Loan Eligibility Prediction Using Machine Learning

Dushyanth Reddy Bommana
Department of Mathematical Sciences
University of Essex

MA902-7-FY: Research Methods
Registration number: 2010727
Supervisor: Dr Lisa Voigt

Date of submission: 6 August 2021
Word count: 4461

University of Essex, United Kingdom.

Contents

1 Abstract

Financial stability depends on a sound banking system, a manageable level of non performing loans, and proper pipelines to avoid these bad loans. Over accumulation of bad debts not only cause financial destabilization, it also hinders the economic growth. In the modern world where every industry is being digitized, millions of gigabytes of data is collected. Using this data to solve problems has become the game of the day. In this study, various Machine Learning models are trained on a dataset to decide the eligibility criteria for loan. These are the attributes that can be used by a bank to take a faster smoother and more accurate decision about processing a loan for any individual. The cost of assets is rising every day, and the amount of money needed to buy an asset is immensely significant. As a result, you won't be able to buy it with your funds. Applying for a loan is the uncomplicated approach to obtain the finances required. Banks' primary business is lending. The main source of profit is the interest on the loan. After an extensive verification and validation process, the loan companies grant a loan. They do not have the assurance that the applicant will be able to repay the loan without difficulty. For banking institutions, loan acceptance is a critical step. The loan applications were either approved or rejected by the system. Loan recovery is a significant contributing factor in a bank's financial statements. It is hard to forecast if the customer will be able to repay the loan. Many researchers have been working on loan approval prediction algorithms in recent years. Machine Learning (ML) approaches are beneficial for predicting outcomes when dealing with enormous amounts of data. In this study, three machine learning methods, namely Logistic Regression (LR), K-Nearest Classifier(KNN), and Decision Tree(DT), are used to predict client loan approval. In terms of accuracy, the experimental results show that the Decision Tree machine learning algorithm outperforms the Logistic Regression and Random Forest machine learning approaches. This can save time and money by lowering the amount of time and employment needed to approve loans and weeding out the best applicants for lending.
Keywords: Loan · Machine Learning · Accuracy · Logistic Regression

2 Introduction:

People currently relying on bank loans to meet their financial necessities. In recent years, the number of loan applications has increased rapidly. When it comes to loan approval, the risk is always involved. The bank executives are highly concerned about the loan amount to pay back by the borrowers. Loan Prediction is exceptionally beneficial to both bank employees and applicants. The purpose of this paper is to have a quick, straightforward, and efficient method of selecting qualified applicants. Once the lender or bank verifies the customer's loan eligibility, the consumer applies for a loan. The company or bank wishes to automate the loan eligibility procedure (real-time) based on the information provided by the customer when filling out the application form. Even after taking numerous measures and thoroughly reviewing the loan applicant data, loan approval decisions are not always correct. This procedure must be automated, so that loan approval is less risky, and banks lose less money.

What is a loan? The term loan refers to a sum of money is lent to another party in exchange for future repayment of the value or principal amount. In many cases, the lender also adds interest and/or processing charges to the principal value which the borrower must repay in addition to the principal balance[1]. Loans come in many different forms including secured, unsecured, commercial, and personal loans. One of the most popular financial institution that provides loans is a bank.

Loans are a part of the financial system for several reasons including major purchases, investing, renovations, debt consolidation, and business ventures. It is because of loans existing companies can expand their operations. This is how competition is created and the industry thrives as a whole. Apart from these, the interest and fees from loans are a primary source of revenue for many banks. There are two types of interest rates on loans. Simple and compound interest. Simple interest is interest on the principal loan. All the banks always charge compound interest only. This is the interest charged on interest accumulated for the periods. This means more money in interest has to be paid by the borrower[1].

What is a bad debt and why is it bad? A bad debt or a non-performing loan (NPL) is a loan for which interest has not been paid in several months. This overdue may be the result of inability to pay or debtors unwillingness to pay or some other misfortune. Such a situation is definitely bad for the lender as both the principle and interest money is lost. It is to be noted that the borrower is also carrying the burden by getting their collaterals tied up in the loan. This also hits their credit score and makes it difficult to continue the cash flow to carry on their financial activities[2].

Why does bad debt occur? Financial institutions work with a certain risk factor and sometimes these risks materialize. This implies that bad debts are, in a way, a part of the banking system and the whole of the financial sector's operation. This does not mean that existence of these bad debts is essential for the financial system to run smoothly. In fact if the total bad debts are not at a moderate and manageable level that can be covered by the reserve amount, the financial stability itself will be at risk. A majority of these bad debts stem from the inefficient risk management and these can be mitigated by taking proper precautions[3].

