Personal loan prediction

A machine learning problem using logistic regression, decision tree, support vector machine and K nearest neighbor

¶

Mijail Dragowski

2023

Introduccion and summary

The objective of this project is to develop a machine learning model to predict whether a passive bank customer will buy a personal loan or not. For that purpose, we will carry out 4 different machine learning classification algorithms (**logistic regression, decision tree, support vector machine and K nearest neighbor**) in order to find which one predicts better. We will find that the machine learning algorithm with the best performance for this problem is decision tree with a 93% of F1_score, followed by suport vector machine with a 91% F1_score, then K nearest neighbor with 80% F1_score, and finally logistic regression with 69% F1_score.

In **section 1** (Data) we will explain the data used in this project, in **section 2** (methodology) we will give a brief theoretical explanation of the models, evaluation metrics and tools used in this project, in **section 3** (results) we will show the main results of each model using a confusion matrix and we will compare the different evaluation metrics, finally in **section 4** (Conclusions) the conclusions of the work and possible recommendations are presented.

This project was programmed and developed in the python language, but for simplicity purposes, this pdf was written and its purpose is to give a summary illustration of what the project is. Those people who are interested in the technical details of programming and coding of data processing, data analysis and machine learning, can see it at the following link: https://github.com/Mijail-Dragowski/Personal-loan-prediction/blob/main/Personal%20loan%20prediction.%20CODE.ipynb

1. Data

The data set used for this project comes from Thera Bank, a U.S. bank. The datavset has the data of 5000 bank clients, where the characteristics of each client are detailed, which will be used as independent and dependent variables to train the models and make predictions.Those variables are:

Dependent variable:

Personal_Loan: Did this customer accept the personal loan offered in the last campaign?

Independet variables:

Age: Customer’s age in completed years

Experience: years of professional experience

Income: Annual income of the customer (in thousand dollars)

ZIP Code: Home Address ZIP code.

Family: the Family size of the customer

CCAvg: Average spending on credit cards per month (in thousand dollars)

Education: Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional

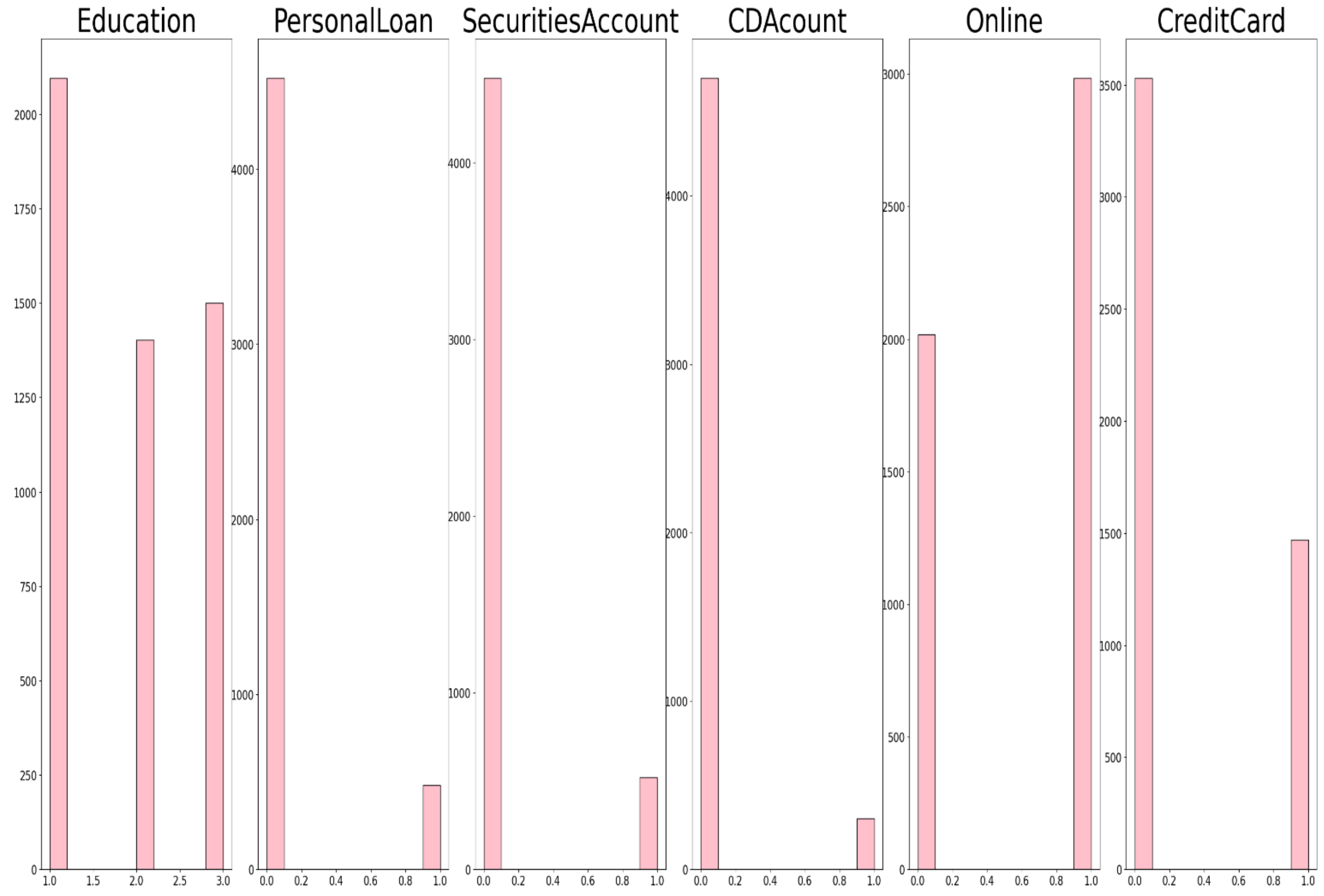
Securities_Account: Does the customer have securities account with the bank?

