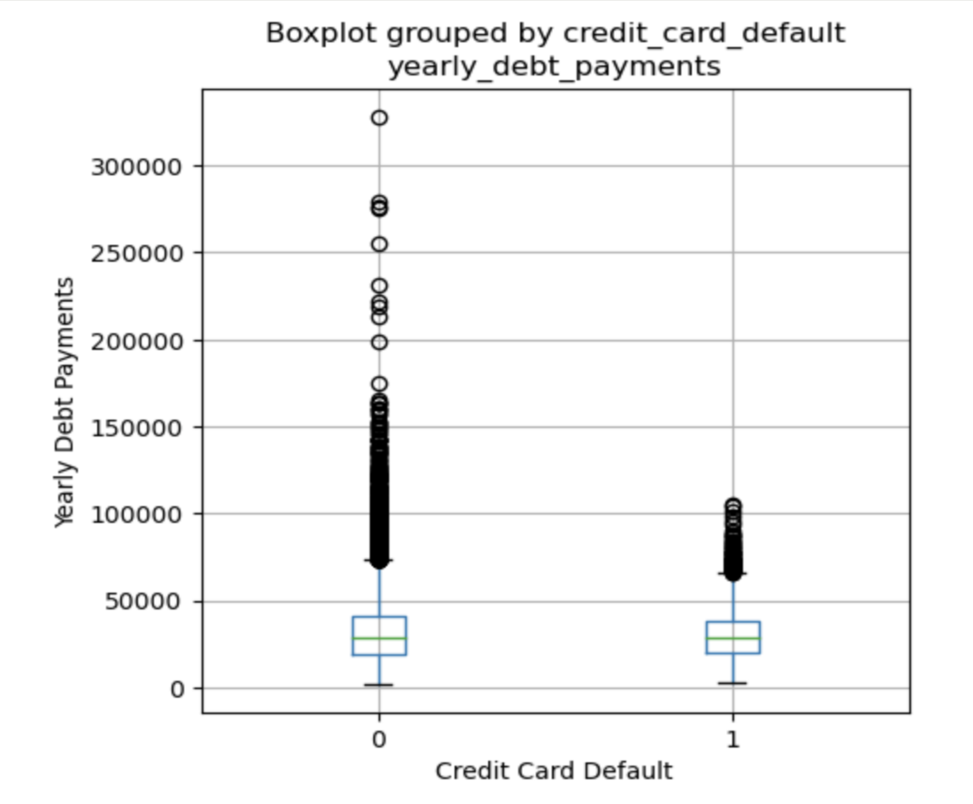
EDA

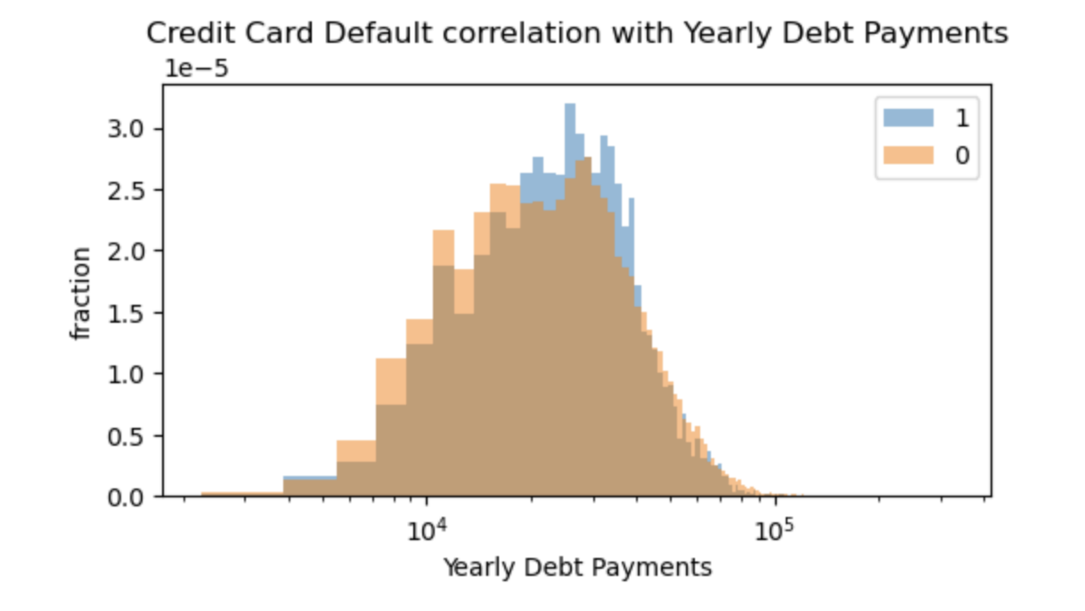
Yearly Debt Payment
(Lower)