What can be done? With development in financial innovations, there is also development in new risks being taken. It is not possible to completely avoid giving out loans as the economy is driven by the credit system. Global economic history shows that these bad loans pile up due to insufficient risk management in the past by financial institutions. Handling the non performing loans after they have become NPL is not an efficient way. Redistribution of risks among various financial institutions through securitization can be done but such a system will eventually pile up more risks as the lenders can get care free about whom they are lending to. Putting in place a better scrutiny process to identify an eligible debtor who can repay the debt can minimize the risk and also keept the financial cycle running[4].

Loan eligibility is defined as a set of criteria basis which a financial institution assesses the creditworthiness of a customer to avail and repay a particular loan amount. Loan eligibility depends on criteria such as age, financial position, credit history, credit score, other financial obligations etc.

Machine Learning in loan eligibility: Machine learning has flipped the script on traditional lending, allowing for more accurate and faster decisions by shifting traditional decision-making from analysis of individuals to analysis of trends and patterns. This has resulted in low operating costs for lenders and repeated business. These outcomes matter in a world where technology-enabled financial service is shaking traditionalists to the core. The leading global investment banking and securities firm *Goldman Sach* says "fintech is poised to gobble a good third of the annual revenue of traditional financial-service companies". Reflecting this transformation, the global digital-lending platform market expected to approach $\$20 billion$ by 2026 for a compound annual growth rate of 19.6% through the seven years prior[5].Although no technique may be favoured above another, logistic regression has been shown to be frequently used as an industrial technique due to its comprehensive simplicity[6].

Quirks of Machine Learning: Advances in artificial Intelligence in banking and machine learning in banking are helping commercial lenders differentiate themselves by[7]:

- Identifying bottlenecks in their operation work flows and bring in significant improvements in process efficiencies and efficiency ratios
- Compensating for a shortage of talent and hiring budgets devoted to regulatory compliance
- Reducing errors and risk
- Using repetitive process automation to free employees to focus on skill-based work and more engaging experiences for customers
- Predict outcomes

How Machine Learning does this?: Machine learning is a subset of artificial intelligence, which is a functionality (some call it a "device," others a "process") that takes into account aspects of its environment to make decisions or predictions, mimicking human cognition. It does so by using algorithms and statistical models to perform many specific *if-this-then-that* type tasks virtually at once, drawing on patterns and inferences rather than explicit case-by-case instructions. In other words, machine learning takes relevant inputs (*"training data"*) and constructs mathematical models that bring "thinking" to nuanced processes requiring multiple inputs from vast datasets — such as determining a would-be borrower's suitability for a loan of a specific size, type, and duration[5].

A deep neural network is a method of machine learning. It functions in the realm of artificial intelligence like a dating app for data. Here's the idea. A neural network sits between multiple datasets coming in (inputs) and decisions or predictions based on that data and the underlying training data (outputs). Its function is to apply the correct formula to the type of data in question[5].

This study uses a dataset to build an Artificial Intelligence system using various various models of supervised learning techniques. These models will take in historic data as training data and build a model to predict whether a particular new individual is eligible for a loan. In a simple term (real time scenario), a financial institution can use this study to automate the Loan Eligibility Process in a real time scenario related to customer's details provided while applying application for a loan.

3 Dataset:

It is important to understand data before getting into any kind of problem solving. The given data for training the model is a labelled data. This implies the training data has a column which refers to the target variable. Since the target here is to determine whether an individual is eligible for a loan or not, the individuals given in the training data are already classified into one or the other category. The training data consists of 614 such rows i.e, it has data of 614 individuals. Each row has 11 attributes of the individual which are relevant to the loan processing and one target column stating if that individual is eligible for a loan. Here are the attributes provided about the individuals.

- Gender → Male/ Female (binary input)
- Married → Applicant married (Y/N) (binary input)
- Dependents → Number of dependents (categorical)
- Education → Applicant Education (Graduate/ Under Graduate) (binary input)
- Self_Employed → Self-employed (Y/N) (binary input)
- ApplicantIncome → Applicant income (numerical)
- CoapplicantIncome → Co-applicant income (numerical)
- LoanAmount → Loan amount in thousands (numerical)
- Loan_Amount_Term → Term of a loan in months (numerical)
- Credit_History → Credit history meets guidelines (binary input)
- Property_Area → Urban/ Semi-Urban/ Rural (categorical)
- Loan_Status → Loan approved (Y/N) (binary output target variable)

There are 7 categorical variables, 4 numerical variables and one binary output target variable in the dataset. The first 5 rows of the dataset are printed below. The columns are in the above order. There is also a loan Id column in the beginning of the table which does not have anything to do with loan eligibility criteria.

