Original Article

Real-Time Approach To Loan Credit Approval And Credit Risk Analysis Using ML

Madhwesha Moudgalya R[1], Kavita Permi[2], Vishesh S[3]

[1]Research Scholar Department of Mathematics. Presidency University.
[2]Associate Professor Department of Mathematics, Presidency University.
[3]BE, Department of Telecommunication, BNMIT, Visvesvaraya Technological University
DOI: 10.47750/pnr.2022.13.S09.533

Abstract

Modern computers' increased computational capabilities and ability to learn on their own have contributed to the emergence of AI. Few ether (dynamic) parameters yield a lot of data. Humans can't explicitly encode all accessible knowledge. Machines that learn this can produce more accurate results/predictions. Time alters environments. Adaptive machines would lessen the need for redesign. Modernized by bots, computers, and automated tools. Massive datasets require automation techniques and computer systems. Machine learning is a branch of AI that educates machines using transactional data.

We employed classification algorithms to develop an ML prediction model. Classification techniques like SVM, Random Forest Classifier, and KNN fit the dataset. During implementation, data patterns must be compared. Regression procedures like linear regression (created from scratch) will improve assignment accuracy (categorical data excluded).

Keywords: Classification algorithms, Logistic regression, SVM, KNN, Random Forest, accuracy, F1 score, Jupyter, Python.

1. INTRODUCTION

Credit risk is the likelihood of a loss if a creditor fails to repay a loan or fulfil other contractual commitments to an investment. It refers to the risk that a lender may not get the outstanding head and premium due to revenue disruptions and increased collection costs. Credit risk may be covered by writing unnecessary cash. Even though it's hard to anticipate who will default on commitments, monitoring credit risk can reduce a loss's severity. The lender or investor earns interest from the borrower or issuer of a debt obligation for risking credit default. When lenders or banks issue mortgages, credit cards, visas, or other loans, the borrower may not repay the debt. If a company gives credit to a client, the client may not pay their solicitations. Credit risk is the chance that a guarantor won't pay.

Insurance business won't pay a claim or pay when asked. Credit risks are based on a borrower's capacity to repay a loan based on its terms. Loan specialists examine credit history, ability to repay, money, loan conditions, and collateral to assess consumer loan credit risk.

2. PROPOSED SYSTEM METHODOLOGY

1. Data Acquisition: A dataset that is applicable to the problem statement is collected from a verified source and cleaned to encourage analysis. Cleaning includes managing missing qualities in any technique most appropriate for the coming steps.
2. Data Pre-Processing: The collected data which is stored in excel sheet may contain unwanted values and noise. Data is pre-processed by removing redundancy, noise, unwanted values. Data file is checked for missing values and treated.
3. Data visualization and analysis: Visualizing data is called data visualisation. Charts, graphs, and maps. Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
4. Data analysis inspects, cleans, transforms, and models data to identify usable information and enhance decision-making.

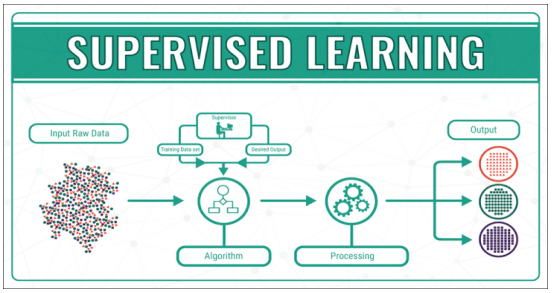
5. Correlation: Correlation describes the relationship between variables. These variables are inputs utilised to forecast our target variable. Correlation, a statistical approach that compares two variables.

Two variables can be positively associated. When one variable grows, the other(s) do too.
Two variables can be negatively linked. When one variable rises, the other(s) fall.
Two variables have no correlation. When one variable increases or decreases, the other(s) don't.

Data Splitting: The training data is used to make sure the machine recognizes patterns in the data, the cross-validation.
data is used to ensure better accuracy and efficiency of the algorithm used to train the machine.
The test data is used to see how well the machine can predict new answers based on its training. Thus, test set and training set is important to make machine learning model better. You want to split the data into training and test datasets. Then the model can take data and predict.
6. Machine Learning Model: Model uses SVM, KNN, Logistic regression, and Random Forest classifier.
7. Verification and Conclusion: The predicted values are well tabulated, precision, recall, F1 score and accuracies are measured and compared.

