LOAN PREDICTION

Omkar Keluskar: ork216 Rushabh Patel: rgp296

Basic Understanding

In current economy, loaning someone money is a complicated business. There are various factors that are needed to be considered to calculate the credibility of the applicant. Another important point to note – is that the applicant capable enough to repay the loan? As banks love to thrive on the interests of customers, but they love it, even more, when customers don't default on loan payments. So, businesses as such would love to have such analysis before taking the decision - whether or not to sanction the loan.

Taking this problem into consideration, data science can aid to predict such outcome and approve the loan based on the historical data we have from businesses. In this way, businesses would know beforehand whether to lend the applicant money or not based on the data on his file compared to analysis from the data science models.

Estimating the probability that an individual would default on their loan, is useful for banks to decide whether to sanction a loan to the individual or not. We introduce an effective prediction technique that helps the banker to predict the credit risk for customers who have applied for loan. A prototype is described in the paper which can be used by the organizations for making the correct or right decision for approve or reject the request for loan of the customers.

By mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not.

Data Understanding

There are two datasets available: train dataset and test dataset. The train data set is now supplied to the model, training of the model is dependent on the date set. The data to the test data set is every new applicant details filled at the time of application. After the operation of testing, model predict whether the new applicant is a perfect case for sanctioning of the loan or not based upon the inference of the training data sets.

Source: https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

Our data set for the problem looks like:

```
Variable          :     Description
Loan_ID           :     Unique Loan ID
Gender            :     Male/ Female
Married           :     Applicant married (Y/N)
Dependents        :     Number of dependents
Education         :     Applicant Education (Graduate/ Undergraduate)
Self Employed     :     Self-employed (Y/N)
ApplicantIncome   :     Applicant income
CoapplicantIncome:     Coapplicant income
LoanAmount        :     Loan amount in thousands
Loan_Amount_Term  :     Term of loan in months
Credit_History    :     credit history meets guidelines
Property_Area     :     Urban/ Semi Urban/ Rural
Loan_Status       :     Loan approved (Y/N)
```

We got the dataset for the problem through this website which has 614 entries. Here are a few inferences, we drew by looking at the output of describe() function show in table 1.

```
In [4]: df.describe()  #Get summary of numerical variables
```

Out[4]:

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

Table 1: Output of describe function

1. LoanAmount has (614 – 592) 22 missing values.
2. Loan_Amount_Term has (614 – 600) 14 missing values.
3. Credit_History has (614 – 564) 50 missing values.
4. We can also look that about 84% applicants have a credit_history. How? The mean of Credit_History field is 0.84 (Remember, Credit_History has value 1 for those who have a credit history and 0 otherwise).
5. The ApplicantIncome distribution seems to be in line with expectation. Same with CoapplicantIncome.

Methodology

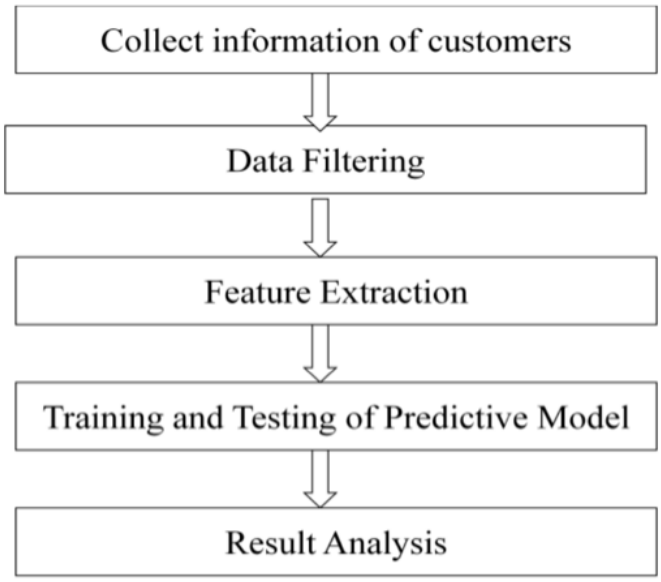

Figure 1: Methodology

Initially the data from costumers is collected, then the process of data filtering takes place where missing values are removed. In the third step, the feature importance is carried out. It makes the model accurate and more efficient. In the fourth step, the data mining algorithms is applied on the train and test dataset with parameters set to default. Finally, evaluation is done on H, Gini, AUC, AUCH, Accuracy etc.[1]

Data Preparation

The data set include 13 attributes such as Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. The data sets contain 615 records which are unfiltered data. The dataset is filtered by removing the missing records. So the filtered Train dataset file contains 479 records. Table 2 show the sample of the filtered dataset.

