Micro Credit Project

Submitted by:

Laxmi Narayan

Acknowledgement

I would like to express my gratitude to my primary SME, Sajid Choudhary, who guided me throughout this project. I would also like to thank my friends and family who supported me and offered deep insight into the study. I wish to acknowledge the help provided by the technical and support staff in the **Data Science of Flib Robo Technologies**. I would also like to show my deep appreciation to my supervisors who helped me finalize my project.

Introduction

In micro-lending markets, lack of recorded credit history is a significant impediment to assessing individual borrowers' creditworthiness and therefore deciding fair interest rates. This research compares various machine learning algorithms on real micro-lending data to test their efficacy at classifying borrowers into various credit categories. We demonstrate that off-the-shelf multi-class classifiers such as random forest algorithms can perform this task very well, using readily available data about customers (such as age, occupation, and location). This presents inexpensive and reliable means to micro-lending institutions around the developing world with which to assess creditworthiness in the absence of credit history or central credit databases.

During the last few decades, credit quality emerged as an essential indicator for banks' lending decisions (**Thomas et al. 2017**). Numerous elements reflect the borrower's creditworthiness, and the use of credit scoring mechanisms could moderate the estimation of the probability of default (PD) while predicting the individual's payment performance.
Lately people depend on bank loans to meet their wishes. The fee of loan packages will increase with a very rapid speed in current years. Risk is constantly involved in approval of loans. The banking officials are very acutely aware of the price of the mortgage quantity by its customers. Even after taking lot of precautions and analyzing the mortgage applicant information, the mortgage approval choices are not continually correct. There is need of automation of this system so that loan approval is much less risky and incur less loss for banks. Since it is a major activity for the banks,to identify whether a loan of the desired amount should be approved to the applicant or not,the Computer Science is capable of making such a system using Artificial Intelligence,which can make this tough decision accurately and quickly.
Using data science,which is responsible to deal with the large amount of data efficiently,and some algorithms of Machine Learning,a prediction system is made,which,on the basis of some training data set,is capable of identifying if the loan applicant is ideal for the loan approval or not.
Machine Learning algorithms like Decision Tree,Logistic Regression,Random Forest,etc. are used for the analysis.These are efficient algorithms that are followed for data analysis and

prediction making.

The system will look into some basic information of the applicant such as his/her profession,age,gender,marital status,etc.,and after analyzing all this information,using visualization and machine learning algorithms,it will come to a decision.

Organizations are overwhelmed with monster measures of information. Accordingly, it's essential to comprehend what to attempt to with this detonating information and the best approach to use it.

