Predict Loan Default for Dream Housing Finance company using Bayes Net, Naive Bayes and Random Forest Algorithm

BUSS-5820-0

Le Ho Thao Nguyen

20194224

Abstract

Dream Housing Finance company has presence across all urban, semi urban and rural areas and deals with all kinds of loans. Customers first apply for a home loan after that company validates the customer eligibility for the loan. This report will explore factors in consumer loan default behaviour through demographic variables, historical loan status or economic variables to understand the loan default behaviour of customers of Dream housing finance company and build a predictive model to predict whether a loan to that customer should be approved or not. This report will present an experimental study on loan applicant customer data of Dream House finance companies and will be using Bayes Net, Naive Bayes and Random Forest algorithm. First, we will briefly discuss the literature review. Second, we will conduct an exploratory data analysis to gain a better understanding of the loan dataset, details of the data set and experimental setup. Third, model evaluation will be presented and compared between each other. Finally, discussion and conclusions are presented.

Keywords
Credit risk, Random Forest, Bayes Net, Naive Bayes, precision, recall, F1-measure, accuracy

Introduction

Credit risk is a major concern for any financial institution that lends money to qualified borrowers and expects them to repay the loan amount in instalments by a certain period of time. For large commercial banks in Canada, such as RBC, Scotia Bank, TD Bank, loans contribute a lot in their total assets and are used to fund different purposes. The rationale of this report is based on the fact as a current student of Cape Breton University having studied and inspired by Capstone Project subject, there was an interest as to explore how Predictive Analytics can help Dream Housing Finance company to predict loan default of its customers.

Related work

There are many studies conducted using different classifiers algorithms in the financial and banking sector. This section would report shortly some of the algorithms used in credit risk management as below.

Barney et al compared and analyzed the performance of logistic regression and neural networks to classify the farmer's loan into good or bad. In his research, he noted that neural networks perform better than logistic regression in this regard [1].

Glorfeld and Hardgrave (2001) implemented a really useful neural network model in predicting credit risk from commercial loans [2]. The model s performance is high (75% correctly classify loan applicants) when using neural network algorithms.

Jozef Zurada and Martin Zurada examined how useful neural networks can be applied in commercial loan classification problems. The developed neural network model was able to classify 75% of the applicants [3].

Research Questions

Would applicants who have repaid their previous debts should have higher chances of loan approval?

Would getting a loan for less time period and less amount should have higher chances of approval?

Would middle-age people have better chance of paying back the loan?

Would loan approval likely to occur in high-income applicant?

How to build a loan prediction model?

How can I evaluate my prediction model?

Methodology

This is a Binary supervised learning classification problem. We will construct a Classification model based on Training data using Bayes Net, Naive Bayes and Random Forest to predict Loan Status, either Yes or No, of forthcoming credit customers. The input to the model is the customer behaviours collected and mentioned in metadata. Based on the output from the classification, a decision on whether to approve or reject the customer request can be made.

The metric for model evaluation that we use are Accuracy, Precision, Recall and F1- score, which are presented as below:

$$\text{Accuracy} = \frac{(TruePositve + TrueNegative)}{(TruePositve + FalsePositive + TrueNegative + FalseNegative)}$$

$$\text{Precision} = \frac{TruePositive}{(TruePositve + FalsePositive)}$$

Recall = $\dfrac{(TruePositive)}{(TruePositive + FalseNegative)}$

F-1 score = $\dfrac{(2 * Precision * Recall)}{(Precision + Recall)}$

Naive Bayes Algorithm

Naive Bayes is simple technique based on the Bayes probability theory which assumes conditional independence to classify data. Naive Bayes develops the model whose predictor variables should be independent [4].

The Naive Bayes Classifier is inspired by Bayes Theorem which states the following equation [8]

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad [8]$$

This can be viewed as

$$Posteriori = \frac{Likelihood\ x\ Prior}{Evidence}$$

Pros and Cons of Naive Bayes [8]

Pros

- The concept is quite easy to understand
- It is easy to implement and performs well in classification problem
- It works well with categorical input variables

Cons

- It requires initial knowledge of many probabilities, involving significant computational cost
- It can encounter the problem when there is a category in the test set which is not in the training set
- The probability estimates are not the most trustworthy from this algorithm
- Naive Bayes holds strong assumptions that predictor variables should be independent and this is not usually the case in real-world problem.

