CUSTOMER LOAN DEFAULT PREDICTION

Interim Progress Report

7COM1039-0109-2021-Advanced Computer Science Masters Project

Student Number: 19056872

Student Name: Vamshi Krishna Thodupunuri

Supervised by: Imran khan (imy)

**INTRODUCTION**

Throughout the 10 years, arising and creating economies experience different monetary framework developments, especially in the financial area. These incorporate portable banking, Internet banking, specialist banking, and other advanced monetary administrations. Yet, a huge piece of the populace is outside the monetary consideration . Most banks monetary foundations actually utilize monetary reimbursement history-based credit choice frameworks where the vast majority in the market don't have a FICO rating. Thus, populaces with low openness to banking administrations face the gamble of being totally barred from monetary environments, mostly from admittance to credit. Credit hazard or advance default risk is one of the huge monetary difficulties in banking and monetary foundations. Credit risk alludes to the vulnerability encompassing their borrowers' capacity to reimburse an advance or satisfy authoritative commitments, bringing about credit default or insolvency. Retail banks (giving advances to people or retail buyers) and business moneylenders (offering credits to associations or ventures) are both presented to credit risk .These days, banks and monetary associations progressively use AI applications in their everyday decision-production. The advance default risk generally relies upon the borrower's profile and credit boundaries, including the dissolvability, credit type, development, the credit sum, and different elements inborn in monetary tasks.

**Feature Selection and Extraction**

Ability choice is a significant and strong portion wherein a helpful information base can be extremely enormous. This makes finding examples and finding connections between components undeniably challenging to register. In Machine learning, the trait of a quantifiable property or capacity, or the quality of a specific article. Picking autonomous, particular, and useful undertakings is a significant stage for productive calculations. Flexible selection and feature selection are marked with function outputs. So most of the machine learning algorithms are tend to give a good accuracy, when they are trained with appropriate data. with regards the loan default analysis having columns like customers credit score, income etc, would be considered as most important data at a customer level that could help algorithm to classify the points correctly.

**Data Set details**

This deals with problems where most of the banks are currently facing.

The data set to build a application have been taken from Kaggle

Url: https://www.kaggle.com/datasets/yasserh/loan-default-dataset?resource=download

There is loan information of 148670 customers in the above data set.

**Data Description**

'

ID' = customer id

'year' = Year of application

'loan\_limit' = loan limit

'Gender' = gender of customer

'approv\_in\_adv' = approved in advance

'loan\_type' = type of loan

'loan\_purpose' = different reason to need a loan

'Credit\_Worthiness' = to determine requirement the credit of customer

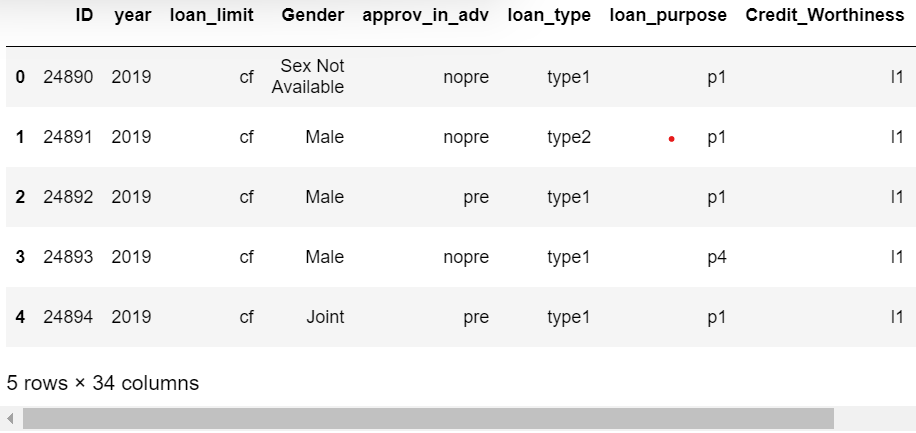
'business\_or\_commercial' = type of loan is business or commercial

'loan\_amount', 'rate\_of\_interest', 'Interest\_rate\_spread', 'Upfront\_charges', 'term', 'Neg\_ammortization',

'interest\_only', 'lump\_sum\_payment', 'property\_value' 'construction\_type', 'occupancy\_type', 'Secured\_by',