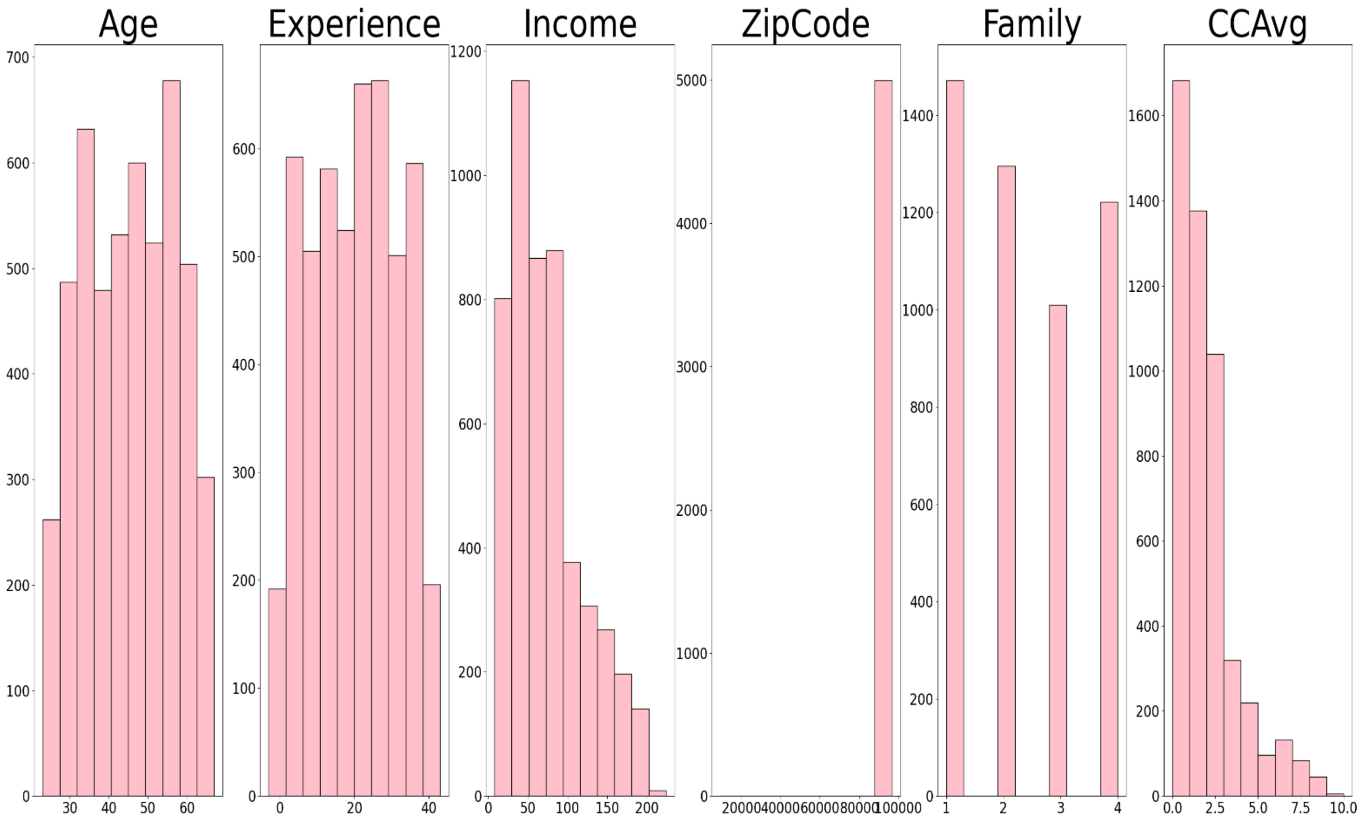
CD_Account: Does the customer have a certificate of deposit (CD) account with the bank?