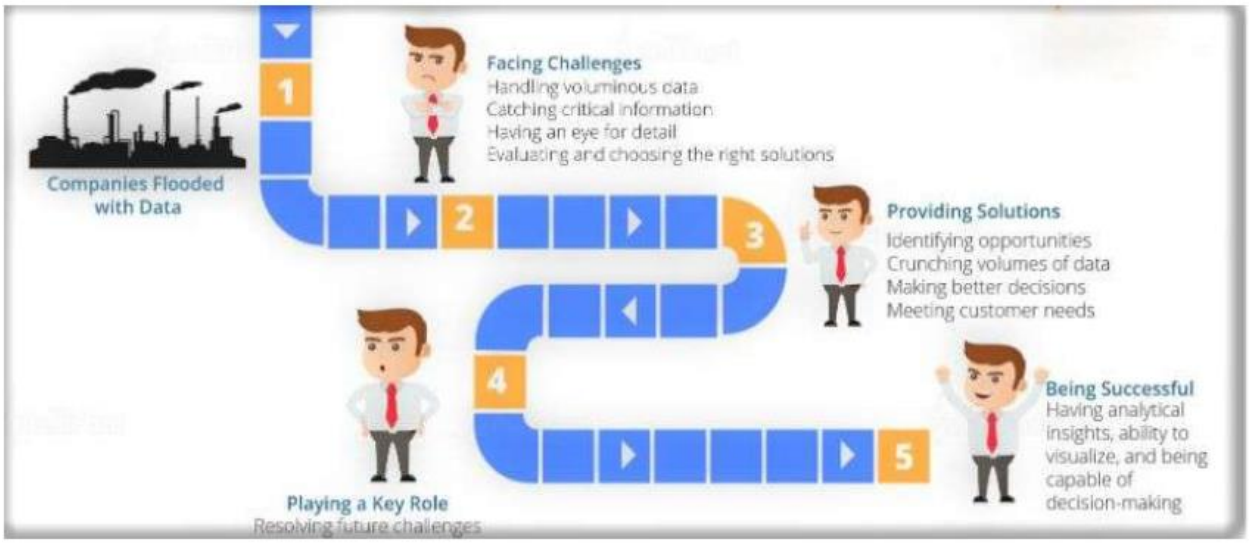

Fig 1: Data flow in the industry

It is here, the idea of Data Science comes into the image. Data science along with arithmetics and organisational information allows the organisation to explore approaches to:

Reduce costs as much as possible

- Explore the new market and get entry into it
- Hold and open to the various demographies
- Get aware of the effective marketing campaign

Traditional database techniques are not acceptable for knowledge discovery due to the fact they're optimized for instant get right of entry to and summarization of statistics, provided what the user wants to ask, or a question, now not discovery of patterns in big swaths of statistics

while customers lack a wellformulated question. Unlike database querying, which asks “What records satisfies this sample (query)?” discovery asks “What patterns fulfill this statistics?” Specifically, our concern is finding interesting and sturdy patterns that fulfill the facts, where “interesting” is usually something sudden and “robust” is a pattern expected to arise within the destiny.

Machine learning uses data science and makes it feasible to generate such models which are able to make accurate predictions,classisfy things into categories,interpretation of images,etc.This makes the machines and robots intelligent as they can learn on their own and there is no need to worry about their accuracy.In today‟s world,where eveything is getting automated,there surely is need of the mechanisms which can be eaily trusted for their accuracy.Machine learning along with data science makes this possible.Machine learning is already helping several industries and is well praised for easing the human effort.

Machine learning is a subset of artificial intelligence (AI) wherein algorithms research by way of instance from historical statistics to expect results and uncover patterns which humans cannot spot easily. For instance,ML can screen clients who are probably to churn, possibly fraudulent coverage claims,etc. While ML has been round since the Fifties, latest breakthroughs in lowvalue compute assets like cloud storage, less complicated collection of data, and the proliferation of information science have made it very a good deal “the next huge thing” in commercial enterprise analytics.

LITERATURE SURVEY

DATA SCIENCE:

Data science is the sphere of study that mixes domain information, programming competencies, and know-how of mathematics and facts to extract significant insights from records. Data scientists follow ML algorithms to numbers, text, pics, video, audio, etc. to generate AI systems to carry out tasks that generally require human intelligence. In turn, those structures generate insights which analysts and users of related field can turn into tangible business value.

Big data is a blanket term for any series of records so large or complicated that it becomes difficult to technique them using traditional records management strategies consisting of, as an instance, the relational database management systems(RDBMS).The extensively adopted RDBMS has lengthy been regarded as a one-length-suits-all solution, but the needs of coping with big records have proven in any other case. Data science involves using techniques to investigate big quantities of information and extract the know-how it carries. You can consider the relationship among big information and data science as being just like the relationship between crude oil and an oil refinery. Data technology and massive facts advanced from facts and traditional facts control but are actually taken into consideration to be different disciplines.

Benefits and uses of data science

Big data and data science are used almost everywhere in both business and noncommercial settings. The variety of use instances is large.Commercial organizations in nearly each industry use big data and data science to gain insights into their customers, tactics, group of workers, completion,and products.Data Science is used by many businesses to offer clients a higher user experience, as well as to cross-sell, up-sell, and customize their offerings.

An example of this is Google AdSense.It generates the advertisements for the users based on their interest and past searches.This makes it easier for the users to get the required items of their interests easily.

Human aid specialists analyse the employees by analysing their moods and behaviours.Relations with the co-workers can also be analysed this way.

Data science is used by financial institutions in prediction of stock markets,decide the risk of lending cash, and discover ways to attract new customers for institution's services. Maximum trade is nowadays takes place with the help of mechanisms which highly operate on the algorithms of machine learning.These are reliable machines and their outputs can be trustred without any question.