0	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y

4 Background Research

Naive Bayes, unpruned C4.5 decision tree, pruned decision tree, and Random forest were created as classification models. Each of these models was ran four times, with assessment metrics recorded each time. Models used unprocessed data in the first iteration, while the other three used processed data with three distinct feature selection techniques[8].In a paper by Dagar[9], four machine algorithms are employed to predict if a person is eligible for a loan or not: Logistic Regression, Random Forest, Support Vector Machine, and XGBoost. Based on the dataset, we discovered that Logistic Regression outperforms other models in terms of accuracy. In an article, the author proposed to lower the risk factor for banks when it comes to choosing the right person for a loan approval thus speeding up the loan approval process[10].He trained the deep learning model using data mining approaches to examine previous records to which the bank has already sanctioned a loan based on the analysis generated out of these records. This study provides a method for forecasting whether a Lending Club peer-to-peer loan will be paid off or defaulted. The following were the primary stages of the process used to find the solution: Data exploration is the process of learning about the properties of a dataset. Data Preprocessing and Classification using different algorithms are two processes that prepare data for analysis[11].Random Forest was shown to be the most effective classification model, with an accuracy of 71.75 percent.The recent development of machine learning and data mining techniques has sparked interest in using these approaches in a variety of sectors[12]. The banking sector is no exception, and the growing pressure on financial organizations to have effective risk management has sparked interest in improving current risk estimating approaches. Machine learning techniques could potentially lead to a better estimate of the financial risks that banks are exposed to. The Basel accords, which establish frameworks for regulatory standards and risk management procedures as a guideline for banks to manage and quantify their risks, have been continuously developed in the credit risk domain. The standardized approach and the internal ratings-based approach (IRB) are two approaches described in Basel II for determining the minimum capital requirement[13]. Different risk metrics are used by banks to predict the potential loss they may face in the future. The expected loss (EL) a bank would bear in the event of a defaulted customer is one of these metrics. The probability of a particular client defaulting is one of the components involved in EL-estimation. Customers who are in default have failed to meet their contractual commitments and may be unable to repay their loans[14] . As a result, acquiring a model that can forecast defaulted consumers is of interest. Logistic Regression[15] is a widely used technique for assessing the likelihood of client default. A set of machine learning approaches will be examined and studied in this thesis in order to see if they can challenge the already used methodologies.It was discovered that no single instrument is clearly superior than others. It is determined that the only way to find an overall better performance model is to use informed tool integration to create a hybrid model. This work contributes to a detailed understanding of the characteristics of the instruments used to create bankruptcy prediction models, as well as its associated flaws[16].In this paper, a new hybrid feature selection algorithm to predict loan eligibility based on the wrapper model and the fisher is presented. The major goal of this paper is to show that the novel hybrid model outperforms the old random forest approach in terms of accuracy[17].Davis et al[18] were one of the first to use machine learning approaches to credit risk. The authors of the paper used two models to test a succession of credit default risk algorithms:a general computational model based on a selection process and a pairing technique, and an artificial neural network (ANN) connective model.Another early work[19] presented an attribute selection metric for building models that significantly reduce the non-monotonicity problem of decision trees while maintaining classification accuracy. In a dataset of mortgage loans[20] compares classification and regression tree (CART) and artificial neural networks (ANN) models with k-nearest neighbour (KNN) models.Shi et al[21] present a credit scoring model for credit risk assessment based on SVM and RF, which establishes a score for the ranking of relevance of a particular attribute. In datasets from German and Australian credit transactions, the authors compare the proposed SVM model to traditional SVM models.

5 Methods used:

This study implements a framework to process this dataset about the details of individuals who have been approved or rejected for loan. The framework includes:

- Initial data study
- Data cleaning and preparation
- Exploratory Data Analysis
- Model building

Python has become the most widespread language to work with Data Science. The *pandas* is a great library to study and build models. So this study is carried out using Python. It offers great flexibility to work with all kinds of data. Since the data provided is a *.csv* file, it can be easily read into the pandas dataframe with a single command *read_csv()*. Further, it can be used to easily print out the data summaries, visualise the data and build models on the data. Our target column is a categorical variable with 2 levels. So this becomes a classification problem. Since the labels are already given, supervised learning models are to be built. This study considers three models to implement on the dataset and compare Coming to the model building, three models are chosen to classify the data into the binary output class.

- Logistic Regression
- kNN Regression
- Decision Tree Classifier

Logistic Regression and kNN Regression are very simple models while decision tree is slightly complex over these two. Intuitively Decision trees gives a lot of interpretability to the model as each step of the process can be visualised.