These 13 attributes are of different types i.e. some are string, some are character and some are integers. For prediction, we need to bring all the attributes to the same integer type. For bringing it to the same type we need to map each of the attributes. For example, Yes is mapped to 1 and No is mapped to 0. After mapping the dataset, the dataset is totally transformed into a new dataset which is then used for modeling. Table 3. shows a newly transformed dataset.

A	B	C	D	E	F	G	H	I	J	K	L	M
Loan_ID	Gender	Married	Dependents	Education	Self_Employ	ApplicantInc	CoapplicantI	LoanAmount	Loan_Amour	Credit_Histo	Property_Ar	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduat	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduat	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001027	Male	Yes	2	Graduate	No	2500	1840	109	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
LP001034	Male	No	1	Not Graduat	No	3596	0	100	240		Urban	Y
LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
LP001038	Male	Yes	0	Not Graduat	No	4887	0	133	360	1	Rural	N
LP001041	Male	Yes	0	Graduate	No	2600	3500	115	360	1	Urban	Y
LP001043	Male	Yes	0	Not Graduat	No	7660	0	104	360	0	Urban	N
LP001046	Male	Yes	1	Graduate	No	5955	5625	315	360	1	Urban	Y
LP001047	Male	Yes	0	Not Graduat	No	2600	1911	116	360	0	Semiurban	N

Table 2: Sample Filtered Dataset

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	P
LP001002	1.0	0.0	0.0	1	0.0	5849	0.0	NaN	360.0	1.0	1
LP001003	1.0	1.0	1.0	1	0.0	4583	1508.0	128.0	360.0	1.0	3
LP001005	1.0	1.0	0.0	1	NaN	3000	0.0	66.0	360.0	1.0	1
LP001006	1.0	1.0	0.0	0	0.0	2583	2358.0	120.0	360.0	1.0	1
LP001008	1.0	0.0	0.0	1	0.0	6000	0.0	141.0	360.0	1.0	1
LP001011	1.0	1.0	2.0	1	NaN	5417	4196.0	267.0	360.0	1.0	1
LP001013	1.0	1.0	0.0	0	0.0	2333	1516.0	95.0	360.0	1.0	1
LP001014	1.0	1.0	3.0	1	0.0	3036	2504.0	158.0	360.0	0.0	2
LP001018	1.0	1.0	2.0	1	0.0	4006	1526.0	168.0	360.0	1.0	1
LP001020	1.0	1.0	1.0	1	0.0	12841	10968.0	349.0	360.0	1.0	2

Table 3: Transformed Dataset

The correlation between each feature is shown in Table 4.

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amo
Gender	1.000000	-0.369612	-0.175970	0.049258	NaN	-0.053989	-0.083946	-0.106947	0.075117
Married	-0.369612	1.000000	0.343417	-0.014223	NaN	0.051332	0.077770	0.149519	-0.103810
Dependents	-0.175970	0.343417	1.000000	-0.059161	NaN	0.118679	0.027259	0.163997	-0.100484
Education	0.049258	-0.014223	-0.059161	1.000000	NaN	0.140760	0.062290	0.171133	0.078784
Self_Employed	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ApplicantIncome	-0.053989	0.051332	0.118679	0.140760	NaN	1.000000	-0.116605	0.570909	-0.045306
CoapplicantIncome	-0.083946	0.077770	0.027259	0.062290	NaN	-0.116605	1.000000	0.188619	-0.059878
LoanAmount	-0.106947	0.149519	0.163997	0.171133	NaN	0.570909	0.188619	1.000000	0.039447
Loan_Amount_Term	0.075117	-0.103810	-0.100484	0.078784	NaN	-0.045306	-0.059878	0.039447	1.000000
Credit_History	-0.016337	0.004381	-0.050082	0.081822	NaN	-0.014715	-0.002056	-0.008433	0.001470
Property_Area	-0.024556	-0.002918	-0.006828	-0.065243	NaN	0.009500	-0.010522	0.045792	0.078748

Table 4: Correlation between each features

The problem is to predict whether the loan will be sanction or not. The question here arises that what should be our target attribute where we can display our prediction after the modelling process. After prediction, the

prediction algorithm will run on the test dataset. Once the dataset is processed completely, a new attribute will be created named as Loan_Status whose values will be Yes if the loan will be sanction else no otherwise. Table 5 shows the Loan_Status attribute in the test dataset.