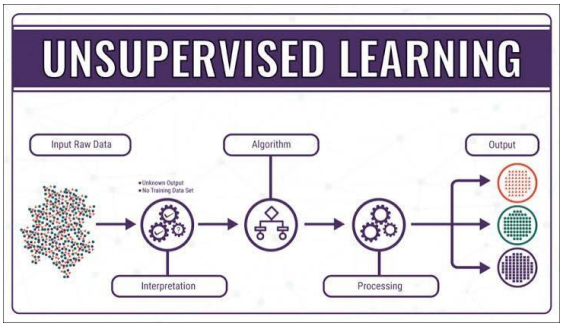
A. Machine Learning
In Machine Learning, a computer programme is assigned some tasks, and its measured performance in these tasks increases as it learns more experience doing them. The machine makes data-based choices and forecasts. Machine Learning solves classification, regression, and clustering problems. Depending on the types and categories of training data, the appropriate machine learning method may be "supervised learning," "unsupervised learning," "semi-supervised learning," or "reinforcement learning.".

3. SUPERVISED MACHINE LEARNING

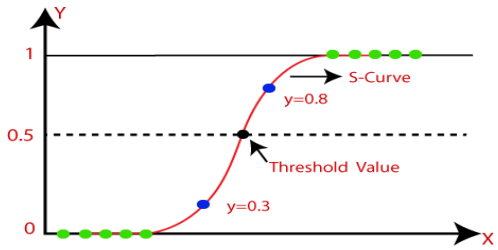


Supervised learning is machine learning in which machines are trained using well-labeled training data and predict output. Some input data has the correct output labelled. In supervised learning, training data trains machines to predict output accurately. It's the same as what a student learns in class. Supervised learning involves giving the machine learning model input and output data. A supervised learning algorithm maps the input variable(x) to the output variable (y). Supervised learning is used for Risk Assessment, Image categorization, Fraud Detection, and spam filtering.They are two types of supervised machine algorithms
1. Regression
2. Classificatin

4. UNSUPERVISED MACHINE LEARNING



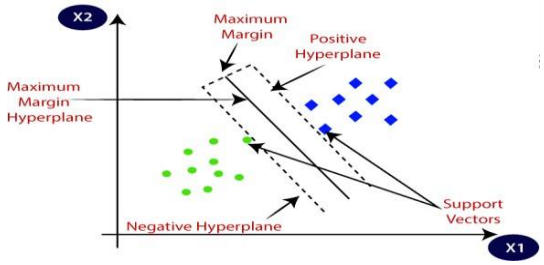
Unsupervised learning is a machine learning technique where models aren't supervised using training datasets. Models detect hidden patterns and insights in data. It's like human brain learning. Definition: Unsupervised learning allows models to learn from unlabeled data without supervision. Unsupervised learning can't be used for regression or classification because we have input data but no output data. Unsupervised learning finds a dataset's underlying structure, groups comparable data, and compresses it.Types of unsupervised learning:
1.Clusstering 2.Association
Here we are using only supervised machine learning algorithm.



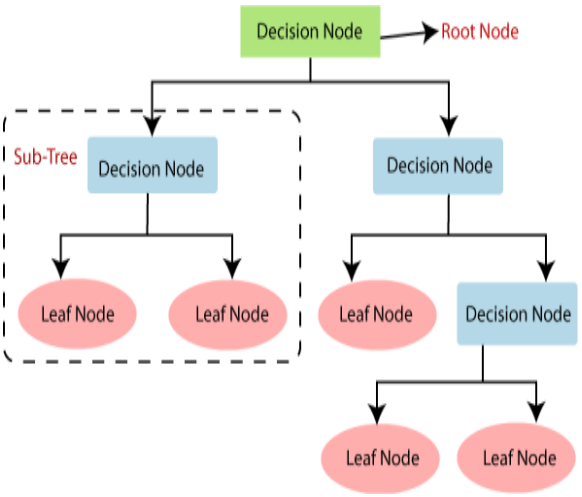
Machine Learning Algorithms

Logistic Regression Logistic regression classifies data. It gives the binomial outcome (in terms of 0 and 1) based on input variable values.