Bayes Net Algorithm

This algorithm also depends on the Bayes theorem of probability, but it differs from Naive Bayes algorithm in the sense that it will allow users to specify which attributes are conditionally independent. The Bayesian Networks model is built after calculating conditional probability to all nodes and forming a graph as an output [4].

The formula is given as below [9]

$$P(X1, X2, \ldots Xn) = P(X1) * P(X2|X1) * P(X3|X2, X1) \ldots P(Xn|X1, X2, \ldots, Xn)$$ [9]

We can see the probability of child nodes depends on its closet parents [9]

Pros and Cons of Bayes Network [10]

Pros

- Bayes Networks are understandable for human and computer.
- Bayesian Network can utilize domain knowledge to determine whether or not to include a certain variable
- Bayesian Networks are more easily scalable than other machine learning methods because every time a new piece of information is added, Bayesian Network would require only the addition of a small number of probabilities and edges in the graph.
- The model's output is a probability, so if we want to change the output of the network from a probability to a predefined output, we will only need to set a threshold.

Cons

- There is no common method to construct a network from data
- Bayes Net requires a large amount of efforts and tend to uncover the casual relationships that are recognized by the person programming it.

Random Forest Algorithm

Random Forest algorithm creates multiple Classification and Regression Trees (CART) based on random samples conducted with replacement of the dataset and then combines predictions from many CART-models, and based on the majority voting for classification problem or averaging for regression problem to make the final prediction [5]. Random Forest usually has better predictive power than CART because of low variance but it is not usually as easy to interpret as CART [6].

Random Forest uses below measures to find the best split which prefers nodes with purer class distribution (lowest entropy and highest information gain as below) [11]

Impurity	Task	Formula	Description
Gini impurity	Classification	$\sum_{i=1}^{C} -f_i(1-f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Entropy	Classification	$\sum_{i=1}^{C} -f_i \log(f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.

Pros & Cons of Random Forest [12]

Pros:

- It has strong ability to identify outliers
- It works well with non-linear data.
- It reduces overfitting compared to decision tree
- It runs efficiently on a large dataset.
- It has better accuracy than other classification algorithms.

Cons:

- Random forests are found to be biased while dealing with categorical variables.
- It can lead to slow training process because it requires a lots of trees to build
- It is not suitable for linear methods

Experiment

Dataset

Variables	Data type	Description
Loan_ID	Nominal	Loan ID of customer
Gender	Nominal	Male/ Female
Married	Nominal	Applicant married (yes/no)
Dependent	Nominal	Number of dependents
Education	Nominal	Graduate/ Undergraduate
Self_Employed	Nominal	Yes/No
ApplicantIncome	Numeric	Applicant Income
CoApplicantIncome	Numeric	CoApplicant Income
LoanAmount	Numeric	Loan Amount in thousands
Loan_Amount_Term	Numeric	Loan Amount duration in months
Credit_History	Numeric	Credit History meet guideline
Property_Area	Nominal	Area where property area is located
Loan_Status	Nominal	Loan Approved (Yes/No)

The dataset is taken from Kaggle [7] and contains demographic and personal information of loan applicants including age, gender, marital status, employment status, income, family dependents and their loan details (amount and term) with loan status. There are 614 rows and 13 columns in the dataset as described below. Loan Status is the target variable and remaining are the predictor variables.

