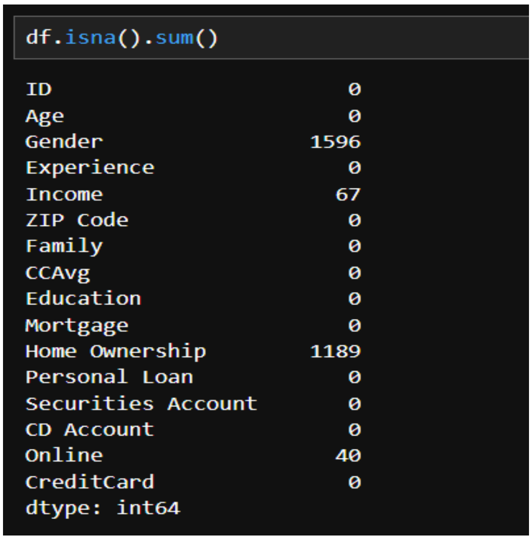
Task 1: Bank loan classification

Description:
You are given a dataset of the bank loans consisting of 15 columns and a corresponding target column. Your task is to build a machine-learning model that can accurately classify whether the personal loan was accepted or not based on the information provided.

Following Process were carried Out:

EDA:

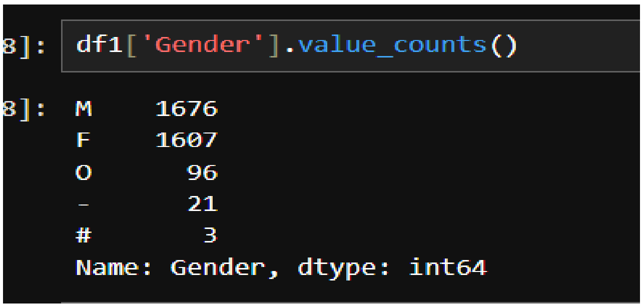
Exploratory Data Analysis was carried out. The data type were checked for each column. Also, the number of unique values were checked out for each column. And during it was observed that the



```
df.isna().sum()

ID                       0
Age                      0
Gender                1596
Experience               0
Income                  67
ZIP Code                 0
Family                   0
CCAvg                    0
Education                0
Mortgage                 0
Home Ownership        1189
Personal Loan            0
Securities Account       0
CD Account               0
Online                  40
CreditCard               0
dtype: int64
```

The Gender, Income and the Home Ownership has null values. As the number of null vales are very large, truncating them would have serious impact upon dataset. SO for Gender there are



```
8]: df1['Gender'].value_counts()

8]: M     1676
    F     1607
    O       96
    -       21
    #        3
    Name: Gender, dtype: int64
```

Data Pre-Processing:

Gender:

There are 5 types of data for gender. So I merged the '#' and '-' into the others as their numbers aren't that significant. And for rest 1596 null values for gender I replaced the nan values in random order for 'M','F' and 'O'. So that the distribution becomes even and isn't biased.

Income:

Similarly, for Income, I replaced the 67 nan values with the mean of Income which is 72.75.

Personal Loan:

And upon checking for Personal Loan(Target) column a mislabeled a row had a ' ' value for Personal Loan. So it was removed.

Age:

On Inspecting the age column there were mistakes. Some data were over 120. There were 6 of them. The values ranges from 120 to 700. So those age data were replaced by the average value of age(46).

Feature Engineering:

The following columns were taken as feature and divided into numerical and categorical features.

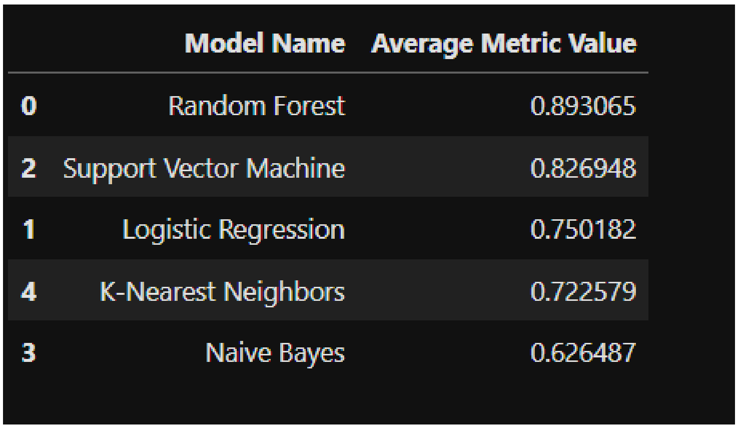
Numerical features: ['Age', 'Experience', ' Income', 'Family', 'CCAvg', 'Mortgage']

Categorical features: ['Gender', 'Education', 'Home Ownership']

The categorical features were one hot encoded and for the numerical features mean were removed and scaled to unit variance using standardscaler().

