VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018

An Internship Project Report
on

Loan Prediction

Submitted in partial fulfillment of the requirements for the VIII Semester of degreeof **Bachelor of Engineering in Information Science and Engineering** of Visvesvaraya Technological University, Belagavi

Submitted By

Akash Anand
1RN18IS010

Under the Guidance of

Mrs. Hema N
Assistant Professor
Department of ISE

ESTD:2001
An Institute with a Difference

Department of Information Science and Engineering

RNS Institute of Technology

Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post,
Channasandra, Bengaluru-560098

2021-2022

RNS INSTITUTE OF TECHNOLOGY

Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post,
Channasandra, Bengaluru - 560098

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the Internship work entitled ***Loan Prediction*** has been successfully completed by **Akash Anand (1RN18IS010)** a Bonafide student of **RNS Institute of Technology, Bengaluru** in partial fulfillment of the requirements of 8th semester for the award of degree in **Bachelor of Engineering in Information Science and Engineering** of **Visvesvaraya Technological University, Belagavi** during academic year **2021-2022**. The internship report has been approved as it satisfies the academic requirements in respect of internship work for the said degree.

Mrs. Hema N	**Dr. Suresh L**	**Dr. M K Venkatesha**
Internship Guide	Professor and HoD	Principal
Assistant Professor	Department of ISE	RNSIT
Department of ISE	RNSIT	

External Viva

Name of the Examiners	**Signature with Date**
1. _____	**1.** _____
2. _____	**2.** _____

DECLARATION

I, **Akash Anand [USN: 1RN18IS010]** student of VIII Semester BE, in Information Science and Engineering, RNS Institute of Technology hereby declare that the Internship work entitled ***Loan Prediction*** has been carried out by me and submitted in partial fulfillment of the requirements for the *VIII Semester degree of* ***Bachelor of Engineering in Information Science and Engineering*** *of Visvesvaraya Technological University, Belagavi* during academic year 2021-2022.

Place: Bengaluru
Date:

Akash Anand (1RN18IS010)

ABSTRACT

Loans are the core business of banks. The main profit comes directly from the loan’s interest. The loan companies grant a loan after an intensive process of verification and validation. However, they still don’t have assurance if the applicant is able to repay the loan with no difficulties.

The two most pressing issues in the banking sector are: How risky is the borrower? Should the bank lend to the borrower given the risk? The response to the first question dictates the borrower's interest rate. Interest rate, among other things (such as time value of money), tests the riskiness of the borrower, i.e. the higher the interest rate, the riskier the borrower. The bank will then decide whether the applicant is suitable for the loan based on the interest rate. Lenders (investors) make loans to creditors in return for the guarantee of interest-bearing repayment. That is, the lender only makes a return (interest) if the borrower repays the loan. However, whether he or she does not repay the loan, the lender loses money. Banks make loans to customers in exchange for the guarantee of repayment.

Some would default on their debts, unable to repay them for a number of reasons. The bank retains insurance to minimize the possibility of failure in the case of a default. The insured sum can cover the whole loan amount or just a portion of it. Banking processes use manual procedures to determine whether or not a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were a large number of loan applications. At that time, making a decision would take a long time. As a result, the loan prediction machine learning model can be used to assess a customer's loan status and build strategies.

ACKNOWLEDGMENT

At the very onset I would like to place my gratefulness to all those people who helped me in making the Internship a successful one.

Coming up, this internship to be a success was not easy. Apart from the sheer effort, the enlightenment of the very experienced teachers also plays a paramount role because it is, they who guided me in the right direction.

First of all, I would like to thank the **Management of RNS Institute of Technology** for providing such a healthy environment for the successful completion of internship work.

In this regard, I express sincere gratitude to my beloved Principal **Dr. M K Venkatesha,** for providing me with all the facilities.

I are extremely grateful to my own and beloved Professor and Head of Department of Information science and Engineering, **Dr. Suresh L**, for having accepted to patronize me in the right direction with all his wisdom.

I place my heartfelt thanks to **Mrs. Hema N** Assistant Professor, Department of Information Science and Engineering for having guided internship and all the staff members of the department of Information Science and Engineering for helping at all times.