5.1 Logistic Regression:

Logistic regression has become an important statistical analysis tool in the discipline of machine learning. It is used to predict a data value based on prior observations of a data set. The logistic regression algorithm classifies incoming data based on the data that the model has been trained on. This is usually called the historic data. More the data fed into training, better are results of the model's performance. Sometime Logistic Regression can also be used as a preprocessing technique to divide the data into predefined categories for further analysis[22].

$$Pr(Y_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 X_i + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)}{1 + exp(\beta_0 + \beta_1 X_i + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)} \tag{1}$$

where β is co-efficient and X_i are the features

Logistic Regression at work. A logistic regression model predicts the target variable of the dataset by extracting the patterns and interactions between one or more existing attributes of the dataset. For example, a logistic regression could be used to predict whether a basketball team will win or lose a match against another team or whether a particular customer will visit a particular supermarket to avail an offer. The resulting analytical model can take into consideration multiple input criteria. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

Purpose of Logistic Regression is to estimate the probabilities of events, including determining a relationship between features and the probabilities of particular outcomes. These are called the log odds. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The logistic function is defined as:

$$Logistic(\eta) = \frac{1}{1 + e^{-}\eta} \tag{2}$$

For classification, probabilities between 0 and 1 are preferred and this is achieved using the logistic function. This forces the output to assume only values between 0 and 1. The figure 1 explains the logistic regression clearly.

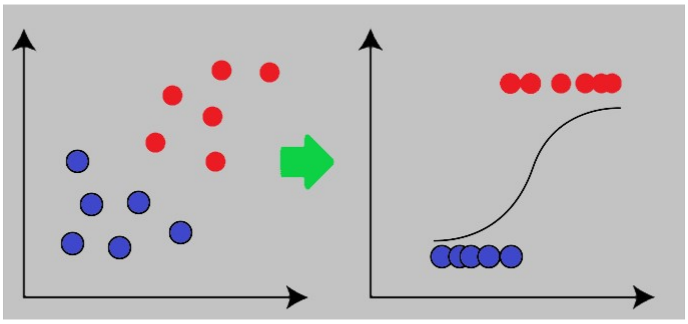

Figure 1: Logistic Regression Explained

5.2 kNN Classifier

K Nearest Neighbor algorithm falls under the Supervised Learning category and is most commonly used for classification. It can also be used for regression. Its basic principle is majority class prediction. It considers a certain number of nearest neighbors for taking this majority. This number of nearest neighbors is the value of k.

The learning method of kNN is an instance based learning. There are no weights or coefficients obtained from the training data. It does not use any pre-defined formulae or mapping functions. It does not remember the training data information. Instead, it calculates the similarity measure between the unseen data and all the training instance every time. This makes it a lazy learning model[23].

$$d\left(p, q\right) = \sqrt{\sum_{i=1}^{n} \left(q_i - p_i\right)^2} \tag{3}$$

where p and q are data points.

The principle behind kNN is identifying the classes of K Nearest Neighbors as mentioned in the name. Here, nearest neighbors are those data points that have minimum distance in feature space from our new data point. Distance is just a figure of speech. It can be any kind of similarity measure. K is the number of such data points we consider in our implementation of the algorithm. Therefore, distance metric and K value are two important considerations while using the KNN algorithm. Euclidean distance is the most popular distance metric[24]. Hamming distance, Manhattan distance, Minkowski distance are used based on the problem requirement. For predicting the class for an unseen entry, it calculates the distance of all the data points in the training dataset from the new entry then identifies the 'K' Nearest Neighbors from dataspace for the new entry and chooses the majority value. The new point in the figure 2 belong to class red.

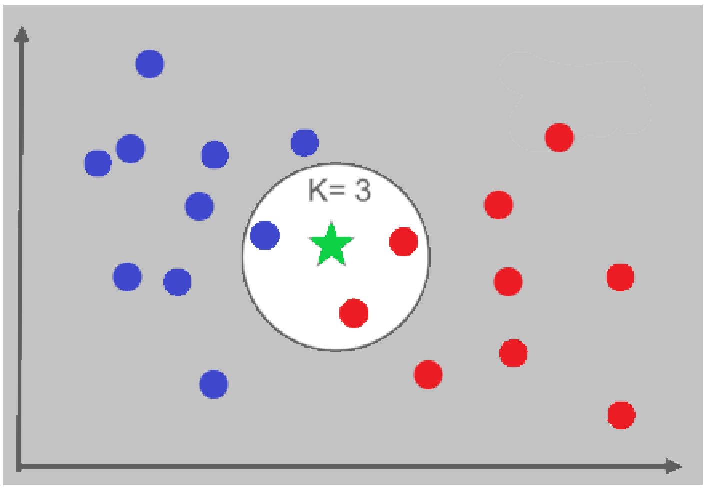

Figure 2: kNN classifying new entry

6 Decision Trees

Decision trees are also one of the widely used algorithms in data science. It is a go to supervised learning algorithm in complex scenarios with high interpretability requirement. It is efficient and has strong algorithms used for predictive analysis. It is also a supervised learning algorithm where data is divided into bins based on decision functions at each node. Decision trees can be used for both classification and regression and is also know as CART (Classification and Regression Tree). In regression, the data is divided based on bin range of the values.