Types of Data:

The main categories of data are these:

- Structured
- Unstructured

- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

Machine Learning

ML is an utility of AI that gives systems the ability to learn mechanically and improve from experience without being programmed explicitly. ML centers around the advancement of PC programs that can get to information and use it learn for themselves.

To accomplish ML, specialists create broadly useful calculations that can be utilized on enormous classes of learning issues. At the point when you need to explain a particular undertaking you just need to take care of the calculation progressively explicit information. As it were, you're customizing by model. By and large a PC will utilize information as its wellspring of data and contrast its yield with an ideal yield and afterward right for it. The more information or "experience" the PC gets, the better it becomes at its assigned activity, similar to a human does.

ML algorithms are categorized as follows:

- Supervised ML algorithms
- Unsupervised ML algorithms
- Semi-supervised ML algorithms
- Reinforcement ML algorithms

Importance of machine learning:

As statistical evaluation depends mostly upon rule-based selection-making, ML excels at responsibilities which are difficult to outline with actual guidelines.ML can be implemented to several enterprise cases in which final results relies on loads of factors.Elements which might be tough or not possible for a person to analyze.Hence, companies use ML of for predicting mortgage defaults, information elements that result in customer churn, figuring out probable

fraudulent transactions, optimizing coverage claims approaches, predicting medical institution readmission, and plenty of different scenarios.

Where machine learning is used in data science process:

Majorly ML is related to the data science‟data modeling step.Still ML somehow can be used in every step of data science.
It is necessary to have some qualitative raw data to get the data modeling phase started.Before this step,ML can be used in the data preparation step.
For instance:Suppose list os strings needs to be cleansed.Comparing the strings to spot the spelling errors can be done by placing the similar strings together,which can be done by ML algorithms.
ML is also useful in data exploration step of data science.
For instance:ML algorithms are capable of finding out patterns in data and bring the required data together,which would have been difficult errand otherwis.

Python tools used in machine learning:

SciPy: A library that is responsible for the integration of some fundamental packages such as NumPy, matplotlib, Pandas, and SymPy.

1.NumPy :It gives access to functions related to array functions and linear algebra.
2.Matplotlib: This library is often called as 2D plotting package with some 3D functionality.
3. Pandas:This library is a very useful high-performance library which is capable of manipulating data easily. It provides dataframes to Python.
4.SymPy: It is a package which helps with computational algebra and symbolic mathematics.
5.StatsModels: It is a package helpful for statistical methods.
6. Scikit-learn:It is a Python library containing a large number of algorithms.
7. RPy2 :This permits us to call functionalities of R from within Python.
8.NLTK (Natural Language Toolkit) : This Python toolkit focuses on text analytics.
It is a wise choice to begin with Python using these libraries.The real performance comes into play when a Python code is run at some regular intervals.

The modeling process:

There are following stes in modeling phase:

1. Feature engineering and model selection
2. Training the model
3. Model validation and selection
4. Applying the trained model to unseen data

Engineering features and selecting a model:

Creation of appropriate predictors for the model is done with engineering features.Since in this step the model recombines these features to attain the required predictions,hence this step is also considered one of the most important step.

Some functions are the variables you get from a records set. In exercise there is a need to discover the features on his/her self, that might be scattered among unique statistics units. In numerous tasks we had to convey collectively more than 20 one of a kind facts resources before we had the raw information we required. Many times there is a need to transform an input until it turns into an accurate predictor or to combine a couple of inputs. An instance of combining more than one inputs would be interaction variables: the impact of both single variable is low, but if both are present their effect becomes high. This is specifically authentic in chemical and environments related to medical science. For instance, even though vinegar and bleach are harmless commonplace family merchandise through themselves, blending them outcomes in poisonous chlorine fuel, a gasoline that killed hundreds all through World War I.

Training the model:

Model training can be done having an idea of efficient modeling technique and using the correct predictors at the correct place.In this step training data is given to the model so that it can learn using this data.The modeling techniques that are famous have all-set implementations ready to be used in nearly all coding languages.Thus,it makes it easier to train the model just by running a few lines of program.Other techniques of data science requires complex calculations and needs to be implemented using modern techniques.After the model gets trained successfully,it is to be checked whether this model is capable to deal with real-word problems or not.

