Data Analysis on Loan Prediction

**Abstract**

Loan prediction is a very common real-life problem that each retail bank faces at

least once in its lifetime. If done correctly, it can save a lot of man hours at the end

of a retail bank. If Company wants to automate the loan eligibility process (real

time) based on customer detail provided while filling online application form.

These details are Gender, Marital Status, Education, Number of Dependents,

Income, Loan Amount, Credit History and others. To automate this process, they

have given a problem to identify the customers segments, those are eligible for

loan amount so that they can specifically target these customers. we have to predict

whether a loan would be approved or not. In a classification problem, we have to

predict discrete values based on a given set of independent variable(s)

**Introduction**

Distribution of the loans is the core business part of almost every banks. The main portion the bank’s assets is directly came from the profit earned from the loans distributed by the banks. The prime objective in banking environment is to invest their assets in safe hands where it is. Today many banks/financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique. The disadvantage of this model is that it emphasize different weights to each factor but in real life sometime loan can be approved on the basis of single strong factor only, which is not possible through this system. Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. 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. This Paper is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities.

**Data sets**

The training data set is now supplied to machine learning model, on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set.

**MACHINE LEARNING METHODS:**

Six machine learning classification models have been used for prediction of android applications .The models are available in R open source software. R is licensed under GNU GPL. The brief details of each model is described below.

**Decision Trees (C5.0)**

The basic algorithm of decision tree [7] requires all attributes or features should be discretized. Feature selection is based on greatest information gain of features. The knowledge depicted in decision tree can represented in the form of IF-THEN rules. This model is an extension of C4.5 classification algorithms described by Quinlan

**Random Forest (RF)**

Random forests [8] are a group learning system for characterization (and relapse) that work by building a large number of Decision trees at preparing time and yielding the class that is the mode of the classes yield by individual trees.

**Support Vector Machine (SVM)**:

Support vector machines are administered learning models that uses association r learning algorithm which analyze features and identified pattern knowledge, utilized for application classification. SVM can productively perform a regression utilizing the kernel trick, verifiably mapping their inputs into highdimensional feature spaces

**Linear Models (LM):**

The Linear Model is numerically indistinguishable to a various regression analysis yet burdens its suitability for both different qualitative and numerous quantitative variables.

**Neural Network (Nnet)**:

Neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs, to find patterns in data, or to capture the statistical structure in an unknown joint probability distribution between observed variables.

**Adaboost (ADB):**

Adaboost short for " Adaptive Boosting ". It is delicate to noisy information data and outliers. It is different from neural systems and SVM because Adaboost preparing methodology chooses just those peculiarities known to enhance the divining power of the model, decreasing dimensionality and conceivably enhancing execution time as potentially features don't have to be processed.

**Project overview**

Loan prediction is a very common real-life problem that each retail bank faces at

least once in its lifetime. If done correctly, it can save a lot of man hours at the end

of a retail bank. If Company wants to automate the loan eligibility process (real

time) based on customer detail provided while filling online application form.

These details are Gender, Marital Status, Education, Number of Dependents,

Income, Loan Amount, Credit History and others. To automate this process, they

have given a problem to identify the customers segments, those are eligible for

loan amount so that they can specifically target these customers. we have to predict

whether a loan would be approved or not. In a classification problem, we have to

predict discrete values based on a given set of independent variable(s)

1. Hypothesis Generation

2. Getting the system ready and loading the data

3. Understanding the data

4. Exploratory Data Analysis (EDA)

○ Univariate Analysis

○ Bivariate Analysis

5. Missing value and outlier treatment

6. Evaluation Metrics for classification problems

7. Model Building

**Project description**

Binary Classification: Loan Granting

This experiment creates a statistical model to predict if a customer will default or fully pay off a loan.

Data

A synthetic data set based on real data was created for the competition. The data set included the following columns.

Loan ID

Customer ID

Loan Status

Current Loan Amount

Term

Credit Score

Years in current job

Home Ownership

Annual Income

Purpose

Monthly Debt

Years of Credit History

Months since last delinquent

Number of Open Accounts

Number of Credit Problems

Current Credit Balance

Bankruptcies

Tax Liens

For this experiment, the Loan Status serves as the label, or attribute to predict. The rest of the columns excluding Customer ID will be used to predict the outcome of the Loan Status for each customer.

