User loan prediction project proposal

by

Raghad Akram

Data Science bootcamp

SDAIA Academy

[October 2021]

1. Question/need

Getting a loan may take quite a time and take effort of bank employees to make the right decision, in our project we want to solve this problem by making a loan prediction system that will predict if user is worth the loan or not.
A banking system that predicts if a user is entitled to receive a disk, based on a set user data stored in the dataset.

2. Data description

In our system, we rely on a database of 14 attributes and more than 5000 rows, below is a table showing the attributes description:

Attribute	Type	Description
ID	Int	User’s id
Age	Int	User’s age
Experience	int	Life experience
Income	Int	User’s income
ZIP Code	Int	Telephone number
Family	Int	Number of family members
CCAvg	Float	The arithmetic average of the credit card
Education	Int	User education level
Mortgage	Int	Kind of foreign payment
PersonalLoan	Boolean	Does the user have a personal loan?
SecuritiesAccount	Boolean	User’s securities account
CD Account	Boolean	Certificate of deposit
Online	Boolean	Is the user account active?
CreditCard	Boolean	Does the user have a credit card?

3. Tools

We have adopted the following tools:

- Python: High-level Software language.
- Jupyter notebook: Open-source web application.
- Sklearn: Software library.
- Pandas: Software library.

We will use machine learning and linear regression.

4. MVP goal

Data preprocessing is a term describing any type of primary processing applied to raw data. Over time, these techniques have evolved to include their users for data preparation to train machine learning models, artificial intelligence, and various data analytics. It can be used with a variety of data source (data stored in databases for example) [1].

As for the data of our system, we will apply it to a preliminary treatment:

- Processing the incomplete data: that is, which contains in some of its cells an empty value. The processing is done by [2]:
    1. Ignoring the entire record.
    2. Filling in the values manually.
    3. Using global constant, missing values are replaced with words like unknown.
    4. Use the mean or median in numerical fields.
- Repeated data: repeated values or records can be detected by ready-made programs such as RapidMiner that contain ready-made tools and cods [2].

Then we explore the dataset by drawing diagrams and displaying the results, through the following link foe the dataset:

https://www.kaggle.com/krantiswalke/bank-personal-loan-modelling

References

[1] Data preprocessing, one of sites “Majara”, https://technologyreview.ae/technodad/%D8%A7%D9%84%D9%85%D8%B9%D8%A7%D9%84%D8%AC%D8%A9-%D8%A7%D9%84%D9%85%D8%B3%D8%A8%D9%82%D8%A9-%D9%84%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA/ , [Accessed3-10-2021].

[2] Rattibha, https://rattibha.com/thread/1187282615289679878?lang=ar , [Accessed4-10-2021].