Validating the model:

There are several data modeling techniques in data science.One just has to chose the correct and efficient one.Basically,there are two distinguishing properties of a good model:

- It has a good prediction making power
- It works efficiently and accurately with new data (test data)

In order to get these properties right,error measure needs to be defined which tells the extent to which the model is inaccurate and also a strategy for its validation.

Some of the validation techniques are:

- Partitioning the given data by taking out some percent of the data as training set,This is very common technique.
- K-folds cross validation:According to this technique the given data is divided into k parts and uses each part one time as a test data set the others as a training data set.The advantage of using this technique is that all the data present in the data set is used.
- eave-1 out:This approach is the similar to k-folds having k=1. One observation is always left out and training is done on the rest of the data. This technique is implemented only on datasets that are not too big.
- Regularization is a famous term in ML. While using regularization, a penalty is givem for using the extra variables during the making of the model.
- A model having minimum possible predictors is ensured with L1 regularization. This ensures the model‟s robustness.
- The variance between the coefficients of the predictors are kept as minimum as possible using L2 regularization.It becomes difficult to see the actual impact of every predictor if the variance between predictors overlap.If there is no such overlapping the variance is interpreted more easily and clearly.
- Basically regularization prevents a model to use several features and this inturn prevents over-fitting.
- Validation checks if the model is actually working properly with the real-world applications also or not.This makes validation step important.

Machine Learning Algorithms For Prediction Making

1) Regression analysis: Regression analysis techniques aim mainly to investigate and estimate the relationships among a set of features. Regression includes many models for analyzing the relation between one target/response variable and a set of independent variables. Logistic Regression (LR) is the appropriate *regression analysis* model to use when the dependent variable is binary. LR is a predictive analysis used to explain the relationship between a dependent binary variable and a set of independent variables. For customer churn, LR has been widely used to evaluate the churn probability as a

2) Decision Tree: Decision Tree (DT) is a model that generates a tree-like structure that represents set of decisions. DT returns the probability scores of class membership. DT is composed of: a) **internal Nodes:** each node refers to a single variable/feature and represents a test point at feature level; b) **branches,** which represent the outcome of the test and are represented by lines that finally lead to c) **leaf Nodes** which represent the class labels. That is how decision rules are established and used to classify new instances. DT is a flexible model that supports both categorical and continuous data. Due to their flexibility they gained popularity and became one of the most commonly

3) Support Vector Machine: Support Vector Machine (SVM) is a supervised learning technique that performs data analysis in order to identify patterns. Given a set of labeled training data, SVM represents observations as points in a high-dimensional space and tries to identify the best separating hyperplanes between instances of different classes. New instances are represented in the same space and are classified to a specific class based on their proximity to the separating gap. For churn prediction, SVM techniques have been widely investigated and evaluated to be of high predictive performance [37]-[41].

4) Bayes Algorithm: Bayes algorithm estimates the probability that an event will happen based on previous knowledge of variables associated with it. Naïve Bayesian (NB) is a classification technique that is based on Bayes' theorem. It adopts the idea of complete variables independence, as the presence/absence of one feature is unrelated to the presence/absence of any other feature. It considers that all variables independently contribute to the probability that the instance belongs to a certain class. NB is a supervised learning technique that bases its predictions for new instances based on the analysis of their ancestors. NB model usually outputs a probability score and class membership. For churn problem, NB predicts the probability that a customer will stay with his service provider or switch to another one [42]-[46].

5) Instance – *based learning*: Also known as ***memory-based learning***, new instances are labeled based on previous instances stored in memory. The most widely used instance based learning techniques for

classification is K-nearest neighbor (KNN). KNN does not try to construct an internal model and computations are not performed until the classification time. KNN only stores instances of the training data in the features space and the class of an instance is determined based on the majority votes from its neighbors. Instance is labeled with the class most common among its neighbors. KNN determine neighbors based on distance using Euclidian, Manhattan or Murkowski distance measures for continuous variables and hamming for categorical variables. Calculated distances are used to identify a set of training instances (k) that are the closest to the new point, and assign label from these. Despite its simplicity, KNN have been applied to various types of applications. For churn, KNN is used to analyze if a customer churns or not based on the proximity of his features to the customers in each classes [17], [51].