Structure of a decision tree is tree like as the name suggests. The tree is built based on conditions. The tree is made up of three structures - internal nodes, branches and a terminal node. At every internal node a condition check is done on one attribute of the entry and a decision is made on the branch to which that entry point belongs to. The terminal nodes define the class labels. When a node gets divided further then that node is termed as parent node whereas the divided nodes or the sub-nodes are termed as a child node of the parent node. In the figure 3, C is the root node, B and D are branches and the nodes with the +ve and -ve in them are the terminal nodes[25].

Decision Tree works on both the type of input and output that is categorical and continuous. In classification problems, the decision tree asks questions, and based on their answers (yes/no) it splits data into further sub branches. It can also be used as a binary classification problem like the one under consideration in this study or multi class classification where the output has more than 2 values. In a decision tree, the algorithm starts with a root node of a tree then compares the value of different attributes and follows the next branch until it reaches the end leaf node. It uses different algorithms to check about the split and variable that allow the best homogeneous sets of population[25].

7 Exploratory Data Analysis

7.1 Descriptive Summary

The following table gives the descriptive statistics of the numerical columns of the dataset. Credit_History is included even though it is categorical because, it is of data type integer.

It can be seen that there are 614 entries for *ApplicantIncome* and *coapplicantincome* but fewer entries for the other 3. This indicates the existence of missing values. Using the *isna().sum()* method, the number

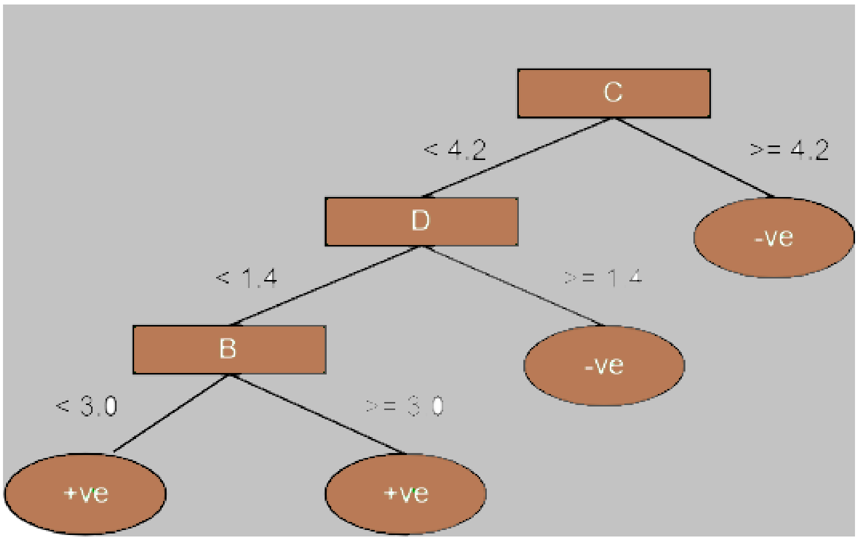

Figure 3: Decision Tree

Table 1: Descriptive Statistics

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614	614	592	600	564
mean	5403.459	1621.246	146.4122	342	0.842199
std	6109.042	2926.248	85.58733	65.12041	0.364878
min	150	0	9	12	0
25%	2877.5	0	100	360	1
50%	3812.5	1188.5	128	360	1
75%	5795	2297.25	168	360	1
max	81000	41667	700	480	1

of missing values are identified in table 2. The column are all heavily skewed to the right. It can be inferred from the max values being far greater than minimum values when compared with the mid value.

7.2 Preprocessing

Two major steps done in cleaning the data are filling the missing values and encoding the categorical columns. The missing values are filled using the mean or mode of that particular column. Machine learning models require all input and output variables to be numeric. This means that if the data contains categorical columns, they must be *encoded* to numbers before we can implement the models. The two most popular techniques are an Label Encoding and a One-Hot Encoding. Label encoding is a simple technique where the categories or classes are replaced with a respective number. A numerical variable can be converted to an ordinal variable by dividing the range of the numerical variable into bins and assigning values to each bin[26].

This is called an Label encoding or an integer encoding and is easily reversible. Often, integer values starting at zero are used. For some variables, an ordinal encoding may be enough. The integer values have a natural ordered relationship between each other and machine learning algorithms may be able to understand and harness this relationship. It is a natural encoding for ordinal variables. For categorical variables, it imposes an ordinal relationship where no such relationship may exist. This can cause problems and a one-hot encoding may be used instead. In this study, only Label encoding is used[26].

