JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CRED CARE - PREDICTION OF CREDIT COMPLIANCE USING MACHINE LEARNING

[1]Krushan Gangurde, [2]Karthik Sridhar, [3]Sharanya Manoharan, [3], [4]Muskan Markam,[5]Rohan Mishra

[1,2,4,5]Student, [3]Professional

[1]Department Of Mechatronics Engineering, [2]Department Of IT Engineering, [4]Department Of Radio Imaging Technology (B.Sc.), [5]Department Of EXTC Engineering

[1,2,5]University Of Mumbai, Mumbai, India, [4]DY Patil University, Navi Mumbai, India

Abstract: The act of borrowing money from banks has become increasingly common in modern times, as banks' primary business is lending. Their profits stem mainly from the interest on loans, but the success or failure of a bank's endeavors is largely dependent on its ability to manage credit, ensuring that borrowers repay their loans rather than default. Consequently, predicting loan defaults has become a critical concern for banks, and as such, a topic of significant research interest. Previous studies have demonstrated numerous methods for managing loan defaults, but accurate prediction is essential for maximizing profits. To this end, predictive analysis, particularly through machine learning, has emerged as a necessary modeling technique for banking. Furthermore, we suggest a recommendation system that anticipates why a loan might be declined and provides advice on how to become financially literate to increase one's chances of being approved. Machine learning models such as K-Nearest Neighbors, Gaussian Naive Bayes, Logistic Regression, XGboost, Random Forest, and SVM are used to predict the likelihood of loan repayment. The system outputs a binary prediction of whether or not the customer will repay, aiming to promote financial literacy, expedite the credit approval process, and reduce non-performing assets (NPAs) for banks

IndexTerms **- Machine learning models, Prediction, XGBoost, Random Forest, Recommendation system.**

I. INTRODUCTION

The banking sector faces a critical question when it comes to issuing loans: How risky is the borrower, and should they lend to the borrower given the risk? Traditional manual procedures for assessing loan applications can be effective but are time-consuming, making it difficult to process a large number of loan applications. The ability to accurately determine an individual's creditworthiness is crucial in the world of finance, as it is a reflection of their likelihood to repay a loan and is often used as a key determinant in whether or not to grant a loan.

In recent years, machine learning models have emerged as a powerful tool for assessing creditworthiness. By analyzing vast amounts of data, machine learning models can identify patterns and correlations that may not be immediately apparent to human analysts. Various research papers have proposed different machine learning models for loan prediction, including MLP [12], which has been shown to outperform other machine learning and statistical techniques for classification and prediction. In particular, models such as XGBOOST [9][10], Random Forest [11][18], and Decision Tree [13] have been found to be implemented for creditworthiness assessment, outperforming traditional statistical techniques.

The goal of this project is to build a machine learning model that can predict the creditworthiness of individuals using various models, such as Random Forest, XGboost, SVM [19][20], Navies Bayes [21], Linear Regression and compare their performance. We will use a dataset that includes financial and demographic features to train and test our models. The aim is to provide financial institutions with an accurate tool for making informed decisions on issuing loans, minimizing risk, and maximizing profit.

Additionally, the model will provide recommendations to help individuals improve their credit score and increase their chances of being approved for a loan. These recommendations will be based on the most important factors identified by our models and tailored to each individual's specific financial situation. The benefits of such a model are clear. For individuals, it provides a more accurate assessment of their creditworthiness, potentially opening up new avenues of credit and allowing them to make more informed financial decisions.

For financial institutions, it offers a more efficient and effective way to assess risk, helping to minimize losses and maximize profits. Ultimately, the development of this machine learning model has the potential to revolutionize the way creditworthiness is assessed and lending decisions are made, assisting financial institutions in making more informed decisions and providing individuals with the tools they need to take control of their financial future.

II. LITERATURE REVIEW

In recent years, the use of machine learning techniques in financial services has gained significant attention. Many studies have been conducted to predict loan defaults and develop recommender systems for stock markets and loan prediction in banks. In this literature review, we will discuss several comparative studies that have used various machine learning techniques to predict loan defaults and develop recommender systems.

