أكاديمية سدايا
SDAIA Academy

User loan prediction final report

by

Raghad Akram

Data Science bootcamp

SDAIA Academy

- ## Abstract

The bank is a financial institution that offers a wide range of various services (money transfer, deposit, loans, etc..), but it suffers from some problems, one of which is the lending process, where one of the difficulties faced by the bank is forecasting in the lending process. Many machines learning-based approaches have been proposed to handle the problem of the loan prediction process. In this project we provide an empirical comparison of the most used machine learning approaches in the loan prediction area. Our experience includes using 3 models (Knn, Naive Bayes and Kmeans). We applied the aforementioned models on a standard dataset that includes personal information about the customer and assists the bank's employs in the decision-making process regarding the granting of loans, as this data is processed through preprocessing data operations. The results show that the that Naive Bayes model provides the best accuracy compared to both Knn and Kmeans models.

- ## Design

Through Preprocessing Data (https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825) and using Machine Learning techniques (https://searchenterpriseai.techtarget.com/definition/machine-learning-ML), this project has been found, which has a set of functional and non-functional advantages for banks located all over the country. Accurate categorization of cases enables banks to easily decide on granting to customers.

- ## Data

The data contains more than 5000 rows and 14 attributes specific to the customer's personal data, this data is processed by Preprocessing operations https://www.kaggle.com/krantiswalke/bank-personal-loan-modelling

- ## Algorithms

The algorithm we followed has two stages:

1. Processing the entered data, the input of this algorithm is our Dataset (70% training and 30% testing). This data processed by Data Preprocessing, which it works to convert raw data into ready-made data, which is the most important process that ensures the coordination of large data sets in a way through which the existing data can be interpreted, one of the most important preprocessing operations is the data cleaning process (filling in missing values, deleting duplicate lines, etc..).
2. After processing, the data is entered into one of the models that we will know in the next paragraph, we are watching the output.

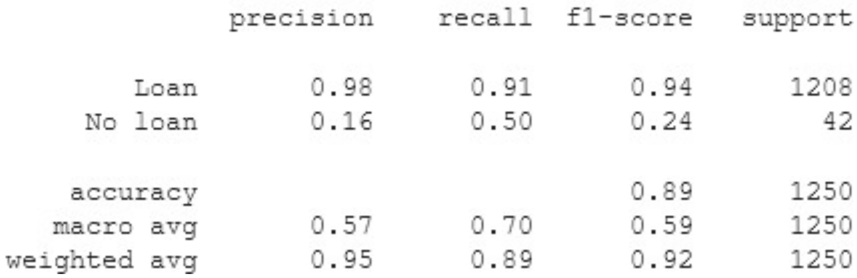
- ## Models

1. Knn: K-Nearest Neighbor, we applied this algorithm to solve classification problems, which is carried out in three basic stages (obtaining data and preparing it, discovering the value of K, and then making predictions.
2. K-Means: we applied this algorithm to solve cluster problems-classify into clusters, where each data point forms a cluster with its centroids close to it (K clusters).

3. Naive Bayes: the method of classification depends on the theory of Bayes and the independent assumption of the distinctive circumstances.

- **Model Evaluation and Selection**

We applied the Knn model results:

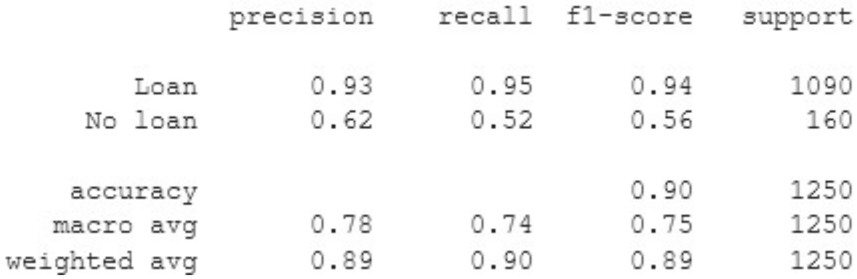


```
              precision    recall  f1-score   support

        Loan       0.98      0.91      0.94      1208
     No loan       0.16      0.50      0.24        42

    accuracy                           0.89      1250
   macro avg       0.57      0.70      0.59      1250
weighted avg       0.95      0.89      0.92      1250
```

Naive Bayes model results:

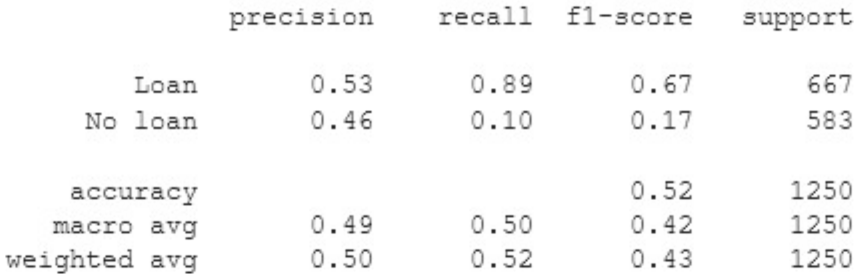


```
              precision    recall  f1-score   support

        Loan       0.93      0.95      0.94      1090
     No loan       0.62      0.52      0.56       160

    accuracy                           0.90      1250
   macro avg       0.78      0.74      0.75      1250
weighted avg       0.89      0.90      0.89      1250
```