Table 2: Missing value counts

Attribute	Missing Values
Gender	13
Married	3
Dependents	15
Education	0
Self_Employed	32
ApplicantIncome	0
CoapplicantIncome	0
LoanAmount	22
Loan_Amount_Term	14
Credit_History	50
Property_Area	0
Loan_Status	0

7.3 Visualisations

Countplots are used to understand the counts of categorical columns. The key categorical column in the target variable. This gives the understanding of the target column distribution (figure 4). There are 422 **Y** and 192 **N** in the target column. This is a heavily uneven distribution.

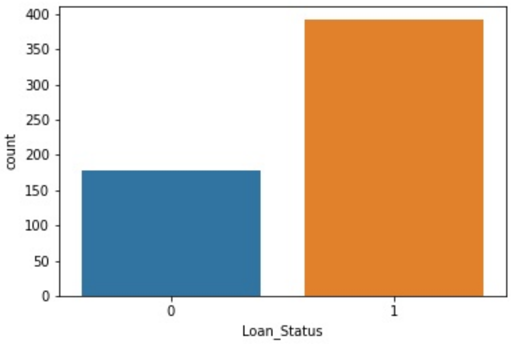

Figure 4: Target column counts

The distributions of the rest of the categorical columns with respect to Loan status are given by the countplots in the figure 5. Certain interesting observations are that males are approved loans more often than females. Married people are approved loans more than unmarried. Very few self employed people apply for loans and are approved. Semiurban population apply more for loans and are also approved more. There is a more probability of Graduates getting loans in comparison. Individuals with 0 dependants apply a lot for loans compared to people with dependants. Loan are rarely approved for people whose credit scores are not matching the required criteria.

8 Model Implementation

Before implementation of any model, the data is split into a training and testing set. This will help us evaluate the model on an unseen data immediately.

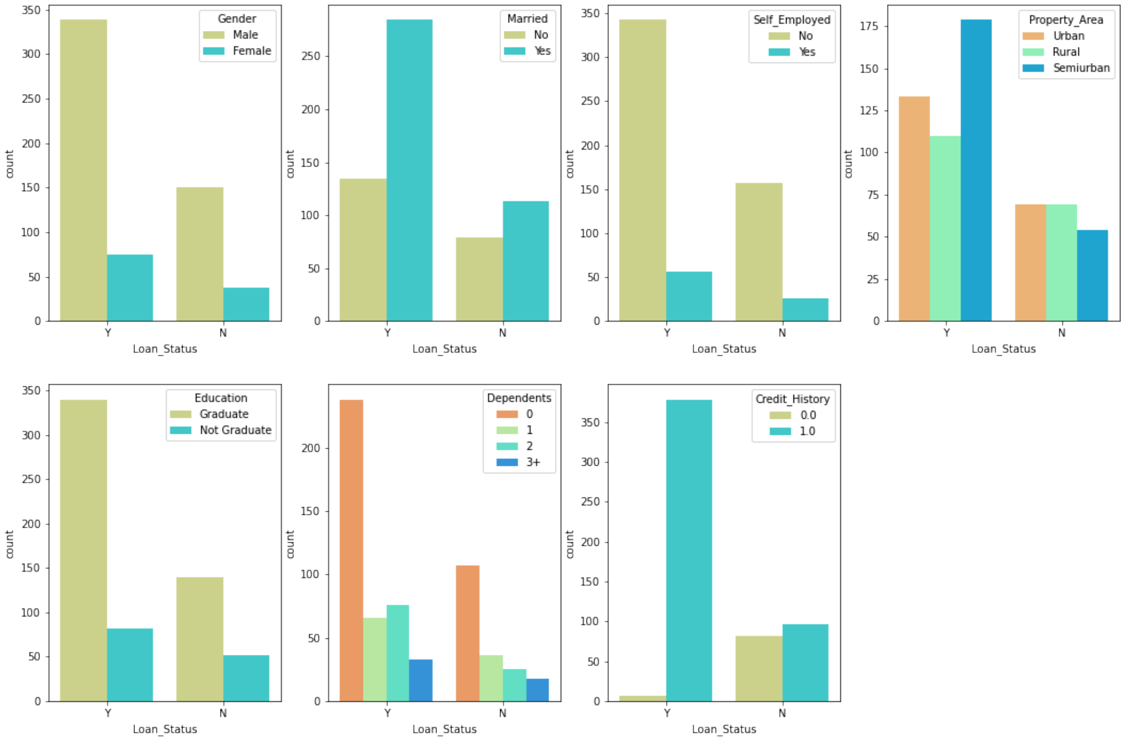

Figure 5: Categorical column distribution

8.1 Logistic Regression

It is a simple model with no hyperparameters to tune. Applying this model directly on the given model on a 75-25 split training and testing data gives an accuracy of 81%. The confusion matrix of the model is printed below.