Debt Burden
(Lower)

≠

Trustworthy





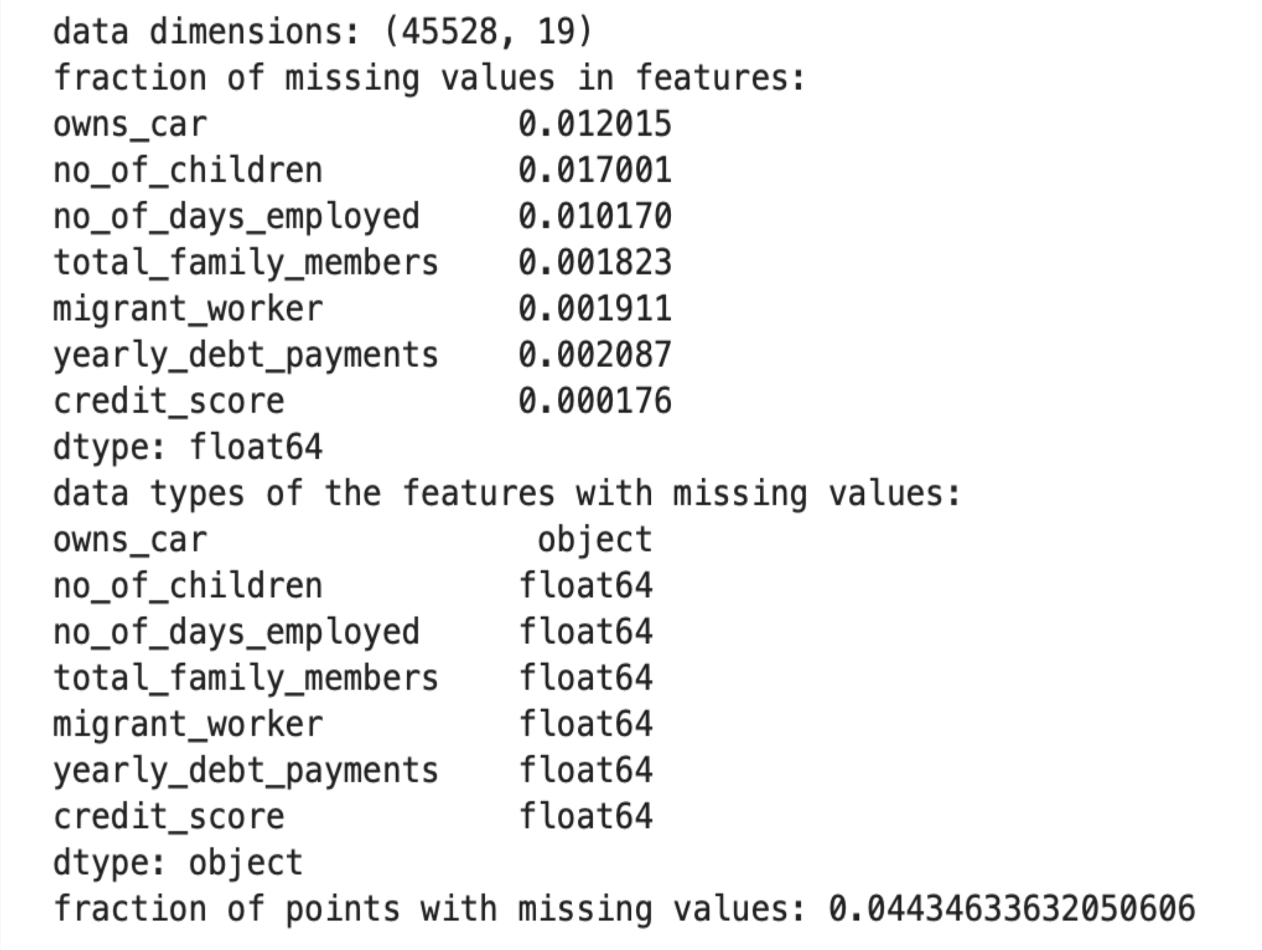
MISSING VALUE

- 7 out of 19 features