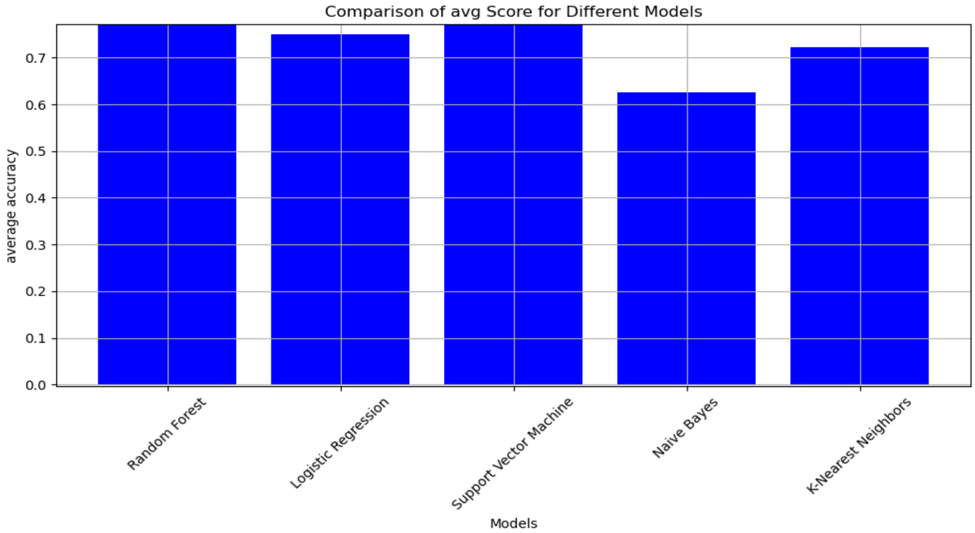
Model Creation:

Now the model was created and fitted. Various models were taken as a pipeline.

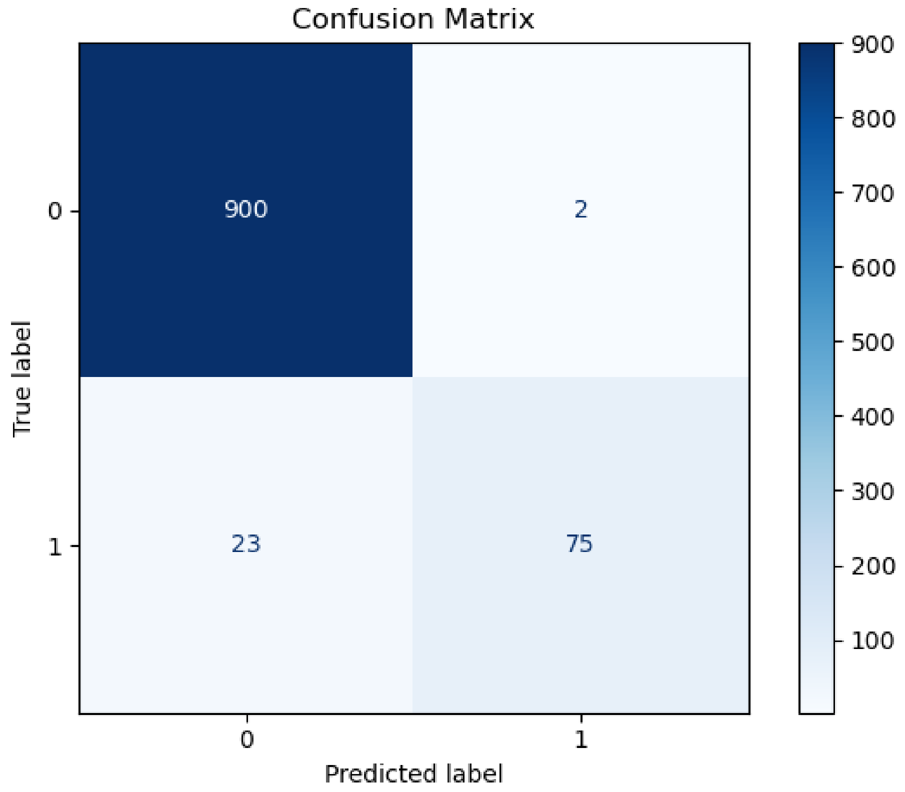


	Model Name	Average Metric Value
0	Random Forest	0.893065
2	Support Vector Machine	0.826948
1	Logistic Regression	0.750182
4	K-Nearest Neighbors	0.722579
3	Naive Bayes	0.626487

Sorting on the basis of Average of F1, Recall, Precision and Accuracy for test dataset.