Several datasets have been used in various studies related to credit scoring using machine learning algorithms. The datasets used include the German credit data set and Australian credit data set [2], the Lending Club Dataset [4], data extracted from the University of Tennessee [6], and data collected from NSE and 10 different sectors [7]. Additionally, the Kaggle dataset has also been used in several studies [1, 4, 5, 8]. These datasets provide a diverse range of information that is crucial for training and testing machine learning models. It is important to choose the appropriate dataset that best represents the target population to ensure that the model is accurate and reliable.

Logistic regression model used to predict lending, achieved an accuracy of 0.66 in [1]. They collected datasets from Kaggle and trained and tested the model using various methods to calculate model accuracy. Another study [2] combined kernel density estimation and SVM algorithms. Their SVM algorithm was then used for training, prediction, and further calculations to calculate accuracy.to achieve an accuracy of 0.699 using German credit data, the UCI Knowledge Base, and the Australian data set. Using the AzureML platform, a third study [3] achieved accuracy ranging from 0.61 to 0.61 with 2-jungle and 2-decision tree algorithms on 10 years of loaned club records. Other studies [4][5][6] used models such as linear discriminant analysis, XGBoost, Random Forest, and SVMs, achieving accuracies ranging from 0.70 to 0.72. These studies compared different supervised machine learning classification algorithms and techniques, such as artificial neural networks, ensemble classifiers, decision trees, and KNN classification, to predict loan defaults.

In addition to loan default prediction, machine learning has also been used in developing recommender systems for stock markets [7] the author used KNN classification to develop a recommender system for stock market data. They used collaborative filtering techniques and compared them to existing models. and product sales in banks [8]. The authors developed a hybrid recommendation system that combines collaborative filtering techniques and a demographics-based approach to product sales in banks. They tested their model on real-world data sets. These studies utilized collaborative filtering techniques and hybrid approaches to achieve their goals.

Paper [9][10] describes a scalable and accurate implementation of gradient boosting machines, demonstrating its effectiveness on several benchmark datasets. [11] introduces random forests, an ensemble learning method that utilizes multiple decision trees to improve prediction accuracy and robustness. [12] investigates the use of multilayer perceptron (MLP) neural networks for credit risk analysis, showing their effectiveness in predicting credit risk when compared to other machine learning algorithms. [13] introduces decision trees as a simple and intuitive machine learning method for decision-making, and explores their applicability in various domains, including finance. Lastly, the [14] paper provides a comprehensive overview of recommendation systems in banking and finance, discussing the challenges and opportunities of their application in this domain, and providing successful implementation examples.

In summary, machine learning techniques have shown great promise in predicting loan defaults and developing recommender systems for financial services. The studies reviewed highlight the importance of selecting appropriate models and algorithms for different financial services tasks and data sets. While these studies have shown promising results, further research is needed to investigate the effectiveness of machine learning techniques in real-world financial services application.

REFERENCES	DATASET USED	MODELS USED	ACCURACY	REMARK
Mohammad Ahmad Sheikh, et.al. IEEE Xplore [1]	Kaggle	Logistic regression	0.66	Single model on a smaller dataset is used, the accuracy can be improved.
Khaldoon Alshouiliy, et.al. IEEE Xplore [2]	German credit data set and Australian credit data set	SVM, KDE	0.699	Performance of KD SVM and find out whether its performance can be improved significantly.
Xingzhi Zhang, et.al. IEEE Xplore [3].	Lending Club Dataset	AzureML platform with Two Jungle algorithm and the Two Decision tree.	0.61	General data has led to leakage of important information or inaccuracy overall.
Evander ET Nyoni, et.al. Research gate [4]	Kaggle	Linear discriminant, Logistic regression, XGBoost, Random forest	0.70	Single model has been used due to which less accuracy has been observed
Aiman Muhammad Uwais, et.al. Ceur WS[5]	Kaggle	Logistic regression, decision tree, random forest, Gradient boosted tree, factorization method, LSVM	0.72	Outlier treatment has not been performed
Sunil Bhatia ,et.al. International Journal of Computer Applications (0975 – 8887) Volume 161 – No 11 [6]	Data extracted from University of Tennessee	ANN-MLP	0.65	Data is biased because of redundant dataset due to which overfitting of model has been observed.
Bhagirathi Nayak. Research gate [7]	Data collected from NSE and 10 different sectors.	KNN	unpredictable.	Hybrid approaches can be used to increase applications in different companies using wider datasets
Oladapo Oyebode,	Kaggle	Collaborative featuring	unpredictable.	Hybrid filtering used to improve

