

Description

Financial institutions suffered significant losses due to default on loans from users. This led to a tightening of loan guarantees and an increase in loan refusal rates. The need for a better credit risk assessment model was also raised by these institutions. This requires a study to estimate the determinants of vehicle loan defaults. The financial institution has hired you to accurately predict the likelihood of a user defaulting on the due date.

This is a synthetic dataset created to be used for academic purposes for beginners who want to practice financial analytics from a simple financial dataset. This dataset contains information on default payments and customer profile.

Task

- Create a model to predict default.
Create loan default prediction model to predict if a user will default or not.

Evaluation

Model are evaluated using F1 between the predicted and actual label.

F1 is calculated as follows:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

True Positive (TP) = your prediction is 1, and the ground truth is also 1

False Positive (FP) = your prediction is 1, and the ground truth is 0

False Negative (FN) = your prediction is 0, and the ground truth is 1

Example calculate F1 using sklearn in python:

```
In [1]: from sklearn.metrics import f1_score  
f1_score(y_true=[0, 1, 0, 1, 1], y_pred=[0, 1, 0, 0, 1], pos_label=1, average='binary')
```

```
Out[1]: 0.8
```

Submission File

For each ID in the test set, you must predict 1 if the user is default, and 0 otherwise. The file should contain a header and have the following format:

```
customer_id, default
```

```
9365, 0
```

```
999, 0
```

```
2835, 1
```

```
..., ...
```

```
3190, 1
```