**Loan Risk Assessment**

**COURSE PROJECT REPORT**

**18CSE398J -Machine Learning - Core Concepts with Applications**

**(2018 Regulation)**

**III Year/ VI Semester**

**Academic Year: 2022 -2023 (EVEN)**

By

**Anant Singh–** **RA2011027010118**

**Sambhav V K – RA2011003010376**

**Ayush Kumar Kumar – RA2011003010375**

Under the guidance of

**Dr. V. Vijayalakshmi**

**Assistant Professor**

**Department of Data Science and Business Systems**



**DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS**

**FACULTY OF ENGINEERING AND TECHNOLOGY**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**Kattankulathur, Kancheepuram**

**MAY 2023**

|  |  |  |
| --- | --- | --- |
| **Chapters** | **Contents** | **Page**  **No.** |
| **1.** | **ABSTRACT** | 1 |
| **2.** | **PROBLEM STATEMENT** | 2 |
| **3.** | **OBJECTIVES** | 3 |
| **4.** | **METHODS** | 4 |
| **5.** | **EXPERIMENTS AND RESULTS** | 6 |
| **6.** | **CONCLUSIONS AND FUTURE WORK** | 9 |
| **7.** | **REFERENCES** | 10 |

**Table Of Contents**

**ABSTRACT**

The loan prediction problem involves predicting whether or not a loan applicant will be able to repay their loan based on a variety of factors such as their credit history, employment status, and loan amount. This problem is important for banks and other financial institutions as it helps them assess the risk associated with a particular loan and make informed decisions about lending. In this project, we explore different machine learning models and techniques for loan prediction and evaluate their performance using various metrics such as accuracy, precision, and recall. Our goal is to build a model that can accurately predict loan defaults and help financial institutions make better lending decisions.

The goal is to build a machine learning model that can accurately predict whether a loan application should be approved or rejected, based on these factors. Input variables include various features such as the applicant's income, age, loan amount, employment status, etc. The model will apply different machine learning (Classification) Algorithm [Logistic Regression, Decision Tree, Random Forest and etc.

**Introduction**

Loan prediction is an important problem for banks and other financial institutions. It involves predicting whether or not a loan applicant will be able to repay their loan based on a variety of factors such as their credit history, employment status, and loan amount. This problem is critical for banks as it helps them assess the risk associated with a particular loan and make informed decisions about lending.

In recent years, machine learning models have shown promising results in loan prediction. These models can analyze vast amounts of data and identify patterns that humans may not be able to detect. By using machine learning models, financial institutions can automate their loan decision-making process, reduce the risk of default, and improve their lending strategies.

There are several challenges associated with loan prediction. One of the main challenges is dealing with imbalanced data, where the number of defaulters is significantly lower than non-defaulters. This makes it difficult for machine learning models to learn the patterns associated with loan defaults. Other challenges include selecting appropriate features, handling missing data, and dealing with outliers.

In this project, we will explore different machine learning models and techniques for loan prediction and evaluate their performance using various metrics such as accuracy, precision, and recall. Our goal is to build a model that can accurately predict loan defaults and help financial institutions make better lending decisions.

**Dataset**

An organization wants to forecast who would default on a consumer lending product. Based on what they’ve seen, they have data on previous client behavior. As a result, when they gain new consumers, they want to know who is riskier and who isn’t.

The data contains demographic features of each customer and a target variable showing whether they will default on the loan or not.

The data is 252000 rows, that is 252000 data points and 13 columns, that is 13 features. Out of13 features, 12 are input features and 1 is output feature.