I thank **Mr. Aman Upadhayaya, Instructor, NASTECH**, for providing the opportunity to be a part of the Internship program and having guided me to complete the same successfully.

I also thank my internship coordinator **Dr. R Rajkumar,** Associate Professor, Department of Information Science and Engineering. I would thank my friends for having supported me with all their strength and might. Last but not the least, I thank my parents for supporting and encouraging me throughout. I have made an honest effort in this assignment.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

AI	artificial intelligence
CNN	convolutional neural network
CSV	comma-separated values
DL	deep learning
FN	false negative
FP	false positive
GAN	generative adversarial network
ML	machine learning
NER	named entity recognition
NLP	natural language processing
OCR	optical character recognition
RNN	recurrent neural network
TN	true negative
TP	true positive

1. INTRODUCTION

Python is developed by Guido van Rossum. Guido van Rossum started implementing Python in 1989. Python is a very simple programming language so even if you are new to programming, you can learn python without facing any issues.

Interesting fact: Python is named after the comedy television show Monty Python's Flying Circus. It is not named after the Python snake.

Merits of Python:

1. Readable: Python is a very readable language.

2. Easy to Learn: Learning python is easy as this is a expressive and high level programming language, which means it is easy to understand the language and thus easy to learn.

3. Cross platform: Python is available and can run on various operating systems such as Mac, Windows, Linux, Unix etc. This makes it a cross platform and portable language.

4. Open Source: Python is a open source programming language.

5. Large standard library: Python comes with a large standard library that has some handy codes and functions which I can use while writing code in Python.

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

1.1 BACKGROUND

There are so many different types of Machine Learning systems that it is useful to classify them in broad categories, based on the following criteria:

• Whether or not they are trained with human supervision (supervised, unsupervised, semi supervised, and Reinforcement Learning)

• Whether or not they can learn incrementally on the fly (online versus batch learning)

• Whether they work by simply comparing new data points to known data points, or instead by detecting patterns in the training data and building a predictive model, much like scientists do (instance-based versus model-based learning)

In finance, a loan is the lending of money by one or more individuals, organizations, or other entities to other individuals, organizations, etc. The recipient (i.e. the borrower) incurs a debt, and is usually liable to pay interest on that debt until it is repaid, and also to repay the principal amount borrowed. To read more check out Wikipedia. The whole process of ascertaining if a burrower would pay back loans might be tedious hence the need to automate the procedure.

2. LITERATURE SURVEY

1. “Loan Prediction using Decision Tree and Random Forest “Author- Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi, Mr. Suresh Kumar Year-2020.

The aim of this paper is to find the nature or background or credibility of client that is applying for the loan. They use exploratory data analysis technique to deal with problem of approving or rejecting the loan request or in short loan prediction. The main focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.

2. “Loan Prediction System Using Decision Tree and Random Forest Algorithms”
Authors- Shubham Chaudhary, Vishal Baliyan, Yatharth Katheria

Most banks are renewing their business models and switching to Machine Learning methodology. In this paper, They have discussed classifiers based on Machine and deep learning models on real data in predicting loan default probability. The most important features from various models are selected and then used in the modelling process to test the stability of Random Forest classifiers and Decision Tree Classifier by comparing their performance on data.

3. A. Goyal and R. Kaur, “Accuracy Prediction for Loan Risk Using Machine Learning Models”.

Estimating the probability that an individual would default on their loan, is useful for banks to make a decision whether to approve a loan to the individual or not. In this paper, They find the accuracy of several models in R language and evaluate it to establish the finest model to forecast the finance status for an organization. They did the experiment five times on the same data set and find the experimental results that show the Tree Model for Genetic Algorithm is the best model for forecasting the finance for costumers.

4. A. Goyal and R. Kaur, “A survey on Ensemble Model for Loan Prediction”, International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.

In this paper they discuss the ensemble model that is combination of two or more algorithms and give better results as compared to stand alone models. The performance is also enhanced through the ensemble model.

3. SYSTEM DESIGN

3.1 CURRENT DESIGN

Banking processes use manual procedures to determine whether or not a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were a large number of loan applications. At that time, making a decision would take a long time.