We see that Random Forest Classification Works Really well!



Confusion matrix for test dataset.

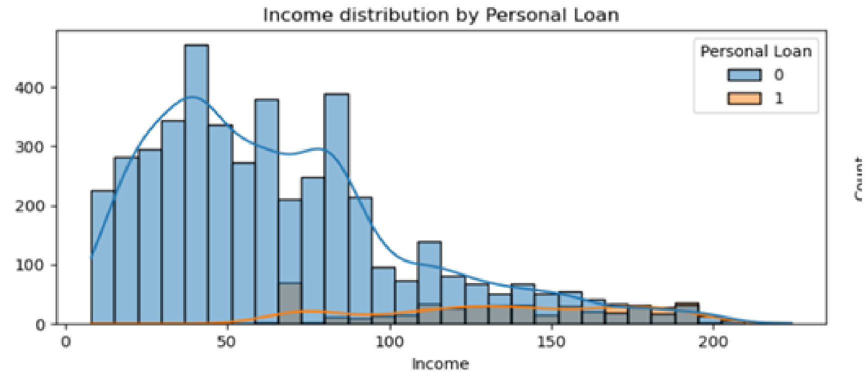
The Random Forest performed best in terms of testing accuracy.

So we will be selecting this model for prediction.

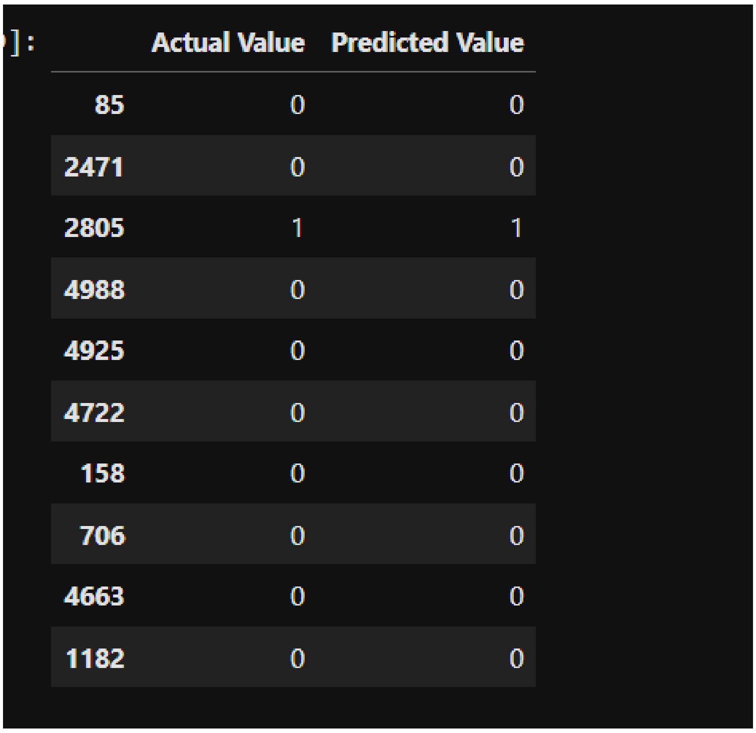
Some Insights about the dataset:

The features were plotted against the Personal Loan for checking the dependency of Target Column (Personal Loan) on these features.

It was found that:



We can see that as income starts to increase, After approximately around 50,000. They seem to have loan. So having high income is essential factor for Loan acceptance.