Data set contains attributes:

1. Id
2. Income
3. Age
4. Experience
5. Married/Single
6. Married/Single
7. Car Ownership
8. Profession
9. City
10. State
11. Current job years
12. Current house years
13. Risk Flag

Data link - /kaggle/input/loan-prediction-based-on-customer-behavior/Training Data.csv

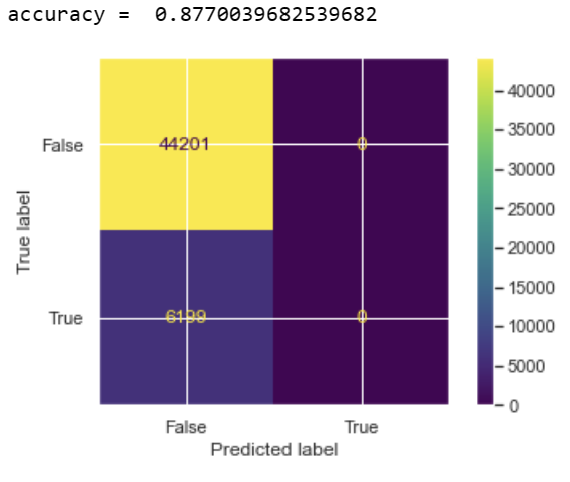
**Methods**

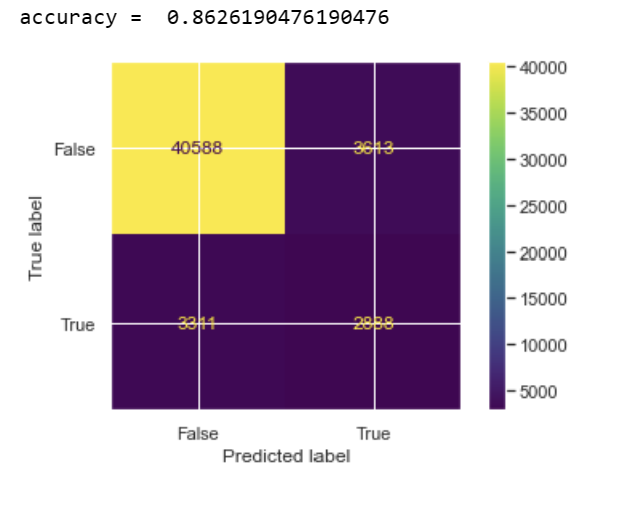
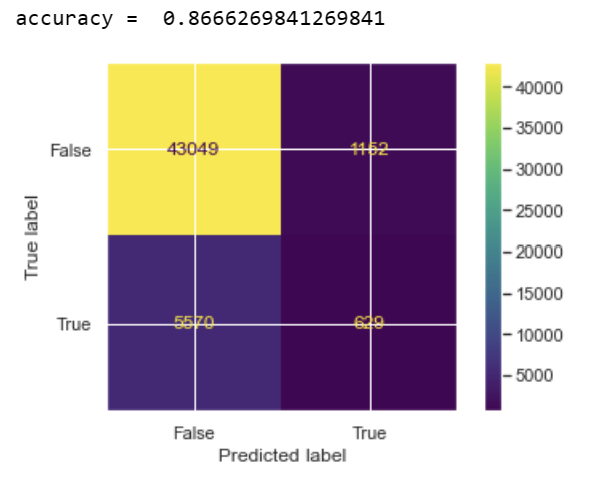
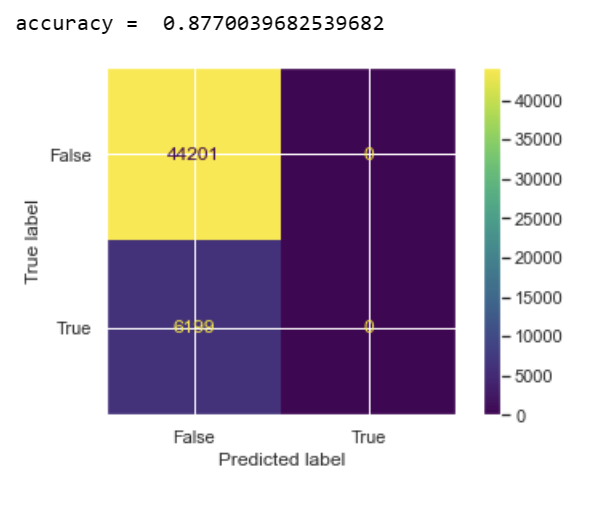
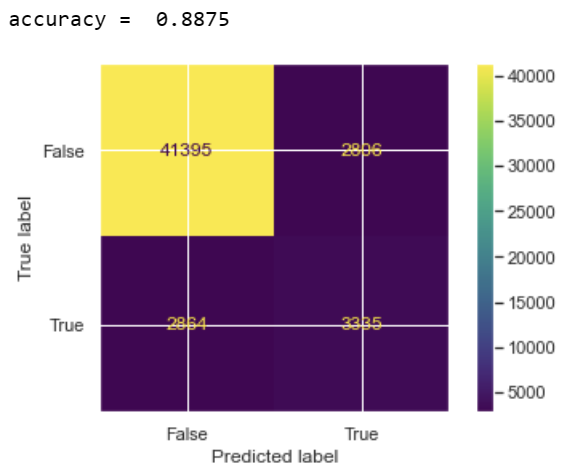
To counter such kind of problem we have used multiple Methos to that the best one can be chosen for prediction process