Data Cleaning

First we check for missing values and replacing them with substituted values

For missing values in Integer and Float type, we are replacing with their median

For missing values in Object type, we are replacing by their mode

The result is 13 records from Gender, 3 records from Married, 15 records from Dependents, 32 records from Self-employed are replaced with their mode. Besides, 22 records from Loan Amount, 14 records from Loan_amount_term, 50 records from Credit History are replaced by their median and no missing values were found as below.

```
Loan_ID              0
Gender               0
Married              0
Dependents           0
Education            0
Self_Employed        0
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount           0
Loan_Amount_Term     0
Credit_History       0
Property_Area        0
Loan_Status          0
dtype: int64
```

We add one more variable "Total Income" which is created by addition of Applicant Income and CoApplicant Income. We also convert categorical into numeric variables before feeding into model.

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term
0	LP001002	1	0	0	0	0	5849	0.0	128.0	360.0
1	LP001003	1	1	1	0	0	4583	1508.0	128.0	360.0
2	LP001005	1	1	0	0	1	3000	0.0	66.0	360.0
3	LP001006	1	1	0	1	0	2583	2358.0	120.0	360.0
4	LP001008	1	0	0	0	0	6000	0.0	141.0	360.0

In order to avoid multicollinearity, we decide to remove CoApplicant_Income variable after adding new feature Total_Income. Besides, we also drop unnecessary columns such as Loan_ID

Since we visualize Loan_amount, Applicant Income and CoApplicant Income variables by distribution plot and box plot, we noted that these variables have positive skewness and there are outliers. So we decided to normalize and scale these with log transformation.

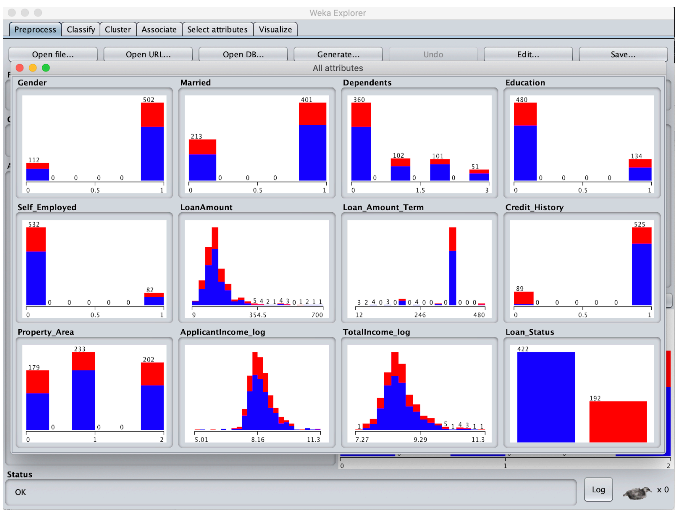
Exploratory analysis

We visualize the dataset using Weka and below is analysis done from the stacked histogram

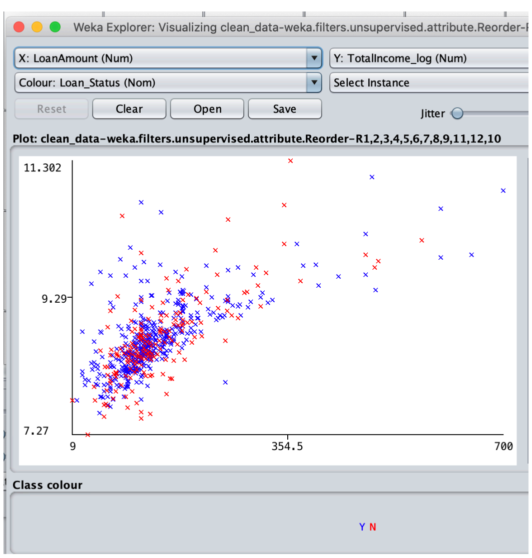
From the dataset visualization below, we can infer that 80% of the applicants in the dataset are male, around 65% of the applicants in the dataset are married, about 15% applicants in the dataset are self-employed, about 85% applicants have paid their debts.

This is an imbalance dataset with about two-thirds of applicants (422 records) is creditworthy and one-third (192 records) is not.

