Loan Eligibility Prediction

INTRODUCTION

Overview

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 main aim of this use-case is to build a predictive model to predict if an applicant is able to repay the lending company or not.

Purpose

Main purpose of this project is to automate the process of checking whether a person is eligible for a loan or not.

LITERATURE SURVEY

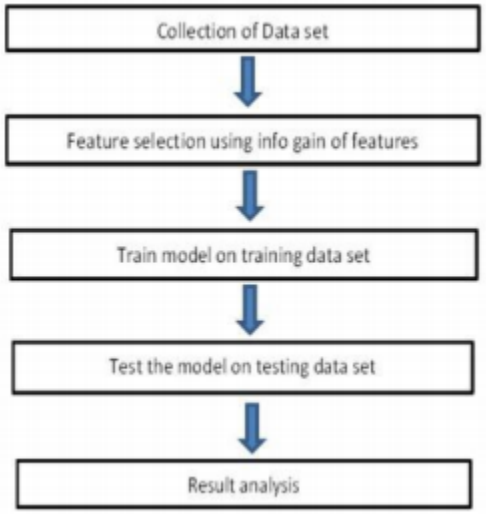
The aim of this project is to provide quick, immediate and easy way to choose the people who are eligible for loans. It can provide special advantages to the bank. The Loan Prediction System can can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight .A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application.

The data set for this project was taken from the website kaggle. Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals.

It provides the data set in the form of a csv file contains various attributes like loan ID, gender, married, dependents, education, loan amount etc.

Working of the Project

The data set obtained was preprocessed. This consists of data cleaning, data integration, data reduction and data transformation. Missing values in the data set were replaced with their means. Unnecessary fields like loan ID, gender, married were removed as they did not affect the final outcome. Logistic regression was used to determine if a person is eligible for loan or not.



Result

The output taken in Weka is as follows,

```
=== Run information ===

Scheme:       weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation:     Loan_predictor
Instances:    614
Attributes:   5
              Married
              Education
              Credit_History
              Property_Area
              Loan_Status
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                               Class
Variable                           Y
=====================================
Married=Yes                   0.5365
Education=Not Graduate       -0.3053
Credit_History='(-inf-0.5]'  -3.8734
Property_Area=Urban          -0.1776
Property_Area=Rural          -0.3859
Property_Area=Semiurban        0.505
Intercept                     1.0807


Odds Ratios...
                               Class
Variable                           Y
=====================================
Married=Yes                    1.71
Education=Not Graduate       0.7369
Credit_History='(-inf-0.5]'  0.0208
Property_Area=Urban          0.8373
Property_Area=Rural          0.6799
Property_Area=Semiurban      1.6569
```

```
Time taken to build model: 0.12 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         497               80.9446 %
Incorrectly Classified Instances       117               19.0554 %
Kappa statistic                          0.4808
Mean absolute error                      0.2963
Root mean squared error                  0.3875
Relative absolute error                 68.8819 %
Root relative squared error             83.5846 %
Total Number of Instances              614

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.983    0.573    0.790      0.983    0.876      0.541    0.762     0.835     Y
                 0.427    0.017    0.921      0.427    0.584      0.541    0.762     0.676     N
Weighted Avg.    0.809    0.399    0.831      0.809    0.785      0.541    0.762     0.785

=== Confusion Matrix ===

   a   b   <-- classified as
 415   7 |   a = Y
 110  82 |   b = N
```