Online: Do customers use internet banking facilities?

CreditCard: Does the customer use a credit card issued by any other Bank

The frequency distribution for each variable





2. Metodology

In this project, 4 classification machine learning models are used, in order to compare them and stay with the one that best predicts. These models are:

Logistic regression: In statistics, the logistic model (or logit model) is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1").

Decision tree: A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Suport vector machine: In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM maps training examples to points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

K nearest neighbor: In statistics, the k-nearest neighbors (k-NN) algorithm is a supervised learning method, where the classification estimation methodology of each observation depends on the characteristics of the other observations that surround it.

The metrics used in this project to evaluate each model are:

Accuracy: It is the ratio of the correct predictions to the total number of predictions. Or more simply, how often is the classifier correct?

Recall: It is the ratio of correct positive predictions to the total number of positive predictions. Or more simply, how sensitive is the classifier to detect positive instances.

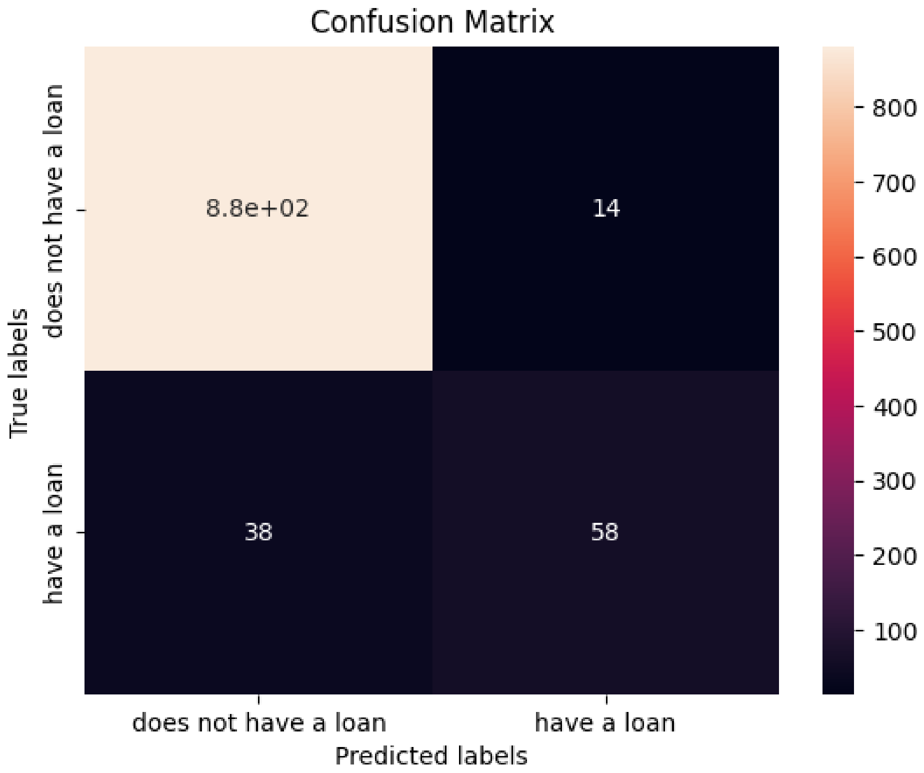
Precision: The ratio of correct predictions to the total number of predicted correct predictions. This measures the accuracy of the classifier in predicting positive cases.