et.al. Research gate [8]				recommendation systems accuracy in which various parameters need to be monitored such as demographic, choice based, collaborative filtering, etc.

III. METHODOLOGY AND RESULTS

The project aims to predict whether a personal loan, home loan, or vehicle loan will be approved based on specific data, and the approach to making a prediction in each of these categories will be customized according to the type of loan required. We started importing the necessary packages like numpy, pandas, seaborn and scikit-learn. Load the dataset into a pandas dataframe and split it into training and test sets using the scikit-learn's train_test_split() function. The split ratio can be set based on the size of the dataset and the needs of the model. We used the default 70:30 split. Data visualization is done by visualizing the dataset to gain insight and understanding of the data. Perform univariate and bivariate analysis using Seaborn's countplot, barplot, scatterplot, heatmap and other visualization techniques. Identifying patterns, trends, and relationships between features in the dataset. Univariate analysis [22] is the most basic form of data analysis technique.

When we want to understand the data contained in only one variable and do not want to deal with the cause or effect relationships, a univariate analysis technique is used. Bivariate analysis [22] is slightly more analytical than univariate analysis. When the dataset contains two variables and researchers want to make comparisons between the two datasets, bivariate analysis is the right type of analysis technique. Data cleansing is done by pre-processing the data by handling missing values, outliers, and irrelevant columns. Used pandas drop() function to remove the unnecessary columns and use techniques like mean, median or mode to fill in missing dataset. Logistic regression [14] is a machine learning classification algorithm used to calculate the probability of a categorical dependent variable to predict. In logistic regression, the dependent variable is a binary variable containing data encoded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X. XGBoost[9][10] is an open source software library that implements optimized distributed gradient boost machine learning algorithms under the gradient boost framework.Random Forest is an ensemble technique capable of performing both regression and classification tasks using multiple decision trees and a technique called bootstrap and aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees to determine the final output instead of relying on individual decision trees. Random Forest [11] has multiple decision trees [13] as basic learning models. We randomly perform row sampling and feature sampling from the dataset that makes up sample datasets for each model.

A decision tree is a flowchart-like tree structure in which an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the result. The top node in a decision tree is called the root node. It learns to partition based on attribute value. Model rating is each model's performance using rating metrics such as accuracy, precision, recall, and F1 score. Use the scikit-learns function category_report() to generate a report on each model's performance. Hyperparameter tuning Selects the best performing model and tunes its hyperparameters using techniques such as GridSearchCV or RandomizedSearchCV to optimize its performance. and finally use the best model to make predictions on the test dataset and evaluate its performance.With best of our knowledge the Random Forest model has shown the highest accuracy of around 89% for predicting personal loans. For home loans, the Random Forest and XGBClassifier models have achieved an accuracy of 78%. In the case of vehicle loans, the Random Forest model with SMOTE [23][24] has achieved an accuracy of 70%.

These results suggest that the Random Forest algorithm is a suitable choice for predicting loan defaults in various domains. However, it is important to note that the accuracy of the models may vary depending on the data sets used and the specific features that are considered for each loan type. Therefore, it is necessary to carefully select and compare different models to obtain the most accurate results.

