Architecture

Abstract

The facilities provided by banks are numerous and help us in different areas of financial needs. It involves granting of loans, mortgage, credit card etc. Credit card is one of the most common utility everyone has today and has been booming for many years. There are also many risks involved with the same. As the number of consumers have been increasing exponentially, the danger of credit card defaulting also has been on the rise. Understanding the characteristics of consumers and their banking details can shed some light on the possibility of them defaulting. This project provides a method to predict credit card defaulters using machine learning models like Logistic regression, KNN, Decision tree, SVM, AdaBoost and Naïve Bayes.

1. Introduction

The purpose of this document is to present a detailed description of the credit card default prediction system. It covers the purposes and features of the system and how the system performs the actions. The objective of the project is to predict if the person is a credit defaulter or not based on the person's characteristics and payment history Predicting if a person will default in their credit card payments depends upon a variety of factors like the balance in their account, age, marital status, gender, education, their repayment status for each month for 6 months, their total bill amount and the amount they have paid back.

1.1 Scope

The credit card default prediction system is a machine learning based model that will help us to detect credit card defaulters. Early detection of defaulters will be key to proper and safe bank management. It is able to make predictions given a number of independent variables like age, education, marital status, repayment status etc.

2. Technical specifications

2.1 Dataset

The dataset used for the study consists of 3000 rows of data and 25 columns which relate to data related to each customer in the bank

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	ID	LIMIT_BAL	SEX	EDUCATIO	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	BILL_AMT	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.payment.next.month		
2	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0	0	0	0	689	0	0	0	0	1		
3	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1		
4	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0		
5	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0		
6	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0		
7	6	50000	1	1	2	37	0	0	0	0	0	0	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800	0		
8	7	5.00E+05	1	1	2	29	0	0	0	0	0	0	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0		
9	8	1.00E+05	2	2	2	23	0	-1	-1	0	0	-1	11876	380	601	221	-159	567	380	601	0	581	1687	1542	0		
10	9	140000	2	3	1	28	0	0	2	0	0	0	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000	0		

2.2 Prediction

The system prompts the user to input the data related to the customer. After the information has been loaded the system will output if the person is a defaulter or not based on the input provided.

2.3 Logging

The system logs all the steps so that the user will be aware of the processes working internally. The type of logging chosen for each depends upon the type of operation that is carried out like info, critical, debug or warning logging. Logging helps to debug issues, so it is a mandatory step.

2.4 Deployment

The machine learning model will be deployed as an API using FastAPI.

3. Proposed Solution

The proposed solution is to build various machine learning models like Logistic regression, KNN, Decision tree, SVM, AdaBoost and Naïve Bayes and find which model is able to best predict the credit card defaulter given a set of training data.

4. Model Training workflow

The different models used in the project will be used to detect if a consumer is a credit card defaulter or not based on the data used in the training stage. The goal is to identify the model which is able to make predictions with the highest accuracy so that defaulters can be identified easily.