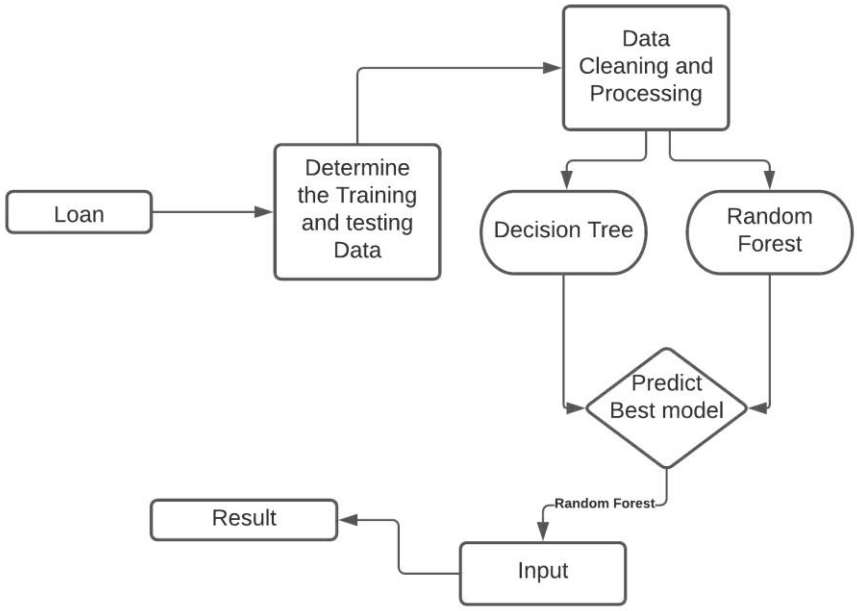
3.2 PROPOSED DESIGN

My aim from the project is to make use of pandas, matplotlib, & seaborn libraries from python to extract insights from the data and scikit-learn libraries for machine learning.

Secondly, to learn how to hyper tune the parameters using grid search cross validation for the Random forest machine learning model.

And in the end, to predict whether the loan applicant can replay the loan or not using voting ensembling techniques of combining the predictions from multiple machine learning algorithms.

3.3 SYSTEM DESIGN



3.3 System design/architecture

Phase 1 – Collection of data:

Through a Kaggle competition.

Phase 2 – Data preprocessing:

Gathered meaningful information

Visualized dataset

Phase 3 – Comparing different models and training :

Trained on Decision Tree and Random Forest

Compared both Models

Phase 4 – Result/Deploying the model

4. IMPLEMENTATION

4.1. EXPLORATORY DATA ANALYSIS

- Purpose : The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other").
- int.rate : The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
- Installment : The monthly installments owed by the borrower if the loan is funded.
- log.annual.inc : The natural log of the self-reported annual income of the borrower.
- Dti : The debt-to-income ratio of the borrower (amount of debt divided by annual income).
- Fico : The FICO credit score of the borrower.
- days.with.cr.line : The number of days the borrower has had a credit line.
- revol.bal : The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
- revol.util : The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
- inq.last.6mths : The borrower's number of inquiries by creditors in the last 6 months.
- delinq.2yrs : The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
- pub.rec : The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

4.2. MACHINE LEARNING METHODS

Decision tree: A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin toss comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing

all attributes).

