PREDICT THE CHANCES OF DEFAULTING INSTALMENTS

♦ INTRODUCTION

Once upon a time, there was a bank offering services to private persons. The services include managing of accounts, offering loans, etc. The bank wants to improve their services by finding interesting groups of clients (e.g. to differentiate between good and bad clients). The bank managers have only vague idea, who is good client (whom to offer some additional services) and who is bad client (whom to watch carefully to minimize the bank looses). Fortunately, the bank stores data about their clients, the accounts (transactions within several months), the loans already granted, the credit cards issued So the bank managers hope to find some answers (and questions as well) by analyzing this data. In this kernel, I am going to follow the basic process of loan default prediction with machine learning algorithms.

Loans default will cause huge loss for the banks, so they pay much attention on this issue and apply various methods to detect and predict default behaviours of their customers.

♦ PROJECT OBJECTIVES

The loan is one of the most important products of the banking. All the banks are trying to figure out effective business strategies to persuade customers to apply their loans. However, there are some customers behave negatively after their application are approved. To prevent this situation, banks have to find some methods to predict customers' behaviours. Machine learning algorithms have a pretty good performance on this purpose, which are widely-used by the banking. Here, we will work on loan behaviours prediction using machine learning models.

♦ TASK DESCRIPTION

In this data challenge, We're going to work with dataset from the bank. As a data scientist, we will need to perform the following tasks:

- Use Python to connect to files and read tables into Pandas dataframes
- Preprocess data for machine learning
- Train a ML model to predict customers who are more likely to default on loans
- Evaluate model performance
- Try to understand the key predictors of default

♦ EXPLORATORY ANALYSIS

To begin this exploratory analysis, first import libraries and prepare the data.

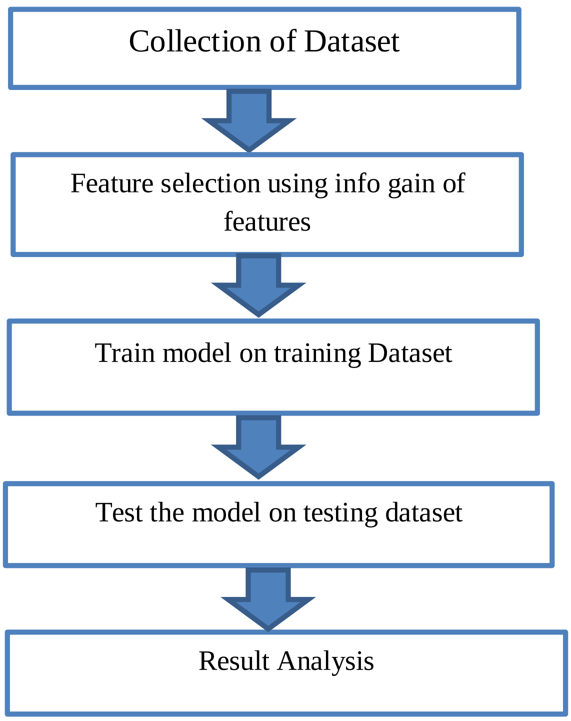
♦ SCOPE

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this project is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight .A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. This project is exclusively for the managing authority of Bank/finance Company. Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities. It is a classification problem. The dataset we will be using undergoes the process of normalisation, missing value treatment, choosing essential columns using filtering, deriving new columns, identifying the target variables and visualising the data in the graphical format. Python is used for easy and efficient processing of data. For this we will use the Pandas library available in Python to process and extract information from the given dataset. The processed data is converted into appropriate graphs for better visualisation of the results and for better understanding. For obtaining the graph Matplot library is used. Our aim from the project is to make use of pandas, Matplotlib, & Seaborn libraries from python to extract insights from the data & Sklearn libraries for machine learning. And in the end, to predict whether the loan applicant can repay the loan or not using techniques of combining the predictions from multiple machine learning algorithms.

The following steps will be applicable to the models:-

- Creating Dummy variables.
- Split the data into training set and data set. The training set be used to fit the model, and test set will be to evaluate the best model to get an estimation of generalization error.

LOAN PREDICTION METHODOLOGY



PREDICT THE CHANCES OF DEFAULTING INSTALMENTS

Data Dictionary:

Number of Instances: 5000 X 20

Number of Attributes: 20 (7 numerical, 13 categorical)

Attribute description

Attribute 1: (Qualitative / Categorical) Status of existing checking account

A11: ... < 0 USD

A12: 0 <= ... < 10000 USD

A13: ... >= 10000 USD

A14: no checking account

Attribute 2: (numerical) Duration in month

Attribute 3: (Qualitative / Categorical) Credit history

A30: no credits taken/all credits paid back duly

A31: all credits at this bank paid back duly

A32: existing credits paid back duly till now

A33: delay in paying off in the past

A34:critical account/other credits existing(not at this bank)

Attribute 4: (Qualitative / Categorical) Purpose

A40: car (new)

A41: car (used)

A42: furniture/equipment

A43: radio/television

A44: domestic appliances

A45: repairs

A46: education

A47: (vacation - does not exist?)

A48: retraining

A49: business

A410: others

Attribute 5: (numerical) Credit amount

Attribute 6: (Qualitative / Categorical) Savings account/bonds

A61: ... < 1000 USD

A62: 1000 <= ... < 5000 USD

A63: 5000 <= ... < 10000 USD

A64: >= 10000 USD

A65: unknown/ no savings account

Attribute 7: (Qualitative / Categorical) Present employment since

A71: unemployed

A72: ... < 1 year

A73: 1 <= ... < 4 years

A74: 4 <= ... < 7 years

A75: .. >= 7 years

Attribute 8: (numerical) Instalment rate in percentage of disposable income

Attribute 9: (Qualitative / Categorical) Personal status and sex

A91: male : divorced/separated

A92: female: divorced/separated/married

A93: male : single

A94: male : married/widowed

A95: female: single

Attribute 10: (Qualitative / Categorical) Other debtors / guarantors

A101: none

A102: co-applicant

A103: guarantor

Attribute 11: (numerical) Present residence since

Attribute 12: (Qualitative / Categorical) Property

A121: real estate

A122: if not A121: building society savings agreement/ life insurance

A123: if not A121/A122: car or other

A124: unknown / no property

Attribute 13: (numerical) Age in years

Attribute 14: (Qualitative / Categorical) Other instalment plans

A141: bank

A142: stores

A143: none

Attribute 15: (Qualitative / Categorical) Housing

A151: rent

A152: own

A153: for free

Attribute 16: (numerical) Number of existing credits at this bank

Attribute 17: (Qualitative / Categorical) Job

A171: unemployed/ unskilled - non-resident

A172: unskilled - resident

A173: skilled employee / official

A174: management/ self-employed/ highly qualified employee/ officer

Attribute 18: (numerical) Number of people being liable to provide maintenance for

Attribute 19: (Qualitative / Categorical) Telephone

A191: none

A192: yes, registered under the customer’s name

Attribute 20: (Qualitative / Categorical) foreign worker

A201: yes

A202: no

Default on Payment due

1 (Defaulted) 0 (No Default)