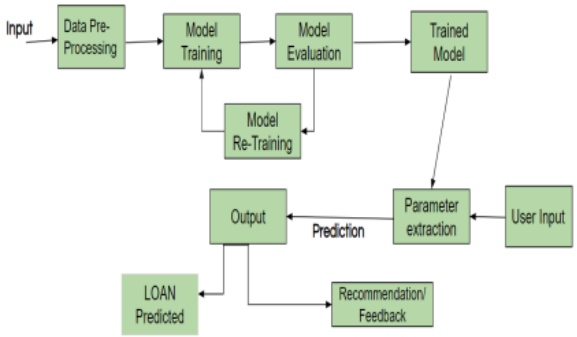

Fig. 1 Proposed Methodology for Personal/Home/Vehicle Loan Prediction Model

3.1 METHODOLOGY FOR PERSONAL LOAN PREDICTION MODEL

A dataset from Kaggle that contains information about loans, including loan amount, interest rate, borrower income, credit score, and loan status (whether the loan was paid in full or defaulted) has been used which contains 252000 rows and 13 columns. Several machine learning models, including logistic regression, decision trees, and random forests are used to predict loan defaults. They compare the performance of these models using various evaluation metrics, such as accuracy, precision, recall.

These models also use feature importance techniques to identify the most significant factors that affect loan defaults. A heatmap to visualize the correlation between these features and loan defaults. which actually helps us in the recommendation part of the project. The results show that the random forest model outperformed the other models in terms of accuracy, precision. Accuracy of the personal loan model we achieved is 88.7619.

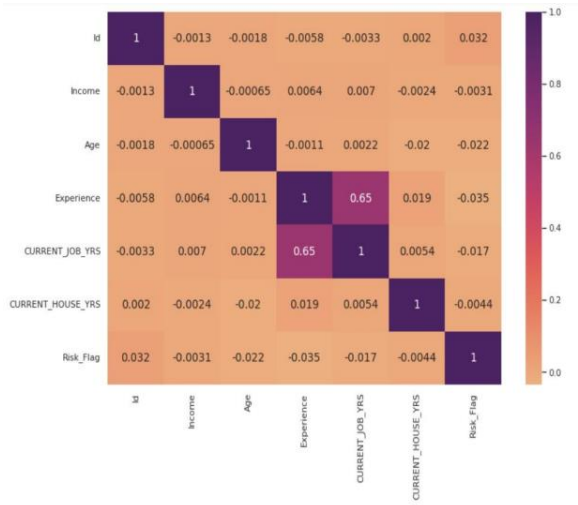

Fig. 2 Heatmap of the personal loan dataset

3.2 METHODOLOGY FOR HOME LOAN PREDICTION MODEL and Sample

This model is designed based on binary classification problems using Python with the main motive of automating the loan eligibility process. Specific customer details are provided while filling online applications and these common details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. A Standard supervised classification task is carried out. A classification problem where we have to predict whether a loan would be approved or not. We used 12 independent variables and 1 target variable, i.e. Loan_Status in the train dataset.

We have 12 independent variables and 1 target variable, i.e. Loan_Status in the training dataset. We have used various model based on the kaggle dataset and after all the data visualization the model accuracy were as follow Decision Tree model gives 71% prediction accuracy, Random Forest model gives 78% prediction accuracy, Random Forest with Grid Search model gives 77% prediction accuracy and XGBClassifier model gives 78% prediction accuracy.