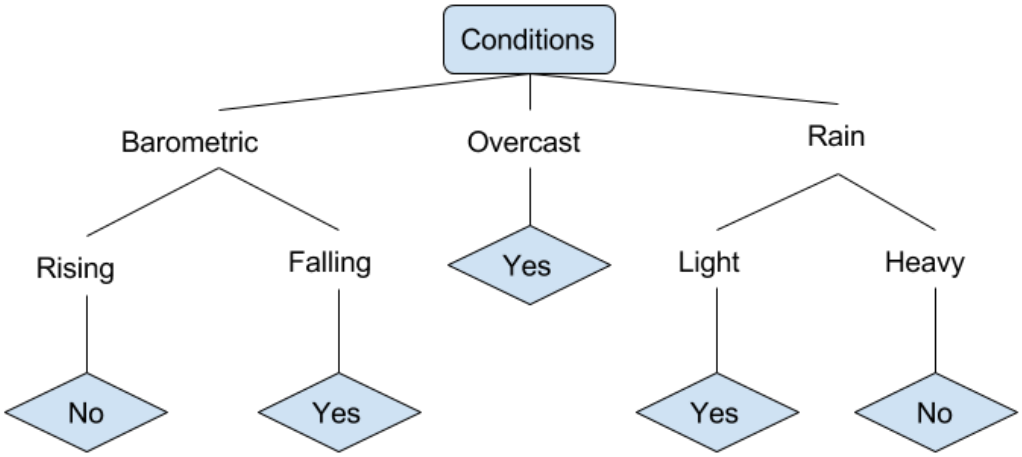


Fig 4.1 Decision Tree.

Random forest: Random forest or random decision forests are an ensemble learning method used for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

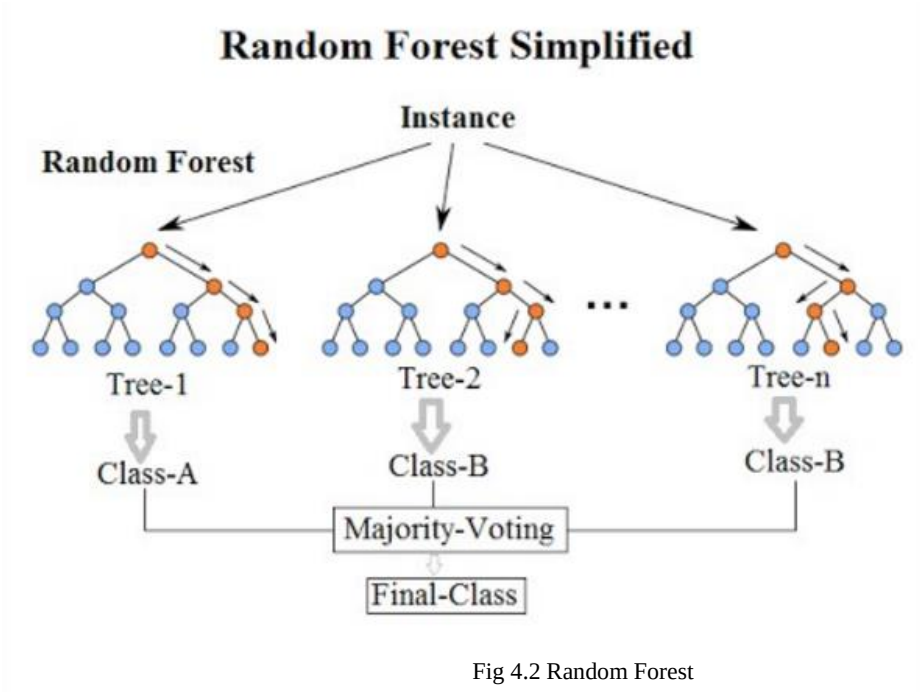


Fig 4.2 Random Forest

4.3 CODE SEGMENT

1. Get the data.

Use pandas to read loan_data.csv as a dataframe called loans.

```
In [3]: df = pd.read_csv('loan_data.csv')

In [4]: df.info()
```

