Name of the project:

Micro Credit Defaulter Project.

Submitted By:

Pooja Rajpal

Internship batch 33

ACKNOWLEDGMENT

I would like to express my appreciation to my institute

(Data Trained Education)

And to my respected teacher(Shankar sir) for his teachings and notes he provided to the students.I would like to thank my institute which is always available to answer our queries.

This project has been source of learning and bring our theoretical knowledge into real time projects .

I would really acknowledge my teacher’s help and guidance.

Introduction

Business problem statement:

Micro credit is a form of finance given to an individual to help them.MFI organization offers loan to low income people so that they can start their own business,home loan,agricultural loans.

Now a days microfinance is a tool accepted for reduction of poverty.Building a model that predicts the probability for each loan that the customer will be paying back the loan amount within 5 days of insurance.

Background of the problem:

To understand the project it is necessary to understand what the project is about.This project is about the banking service provide to unemployed or low income individuals to allow the people to take small business loans safely .Its goal is to provide financial services to encourage enterprenuers to become self sustainable.

Micro credit is important as it provides resources and access to capital to those who are unable to take credit.

Some micro finance instittions provide small loans and resources to entrepreneurs to help them to get business.However the interest rates are higher as because of high risk.

- Review of Literature:
  Micro credit has been proven to be an essential tool for reducing the poverty .Actually microcredit provides credit to low income people such as villagers who are unable to run even a small business.MFI is an organization that offers these financial services to these people .This project is about the telecommunication services who collaborated with MFI to provide micro credit on mobile balances whichn is to be paid back within 5 days .

- Motivation for the Problem Undertaken:

  Building project will improve my skills and make things better.Building successful projects gives the feeling of pride .And for that I need to build more and more projects passionately and try to do it with more accuracy.

  However completing the project requires complex steps.Things get really difficult at some stages.but

we need to adjust our vision .And try once again.Also its my passion to build a model with high

accuracy and will definitely try to learn more n more for that.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

The first step here is to import necessary libraries.And load the dataset in jupyter notebook to analyse it. Statistical Modelling is necessary because it summarizes the results can be observed by the evaluators .It is relationship between variables.

Checking for nulls is another step in this dataset and impute them using imputers to clean the data.

Checking for zeros and treating them

Statistical techniques such as mean,median ,standard deviation ,interquartile ranges .These are simple and we can start it as starting point for EDA.
The goal is to collect the data and make predictions about real world.

Two types statistics (Descriptive and inferential)

Eg(weather forecasting ,stock market,loan approvaland fraud detection ,housingetc)
Mean ,median and mode are the measures of central tendency and is descriptive statistics.

Sampling of data and infer the results to describe entire population.

Central limit theorem that normalises the non normal distribution.If the sample size >30 distribution starts looking normal.

Normal distribution, Bernoulli distribution and binomial distribution, uniform distribution are types of distribution

As the label is discrete this is a classification problem.and the model used are classification models.According to me DECISIONTREE CLASSIFIER IS GOOD FOR THIS DATASET.

- Data Sources and their formats
  The data collected is raw data which is not useful ,cleaning of raw data and utilizing the data for further analysis.Data collection is the process of collecting data whether in structured format or unstructured format

  Sources of data collection:
  Primary data:includes interview and surveys

Secondry data:types(internal data includes organization and external data includes government agencies)

This data is internal data in the form of CSV file given to us by our organization which need to be cleaned to make a model.

```
In [1]: import pandas as pd
        import numpy as np

In [2]: df=pd.read_csv("D:\Micro Credit Project\Data file.csv")
        df.head()
```

Out[2]:

	Unnamed: 0	label	msisdn	aon	daily_decr30	daily_decr90	rental30	rental90	last_re
0	1	0	21408I70789	272.0	3055.050000	3065.150000	220.13	260.13	
1	2	1	76462I70374	712.0	12122.000000	12124.750000	3691.26	3691.26	
2	3	1	17943I70372	535.0	1398.000000	1398.000000	900.13	900.13	
3	4	1	55773I70781	241.0	21.228000	21.228000	159.42	159.42	
4	5	1	03813I82730	947.0	150.619333	150.619333	1098.90	1098.90	

5 rows × 37 columns

```
In [3]: df.shape
```

Out[3]: (209593, 37)

- Data Preprocessing Done

Step1:collection of data and load the dataset in jupyter notebook using pandas and for that we need to import pandas library .