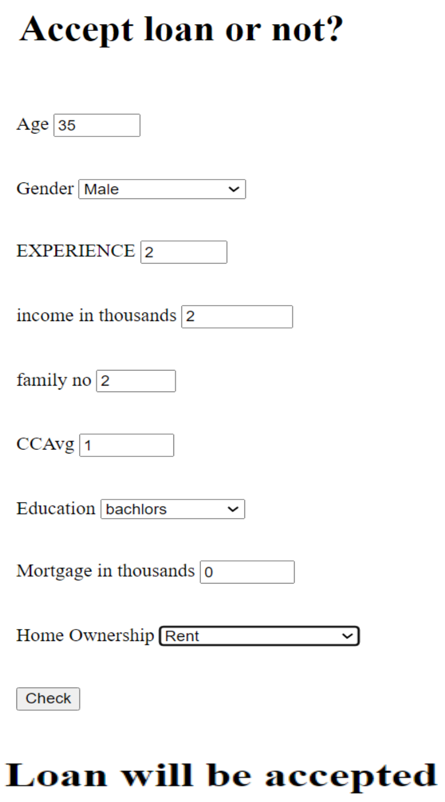
]:

	Actual Value	Predicted Value
85	0	0
2471	0	0
2805	1	1
4988	0	0
4925	0	0
4722	0	0
158	0	0
706	0	0
4663	0	0
1182	0	0

Output

# Run App.py

User can provide data from the webpage.



Accept loan or not?

Age 35

Gender Male

EXPERIENCE 2

income in thousands 2

family no 2

CCAvg 1

Education bachlors

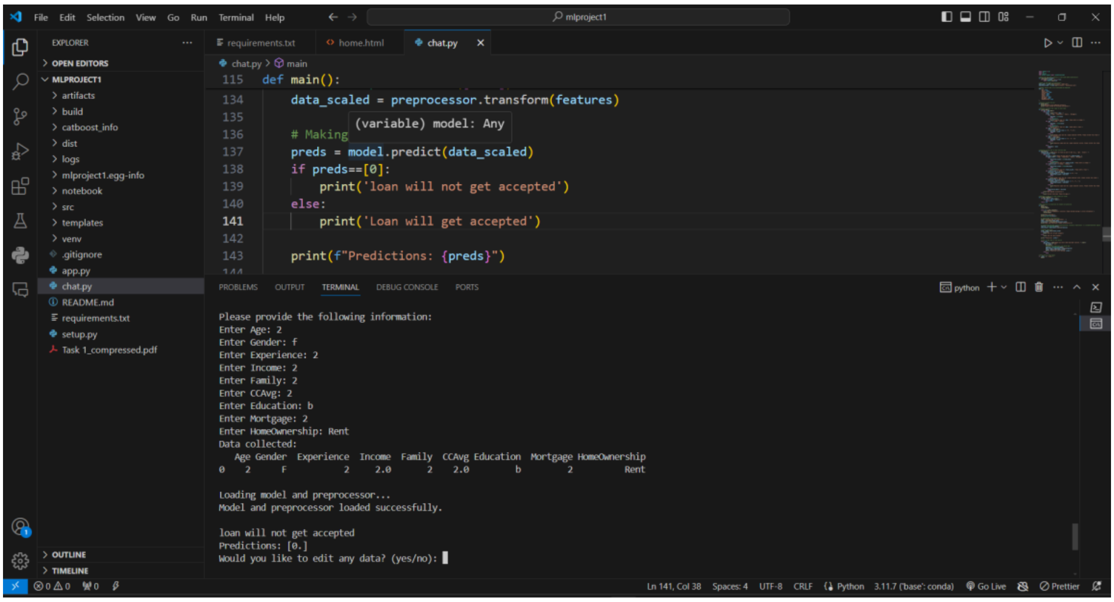
Mortgage in thousands 0

Home Ownership Rent

Loan will be accepted

# Run Chat.py

FOR Chat.py, user can provide data interactively and it will provide the predictions.



```python
def main():
    data_scaled = preprocessor.transform(features)

    # Making
    preds = model.predict(data_scaled)
    if preds==[0]:
        print('loan will not get accepted')
    else:
        print('Loan will get accepted')

    print(f"Predictions: {preds}")
```

```
Please provide the following information:
Enter Age: 2
Enter Gender: f
Enter Experience: 2
Enter Income: 2
Enter Family: 2
Enter CCAvg: 2
Enter Education: b
Enter Mortgage: 2
Enter HomeOwnership: Rent
Data collected:
   Age Gender  Experience  Income  Family  CCAvg Education  Mortgage HomeOwnership
0    2      F           2     2.0       2    2.0         b         2          Rent

Loading model and preprocessor...
Model and preprocessor loaded successfully.

loan will not get accepted
Predictions: [0.]
Would you like to edit any data? (yes/no): 
```