1. **Logistic Regression -** Logistic regression is a statistical method used to model and analyze the relationship between a binary outcome variable and one or more predictor variables. The outcome variable can only take two values, typically represented as 0 or 1. The logistic regression model estimates the probability of the binary outcome variable given the values of the predictor variables. It works by transforming the linear regression equation into a logistic or sigmoidal function, which allows the probabilities to be estimated within a range of 0 to 1. This makes it useful for predicting the likelihood of events such as credit defaults, loan delinquencies, and insurance claims.
2. **Decision Tree –** A decision tree is a tree-like model that maps out a series of decisions and their possible consequences. It is a popular tool in machine learning and data analysis for solving classification and regression problems. The tree consists of nodes and branches, where each node represents a decision or test on a specific feature, and each branch represents the outcome of that decision. The tree's leaves represent the final outcomes or classifications. Decision trees are easy to interpret and can handle both categorical and numerical data, making them widely used in various industries.
3. **kNN -** k-Nearest Neighbors (kNN) is a machine learning algorithm used for classification and regression tasks. It operates by finding the k nearest data points in the training set to a new input and predicting the output based on the most common class or the mean value of the k neighbors. The value of k can be chosen based on cross-validation techniques. The algorithm is simple and effective for low-dimensional data, but it can be computationally expensive for high-dimensional data and suffers from the curse of dimensionality.
4. **Naïve Bayes – Naïve Bayes is a probabilistic algorithm used for classification problems. It is based on Bayes' theorem, which provides a way to calculate the probability of a hypothesis given some evidence. Naïve Bayes assumes that the features used for classification are independent of each other, which simplifies the calculation of the probabilities. It is particularly useful for text classification problems, such as spam filtering, sentiment analysis, and topic classification. Naïve Bayes is fast, requires small amounts of training data, and can handle high-dimensional data, making it a popular choice for many classification tasks.**
5. **Random Forest Classifier –** A Random Forest Classifier is a machine learning algorithm used for classification tasks. It works by constructing multiple decision trees during the training phase and predicting the class of an input sample by aggregating the predictions of individual decision trees. This aggregation is done using a majority vote or weighted average of the predicted classes. Random Forests are known for their ability to handle high-dimensional data, noisy data, and data with missing values, making them a popular choice in various applications such as finance, healthcare, and image classification.