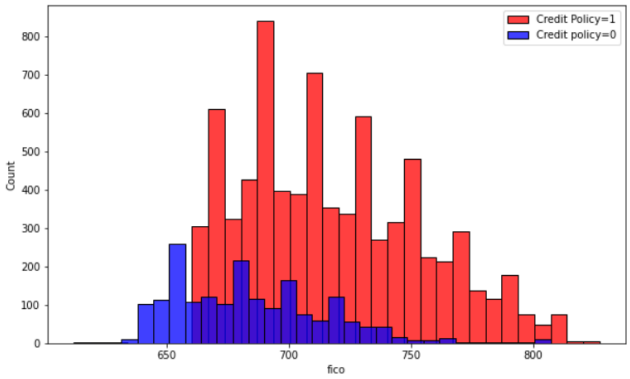
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9578 entries, 0 to 9577
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   credit.policy      9578 non-null   int64
 1   purpose            9578 non-null   object
 2   int.rate           9578 non-null   float64
 3   installment        9578 non-null   float64
 4   log.annual.inc     9578 non-null   float64
 5   dti                9578 non-null   float64
 6   fico               9578 non-null   int64
 7   days.with.cr.line  9578 non-null   float64
 8   revol.bal          9578 non-null   int64
 9   revol.util         9578 non-null   float64
 10  inq.last.6mths     9578 non-null   int64
 11  delinq.2yrs        9578 non-null   int64
 12  pub.rec            9578 non-null   int64
 13  not.fully.paid     9578 non-null   int64
dtypes: float64(6), int64(7), object(1)
memory usage: 1.0+ MB
```

Fig 4.3 importing the data.

2. Exploratory data analysis.

```
In [21]: plt.figure(figsize=(10,6))