The data set was available as it was stored in a web URL in csv form by Azure.

**Problem definition**

**Existing system**

Loan prediction is a very common real-life problem that each retail bank faces at

least once in its lifetime. If done correctly, it can save a lot of man hours at the end

of a retail bank. If Company wants to automate the loan eligibility process (real

time) based on customer detail provided while filling online application form.

These details are Gender, Marital Status, Education, Number of Dependents,

Income, Loan Amount, Credit History and others. To automate this process, they

have given a problem to identify the customers segments, those are eligible for

loan amount so that they can specifically target these customers. we have to predict

whether a loan would be approved or not. In a classification problem, we have to

predict discrete values based on a given set of independent variable(s)

**Proposed system**

In Machine Learning, we are using semi-automated extraction of knowledge

of data for identifying whether a loan would be approved or not.

Classification is a supervised learning in which the response is categorical

that is its values are in finite unordered set. To simply the problem of

classification, scikit learn tools has been used.

**ADVANTAGES OF PROPOSED SYSTEM:**

The company need not has to maintain a ground team to validate and verify

the customer records. They can easily check whether the loan has to be

approved or not by this prediction model.

**SYSTEM REQUIREMENTS:**

Anaconda Software to be installed.

● Microsoft Windows 10/8/7/Vista/2003/XP.

● For Anaconda—Minimum 3 GB disk space to download and install.

● 4 GB RAM recommended.

● pandas = 0.20.3

● seaborn = 1.0.0

● sklearn = 0.19.1

**Data flow diagrams**

**code**

**Data Preparation**

First the Loan Status columns were altered slightly, making the "Charged Off" status 1 and the "Fully Paid" status 0. The next R script individually address most columns. Many of the numeric columns had extra non numerical characters in them suck as "$". With the excepction of "." non numerical characters were removed from numerical columns. In addition, the columns meant to be numerical were all typecasted to numeric. To add, the "Years in current job" column was condensed. Instead of indivudal years ranging from 1-9, the years of experience were placed into bins. For example, people who have 1-5 years were all placed into one bin. Finally, some credit scores were over 1000. After many iterations of testing, the credit scores of that nature were lowered to 650 as they all consisted of people who were Charged Off.

Next, extreme outliers were clipped and replaced with the median. Some columns contained data entries with extremely high values such as 99999999. These values are most likely not valid and were put into place to signify that the value is missing.

Duplicate customer ID's were removed.

In the next execute R script, feature engineering was used to create the dti column (debt to income ratio) by dividing annual salary by the amount of debt.

The next few modules make the string columns into categorical columns and the column Purpose has been condensed into two main groups (debt, other) from the several that previously existed.

From here the model is created. The label (Loan Status) is split off from the other columns before creating the model. A boosted decision tree is used to build the model. In addition, the Tune Model Parameters module is used to find the most efficient parameters to use for the model. A random sweep was used, and module chooses to maximize AUC (Area Under the Curve). The best parameters from the Tune Model Parameters is used as the model to train the data.

The data is split into training and test datasets using the Split Data module. The training data is used to train the model and test data is used in the score model to evaluate performance. The confusion matrix can be seen in the Evaluate Model module

**Final Thoughts:**

Creating a model to accurately predict the hidden dataset for this competition was very tough for several reasons. Many of the participants including myself tried dealing with the imbalanced data. About 75% of the data set consisted of people who fully paid off their loan, and the remainder were people who were charged off. This posed many challenges since the models became biased towards people who fully paid off their loans. In addition, the distributions of credit score, current loan amount, and many other columns were very similar between customers who fully paid off their loans and people who did not. If there was some more variability between the two types of customers this may have been easier. Overall this was a fun model to create and hope to tackle so more fun data science challenges in the future.

**Output screens**

**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 application is working properly and meeting to all Banker requirements. There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system, So in the near future the so –called software could be made more secure, reliable and dynamic weight adjustment .In near future this module of prediction can be integrate with the module of automated processing system. the system is trained on old training dataset in future software can be made such that new testing date should also take part in training data after some fix time.