F1_score: is a weighted average of recall and precision, with a higher score for best model. The f1_score will be our main reference metric, since it makes a joint representation of what is the precision and the recall.

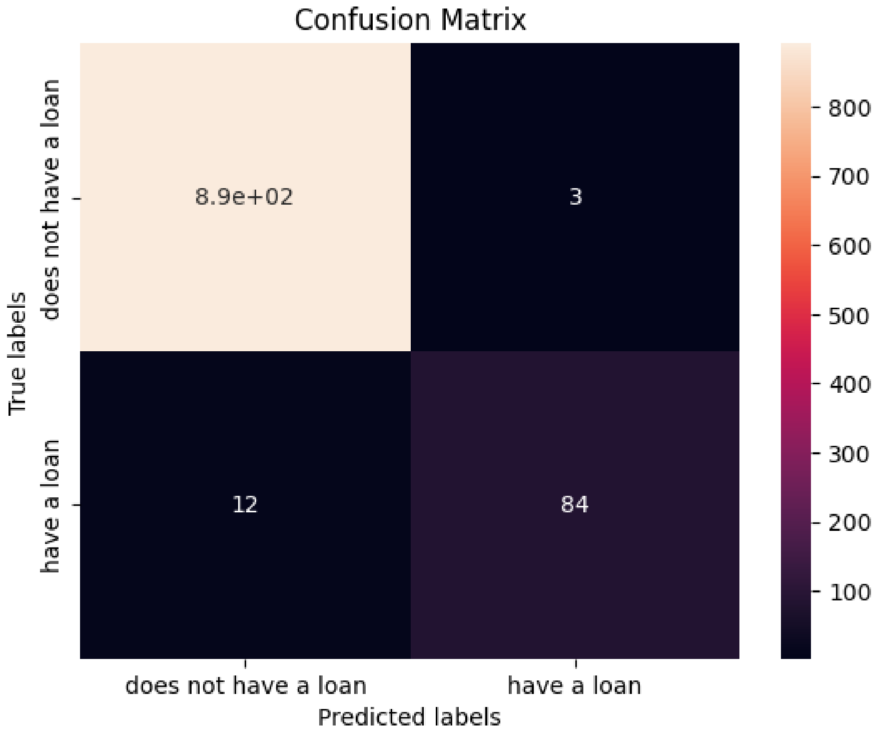
3. Results

To represent the results of each model in a summarized and easy to understand way, we will use the confusion matrix.**Confusion matrix**: Is a table that describes the performance of a classification model on a set of test data whose true values are known. A confusion matrix is highly interpretive and can be used to estimate all evaluation metrics for classification models:

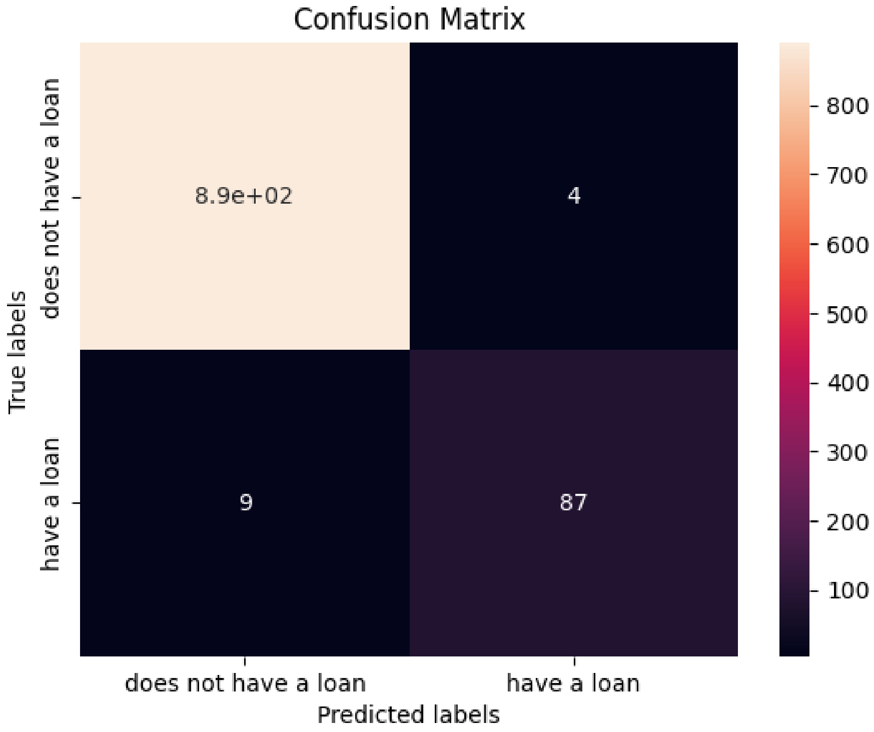
Logistic regression:



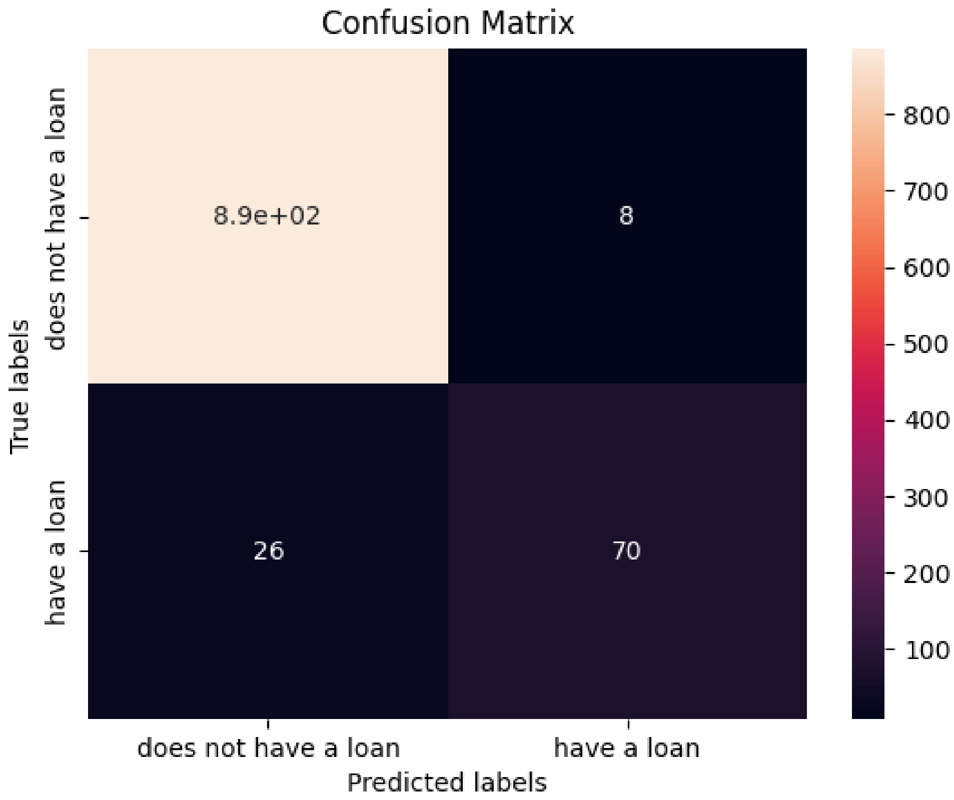
Suport vector machine



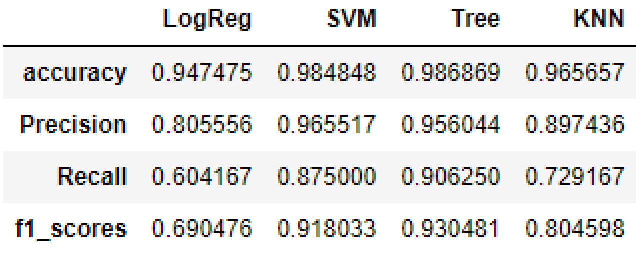
Decision Tree



K nearest neighbor



Comparison of each model according to its metrics



	LogReg	SVM	Tree	KNN
accuracy	0.947475	0.984848	0.986869	0.965657
Precision	0.805556	0.965517	0.956044	0.897436
Recall	0.604167	0.875000	0.906250	0.729167
f1_scores	0.690476	0.918033	0.930481	0.804598

4. Conclusion

After analyzing the results of the different models, we are now in a position to answer the initial question: ¿Which model is better to predict whether a client will take a loan or not? For this case, as we can see in the comparative evaluation metrics table from the results section, the model with the best predictive performance was the decision tree.
Once the predictive model has been carried out, as data scientists we are already in a position to advise the bank regarding what type of clients are more likely to request a loan