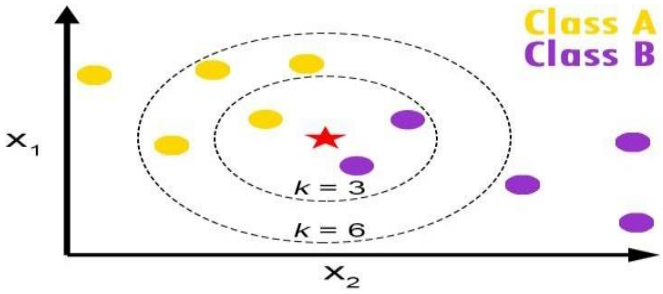
Support Vector Machine: SVM can handle classification and regression. Hyperplane is the decision boundary in this method. Decision plane separates classes of items. If the objects aren't linearly separable, complicated mathematical functions called kernels are needed to separate them. SVM classifies objects based on training data. It can handle both semi structured and structured data, it can handle complex function if the appropriate kernel function can be derived. As generalization is adopted in SVM so there is less probability of over fitting. It can scale up with high dimensional data. It does not get stuck in local optima.



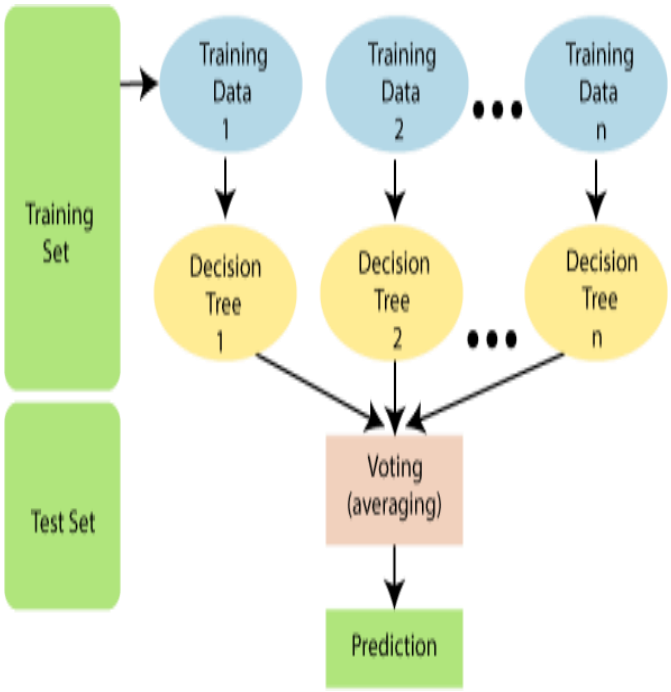
K-Nearest Neighbour: K-Nearest Neighbor is a Supervised Learning algorithm. K-NN algorithm assumes similarity between new case/data and available cases and puts new case in related category. K-NN maintains all available data and classifies new data on similarity. Using K-NN, new data can be easily sorted into a well-suited category. K-NN can be used for Regression and Classification, but predominantly Classification. K-NN is a non-parametric algorithm, hence it makes no data assumptions. It's dubbed a lazy learner algorithm because it doesn't



learn from the training set instantly. Instead, it saves the dataset and classifies it later. KNN algorithm saves the training dataset and classifies incoming input into a similar category. It measures k neighbours' distances. It preserves all training data without generalisation. It handles expensive huge data sets. Higher-dimensional data reduces region accuracy. KNN is used in recommendation systems, medical diagnosis of various diseases with similar symptoms, credit rating utilising feature similarity, handwriting detection, financial institution analysis before loan approval, video recognition, and political vote predicting.