- 4.43% of points

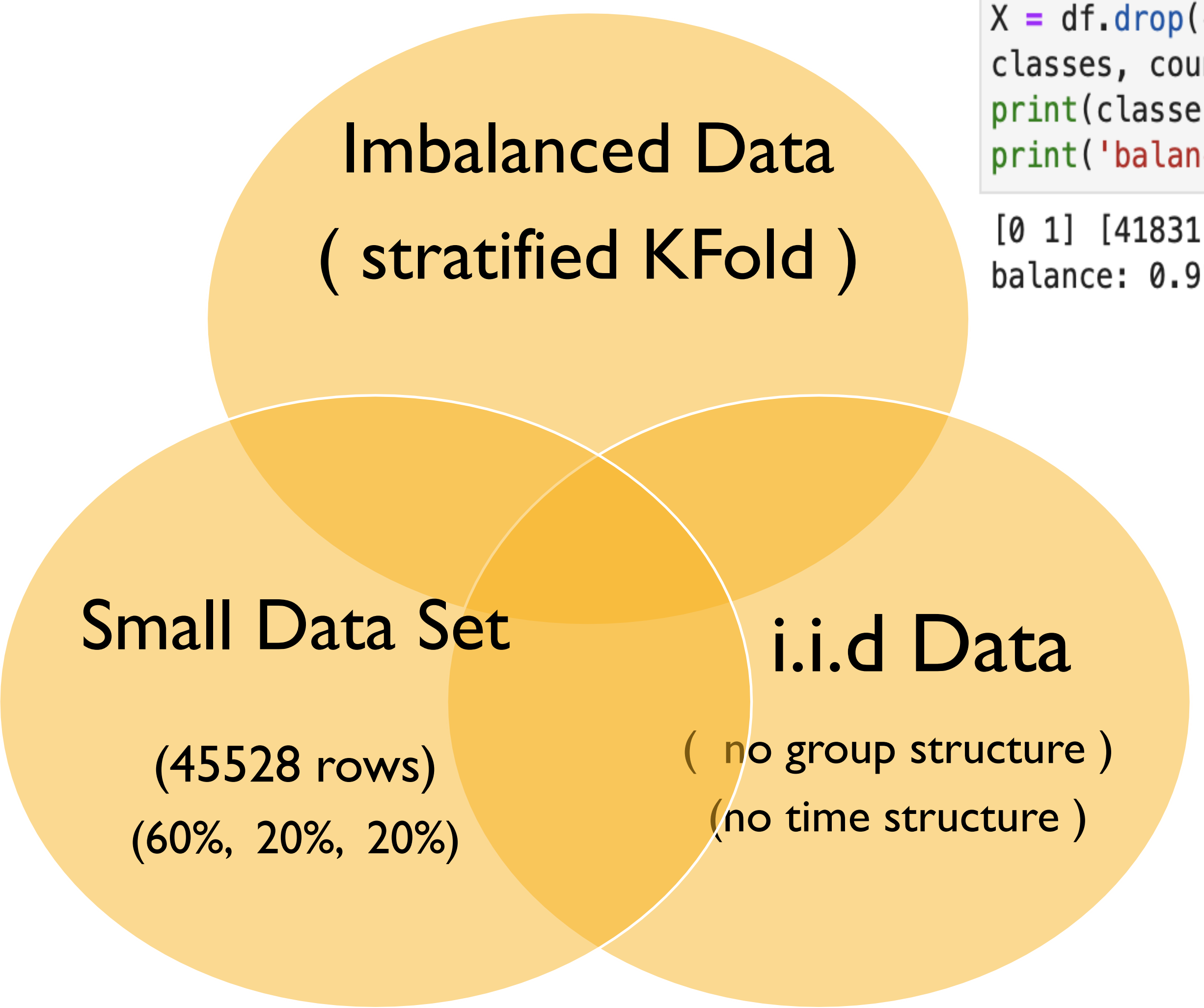
- 4 categorical features
- 3 continuous features
- 0.01% - 1.70%

