**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.Vijayalakshmi V**

**Associate 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**

**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

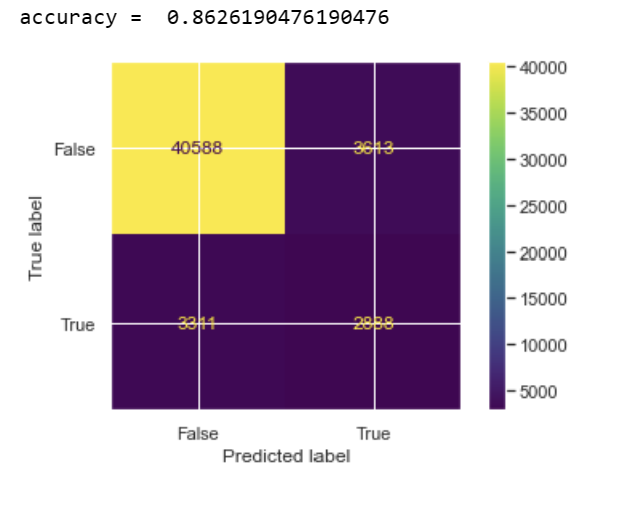
**Methods**

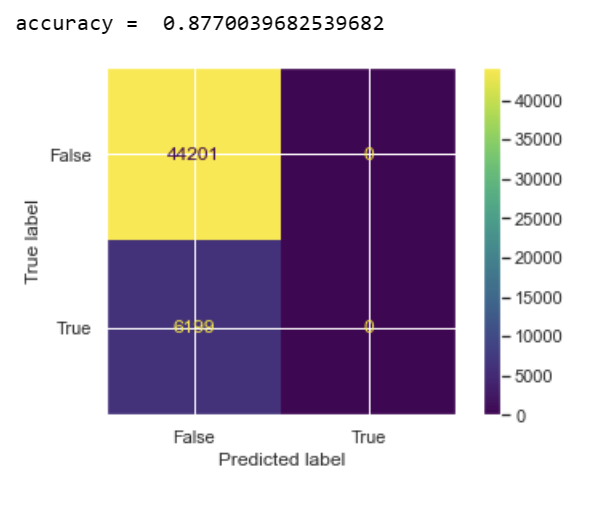
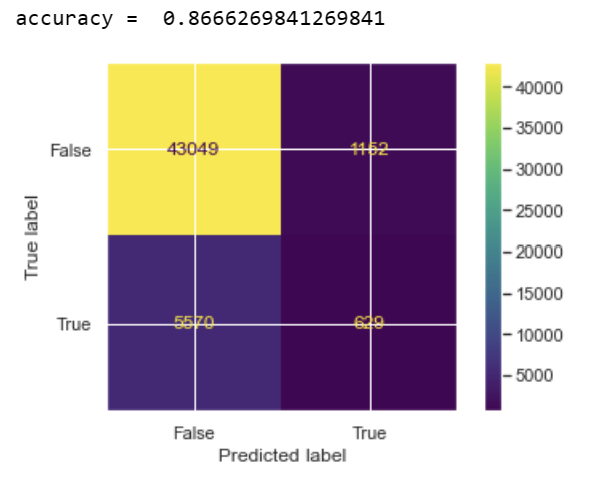
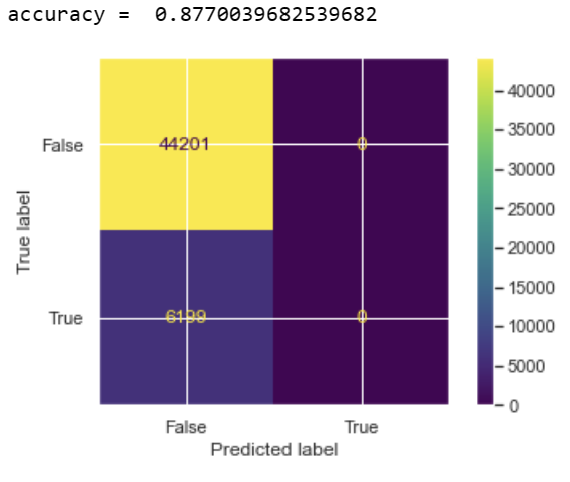
To counter such kind of problem we have used multiple Methos to that the best one can be chosen for prediction process

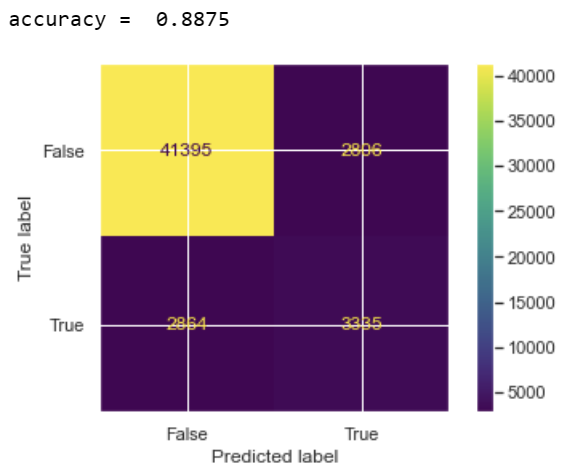
1. **Logistic Regression –** It’s a method in which sigmoid function is a mathematical function used to map the predicted values to probabilities.
2. **Decision Tree –** It’s a method in whichinternal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
3. **kNN -** Algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories
4. **Naïve Bayes –** It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
5. **Random Forest Classifier – is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.**

**Experiments and results**

* **Logistic Regression**

****

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

****

**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.