Loan Amount has positive skewness, minimum at 9, maximum at 700, mean at 145.466 and standard deviation at 84.181

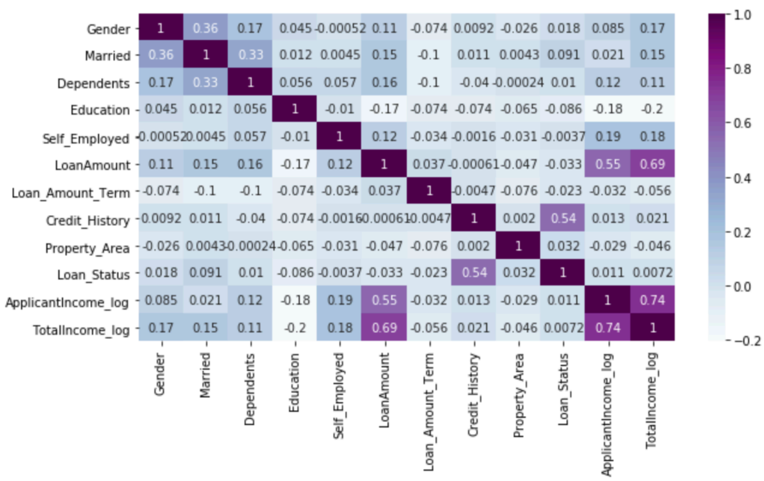


Besides, we visualize the relationship between Loan Amount and Total Income by Loan Status by scatterplot as below. We can see that Loan Amount and Total Income has positive relationship



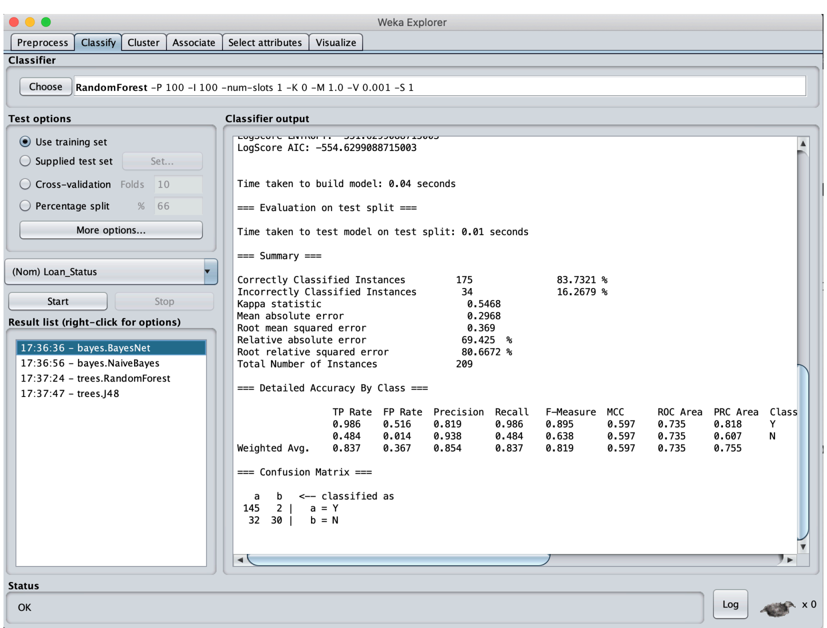
Model implementation

There are several popular classification algorithms such as Naive Bayes Classifier, Neural Network Classifier, Decision Tree Classifier etc. In this paper, we use Bayes Net, Naive Bayes and Random Forest algorithms to build a model for predicting loan default. The attributes taken to build those three models are presented as below since we plot correlation heat-map and find out that these variables have correlation to target variable. These include: Gender, Married, Dependents, Education, Self-Employed, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area, ApplicantIncome_log, TotalIncome_log and Loan_Status