'total\_units', 'income', 'credit\_type', 'Credit\_Score', 'co-applicant\_credit\_type', 'age', 'submission\_of\_application',

'LTV', 'Region', 'Security\_Type' ,'Status', 'dtir1'



**Literature review**

(Vinay Padimi, et al.., 2022) proposed In distributed (P2P) loaning, borrowers would get to advances with lower financing costs than what they typically got from conventional loan specialists. Individuals can straightforwardly acquire from the P2P stage with the guidelines that make them simple to acquire advances and put free assets into P2P, which can help the two borrowers and loan specialists. Be that as it may, the simple way to acquire advances accompanies chances. One of the significant issues is that borrowers might default on the advance taken. In such cases, they can get advances rapidly from P2P online stages with practically no bank obstructions. Along these lines, the loan specialist can ascertain his gamble for advance default. In this undertaking, we consider the P2P loaning information to anticipate the advance default consoling the moneylender to keep giving advances from here on out. In our investigation, we think about the Logistic Regression, naive Bayes, Random Forest, K Nearest Neighbor, and Decision tree to characterize advance information in light of their probability.

Here we are going to try various machine learning algorithms like decision tree, random forest, naïve bayes, with best metrics that could enable algorithms giving best accuracy. We may be also performing hyper parameter tuning for the necessary parameters of the above algorithms for better accuracy. Then we compare results of our proposed model with existing application.

**RESEARCH QUESTION**

1.What are the most correlated features among the customer inputs used in data?

2. Which machine learning algorithm that provides higher accuracy in this project?

3.Choosing right metrics and choosing which metric giving better accuracy?

**OBJECTIVES**

1. Understand the patterns using the exploratory data analysis.

2. feature engineering - handling null values (null imputation) - handling categorical variables. - detecting and removing anomalies(outliers) - converting imbalanced data set into balanced data set

3. feature selection - process to identify which variables are highly significant to perform analysis, and giving Better accuracy.

4. model building - will build various models and find which model provides a better accuracy. 5. hyper metric tuning.

**Legal, ethical and social issues**

The ethical issues that pose a risk in the perspective of this project are identified using the Ethical OS Toolkit. Sharing the details of the customers to other department even with the company would be considered as breach of law. The banks need to take consent from the customers and passing information to the organization to input data for the algorithms.

Where as legal issues are issues that should not violate the guidelines that are given by government, Prior to sending any calculation, the calculation should be tried utilizing numerous exhibition measurements and just when the calculation shows higher correctnesses with low blunder rates then the calculations should be sent to stay away from any legitimate issues because of one-sided expectations.

Social issues can be characterized as the issues that would influence the citizenry because of the undertaking. Once algorithm states that the customer is default he or she could not able to get loan sanction from other banks, they left with an option p2p which does not have any bank interference .

**Project Progress**

This is the weekly progess of my work on project

Week 1: module lectures are started and come to know what need to be done during project.

Week 2: we will gain from the talks and find out about the assumptions, as well as getting a superior feeling of how to move toward the project work.

Week 3: Understanding of project arranging and risk management the board from the talks during the third week.

Week 4: Conducting research for the project proposition on regions that are associated with the course educational plan.

Week 5: During the fifth week, I dealt with creating and presenting a project proposition.

Week 6: Gaining new bits of knowledge into the task issue, doing additionally investigate on subjects associated with the project, and fostering a decent arrangement of examination inquiries for my undertaking were all needs.

Week 7: Conducting a survey of the writing by perusing the exploration articles that are available on the web and taking notes on what I have gained from the distributions.

Week 8: Work on the literature review.

**Project plan**

I need to conduct furthermore research to understand entirely about the data set, I almost done with EDA(exploratory data analysis) which will be finished by next week. And I will start performing data analysis once my research was done. I will finish half of the data analysis by first week of June.