Step2:data cleaning that include checking for nulls and zeros and treat them using imputers.

Step3:checking for datatypes and if object datatype convert them into integer as computers only understand numeric data

Step 4:data visualisation

Step 5:plotting boxplot and distplot for columns to check for outliers and treating them using outlier detection

Step6:checking for skewness and treat them using log method

At the end split the dataset into features and label

- Data Inputs- Logic- Output Relationships
  All the other columns are features except for label.columns such as :
  Msisdn:mobile number of users
  Aon:age on cellular network
  Daily _decr90
  Daily_decr30
  Rental 30

Rental 90

- Hardware and Software Requirements and Tools Used
  Laptop and Anaconda navigator is desktop graphical user interface that allows to launch applications
  Jupyter notebook is open source software.

Libraries used are pandas ,numpy ,matplotlibpyplot and seaborn for visualizations, Sklearn.preprocessing to import (label encoderto encode the columns,power transformer to remove the skewness),sklearn.model selection to import train_test_ split to split the dataset into training and testing data,sklearn .metrics to know the acuuracy score ,classification report and libraries to import models

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods):

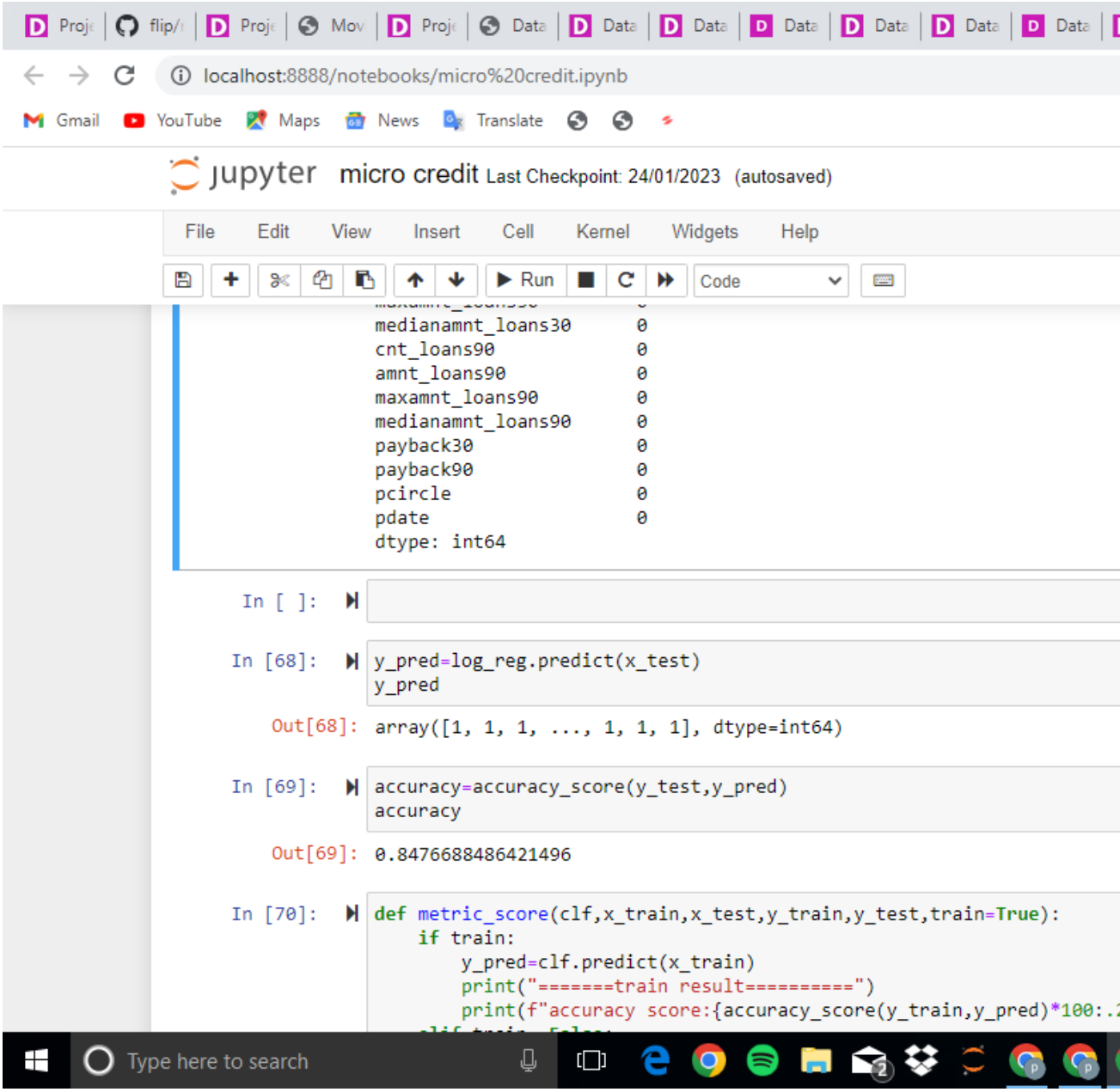
Two main statistical methods are used in data analysis ie descriptive statistics and inferential statistics.I used descriptive statistics which summarizes data using mean,median.
Five basic methods of statistical analysis are mean standard deviation,regression ,hypothesis testing and sample size determination.