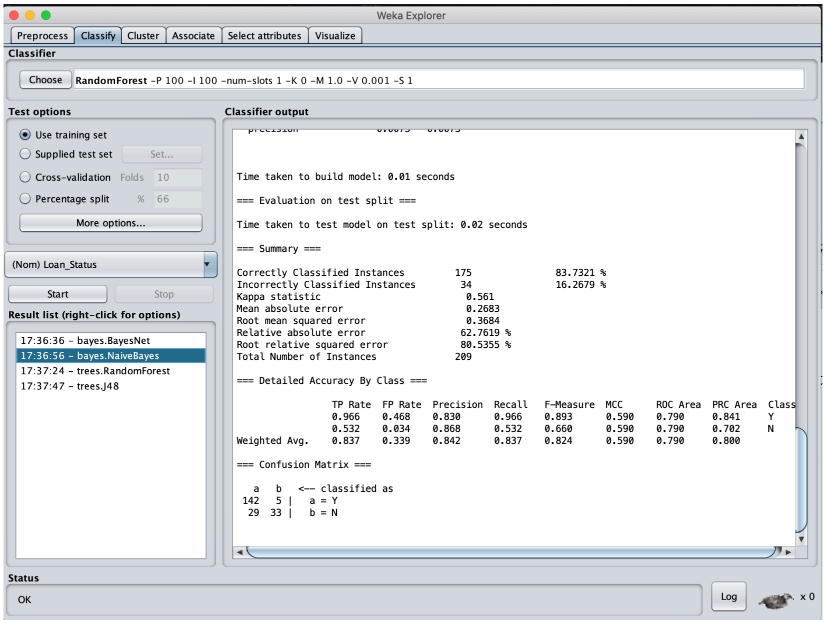


We divide the original data set into two groups, training set which represent 66% from all data and cross-validation set which represent 34% of the data set.

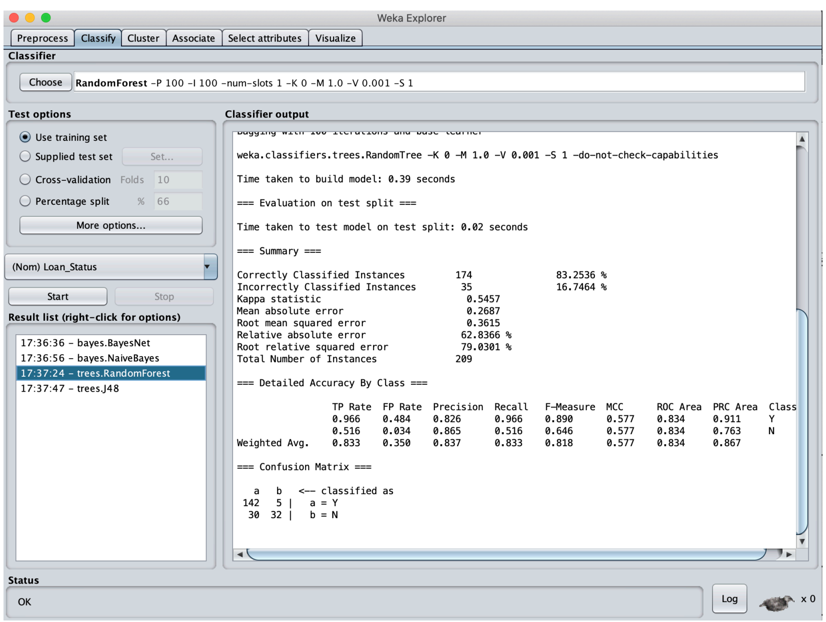
Bayes Net Model -The model built from using Bayes Net algorithm has been presented as below



Naives Bayes Model -The model built from using Naives Bayes algorithm has been presented



Random Forest Model - The model built from using Random Forest algorithm has been presented as below.



Model result and Discussion

Technique	Accuracy	Precision (Class = No)	Recall (Class = No)	F1-score (Class = No)	Mean absolute error	Executed time
BayesNet	83.7%	93.8%	48.4%	63.8%	0.2968	0.01 secs
Naive Bayes	83.7%	86.6%	53.2%	66%	0.2683	0.04 secs
Random Forest	83.2%	86.5%	51.6%	64.6%	0.2687	0.39 secs

After applying classification's algorithms including Bayes Net, Naive Bayes and Random Forest algorithms, the results from experiments are presented in the table above. We compare the correctly classified instance percent and note that the best algorithm for loan classification for this dataset is Naive Bayes. The reasons are high accuracy and low mean absolute error as shown in the result in the table. Besides, it also performs best in F1-Measure compared to two remaining techniques as shown in the confusion matrix of the three algorithms, even though the precision and recall of the prediction model based on above models are only about 50% which indicates that the models would not have strong ability to generalize in the future. However, Naive Bayes model would still be helpful the most for decision makers from Dream House finance company to accept or reject loan applications by predicting the credibility of loan borrowers.