A	B	C	D	E	F	G	H	I	J	K	L	M
Loan_ID	Gender	Married	Dependents	Education	Self_Employ	ApplicantInc	CoapplicantI	LoanAmount	Loan_Amou	Credit_Histo	Property_Ar	Loan_Status
LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban	??
LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban	??
LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban	??
LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban	??
LP001051	Male	No	0	Not Graduat	No	3276	0	78	360	1	Urban	??
LP001054	Male	Yes	0	Not Graduat	Yes	2165	3422	152	360	1	Urban	??
LP001055	Female	No	1	Not Graduat	No	2226	0	59	360	1	Semiurban	??
LP001056	Male	Yes	2	Not Graduat	No	3881	0	147	360	0	Rural	??
LP001059	Male	Yes	2	Graduate	NO	13633	0	280	240	1	Urban	??
LP001067	Male	No	0	Not Graduat	No	2400	2400	123	360	1	Semiurban	??
LP001078	Male	No	0	Not Graduat	No	3091	0	90	360	1	Urban	??
LP001082	Male	Yes	1	Graduate	No	2185	1516	162	360	1	Semiurban	??
LP001083	Male	No	3+	Graduate	No	4166	0	40	180		Urban	??
LP001094	Male	Yes	2	Graduate	No	12173	0	166	360	0	Semiurban	??
LP001096	Female	No	0	Graduate	No	4666	0	124	360	1	Semiurban	??
LP001099	Male	No	1	Graduate	No	5667	0	131	360	1	Urban	??
LP001105	Male	Yes	2	Graduate	No	4583	2916	200	360	1	Urban	??
LP001107	Male	Yes	3+	Graduate	No	3786	333	126	360	1	Semiurban	??
LP001108	Male	Yes	0	Graduate	No	9226	7916	300	360	1	Urban	??

Table 5: Loan_Status that needs to be predicted.

Modeling

Following are the models have been used for prediction of this dataset. 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: The basic algorithm of decision tree 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. [2]

Random Forest: Random forests 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. [3]

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 high dimensional feature spaces. [4]

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. [5]

Table 6 shows the parameter settings of the techniques used.

Models	Method Used	Packages	Tuning Paramters
Decision Trees	Rpart	Rpart	Min Split = 20, Max Depth = 30, Min Bucket = 7
Random Forest	Random Forestb	Random Forest	Number of tree = 500, Number of variables=8
SVM	Ksvm	kernlab	Kernel Radial Basis
Linear Model	Multinom	Car, nnet	Multinomial

Table 6: Parameter setting for machine learning models

Evaluation

Models are evaluated based on the performance of prediction. The measure that are used for evaluation are [6]:

Accuracy: Accuracy is determined on the basis of how data is collected, and judged on basis of comparing of several parameters. True positive (TP) describes the amount of predictions which are positive, the actual value being positive. Similar in the case of true negative (TN). The accuracy is calculated as:

$$\text{Accuracy} = (TP+TN\ /\text{Total Data}) * 100$$

AUC: AUC or Area under Curve is a metric for binary calculation. It's a probably the second most popular parameter after Accuracy. It computes the area under the curve of a given performance measure. Its value lies between 0.5 and 1. It depicts the quality of models used for classification problems.

Gini Coefficient: The disparity of a distribution is calculated by using Gini coefficient and its values lies between 0 and 1. These are mainly used for model comparison.

$$\text{Gini}=2\text{AUC}-1$$

ROC: Curve A receiver operating characteristic (ROC) curve is used to classify problem of binary type. The function is included in pROC package.

K-S chart: K-S or Kolmogorov-Smirnov chart measures performance of classification models. More accurately, K-S is a measure of the degree of separation between the positive and negative distributions.

MER: MER metrics represents the Minimum Error Rate. Here threshold value act as a free parameter

MWL: MWL metrics represents the Minimum Cost- Weighted Error Rate. It is related to the KS statistics. Cost guides the threshold value in this measure.

K-Fold Cross Validation: Cross-validation is a technique to evaluate predictive models by partitioning the data into a training set to train the model, and a test set to evaluate it. In k-fold cross validation, the original sample is arbitrarily divided into k equal size subsamples. The advantage of this method is that it matters less how the data gets divided.