- Testing of Identified Approaches (Algorithms)

  As it is a classification problem I used logistic regression, decision tree classifier and kneighbors classifier.

- Run and Evaluate selected models

  With logistic regression



With KNeighborsClassifier

jupyter micro credit Last Checkpoint: 24/01/2023 (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Truste

Run Code

```
pred=clf.predict(x_test)
print("=======test result==========")
print(f"accuracy score:{accuracy_score(y_test,pred)*100:.2f}%")
```

In [71]:
```
#score with k neighbors classifier
```

In [72]:
```
knn=KNeighborsClassifier()
knn.fit(x_train,y_train)
```

Out[72]:
```
KNeighborsClassifier()
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [73]:
```
#calling the function and checking the score
```

In [74]:
```
metric_score(knn,x_train,x_test,y_train,y_test,train=True)
metric_score(knn,x_train,x_test,y_train,y_test,train=False)
```

```
=======train result==========
accuracy score:88.35%
=======test result==========
accuracy score:85.87%
```

In [75]:
```
# hyper parameter tuning
```

In [76]:
```
from sklearn.model_selection import GridSearchCV
```

e here to search

With Decision Tree Classifier

jupyter micro credit Last Checkpoint: 24/01/2023 (autosaved) Logout

le Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
                                'min_samples_leaf': range(2, 6),
                                'min_samples_split': range(3, 8)})
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [86]:
```
best_parameters=grd_srch.best_params_
print(best_parameters)
```

```
{'criterion': 'entropy', 'max_depth': 10, 'max_leaf_nodes': 8, 'min_samples_leaf': 2, 'min_samples_split': 3}
```

In [87]:
```
clf=DecisionTreeClassifier(criterion='entropy',max_depth=10,max_leaf_nodes=8,min_samples_leaf=2,min_samples_split=3)
clf.fit(x_train,y_train)
```

Out[87]:
```
DecisionTreeClassifier(criterion='entropy', max_depth=10, max_leaf_nodes=8,
                       min_samples_leaf=2, min_samples_split=3)
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [88]:
```
metric_score(clf,x_train,x_test,y_train,y_test,train=True)
metric_score(clf,x_train,x_test,y_train,y_test,train=False)
```

```
=======train result==========
accuracy score:90.08%
=======test result==========
accuracy score:90.17%
```

In [ ]:

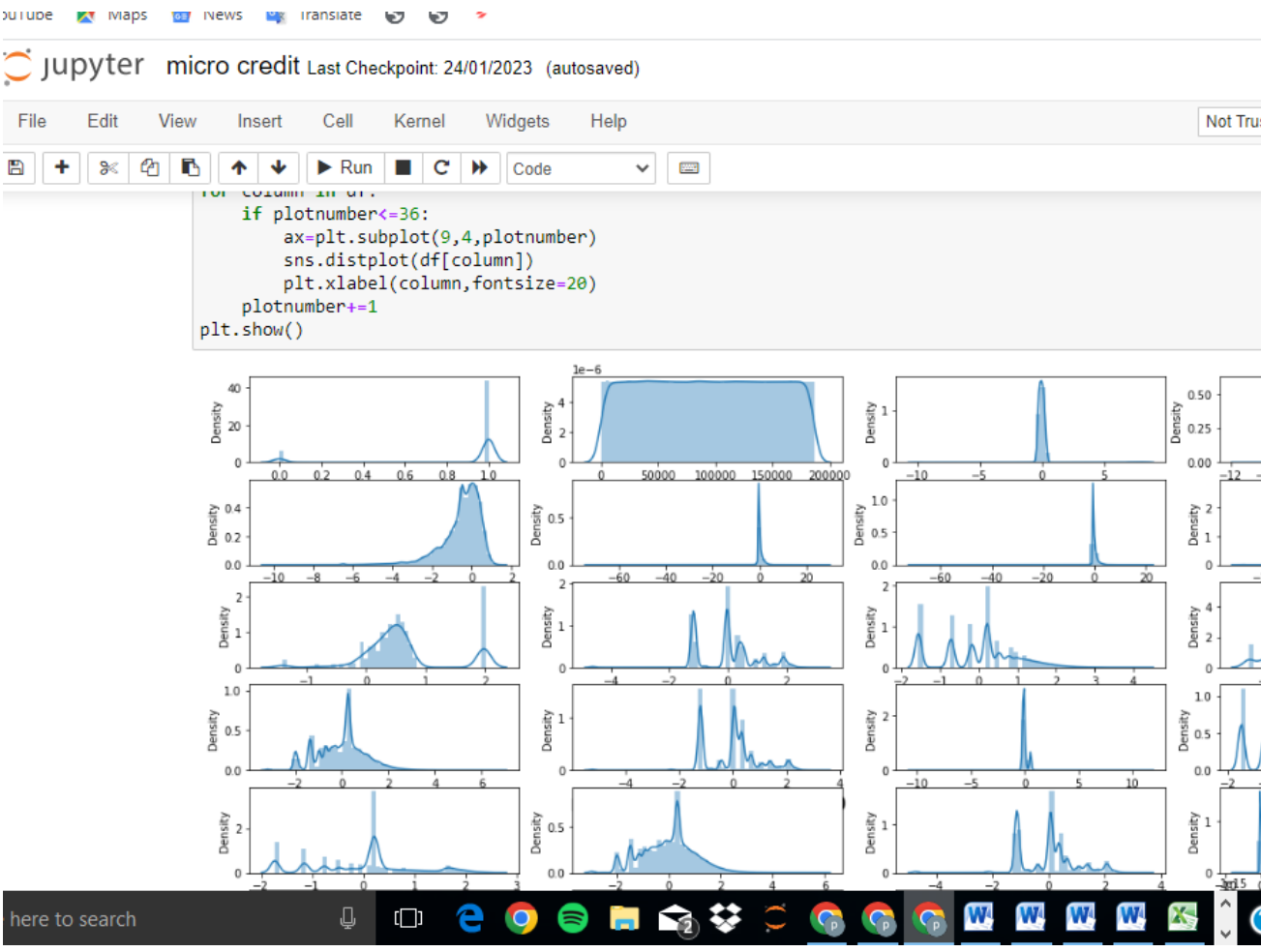
e to search ENG 01:48 01-02-2023

- Key Metrics for success in solving problem under consideration

  Metrics are used to measure the performance of the model. I used the accuracy_score to find the accuracy of the model.because according to me it gives most reliable results and its easy to determine what is the level of accuracy of the model built.

- Visualizations

  Distplot to visualise whether the data is right, left skewed to make it



Dist plot or distribution plot represents the overall distribution of continuous data variables .The seaborn module along with matplotlib module is usedt depict the dist plot.

- Interpretation of the Results

  DecisionTree classifier works well with this dataset .As this is a classification problem because the label is discrete data .Preprocessing is necessary to clean the data

Visualization because it is easily understood by us to observe the data in graphical format.

CONCLUSION

This is micro credit defaulter dataset which required data cleaning and Exploratory Data Analysis to analyse the data.After that checking the correlation between the features because it can affect the accuracy of the model .

Hyperparameter tuning is necessary to increase the accuracy.

Then building model and checking the accuracy.

- Learning Outcomes of the Study in respect of Data Science

  I faced lot of challenges in making this project

  Also learnt lot of things As it has lot of columns found some difficulties .

  In the regression problem the output variable must be continuous and in classification it must be discrete.

    - Limitations of this work and Scope for Future Work

      I think I should have worked harder to increase the accuracy .