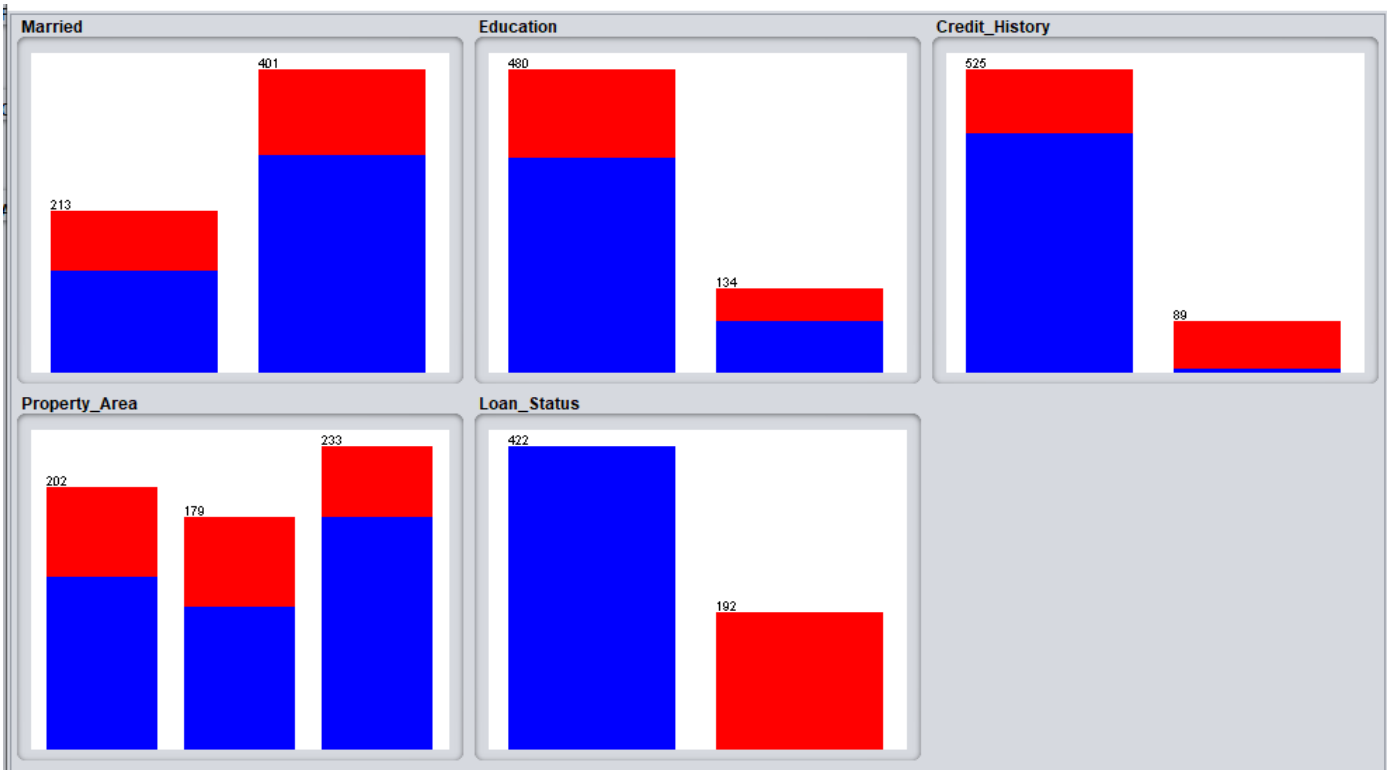
The output of the Java code which is written to emulate Logistic regression is as follows:

```
614
** Logistic Regression Evaluation with Datasets **

Correctly Classified Instances         497               80.9446 %
Incorrectly Classified Instances       117               19.0554 %
Kappa statistic                          0.4808
Mean absolute error                      0.2935
Root mean squared error                  0.3832
Relative absolute error                 68.2518 %
Root relative squared error             82.6556 %
Total Number of Instances              614

Confusion matrix:
[415.0, 7.0]
[110.0, 82.0]
Area under the curve
0.7809661334913112
[Correct, Incorrect, Kappa, Total cost, Average cost, KB relative, KB information, Correlation, Complexity 0, Complexit
Recall :
0.43
Precision:
0.92
F1 score:0.58
Accuracy:0.81
```

The logistic regression model built for the loan predictor has 81% accuracy with an error rate of approximately 19%.



Advantages

- One of the key advantages is that missing values are allowed
- It Has an accuracy rate of 81%
- The entire process is automated

Disadvantages

- One of the biggest disadvantage is that it is a supervised learning algorithm so it needs the inputs and outputs of a data set.
- Outliers can become a problem

Conclusion

From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems.
As the training data improves, the efficiency of the algorithm will also improve, thus giving out better results.

Bibilography

1. https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset
2. https://www.tutorialspoint.com/weka/weka_preprocessing_data.htm
3. https://towardsdatascience.com/ml-basics-loan-prediction-d695ba7f31f6
4. https://www.knowledgehut.com/blog/data-science/logistic-regression-for-machine-learning