- OneHotEncoder



```
data dimensions: (45528, 19)
fraction of missing values in features:
owns_car                0.012015
no_of_children          0.017001
no_of_days_employed     0.010170
total_family_members    0.001823
migrant_worker          0.001911
yearly_debt_payments    0.002087
credit_score            0.000176
dtype: float64
data types of the features with missing values:
owns_car                 object
no_of_children          float64
no_of_days_employed     float64
total_family_members    float64
migrant_worker          float64
yearly_debt_payments    float64
credit_score            float64
dtype: object
fraction of points with missing values: 0.04434633632050606
```

SPLIT



Imbalanced Data

(stratified KFold)

Small Data Set

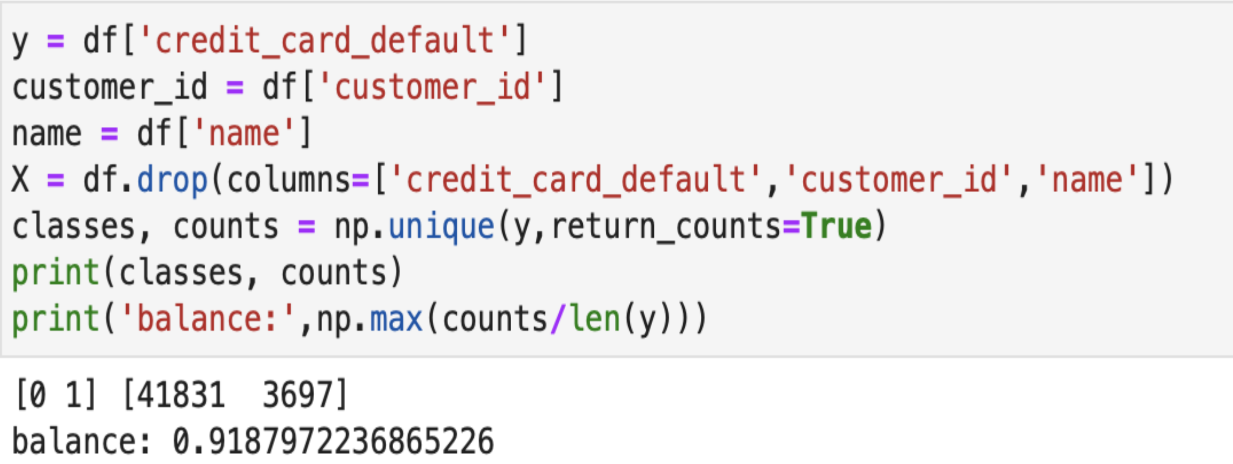
(45528 rows)

(60%, 20%, 20%)

i.i.d Data

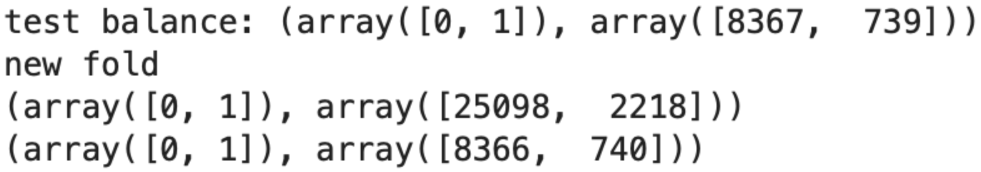
(no group structure)

(no time structure)