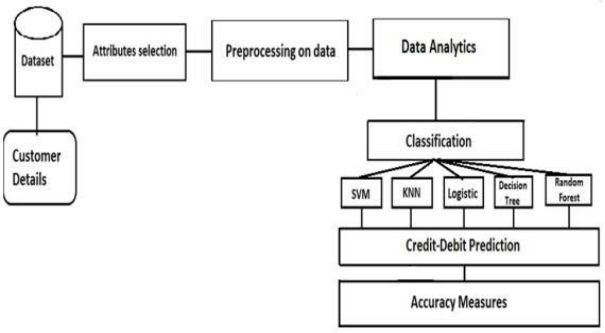


Decision Tree: Decision Tree is a supervised learning technique used for classification and regression issues, but predominantly classification. It's a tree structured classifier where internal nodes represent dataset attributes, branches represent decision rules, and leaf nodes reflect outcome. Decision trees have two nodes: Decision and Leaf .Decision nodes make decisions and have several branches, while Leaf nodes provide the output and have no branches. Decisions or tests are based on dataset attributes. It shows all viable answers to a problem/decision based on provided conditions. It's termed a decision tree because, like a tree, it starts with a root node and branches out. Classification and Regression Tree algorithm (CART) is used to generate trees. Based on the answer (Yes/No), a decision tree splits into subtrees. Decision trees can predict library book use and tumour prognosis. Diagram below shows decision tree structure.



Random Forest: Random Forest is a supervised learning system. ML Classification and Regression can use it. It uses ensemble learning to solve complicated problems and improve model performance. Random Forest is a classifier that uses a number of decision trees on subsets of a dataset to increase predicted accuracy. The random forest forecasts the final output based on the majority of predictions from each tree. More trees improve accuracy and prevent overfitting. Since random forest uses many trees to determine dataset class, certain decision trees may predict the proper output. All the trees predict the right output. For a better Random forest classifier, assume the following: I The dataset's feature variable needs actual values so the classifier can predict accurate outcomes. (ii) Each tree's predictions must be uncorrelated. Diagram of Random Forest algorithm.

5. BLOCK DIAGRAM OR FLOW DIAGRAM



Above figure illustrates the block diagram of the system used in this paper. Customer details are collected and stored as dataset. From the available dataset attribute selection is made. Data needs to be examined for pattern recognition and data pre-processing needs to be carried out to –

1) Fill missing/null values.
2) Remove duplicates.
3) Treat NaNs.
4) Replace string values with numbers.

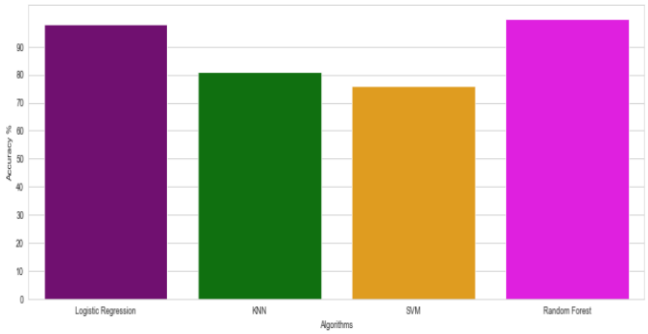
Graphs and maps help visualise data for greater understanding. Training the machine requires preprocessed data. The patterns would teach the system to predict all outcomes. Apply SVM, Random forest classifier, KNN, and logistic regression. For accuracy and F1 score, linear regression is modelled without libraries.

6. SOFTWARE USED

Jupyter Notebook is an open-source web software for creating and sharing documents with live code, equations, visualisations, and text. Data cleaning, transformation, numerical simulation, statistical modelling, data visualisation, machine learning, etc. JupyterLab is a web-based Jupyter notebook, code, and data development environment. JupyterLab's configurable user interface supports data research, scientific computing, and machine learning workflows. JupyterLab is modular and extendable; plugins provide new components. Jupyter Notebook is an opensource web software for creating and sharing documents with live code, equations, visualisations, and text. Project Jupyter runs Jupyter Notebook. I Python Notebooks spun out into Jupyter Notebooks. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the I Python kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

7. CONCLUSIONS

A 21st-century bank has many day-to-day transactions. Historical and current data analytics were used to generate conclusions. Create or improve the ML model and compare accuracy. To analyse and draw conclusions, python code was ran in Jupyter. Logistic Regression, SVM, Random Forest Classifier, and KNN are used to fit the dataset. During implementation, data patterns must be compared. Logistic Regression is 98.16% accurate, K-Nearest Neighbor is 80.89% at K=3, Support Vector Machine is 75.87% accurate, and Random Forest Classifier is 100% accurate. Random Forest Classifier and Logistic Regression fit this dataset best.