sns.histplot(df[df['credit.policy']==1]['fico'],color='red',label='Credit Policy=1',bins=30)
sns.histplot(df[df['credit.policy']==0]['fico'],color='blue',label='Credit policy=0',bins=30)
plt.legend()
plt.xlabel = 'FICO'
```



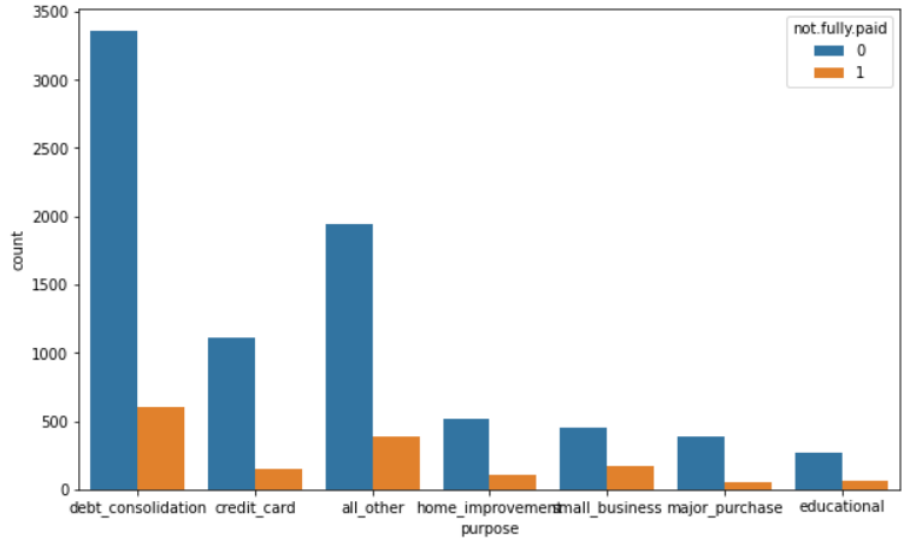
4.4 plotting histogram for credit policy

-

A countplot using seaborn showing the counts of loans by purpose, with the color hue defined by not.fully.paid.

In [24]:
```
plt.figure(figsize=(10,6))
sns.countplot(x=df['purpose'],hue=df['not.fully.paid'])
```

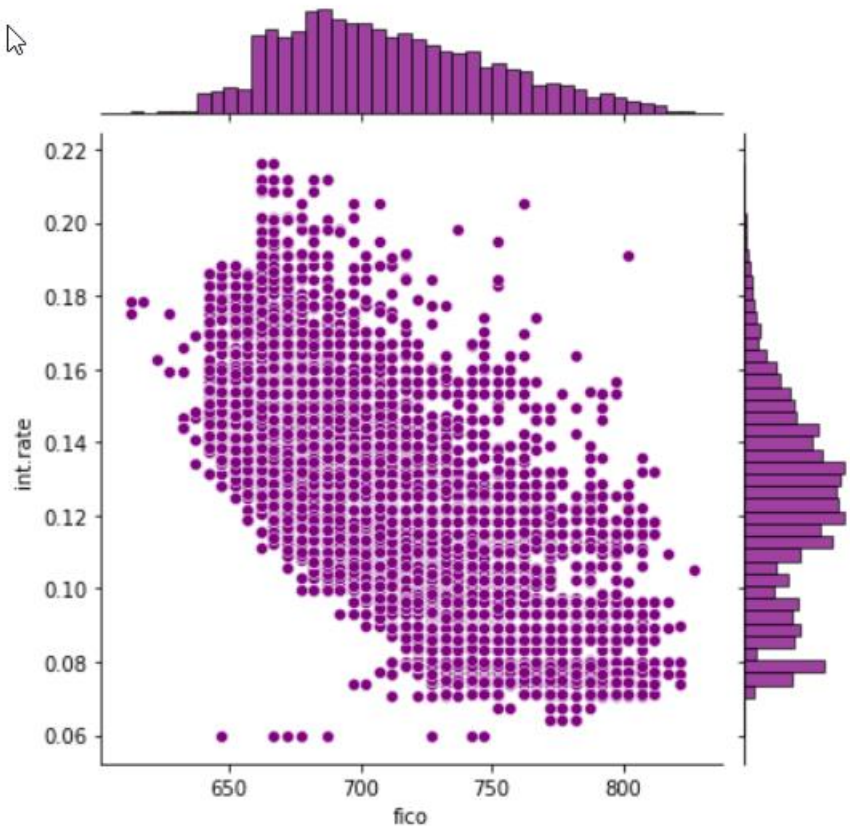
Out[24]: <AxesSubplot:xlabel='purpose', ylabel='count'>



4.5 plotting histogram for purpose of loan

In [33]:
```
sns.jointplot(data=df,x='fico',y='int.rate',color='purple')
```

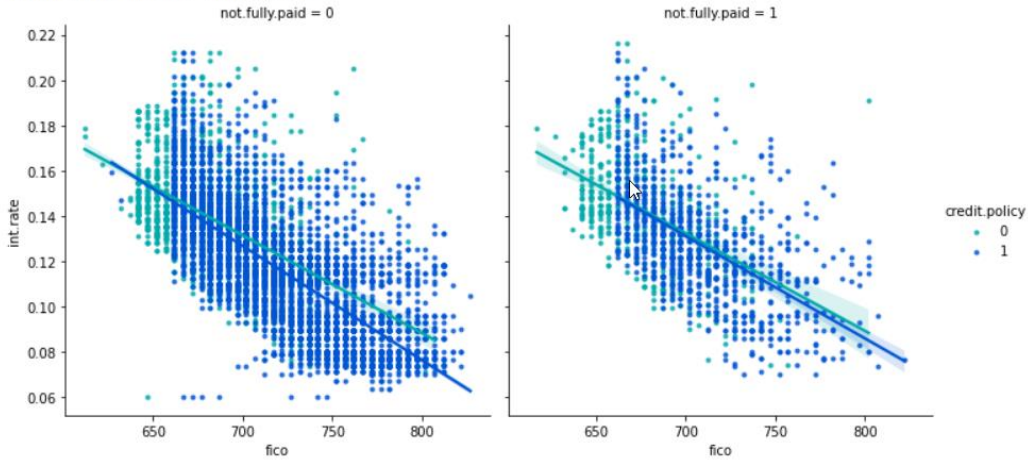
Out[33]: <seaborn.axisgrid.JointGrid at 0x7f23061bea60>



4.6 graph for fico Vs interest rate

```
In [43]: plt.figure(figsize=(11,7))
         sns.lmplot(data=df,x='fico',y='int.rate',col='not.fully.paid',palette='winter_r',hue='credit.policy',markers='.')
```

```
Out[43]: <seaborn.axisgrid.FacetGrid at 0x7f22fee408e0>
<Figure size 792x504 with 0 Axes>
```



4.7 graph on fico Vs Int_rate w.r.t fully paid

3. Categorical Features

Categorical Features

