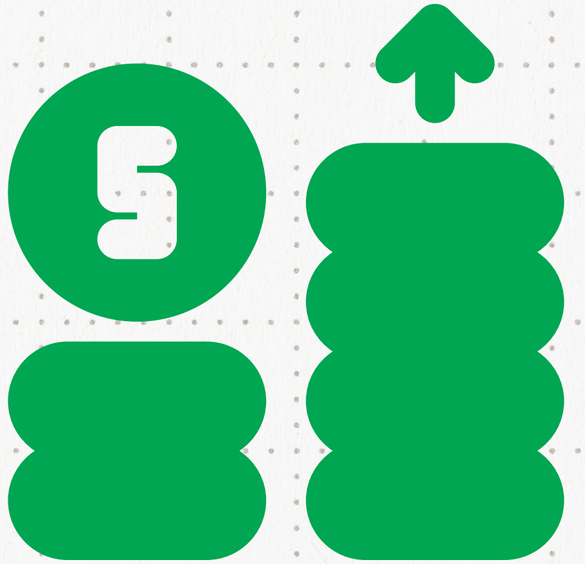
Supervised Learning

USING INDUSTRY DATA FROM BANKING SECTOR

Bhumika Mittal

ANALYSING GIVEN DATA



What is given?

The given dataset contains the various parameters and details about 10k customers, which focuses on the default date, payment due by the customer and several other details - some of which could be related. In contrast, others would be irrelevant for our analysis.

Relevant Observation

Using the given features, we can create a model to predict whether a customer will default. We will use various methods learnt during this course to create a supervised ML model for the same.

Assumptions and Variable Selection

Assumption: We want to predict using the given data if the customer will Default or not.

01

Since the customer ID is unique for all customers, it is assigned by the bank and does not correlate with their chances of defaulting.

02

The date of Default doesn't add any information to the chances of defaulting, as it is after the customer has already defaulted.

How is the model used in predicting Default?

01

Import libraries & Data and clean data

Import all the required libraries. Then load the data from CSV using pandas. Since not all columns are required, we remove unnecessary columns and assign x and y. Some columns contain NA values which are estimated using multivariate feature imputation

02

Standardize the data across different columns

Since the data range varies from dimension to dimension, we must standardise the data before working with it. Some values need to be converted using LabelEncoder and OneHotEncoder.

03

Check p-value and adjusted R-squared value

Observe that according to the p-value, none of the variables is positively correlated to the default chances, which Is an interesting observation. Adjusted R-squared is also coming to be -0.

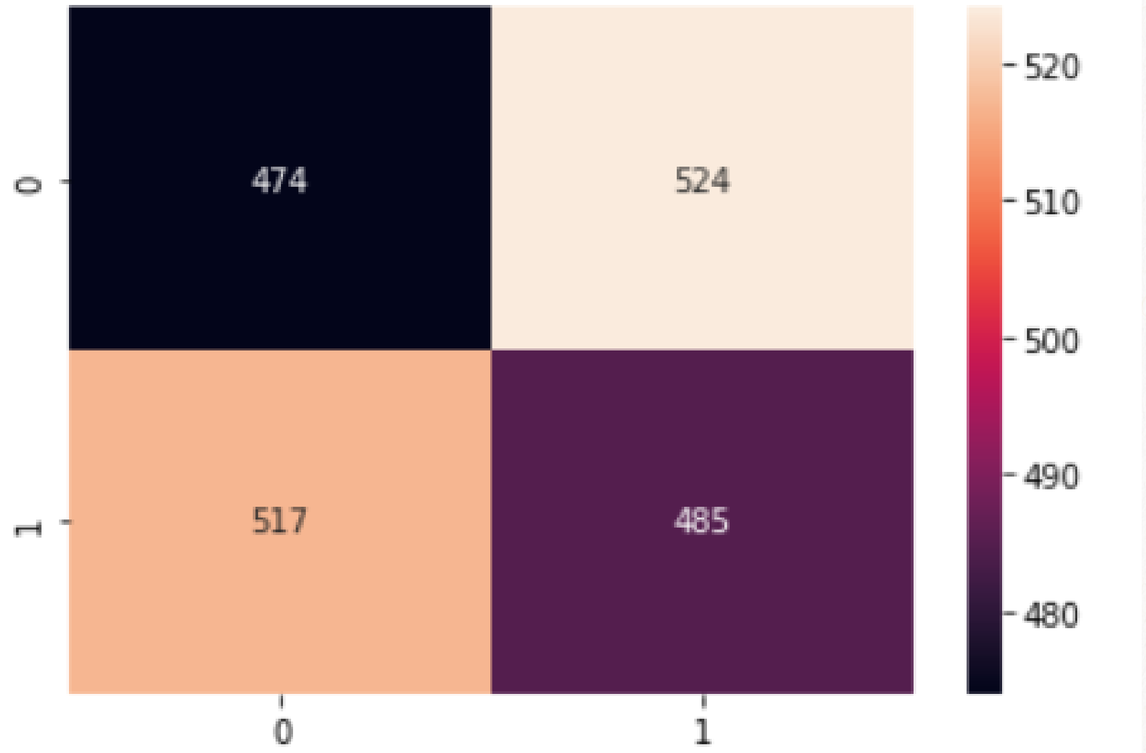
04

Making the model and checking accuracy

Dividing the data into the test and training data, we run the DecisionTreeClassifier model and get the accuracy of ~49%

Result of the model

```
Accuracy: 0.4795
[[474 524]
 [517 485]]
<matplotlib.axes._subplots.AxesSubplot at 0x7f1(
```



Accuracy ~ 49%

The accuracy of the model is ~49%, which means the prediction model is not very good. One of the major reasons for this is the number of variables and the low correlation between them.

Other attempts and why did they didn't work?

Using Dimension Reduction

I tried using dimension reduction on the model to reduce the number of variables so that the classifier tree can be visualised, but due to low correlation, the accuracy drops. Hence, using the dimension reduction on the given industry data is not helpful.

THANK YOU