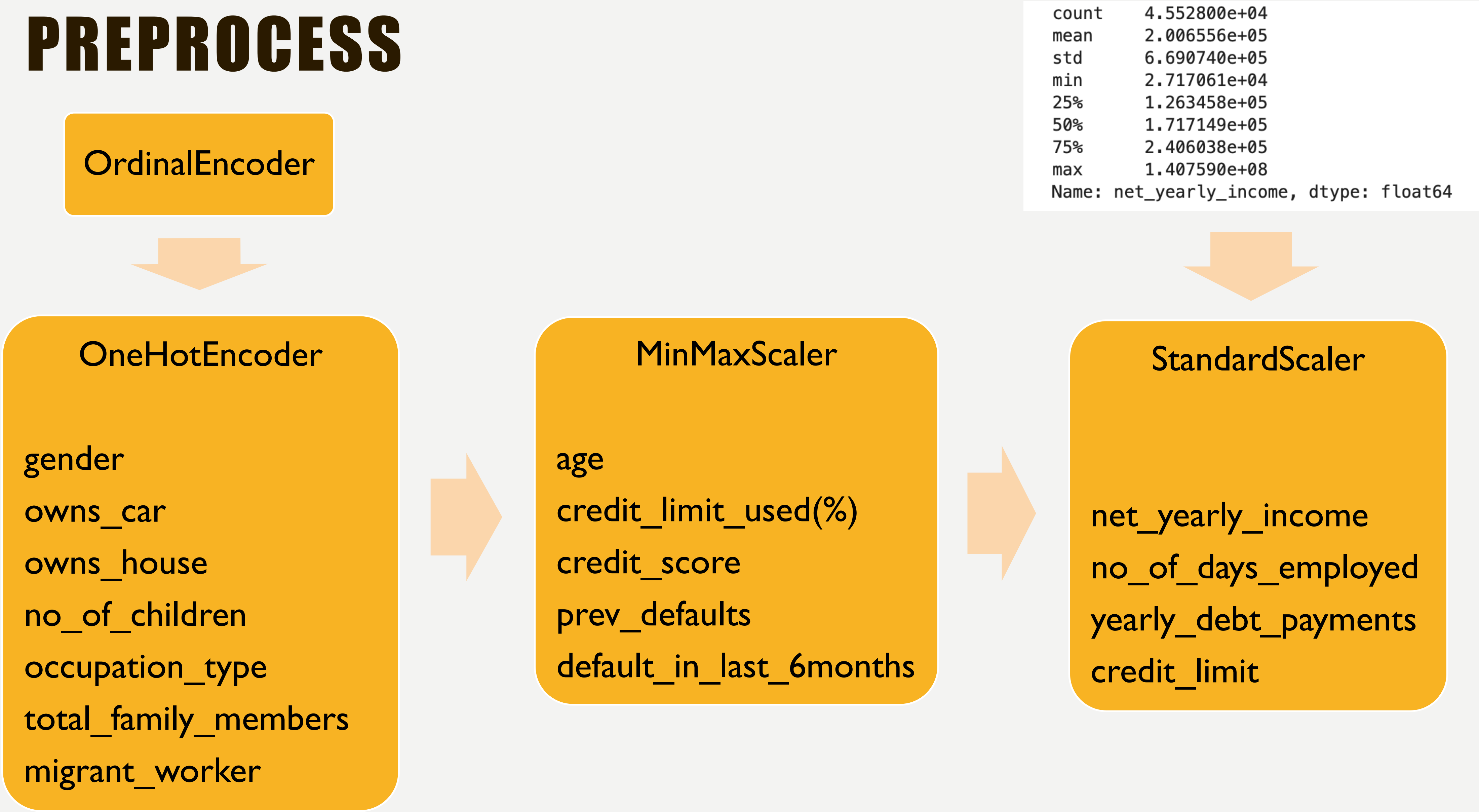
```
y = df['credit_card_default']
customer_id = df['customer_id']
name = df['name']
X = df.drop(columns=['credit_card_default','customer_id','name'])
classes, counts = np.unique(y,return_counts=True)
print(classes, counts)
print('balance:',np.max(counts/len(y)))
```

```
[0 1] [41831  3697]
balance: 0.9187972236865226
```



```
test balance: (array([0, 1]), array([8367,  739]))
new fold
(array([0, 1]), array([25098,  2218]))
(array([0, 1]), array([8366,  740]))
```

PREPROCESS



```
count     4.552800e+04
mean      2.006556e+05
std       6.690740e+05
min       2.717061e+04
25%       1.263458e+05
50%       1.717149e+05
75%       2.406038e+05
max       1.407590e+08
Name: net_yearly_income, dtype: float64
```

OrdinalEncoder

OneHotEncoder

gender
owns_car
owns_house
no_of_children
occupation_type
total_family_members
migrant_worker