Fig.3. Heatmap of the home loan dataset

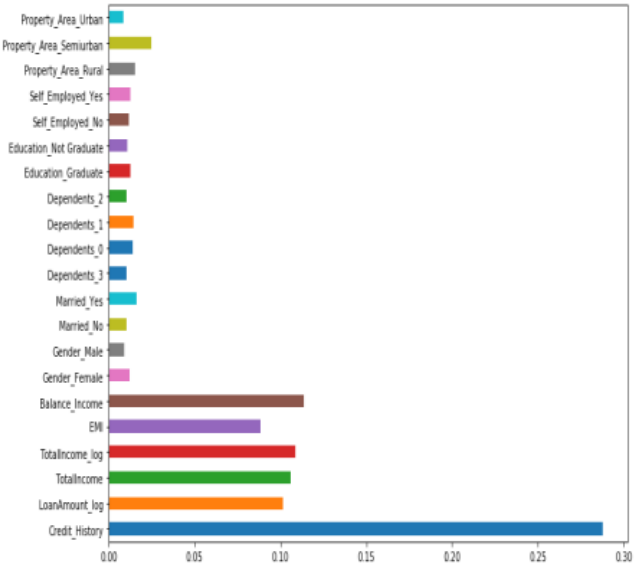

Fig. 4 Important features

3.3 METHODOLOGY FOR VEHICLE LOAN PREDICTION MODEL

A dataset from Kaggle that contains information about loans, including loan amount, interest rate, borrower income, credit score, and loan status (whether the loan was paid in full or defaulted) there are 233154 Rows and 41 columns. We have used several machine learning models, including logistic regression, decision trees, and random forests, to predict loan defaults. They compare the performance of these models using various evaluation metrics, such as accuracy, precision, recall, and F1 score. They also use feature importance techniques to identify the most significant factors that affect loan defaults. Which helps us in the recommendation part of the model. The results show that the random forest model outperforms the other models in terms of accuracy, precision, and F1 score.

The most important features that affect loan defaults are credit history, loan amount, and applicant income. The model creates visualizations to show the correlation between these features and loan defaults. We have also used various methods to handle the imbalanced data, SMOTE (Synthetic Minority Over-sampling Technique) working with Random-forest gave us the best results. Accuracy of the model is 0.708040968278749. Its F1 Score is 0.2505064740597199, Recall Score is 0.2267219387755102 and Balanced Accuracy Score is 0.5333740855753166.

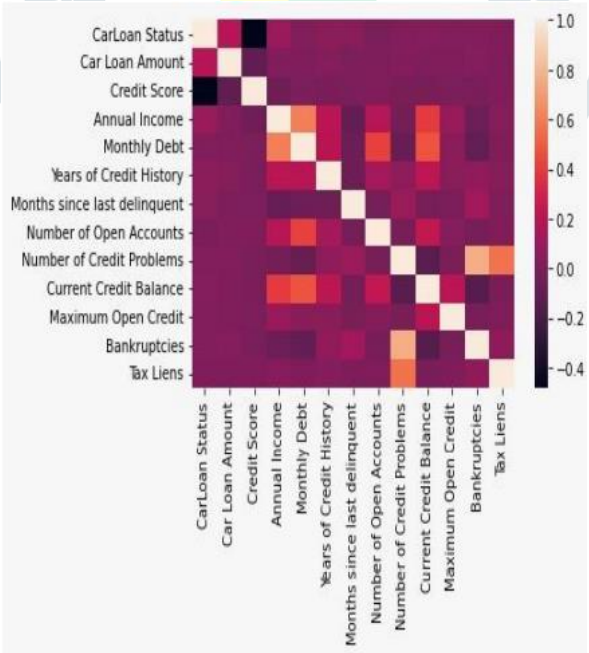

Fig. 5 Heatmap of the vehicle loan dataset

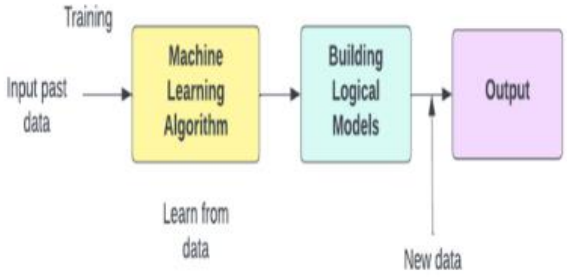

Fig. 6 Model Flowchart

3.4 RECOMMENDATION SYSTEM and Sample

A recommendation system provides personalized suggestions. One approach to building a recommendation system is to use Pearson's correlation theorem [25], which is a statistical method for measuring the linear correlation between two variables. The first step in building a recommendation system using Pearson's correlation is to gather data on user behavior, such as their past purchases, ratings, or reviews. This data is used to calculate the correlation between users based on their shared preferences. For example, if two users have rated similar products highly, their correlation coefficient will be high, indicating that they have similar preferences. Once the correlation coefficients have been calculated, they can be used to create a heatmap that visualizes the similarities between users.