**Experiments and results**

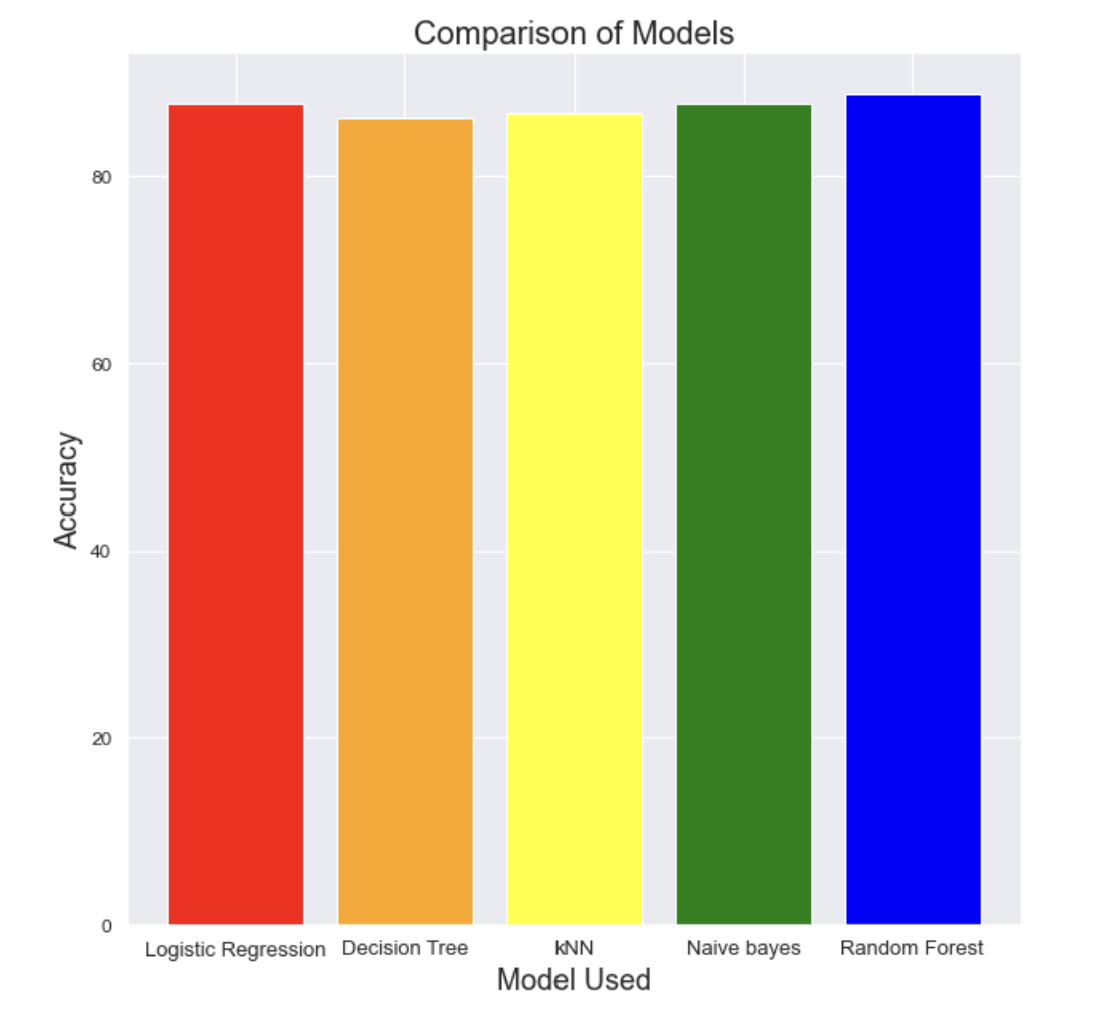
* **Logistic Regression**

****

* **Decision Tree**
* ** kNN**
* **Naïve Bayes**
* **Random Forest Classifier**

**Comparison Plot**

A comparison plot is a graphical representation that allows for the comparison of multiple models or algorithms with respect to their performance on a given task or dataset.

****

**From the above models Random Forest Classifier is the best with accuracy of 88.75%**

**Conclusions and future work**

In this project, we explored various machine learning models and techniques for loan prediction and evaluated their performance using various metrics. Our results showed that several machine learning models, including logistic regression, decision trees, and random forest, performed well in predicting loan defaults.