Results

In it we calculate the results of prediction of all models on the training dataset. The performance is calculated on basis of its Accuracy, H, Gini, AUC, AUCH, KS, MER, MWL, and ROC. The results are shown in two sections. First part of this section displays accuracy rate for each of model (Table 7). Second part of this section shows the cross validation of predicted values of top two individual models having high accuracy and the ensemble model (Figure 2).

Models	Accuracy	H	Gini	AUC	AUCH	KS	MER	MWL	ROC
Decision Trees	78.23	0.28	0.56	0.78	0.78	0.55	0.25	0.2	0.79
Random Forest	81.52	0.36	0.66	0.82	0.82	0.63	0.23	0.15	0.82
SVM	81.52	0.36	0.66	0.82	0.82	0.63	0.23	0.15	0.84
Linear Model	79.92	0.32	0.63	0.82	0.82	0.63	0.22	0.14	0.82

Table 7: Training Dataset Results

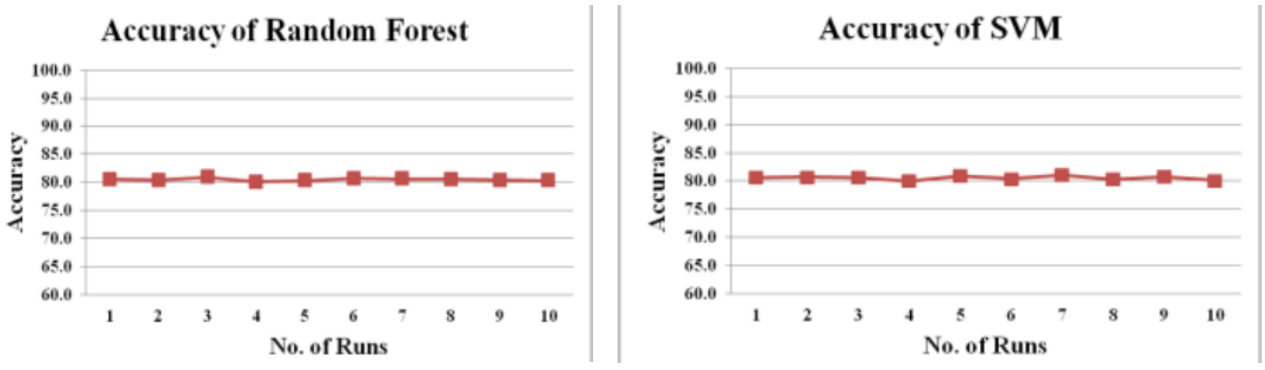

Figure 2: Cross Validation graph

Deployment

In the proposed work, four models are constructed which have nine properties that are used to predict the credit risk of costumer's who have applied for loan. Under different training algorithms, this models for loan predications by using several parameters like Accuracy, Gini, Auc, Roc, etc. to do the comparison.

The main purpose of this paper is to test the accuracy of model to predict the loan of costumers. After completing a proper analysis of all the key points and constraints, we came in to a conclusion that this application is a highly efficient component. This application can work properly by meeting all the requirements that are defined in the problem statement.

One of the key point to remember while deploying this application, each and every feature are evaluated for prediction. So whenever a new customer enters his/her information, data of all the attributes must be specified by the customer. So there should be a constraint placed by the bank if any of the value founds to missing.

This models may go through from over fitting problem. To overcome this over fitting problem, all models is set to run on their default parameters and the data is distributed among training and testing set are 70% and 30% correspondingly for all the models.

There have been numbers 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.

Analysis of this project is concerned with bank loan. Our analysis gives better results. Similarly, this analysis is useful for other problems related with medical studies like predicting cancer survivability.

For this project we have data on 12 attributes listed as above. We developed model only with available variables but if add other important variables then we expect that our models give better results.

So in this paper we developed a model which helps to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets on the basis of their records.

Reference

[1] http://www.iosrjournals.org/iosr-jce/papers/Vol18-issue3/Version-1/O1803017981.pdf

[2] J.R. Quinlan. Induction of decision trees. MachinelearningnSpringer, 1(1):81–106, 1086.

[3] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. R News(http://CRAN.R-project.org/doc/Rnews/), 2(3):9–22, 2002.

[4] S.S. Keerthi and E.G. Gilbert. Convergence of a generalize SMO algorithm for SVM classifier design. Machine Learning, Springer, 46(1):351–360, 2002.

[5]. J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077.

[6] http://www.ijarcce.com/upload/2016/march-16/IJARCCE%20128.pdf

Appendix

Work was equally distributed among the members. Both the members sat together and completed this paper.