The heatmap can be used to identify clusters of users with similar preferences, which can then be used to make personalized recommendations. For example, if a user belongs to a cluster of users who prefer romantic comedies, the recommendation system might suggest similar movies to that user. In addition to the heatmap, the recommendation system can also identify important features that are strongly correlated with user preferences. For example, if users who enjoy action movies also tend to enjoy movies with a high level of special effects, this feature can be used to improve the accuracy of the recommendations. Other important features might include the actors, directors, or genres of movies, or the specific products that users have purchased or rated highly.

Overall, a recommendation system based on Pearson's correlation theorem and a heatmap can provide a powerful tool for making personalized recommendations to users. By identifying clusters of users with similar preferences and important features that are strongly correlated with user behavior, the system can provide recommendations that are tailored to each individual user, improving their satisfaction and engagement with the platform.

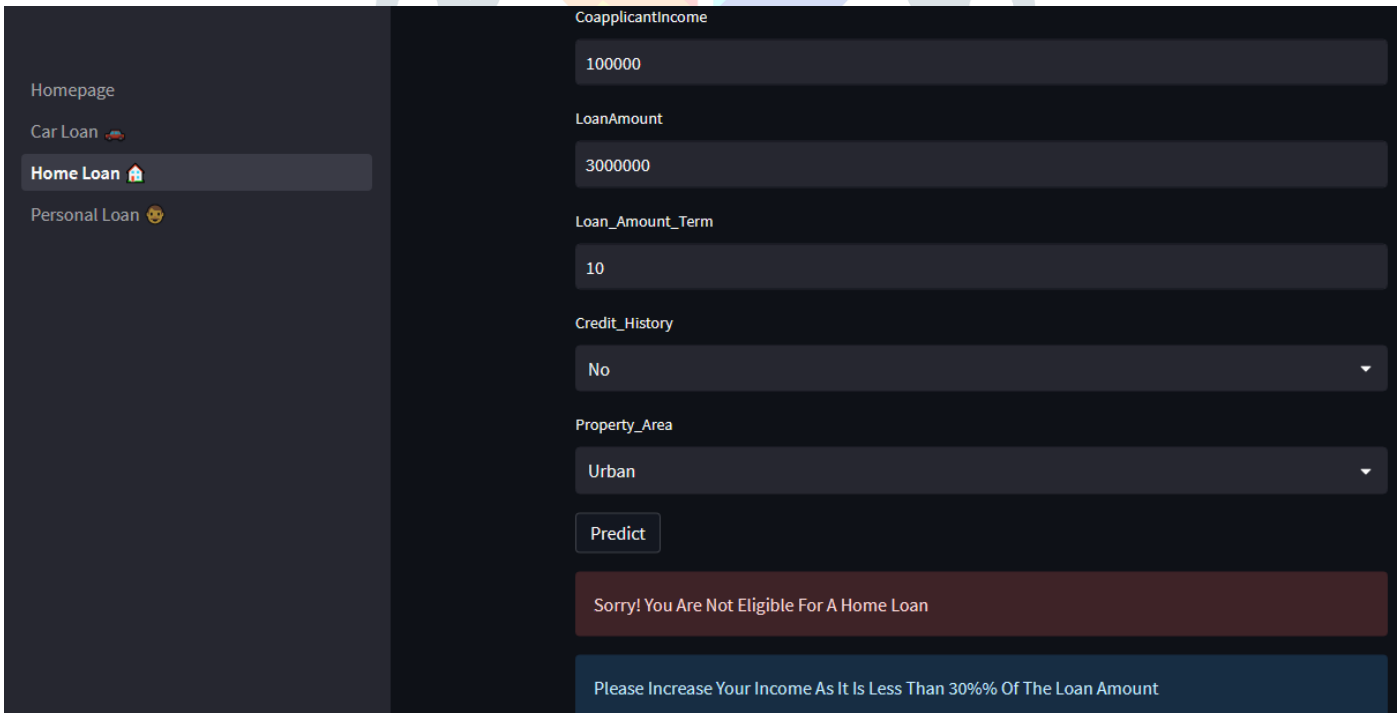

Fig.7. Recommendation System Output

IV. CONCLUSION

In this document, our goal is to provide a one-stop solution for all credit-related financial skills, credit prediction and systems to improve credit scores. Here we used machine learning algorithms to prognosticate whether borrowers will pay or not. It allows financial institutions (lenders) to be notified in advance of defaults on originated loans, which helps them reduce financial losses and the costs associated with loan recovery, thereby increasing profits. The most important part of the project is to optimize the trained models to maximize profit. Classification thresholds are malleable to change the rigor of the prediction results: lower

thresholds make the model more aggressive, allowing more credit to be made; with higher thresholds, it becomes more conservative and won't issue the loans unless there is a high probability that the loans can be repaid.

Of the 6 classification models used are KNN, Gaussian Naive Bayes, Logistic Regression, Linear SVM, Random Forest and XGBoost. The ensemble model from Random Forest and XGBoost model showed the highest accuracy of 89%. We performed exploratory data analysis on the features of this dataset and saw how each feature is distributed. Then we did a bivariate and multivariate analysis to see the impact of one another on their features using maps. Further we also analyzed each variable to check that the data is gutted and typically distributed. We cleaned the data. We also generated a hypothesis to prove an association between the independent variables and the target variable. And based on the results, we assumed whether there was a connection or not. We calculated the correlation between independent variables and found that the applicant's income and the loan quantum have significant relation. Also, we created dummy variables to build the