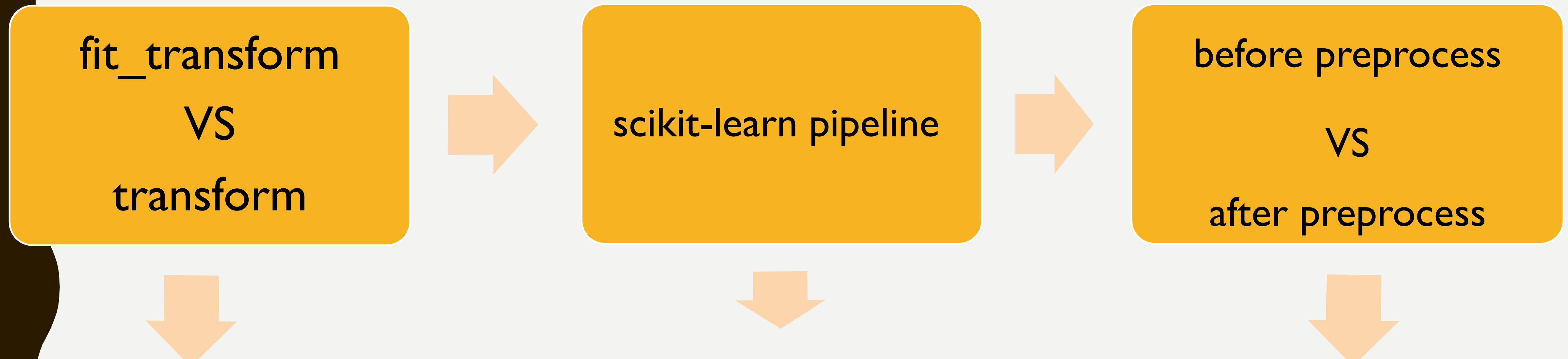
MinMaxScaler

age
credit_limit_used(%)
credit_score
prev_defaults
default_in_last_6months

StandardScaler

net_yearly_income
no_of_days_employed
yearly_debt_payments
credit_limit

PREPROCESS



fit_transform

VS

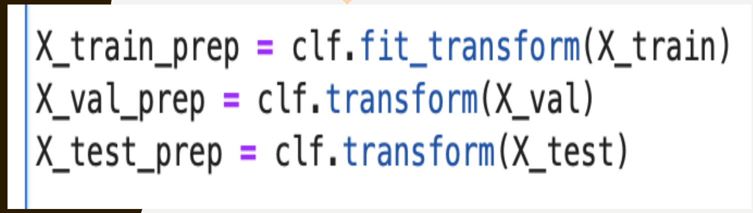
transform

scikit-learn pipeline

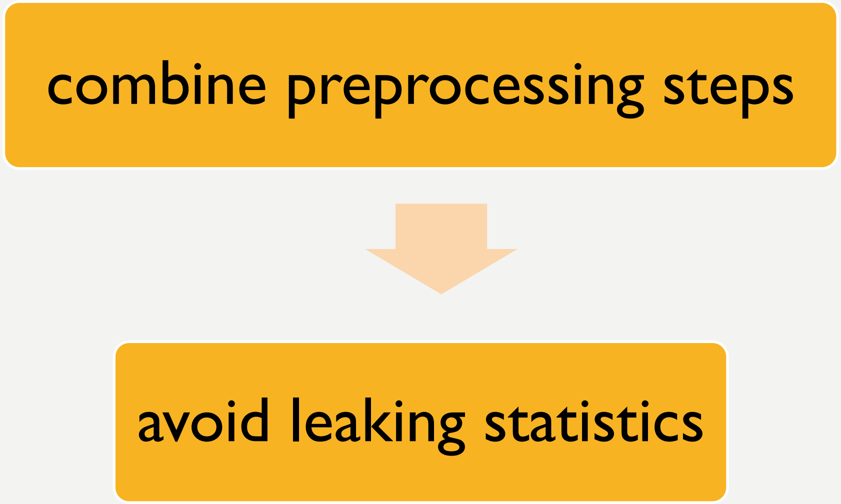
before preprocess

VS

after preprocess

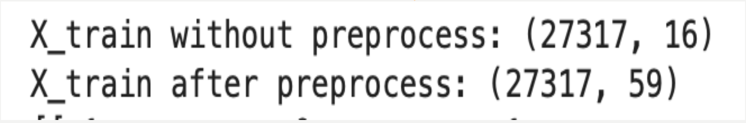


```
X_train_prep = clf.fit_transform(X_train)
X_val_prep = clf.transform(X_val)
X_test_prep = clf.transform(X_test)
```



combine preprocessing steps

avoid leaking statistics



```
X_train without preprocess: (27317, 16)
X_train after preprocess: (27317, 59)
```

Q&A

EMMA SUN DONGYAN_SUN@BROWN.EDU