Future studies can add real-time data for more accurate predictions. Sentiment analysis of news items and economic data from the consumer's region can help assess loan risk. A user interface can allow the client to enter consumer details and return if the loan is doable. Larger datasets enhance accuracy. Better predictions can be made with more robust machine learning algorithms.

8. REFERENCES

1. Youness Abakarim, Mohamed Lahby, and Abdelbaki Attitoui. “Towards an Efficient Real Time Approach To Loan Credit Approval Using Deep learning”. In 2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC). IEEE, 2018 DOI 10.1109/ISIVC.2018.8709173.
2. Trilok Nath Pandey, Alok Kumar Jagadev, Suman Kumar Mohapatra, and Satchidananda Dehuri. “Credit risk analysis using machine learning classifiers”. In 2017,International Conference on Energy.Communication, Data Analytics and Soft Computing (ICECDS), pages 1850–1854. IEEE, 2017 DOI: 10.1109/icecds 2017.8389769.
3. Mohammad Ahmad Sheikh, Amit Kumar Goel and Tapas Kumar. “An Approach for Prediction of Loan Approval using Machine Learning Algorithm”. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020 DOI: 10.1109/ICESC48915.2020.9155614.
4. Regina Esi Turkson, Edward Yeallakuor Baagyere and Gideon Evans Wenya. “A machine learning approach for predicting bank credit worthiness”. In 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR). IEEE, 2016 DOI 10.1109/ICAIPR.2016.7585216.
5. Ashlesha Vaidya. “Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval”. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2017 DOI 10.1109/ICCCNT.2017.8203946.
6. TUNG, Hui-Hsuan, CHENG, Chiao-Chun, CHEN, Yu- Ying, et al, “Binary Classification and Data Analysis for Modeling Calendar Anomalies in Financial Markets”. In Cloud Computing and Big Data (CCBD), 2016 7th International Conference on. IEEE, (2016). p. 116-121.
7. Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., ’Support vector machines’, IEEE Intelligent Systems, Vol. 13 (4), pp. 18– 28, 2008.
8. R.Karthiban, M.Ambika, Dr. K E Kannammal, “A Review on Machine Learning Classification Technique for Bank Loan Approval”. In 2019 International Conference on Computer Communication and Informatics (ICCCI -2019), IEEE Jan. 23 – 25, 2019, Coimbatore, INDIA.
9. Susmita Ray from Department of Computer Science & Technology Manav Rachna University, “A Quick Review of Machine Learning Algorithms”. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019.
10. Ruttala Sailusha, V. Gnaneswar, R. Ramesh, G. Ramakoteswara Rao, “Credit Card Fraud Detection Using Machine Learning”. Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2.
11. Kedar Potdar, Rishab Kinnerkar, “A Comparative Study of Machine Algorithms applied to Predictive Breast Cancer Data”. International Journal of Science & Research, Vol. 5, Issue 9, pp. 1550-1553, September 2016.
12. Sonal S. Ambalkar, S. S. Thorat2, “Bone Tumor Detection from MRI Images using Machine Learning: A Review”. International Research Journal of Engineering & Technology”, Vol. 5, Issue 1, Jan -2018.
13. Heta Naik, Prashasti Kanikar: “Credit card Fraud Detection based on Machine Learning Algorithms”, International Journal of Computer Applications (0975 – 8887) Volume 182 – No. 44, March 2019.
14. C. Phua, V. Lee, K. Smith, R. Gayler (2010); “Comprehensive Survey of Data Mining-based Fraud Detection Research”, ICICTA ’10 Proceedings of the 2010 International Conference on Intelligent Computation Technology and Automation Volume 1, pp. 50-53.
15. Rajat Raina, Alexis Battele, Honglak Lee,Benjamin Packer, Andrew Y. Ng , “Self-taught Learning : Transfer of Learning from Unlabeled Data”, Computer Science Department, Stanford University, CA, USA, Proceedings of 24th International Conference on Machine Learning Corvallis, OR, 2007.