	Actual 0	Actual 1
Predicted 0	19	24
Predicted 1	2	98

8.2 kNN Classifier

It is a simple model and depends on the number k provided to choose the nearest neighbors. A rule of thumb is to choose the square root of number of entries. But this model need scaling to be done on the dataset as it is dependent on the distances between datapoints. So with a k value of 24, the model is implemented on a 75-25 split training and testing data gives an accuracy of 79%. The confusion matrix of the model is printed below.

	Actual 0	Actual 1
Predicted 0	17	26
Predicted 1	2	98

8.3 Decision Tree

Decision tree can work with any kind of data. But it has hyper-parameters to tune. With entropy as the criterion and max_depth as 3, the model has obtained an accuracy of 81%. The confusion matrix is printed below. The tree is also visualized in the figure 3.

	Actual 0	Actual 1
Predicted 0	17	26
Predicted 1	2	98

9 Conclusion

All the three models performed very well on the given dataset. Both the training and testing accuracies are aroun 80% only. This means the models are not over fir either.

Model	Training Accuracy	Testing Accuracy
Logistic Regression	81.81	80.82
KNN	80.42	80.2
Decision Tree	81.81	81.45

Since all the models are producing the same result, Decision Tree is the better model to choose as it clearly demonstrates why a particular individual is approved or rejected for the loan.

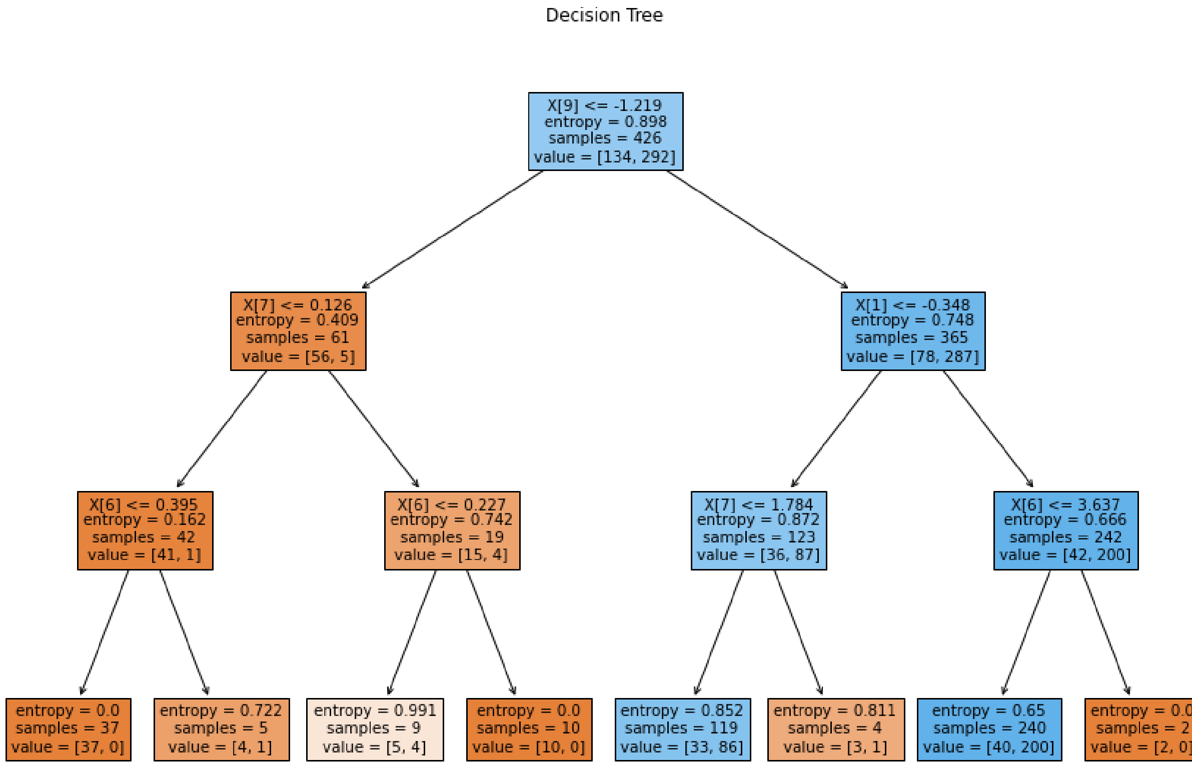

Figure 6: Decision Tree

References

[1] J. Kagan, "Loan," https://www.investopedia.com/terms/l/loan.asp, May 2021, accessed: 2021-8-5.

[2] M. Balgova, M. Nies, and A. Plekhanov, "The economic impact of reducing non-performing loans," *SSRN Electron. J.*, 2016.