model. We constructed the model considering various variables and using the odds ratio, found that credit history has the most impact on the lending decision. After the analysis, the following conclusions are drawn that those applicants whose creditworthiness was the worst will not receive a due loan approval at a higher probability of not repaying the loan amount. In most cases, applicants with high incomes and requests for a smaller loan amount are more likely to get approval, which makes sense, and are more likely to pay off their loans. Some other characteristics such as gender and marital status do not seem to be taken into account by the company. Finally, we obtained a model with the co-applicant's income and credit history as the independent variable with the highest accuracy.

The other aim of this paper was to introduce a recommendation system into the credit approval process of banks that uses a machine approach. In our project we used a hybrid approach between context-based filtering and collaborative filtering to implement the system. This approach overcomes the downsides of each algorithm and improves the performance of the system. Techniques such as clustering, similarity, and classification are used to provide better recommendations, thereby increasing precision and accuracy. The project on loan default prediction and recommender system development for financial services presents numerous future scope and applications. Its key advantage lies in reducing the possibility of non-performing loans. By identifying loans that are likely to default, banks can prevent such events from occurring, thereby minimizing their financial losses.

Moreover, the referral system developed as part of this project can improve financial literacy and creditworthiness, leading to better loan terms and rates for individuals and businesses. This can enhance financial inclusion and contribute to economic growth. Another benefit is the reduced paperwork for banks, as machine learning can streamline and optimize loan application processes, saving time and resources.

In summary, this project has the potential to transform financial services, providing faster, more efficient issuing processes, improved creditworthiness, and reduced paperwork. This, in turn, can lead to a better customer experience and help more people achieve their financial objectives. Ongoing research and development in this area will result in further innovative applications and benefits for the financial industry.

REFERENCES

[1] Mohammad Ahmad Sheikh, Amit Kumar Goel and Tapas Kumar. An Approach for Prediction of Loan Approval using Machine Learning Algorithm. International Conference on Electronic and Sustainable Communication System, IEEE Xplore 2020.

[2] Khaldoon Alshouiliy and Ali AlGamdi. AzureML Based Analysis and Prediction Loan Borrower Creditworthy.International Conference on Information and Computer Technology(ICICT), IEEE Xplore 2020.

[3] Xingzhi Zhang and Zhurong Zhou. Credit Scoring Model based on Kernel Density Estimation and Support Vector Machine for Group Feature Selection.IEEE Xplore 2018.