```
In [47]: cat_feats = ['purpose']
```

```
In [51]: final_data = pd.get_dummies(data=df,columns=cat_feats)
         final_data
```

Fig 4.8 Categorical Features

credit.policy, int.rate, installment, log.annual.inc, dti, fico, days.with.cr.line, revol.bal, revol.util, inq.last.6mths, delinq.2yrs, pub.rec, not.fully.paid, purpose_all_other, purpose_credit_card, purpose_debt_consolidation, purpose_educational, purpose_home_improvement , purpose_major_purchase, purpose_small_business

4. Training

Train Test Split

```
In [53]: from sklearn.model_selection import train_test_split
```

```
In [77]: x= final_data.drop('not.fully.paid',axis=1)
         y= final_data['not.fully.paid']
         x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.3,random_state=101)
```

4.9 train test splitting

Training a Decision Tree Model

In [56]:
```
from sklearn.tree import DecisionTreeClassifier
```

In [58]:
```
dt = DecisionTreeClassifier()
```

In [61]:
```
dt.fit(x_train,y_train)
```

Out[61]:
```
DecisionTreeClassifier()
```

Training the Random Forest model

In [65]:
```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(n_estimators=600)
```

In [78]:
```
rfc.fit(x_train,y_train)
```

Out[78]:
```
RandomForestClassifier(n_estimators=600)
```

4.10 Training for decision and training random forest

5. RESULTS

Predictions and Evaluation of Decision Tree

In [63]:
```
predict = dt.predict(x_test)
```

In [62]:
```
from sklearn.metrics import confusion_matrix,classification_report
```

In [64]:
```
print(confusion_matrix(y_test,predict))
print('\n')
print(classification_report(y_test,predict))
```

```
[[2012  384]
 [ 374  104]]


              precision    recall  f1-score   support

           0       0.84      0.84      0.84      2396
           1       0.21      0.22      0.22       478

    accuracy                           0.74      2874
   macro avg       0.53      0.53      0.53      2874
weighted avg       0.74      0.74      0.74      2874
```

5.1 Prediction and evaluation of Decision Tree

Decision Tree model has predicted about accuracy of 74%.

Confusion matrix showed TN = 2012, FP = 384, FN = 374, TP = 104.

It had precision of 84% , recall 84%, f1-score 84%.

Predictions and Evaluation of Random Forest

In [79]:
```
predict = rfc.predict(x_test)
```

In [71]:
```
from sklearn.metrics import confusion_matrix, classification_report
```

In [80]:
```
print(confusion_matrix(y_test,predict))
print('\n')
print(classification_report(y_test,predict))
```

```
[[2424    7]
 [ 433   10]]


              precision    recall  f1-score   support

           0       0.85      1.00      0.92      2431
           1       0.59      0.02      0.04       443

    accuracy                           0.85      2874
   macro avg       0.72      0.51      0.48      2874
weighted avg       0.81      0.85      0.78      2874
```

5.2 Prediction and evaluation of Random Forest

Random Forest model has predicted about accuracy of 74%.

Confusion matrix showed TN = 2424, FP = 433, FN = 19, TP = 10.

It had precision of 85% , recall 100%, f1-score 92%.

6. CONCLUSION AND FUTURE ENHANCEMENT

The main purpose of this work is to classify and analyze the nature of the loan applicants. From a proper analysis of available data and constraints of the banking sector, it can be concluded that by keeping safety in mind that this modelling is much effective or highly efficient. This application is operating efficiently and fulfilling all the major requirements of Banker. Although the application is flexible with various systems and can be plugged effectively.

The module can be enhanced to cover below business scenarios

This work can be extended to higher level in future so the software could have some better changes to make it more reliable, secure, and accurate.

Thus, the system is trained with a present data sets which may be older in future so, it can also take part in new testing to be made such as to
pass new test cases.

7. REFERNCES

- Loan Prediction using Decision Tree and Random Forest Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi, Mr. Suresh Kumar, IRJET
- Loan Prediction System Using Decision Tree and Random Forest Algorithms ,Shubham Chaudhary, Vishal Baliyan, Yatharth Katheria, IJETIE
- A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models".
- A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.
- https://en.wikipedia.org/wiki/Exploratory_data_analysis

- https://www.experian.com/blogs/ask-experian/credit-education/score- basics/what-is-a-good-credit-score/