6) Ensemble – *based Learning*: Ensemble based learning techniques produce their predictions based on a combination of the outputs of multiple classifiers. Ensemble learners include bagging methods (i.e. Random Forest) and boosting methods (i.e. Ada Boost, stochastic gradient boosting).

a) **Random Forest**

Random forests (RF) are an ensemble learning technique that can support classification and regression. It extends the basic idea of single classification tree by growing many classification trees in the training phase. To classify an instance, each tree in the forest generates its response (vote for a class), the model choses the class that has receive the most votes over all the trees in the forest. One major advantage of RF over traditional decision trees is the protection against overfitting which makes the model able to deliver a high performance [47]-[50].

b) **Boosting – based techniques (Ada Boost and Stochastic Gradient Boosting**)

Both AdaBoost (Adaptive Boost) and Stochastic Gradient Boosting algorithms are ensemble based algorithms that are based on the idea of boosting. They try to convert a set of weak learners into a stronger learner. The idea is that having a weak algorithm will perform better than random guessing. Thus, Weak learner is any algorithm that can perform at least a little better than random solutions. The two algorithms differ in the iterative process during which weak learners are created. Adaboost filters observations, by giving more *weight* to problematic ones or the ones that the weak learner couldn't handle and decrease the correctly predicted ones. The main focus is to develop new weak learns to handle those misclassified observations. After training, weak learners are added to the stronger learner based on their alpha weight (accuracy**)**, the higher alpha weight, the more it contributes to the final learner. The weak learners in AdaBoost are decision trees with a single split and the label assigned to an instance is based on the combination of the output of all weak learners weighted by their accuracy [56].

7) Artificial neural network: Artificial Neural Networks (ANNs) are machine-learning techniques that are inspired by the biological neural network in human brain. ANNs are adaptive, can learn by example, and are fault tolerant. An ANN is composed of a set of connected nodes (neurons) organized in layers. The input layer communicates with one or more hidden layers, which in turn communicates with the output layer. Layers are connected by weighted links. Those links carry signals between neurons usually in the form of a real number. The output of each neuron is a function of the weighted sum of all its inputs. The weights on connection are adjusted during the learning phase to represent the strengths of connections between nodes. ANN can address complex problems, such as the churn prediction problem. Multilayer perceptron (MLP) is an ANN that consists of at least three layers. Neurons in each layer use supervised learning techniques [52]. In the case of customer churn problem, MLP has proven better performance over LR .

SYSTEM DEVELOPMENT

The system requirements for the algorithms to run efficiently and for the implementation of the whole idea are:

- Windows 10 (64-bit)
- 8 GBRAM
- Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz
- ANACONDA
- Python

Performance Analysis

In this project:

1. The data set needs to be explored initially.
2. Certain models will be created toknow whether the loan should be approved or not.
3. Accuracy scores for the clients will be generated.

The first step is of this project development is to import the required libraries:

```
In [1]: # importing libraries
        import pandas as pd
        import numpy as np              # For mathematical calculations
        import seaborn as sns           # For data visualization
        import matplotlib.pyplot as plt
        import seaborn as sn            # For plotting graphs
        %matplotlib inline
        import warnings                 # To ignore any warnings
        warnings.filterwarnings("ignore")
```

The next step is to import the training data set in the object named train and the testing data set in the object named test.

This import is done by pandas method of reading .csv files into the system.

```
In [2]: #loading the data
        data=pd.read_csv('Data file.csv')
```

Next we check the the features present in the loaded data sets.

This is done by .columns .

```
data.columns
```

```
Index(['Unnamed: 0', 'label', 'msisdn', 'aon', 'daily_decr30', 'daily_decr90',
       'rental30', 'rental90', 'last_rech_date_ma', 'last_rech_date_da',
       'last_rech_amt_ma', 'cnt_ma_rech30', 'fr_ma_rech30',
       'sumamnt_ma_rech30', 'medianamnt_ma_rech30', 'medianmarechprebal30',
       'cnt_ma_rech90', 'fr_ma_rech90', 'sumamnt_ma_rech90',
       'medianamnt_ma_rech90', 'medianmarechprebal90', 'cnt_da_rech30',
       'fr_da_rech30', 'cnt_da_rech90', 'fr_da_rech90', 'cnt_loans30',
       'amnt_loans30', 'maxamnt_loans30', 'medianamnt_loans30', 'cnt_loans90',
       'amnt_loans90', 'maxamnt_loans90', 'medianamnt_loans90', 'payback30',
       'payback90', 'pcircle', 'pdate'],
      dtype='object')
```

Shape of the data sets are checked further.