In addition, another metric that should be taken into consideration is the execution time for each model. As we can see from the result table, it took more time for Naive Bayes and Random Forest to execute the model compared to Bayes Net. Bayes algorithm seem to be simple and more robust for classification problems in this dataset because it uses probability theory to classify data instead of choosing the split point that generates highest information gain and lowest entropy from multiple trees like Random Forest. Therefore, we can conclude that the Bayes algorithm generally outperforms Random Forest in terms of speed and should be confidently used as the prediction tool for Dream Housing finance company.

In further study, we would try to conduct experiments on a larger scale of dataset, or try splitting the dataset with different ratios, or try to tune the settings of the model so as to achieve the most optimal performance of the model.

Conclusion

In this paper, three algorithms - Bayes Net, Naive Bayes and Random Forest algorithms were used to build predictive models in order to predict and classify the credibility of forthcoming applicants. And after applying above-mentioned classification algorithms, we find out that Naive Bayes performs slightly better than the other two techniques because of its higher accuracy, higher F1-Measure, lowest mean absolute error with acceptable speed and as shown in the table result.

Bibliography

1. Zurada, Jozef, and Martin Zurada. "How Secure Are Good Loanss: Validating Loan-Granting Decisions And Predicting Default Rates On Consumer Loans."Review of Business Information Systems (RBIS) 6.3 (2011): 65-84.

2. Abhijit A. Sawant and P. M. Chawan, "Comparison of Data Mining Techniques used for Financial Data Analysis," International Journal of Emerging Technology and Advanced Engineering, june 2013.

3. Sudhakar M and Dr. C. V. Krishna Reddy, "CREDIT EVALUATION MODEL OF LOAN PROPOSALS FOR BANKS USING DATA MINING," International Journal of Latest Research in Science and Technology, pp. 126-131, july 2014.

4. Jafar, H. (Mar 2016). Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal (MLAIJ)*, *3*(1), 1-9. DOI:10.5121/mlaij.2016.3101

5. Sonia Singh. (2014). International Journal of Advanced Information Science and Technology (IJAIST). *Comparative Study Id3, Cart And C4.5 Decision Tree Algorithm: A Survey*, *3*(7). DOI:10.15693/ijaist/2014.v3i7.47-52

6. A, E. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*, 1772-1778.

7. Kaggle. (2020, 5 17). *Dream House Finance Company*. Finance Company Loan data. https://www.kaggle.com/sethirishabh/finance-company-loan-data?select=train_ctrUa4K.csv

8. Bastin, A. M. (2019, July 4). *The Math of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark*. Datasciencecentral. https://www.datasciencecentral.com/profiles/blogs/the-mathematics-of-decision-trees-random-forest-and-feature

9. McCloskey, S. (2000). *Probabilistic Reasoning and Bayesian Networks*. CIM. http://www.cim.mcgill.ca/~scott/RIT/researchPaper.html

10. Mitra, A. (2019, Aug 19). *Basics of Bayesian Network*. Towardsdatascience. https://towardsdatascience.com/basics-of-bayesian-network-79435e11ae7b

11. S, T. (2020, June 5). *A Mathematical Explanation of Naive Bayes in 5 Minutes*. Towardsdatascience. https://towardsdatascience.com/a-mathematical-explanation-of-naive-bayes-in-5-minutes-44adebcdb5f8

12. Trehan, D. (2020, July 2). *Why Choose Random Forest and Not Decision Trees*. TowardsAI. https://towardsai.net/p/machine-learning/why-choose-random-forest-and-not-decision-trees