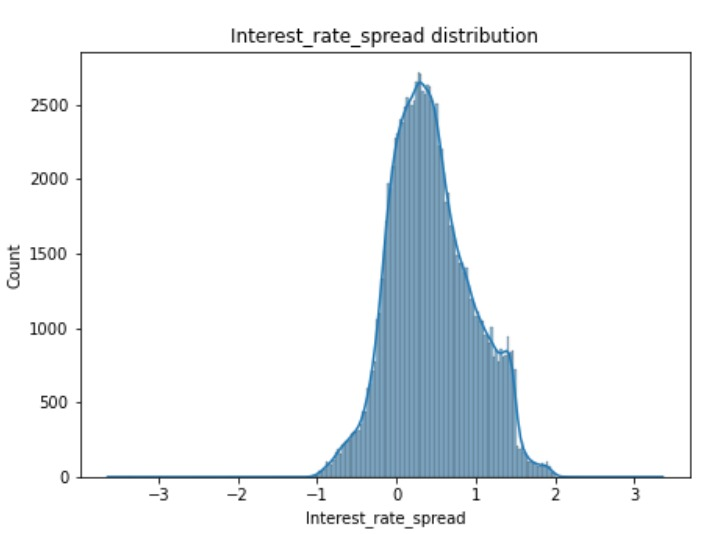
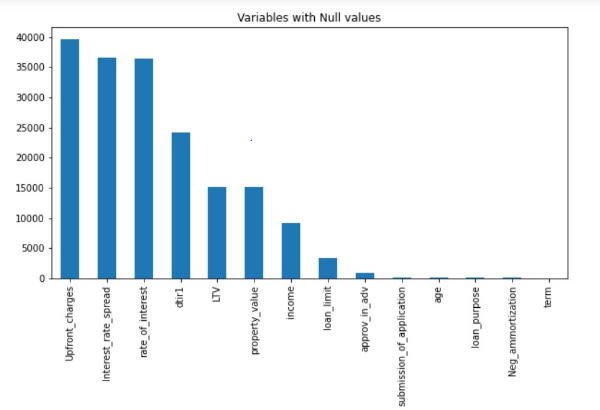
Later on I also need to perform various machine learning algorithms, where I will train various machine learning algorithms to my data, some of the machine learning algorithms such as decision tree, random forest ,naïve bayes , logistic regression.

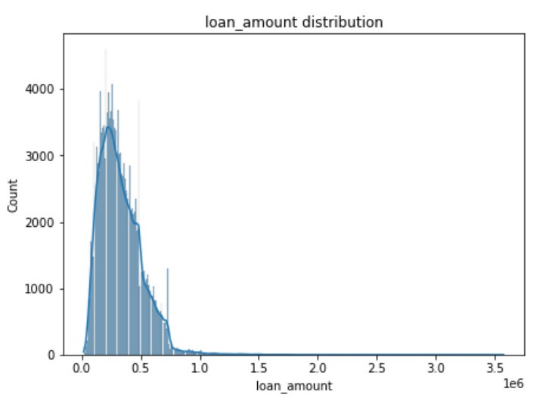
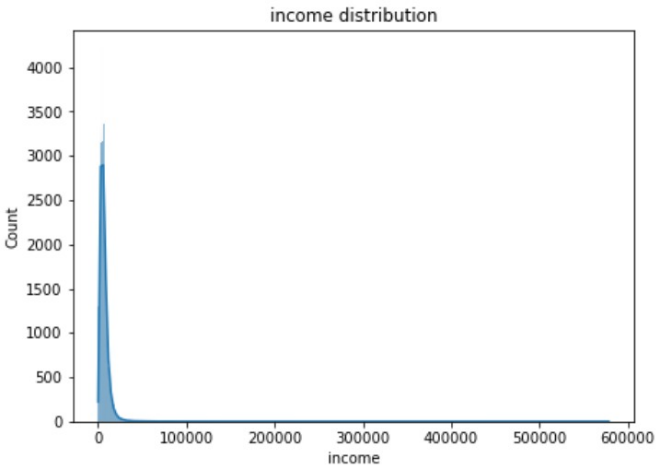
And also I will perform hyperparameter tuning to the algorithms to determine optimal parameters for each algorithm. Later I will chose best algorithms that gives better accuracy to the data.