Kmeans model results:



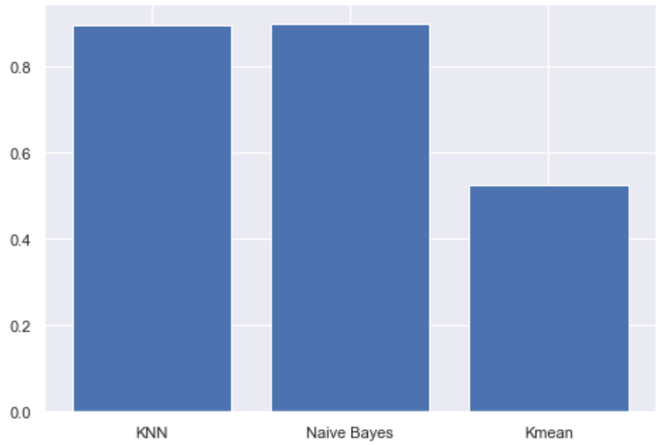
```
              precision    recall  f1-score   support

        Loan       0.53      0.89      0.67       667
     No loan       0.46      0.10      0.17       583

    accuracy                           0.52      1250
   macro avg       0.49      0.50      0.42      1250
weighted avg       0.50      0.52      0.43      1250
```

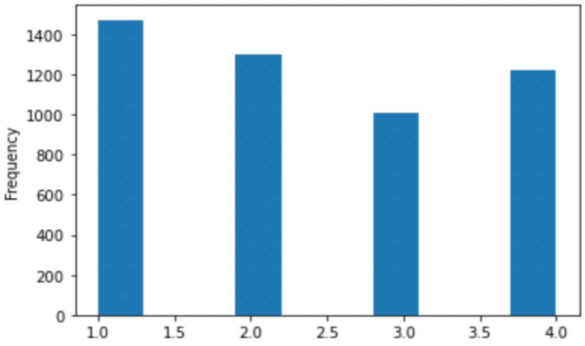
- Communication

Through the results represented in the following diagram, we find that the Naive Bayes is the best algorithm.

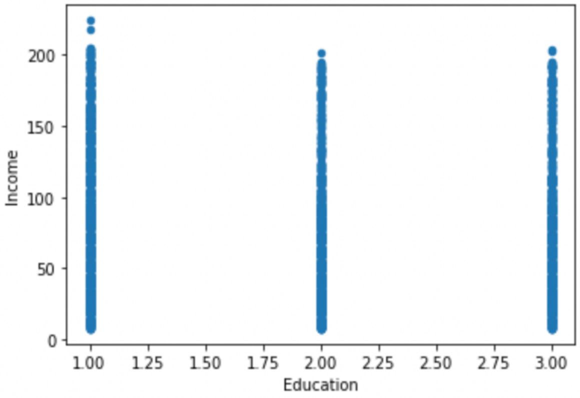


The finding from EDA:

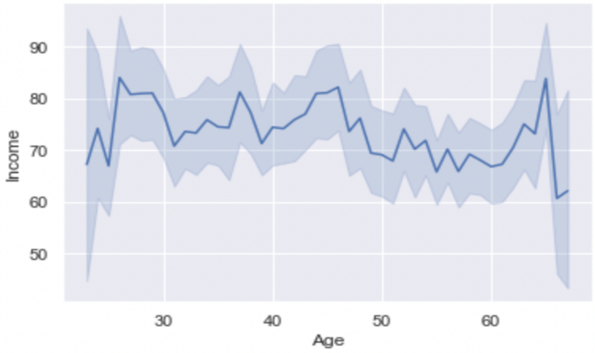
Fix and remove any noisy rows and Preparing the data to train a model
Through the results represented in the following diagrams, we find this results:



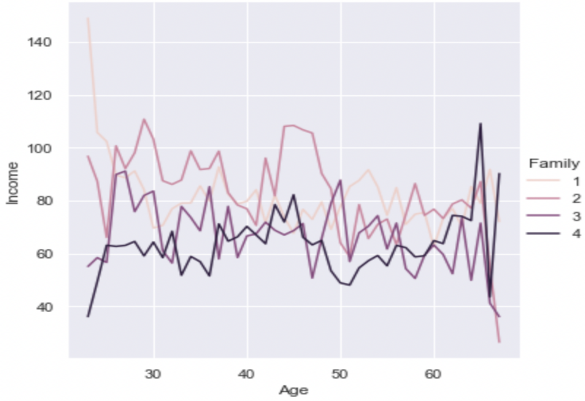
- We can see that the size of the family of customers consisting of 1 and 2



- Here the customers whose education level is 1 is having more income than the others



- We note that the ages of customers between 30 and 50 have higher incomes
- We also note that between the ages of 50 and 60, and before the age of 25, they have less income



- We note that the smallest family size and the youngest age have the higher income more than 100K

- **Tools**

1. Python: High-level Software language.
2. Jupyter notebook: Open-source web application.
3. Sklearn: Software library for modeling.
4. Pandas: Software library for data manipulation.
5. Matplotlib and Seaborn for plotting.

- **Conclusion**

In this project, we solved the loan prediction problem using Machine Learning techniques. We chose the three most used models and conducted the experiment on the existing dataset, in order to test which of the three techniques are better in terms of results. Naive Bayes algorithm was the best of the three techniques.