The .shape tells us how many total number of rows and columns are present in a dataset

```
data.shape
```

```
(209593, 37)
```

Now let us check out the datatype of each feature:

The .dtypes is used for this.

```
data.dtypes
```

```
Unnamed: 0                int64
label                     int64
msisdn                   object
aon                     float64
daily_decr30            float64
daily_decr90            float64
rental30                float64
rental90                float64
last_rech_date_ma       float64
last_rech_date_da       float64
last_rech_amt_ma          int64
cnt_ma_rech30             int64
fr_ma_rech30            float64
sumamnt_ma_rech30       float64
medianamnt_ma_rech30    float64
medianmarechprebal30    float64
cnt_ma_rech90             int64
fr_ma_rech90              int64
sumamnt_ma_rech90         int64
medianamnt_ma_rech90    float64
medianmarechprebal90    float64
cnt_da_rech30           float64
fr_da_rech30            float64
```

To display some of the records of a dataset for reference, .head() is used.If we specify a number in the parameter of this function,then that many rows are displayed otherwise top five records are displayed.

	Unnamed: 0	label	msisdn	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	...	maxamnt_loans30	mediana
0	1	0	21408I70789	272.0	3055.050000	3065.150000	220.13	260.13	2.0	0.0	...	6.0	
1	2	1	76462I70374	712.0	12122.000000	12124.750000	3691.26	3691.26	20.0	0.0	...	12.0	
2	3	1	17943I70372	535.0	1398.000000	1398.000000	900.13	900.13	3.0	0.0	...	6.0	
3	4	1	55773I70781	241.0	21.228000	21.228000	159.42	159.42	41.0	0.0	...	6.0	
4	5	1	03813I82730	947.0	150.619333	150.619333	1098.90	1098.90	4.0	0.0	...	6.0	

5 rows × 37 columns

Model Building

Now,the development of a model for predicting if the user will apply for a loan or not will start. Dummies will be used for converting categorical variables into numerical variables because sklearn models allows only numerical inputs.

The train data set will be divided into two parts,70% of the data will act as training data and the remaining 30% data will be the validation data.

```
[69]: from sklearn.ensemble import RandomForestClassifier
      from sklearn.model_selection import train_test_split

[70]: x_train,x_test,y_train,y_test=train_test_split(datafs,Y,random_state=7)
```

Logistic Regression:

We will first build a Logistic Regression model since logistic regression is used for classification problems.

```
# For Logistic Regression
lg = LogisticRegression()
lg.fit(x_train, y_train)
pred_lg = lg.predict(x_test)
print("Accuracy Score of Logistic Regression model is", accuracy_score(y_test, pred_lg)*100)
```

```
Accuracy Score of Logistic Regression model is 87.72861732243392
```

Now the accuracy of the predictions will be checked. Calculating accuracy on the validation set.

```
from sklearn.model_selection import cross_val_score

lg_scores = cross_val_score(lg, x, y, cv = 10) # cross validating the model
print(lg_scores) # accuracy scores of each cross validation cycle
print(f"Mean of accuracy scores is for Logistic Regression is {lg_scores.mean()*100}\n")
```

```
[0.87800573 0.8798187  0.87781489 0.87943127 0.87790448 0.87900186
 0.87842931 0.87804762 0.87881101 0.88048094]
Mean of accuracy scores is for Logistic Regression is 87.87745806607519
```

Accuracy on the validation set came out to be 87.87%.

Decision Tree algorithm:

Let's try decision tree algorithm now to check if we get better accuracy with that.