[3] "Bad debts and economic growth," https://econs.online/en/articles/economics/bad-debts-and-economic-growth/, accessed: 2021-8-5.

[4] M. C. Aniceto, F. Barboza, and H. Kimura, "Machine learning predictivity applied to consumer creditworthiness," *Future Business Journal*, vol. 6, no. 1, pp. 1–14, December 2020. [Online]. Available: https://ideas.repec.org/a/spr/futbus/v6y2020i1d10.1186_s43093-020-00041-w.html

[5] V. Sechkin, "How machine learning is used in the lending industry," https://www.turnkey-lender.com/blog/how-machine-learning-is-used-in-the-lending-industry/, Feb. 2020, accessed: 2021-8-6.

[6] K. Eria and P. Subramanian, "Decision support credit scoring model to improve loan default prediction in financial institutions," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 8, pp. 3514–3518, 2019.

[7] S. Guttman, "Machine learning in banking and construction loan administration," https://

rabbet.com/blog/machine-learning-in-banking-and-construction-loan-administration/, Apr. 2018, accessed: 2021-8-31.

[8] A. Al-qerem, G. Al-Naymat, and M. Alhasan, "Loan default prediction model improvement through comprehensive preprocessing and features selection," in *2019 International Arab Conference on Information Technology (ACIT)*. IEEE, 2019, pp. 235–240.

[9] A. Dagar, "A comparative study on loan eligibility," 2021.

[10] A. Kumar, R. Dugyala, and P. Bhattacharya, "Prediction of loan scoring strategies using deep learning algorithm for banking system," in *Innovations in Information and Communication Technologies (IICT-2020)*. Springer, 2021, pp. 115–121.

[11] Z. Alomari and D. Fingerman, "Loan default prediction and identification of interesting relations between attributes of peer-to-peer loan applications," *New Zealand Journal of Computer-Human*, 2017.

[12] T. M. Mitchell, "Machine learning and data mining," *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[13] I. Basel, "International convergence of capital measurement and capital standards, a revised framework. comprehensive version, basel committee on banking supervision, bank for international settlements, basel, june 2006. 4," *Basel III: A global regulatory framework for more resilient banks and banking systems, Basel Committee on Banking Supervision, Bank for International Settlements, Basel, December*, vol. 5, pp. 2002–2010, 2010.

[14] P. Mishra, "Size of farm and productive efficiency: A review," *International Journal of Research in Social Sciences And Humanities*, vol. 3, no. 1, 2014.

[15] E. N. Tong, C. Mues, and L. C. Thomas, "Mixture cure models in credit scoring: If and when borrowers default," *European Journal of Operational Research*, vol. 218, no. 1, pp. 132–139, 2012.

[16] H. A. Alaka, L. O. Oyedele, H. A. Owolabi, V. Kumar, S. O. Ajayi, O. O. Akinade, and M. Bilal, "Systematic review of bankruptcy prediction models: Towards a framework for tool selection," *Expert Systems with Applications*, vol. 94, pp. 164–184, 2018.

[17] B. R. Jawale, P. A. Badgujar, R. D. Talele, and D. D. Patil, "Loan amount prediction using machine learning."

[18] R. H. DAVIS, D. Edelman, and A. Gammerman, "Machine-learning algorithms for credit-card applications," *IMA Journal of Management Mathematics*, vol. 4, no. 1, pp. 43–51, 1992.

[19] A. Ben-David, "Monotonicity maintenance in information-theoretic machine learning algorithms," *Machine Learning*, vol. 19, no. 1, pp. 29–43, 1995.

[20] J. Galindo and P. Tamayo, "Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications," *Computational Economics*, vol. 15, no. 1, pp. 107–143, 2000.

[21] J. Shi, S.-y. Zhang, and L.-m. Qiu, "Credit scoring by feature-weighted support vector machines," *Journal of Zhejiang University SCIENCE C*, vol. 14, no. 3, pp. 197–204, 2013.

[22] C. Molnar, "4.2 logistic regression," https://christophm.github.io/interpretable-ml-book/logistic.html, Aug. 2021, accessed: 2021-8-6.

[23] Wikipedia contributors, "K-nearest neighbors algorithm," https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=1031966276, Jul. 2021, accessed: 2021-8-6.

[24] sai, "KNN algorithm," https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/, Apr. 2021, accessed: 2021-8-31.

[25] R. Dwivedi, "Introduction to decision tree algorithm in machine learning," https://www.analyticssteps.com/blogs/introduction-decision-tree-algorithm-machine-learning, accessed: 2021-8-6.

[26] J. Brownlee, "Ordinal and one-hot encodings for categorical data," https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/, Jun. 2020, accessed: 2021-8-31.