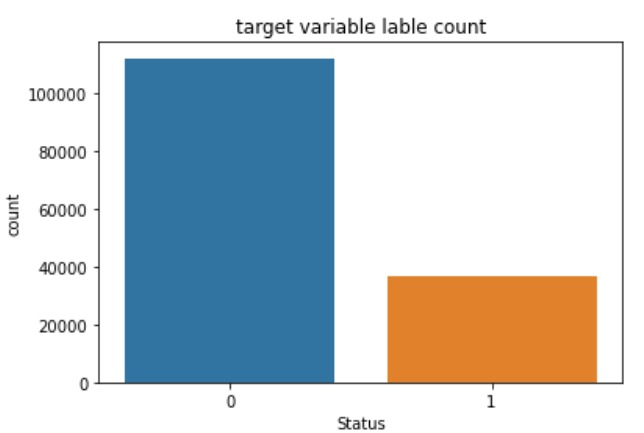
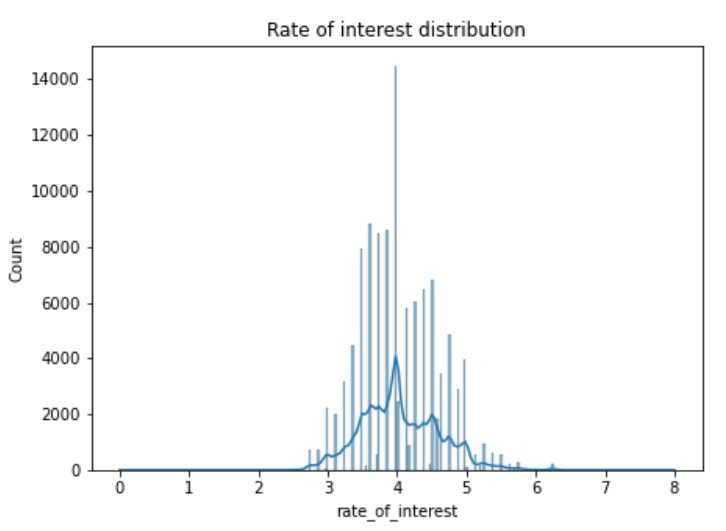
And at last final project report will be done by through out the semester, I expect to complete final project report by first week of July and will submit final project report by mid July with all required modifications.

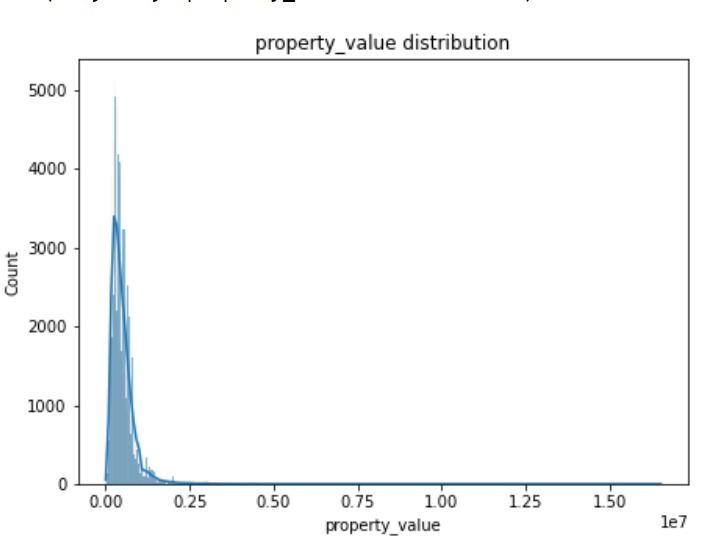
**Appendices**

Some screen snippets have given in the following:





****

****

**References**

1. (Vinay padimi, et al.., 2022), Journal of Neutrosophic and Fuzzy Systems (JNFS)

Vol. 2, No. 2, PP. 44-56, 2022.

2. Woo-Young Lee, Seung-Min Park, Kwee-Bo Sim, “Optimal hyperparameter tuning of convolutional neural networks based on the parametersetting-free harmony search algorithm," Optik,Volume 172, 2018, pp. 359-367, ISSN 0030-4026, <https://doi.org/10.1016/j.ijleo.2018.07.044>.

3. Lai, L.: Loan default prediction with machine learning techniques. In: 2020 International Conference on Computer Communication and Network Security (CCNS). pp. 5–9. IEEE (2020)

4. Meer, K.: Machine learning models for mortgage default prediction in pakistan. In: 2021 International Conference on Artificial Intelligence (ICAI). pp. 164–169. IEEE(2021)

5. Y. -R. Chen, J. -S. Leu, S. -A. Huang, J. -T. Wang and J. -I. Takada, "Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets," in IEEE Access, vol. 9, pp. 73103-73109, 2021, doi: 10.1109/ACCESS.2021.3079701.