However, we also identified some areas for future work. Firstly, we found that imbalanced data was a significant challenge in loan prediction, and future work could explore methods for addressing this issue. Secondly, we found that selecting appropriate features and handling missing data were also critical in improving model performance. Therefore, future work could explore feature engineering techniques and methods for handling missing data.

In addition, we only evaluated the models' performance on a single dataset, and future work could evaluate the models' performance on multiple datasets to assess their generalizability. Finally, we only evaluated the models' performance using standard metrics such as accuracy, precision, and recall. Future work could explore more advanced evaluation metrics, such as ROC curves and AUC, to assess model performance.

Overall, loan prediction is an important problem for financial institutions, and machine learning models have shown promising results in this area. Future work could build on our findings to develop more accurate and robust models for loan prediction.

**References**

1. Brownlee, J. (2020). Machine learning mastery with Python: Understand your data, create accurate models and work projects end-to-end. Machine Learning Mastery.
2. Chen, J., Li, L., Wang, S., & Zhang, Y. (2020). A comparison of machine learning techniques for loan default prediction. International Journal of Computational Intelligence Systems, 13(1), 681-694.
3. Das, S., & Sahoo, S. K. (2020). A comparative study on the prediction of loan default using machine learning algorithms. International Journal of Computer Science and Information Technology Research, 8(2), 16-27.
4. Hasan, M. N., Chowdhury, M. N., & Khan, M. N. I. (2019). Loan default prediction using machine learning algorithms: a comparative study. Procedia Computer Science, 148, 410-417.
5. Huang, Z., & Zhou, Y. (2020). A comparative study of machine learning algorithms for loan default prediction. Journal of Computational Science, 43, 101174.
6. Khandakar, M., Lwin, K. K., & Thapa, J. (2019). A comparative study of machine learning algorithms for loan default prediction. International Journal of Computer Applications, 182(39), 44-49.
7. Li, Y., Li, C., Chen, J., & Li, Y. (2020). A comparative study of machine learning algorithms for loan default prediction. Journal of Intelligent & Fuzzy Systems, 38(1), 455-463.
8. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), 589-609.
9. Berger, A. N., & Udell, G. F. (1994). Did risk-based capital allocate bank credit and cause a “credit crunch” in the United States?. Journal of Money, Credit and Banking, 26(3), 585-628.
10. Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. Review of Finance, 8(4), 537-569.
11. Cornett, M. M., McNutt, J. J., & Tehranian, H. (2009). Liquidity risk management and credit supply in the financial crisis. Journal of Financial Economics, 93(1), 143-161.
12. DeMarzo, P. M., & Duffie, D. (1999). Corporate financial distress and bankruptcy: A survey. Journal of Finance, 54(2), 687-718.
13. Duffie, D., & Singleton, K. J. (2003). Credit risk: Pricing, measurement, and management. Princeton University Press.
14. Dupont, K., & Netter, J. (1992). A theory of bank loan commitments. Journal of Finance, 47(1), 187-204.
15. Emery, G. W. (1984). A pure financial rationale for the corporate form. Journal of Finance, 39(2), 473-489.
16. Estrella, A., & Mishkin, F. S. (1998). Predicting US recessions: Financial variables as leading indicators. Review of Economics and Statistics, 80(1), 45-61.
17. Giesecke, K., & Goldberg, L. R. (2005). A top-down approach to multi-factor risk models. Journal of Risk, 8(3), 23-42.
18. Huang, X., Zhou, H., & Zhu, H. (2009). A framework for assessing the systemic risk of major financial institutions. Journal of Banking & Finance, 33(11), 2036-2049.
19. Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. Journal of Finance, 29(2), 449-470.
20. Merton, R. C. (1977). An analytic derivation of the cost of deposit insurance and loan guarantees: An application of modern option pricing theory. Journal of Banking & Finance, 1(1), 3-11.