Fitting Decision Tree Model

```
# For Decision Tree Classifier
dtc = DecisionTreeClassifier()
dtc.fit(x_train, y_train)
pred_dtc = dtc.predict(x_test)
print("Accuracy Score of Decision Tree Classifier model is", accuracy_score(y_test, pred_dtc)*100)
```

```
Accuracy Score of Decision Tree Classifier model is 86.11597060975222
```

```
dtc_scores = cross_val_score(dtc, x, y, cv = 10)
print(dtc_scores)
print(f"Mean of accuracy scores is for Decision Tree Classifier is {dtc_scores.mean()*100}\n")
```

```
[0.86149809 0.86469466 0.8634542  0.86487905 0.86277971 0.86421108
 0.86769407 0.86464049 0.86220717 0.86516532]
Mean of accuracy scores is for Decision Tree Classifier is 86.41223834775202
```

We got an accuracy of more than 90.4% on the validation set.

Random Forest Algorithm:

Now let us calculate the accuracy using another algorithm called Random Forest.

First we have to set up the model for this algorithm.

```
# For Random Forest Classifier
rfc = RandomForestClassifier()
rfc.fit(x_train, y_train)
pred_rfc = rfc.predict(x_test)
print("Accuracy Score of Random Forest model is", accuracy_score(y_test, pred_rfc)*100)
```

```
Accuracy Score of Random Forest model is 90.914151213461
```

```
rfc_scores = cross_val_score(rfc, x, y, cv = 10)
print(rfc_scores)
print(f"Mean of accuracy scores is for Random Forest Classifier is {rfc_scores.mean()*100}\n")
```

```
[0.91025763 0.91116412 0.90782443 0.91244811 0.9108259  0.91096903
 0.91297295 0.9108259  0.90834486 0.91387948]
Mean of accuracy scores is for Random Forest Classifier is 91.09512417282161
```

Next we will generate a confusion matrix and classification report.

Model Evaluation

```
from sklearn.metrics import plot_roc_curve
from sklearn.metrics import confusion_matrix, classification_report

print("Accuracy Score of RFC model is", accuracy_score(y_test, pred_rfc)*100)
print("Confusion matrix for RFC Model is")
print(confusion_matrix(y_test, pred_rfc))
print("Classification Report of the RFC Model is")
print(classification_report(y_test, pred_rfc))

plot_roc_curve(rfc, x_test, y_test) # arg. are model name, feature testing data, label testing data.
plt.title("Recevier's Operating Characteristic")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.show()
```

```
Accuracy Score of RFC model is 90.914151213461
Confusion matrix for RFC Model is
[[ 3345  4628]
 [ 1085 53820]]
Classification Report of the RFC Model is
              precision    recall  f1-score   support

           0       0.76      0.42      0.54      7973
           1       0.92      0.98      0.95     54905

    accuracy                           0.91     62878
   macro avg       0.84      0.70      0.74     62878
weighted avg       0.90      0.91      0.90     62878
```

Conclusion

1 Lately people depend on bank loans to meet their wishes. The fee of loan packages will increase with a very rapid speed in current years. Risk is constantly involved in approval of loans. The banking officials are very acutely aware of the price of the mortgage quantity by its customers. Even after taking lot of precautions and analyzing the mortgage applicant information, the mortgage approval choices are not continually correct. There is need of automation of this system so that loan approval is much less risky and incur less loss for banks.

2.Artificial Intelligence AI is a rising technology. The utility of AI solves many real world troubles. Machine Learning is an AI method which could be very useful in prediction systems.A model is created from a training data. While making the prediction the model that is evolved by way of training algorithm(ML) is used. The ML algorithm trained the machine the usage of a fragment of the statistics available and the remaining data is tested.

References

[1]. Data science, “Benefits and uses of data science” https://www.simplilearn.com/why-andhow- data-science-matters-to-business-article

[2]. Data science, “Types of data” https://www.wintellect.com/beginning-statistics-for-datascience- types-of-data

[3]. Data science, “The data science process” https://www.kdnuggets.com/2016/03/datascience-process.html

[4].Machinelearning,“Introduction” https://www.digitalocean.com/community/tutorials/anintroduction- to-machine-learning

[5]. Machine learning,“Application in data science”https://www.simplilearn.com/importanceof-machine-learning-for-data-scientists-article

[6]. Python, “Libraries used in machine learning” https://www.geeksforgeeks.org/bestpython-libraries-for-machine-learning/

[7]. Machine learning “Machine learning algorithms”https://www.geeksforgeeks.org/bestpython-libraries-for-machine-learning/