[4] Evander ET Nyoni,Ntandoyenkosi Matshisela.Credit scoring using machine learning algorithms, Research gate Nov 2018.

[5] Aiman Muhammad Uwais and Hamidreza Khaleghzadeh. Loan Default Prediction Using Spark Machine Learning Algorithms,https://ceur-ws.org/Vol-3105/paper30 .pdf,2021

[6] Sunil Bhatia ,Pratik Sharma ,Rohit Burman ,Santosh Hazari ,Rupali Hande. Credit Scoring using Machine Learning Techniques ,International Journal of Computer Applications (0975 – 8887) Volume 161 – No 11, March 2017

[7] Bhagirathi Nayak. Machine Learning Finance: Application of Machine Learning in Collaborative Filtering Recommendation System for Financial Recommendations, Research gate July 2019.

[8] Oladapo Oyebode, Rita Orji. A hybrid recommender system for product sales in a banking environment, Research gate Jan 2020.

[9] Hongwei Chen1, He Ai, Zhihui Yang, Weiwei Yang , Zhiwei Ye, Dawei Dong. An Improved XGBoost Model Based on Spark for Credit Card Fraud Prediction, IEEE Xplore 2020.

[10] Tianqi Chen, Carlos Guestrin.Xgboost: A Scalable Tree Boosting System,22nd Acm Sigkdd International Conference On Knowledge And Data Mining 2016.

[11] Leo Breiman.Random Forests,Springer Link 2001.

[12] Javier Bajo, María L. Borrajo , Juan F. De Paz , Juan M. Corchado , María A. Pellicer.A Multi-Agent System For Web-Based Risk Management In Small And Medium Business, ScienceDirect 2012..

[13] J.R. Quinlan.Decision Trees And Decision-Making, IEEE Xplore 1990.

[14] Xiaonan Zou; Yong Hu; Zhewen Tian; Kaiyuan Shen. Logistic Regression Model Optimization And Case Analysis, IEEE Xplore 2020.

[15] Johan Eko Purnomo, Sukmawati Nur Endah. Rating Prediction On Movie Recommendation System: Collaborative Filtering Algorithm (Cfa) Vs. Dissymmetrical Percentage Collaborative Filtering Algorithm (Dspcfa), IEEE Xplore 2019.

[16] Ya-Qi Chen, Jianjun Zhang, Wing W. Y. Ng. Loan Default Prediction Using Diversified Sensitivity Undersampling, IEEE Xplore 2018.

[17] Anshika Gupta1 , Vinay Pant2 , Sudhanshu Kumar3 And Pravesh Kumar Bansal. Bank Loan Prediction System Using Machine Learning, IEEE Xplore 2020

[18] Du Shaohui, Guanwen Qiu, Huafeng Mai, Hongjun Yu3. Customer Transaction Fraud Detection Using Random Forest, IEEE Xplore 2021.

[19] Wang Zhen, Sun Wenjuan. Commercial Bank Credit Risk Assessment Method Based On Improved Svm, IEEE Xplore 2016.
[20] Sourish Ghosh; Anasuya Dasgupta; Aleena Swetapadma. A Study On Support Vector Machine Based Linear And Non-Linear Pattern Classification, IEEE Xplore 2019.
[21] Yuguang Huang; Lei Li.Naive Bayes Classification Algorithm Based On Small Sample Set, IEEE Xplore 2016.
[22] A.B. Badiru.Computational Survey Of Univariate And Multivariate Learning Curve Models, IEEE Xplore 1992.
[23] Widi Satriaji; Retno Kusumaningrum,Effect Of Synthetic Minority Oversampling Technique (Smote), Feature Representation, And Classification Algorithm On Imbalanced Sentiment Analysis, IEEE Xplore 2019).
[24] Linmao Feng. Research On Customer Churn Intelligent Prediction Model Based On Borderline-Smote And Random Forest, IEEE Xplore 2022..
[25] Bin Wang; Qing Liao; Chunhong Zhang. Weight Based KNN Recommender System, IEEE Xplore 2013.