College of Engineering

Dept of Computer Science and Engineering

Professor: Salam Dhou

CMP466 **Project Proposal**

Title: ***Credit Score Classification***

Omar Odeh - B00082635

Mohamed AlMarzooqi - B00090375

Khaled Mohamed - B00087968

Koushal Parupudi - B00087520

Date of submission: 9th February 2024

Spring 2024

American University of Sharjah

Introduction and Problem Statement

Access to credit plays a pivotal role these days in enabling individuals and businesses realize their aspirations. However, inaccuracies in determining creditworthiness using legacy statistical methods can exclude demographics from accessing financial resources. Our machine learning project aims to address this issue by building more precise and fair credit scoring models leveraging advanced algorithms. The motivation is that improved assessment of default risk can empower wider sections of society by increasing inclusion. Sophisticated analytics on heterogeneous data sources can potentially capture evolving consumer financial patterns better and enable prudent risk management as well. Overall, the promise of ML in this domain is to develop customized insights from transactions data that maintain accountability while providing opportunities. Through continued model evaluation and refinement, we hope to responsibly expand this capability at scale. We believe this data-driven approach can add value for all stakeholders ranging from consumers to policy makers in building sustainable credit frameworks.

Brief Overview of the literature

Accessing credit-worthiness accurately has seen a rising research interest owing to the limitations of legacy credit score classification models. A review of the modern machine learning techniques has enhanced predictive accuracy and fairness. For instance, algorithms like regression, decision trees, random forests, support vector machines and neural networks have been experimented with using features spanning credit records, demographics, repayment

behavior and alternative data like transactions. Overall, academic literature highlights the need for reliable and interpretable credit scoring to uphold financial inclusion policies. As we advance credit modeling capabilities, maintaining public trust via ethical design principles is equally crucial. Fostering access while managing risk requires a balanced socio-technical approach.

Federated learning is an emerging approach that enables building machine learning models in a distributed manner without directly sharing or centralizing private data from end-devices like phones or hospitals. The key idea is to cooperatively learn a shared prediction model while keeping local training data decentralized on respective devices to address critical issues of privacy, security and access rights (McMahan & Ramage, 2017). Research in federated learning has taken off as a way to unlock the untapped data in devices for improved models while retaining user benefits. For this reason, federated learning can be applied to the problem of credit score classification due to the importance of retaining the privacy of credit card holders' sensitive information.

For credit scoring issues, feature selection strategies along with various classification approaches have been thoroughly studied in the past. For example, researchers have used a range of classifiers, including LDA, NB, TDNN, KNN, DT, ELM, RF, and SVM, in conjunction with nine different feature selection approaches: ILFS, ECFS, Relief, FSV, LS, MFCS, UDFS, LLCFS, and CFS (Tripathi et al., 2021). Moreover, pre-processing procedures, such as data cleansing and transformation, are often used in these investigations to address missing values and categorical features. In particular, CFS consistently yields optimal results across various classification algorithms, whereas TDNN and RF have emerged as the top performers across several datasets. Furthermore, improvements in classification accuracy are seen when compared

to earlier studies, indicating the potential efficacy of ensemble classification techniques (Tripathi et al., 2021).

Our Contribution

Credit score classification models are trained on sensitive financial data like income, credit history etc. Centralizing such data raises risks of privacy violations and security breaches. Federated Learning allows institutions to collaboratively build models without sharing raw data. Despite the nature of federated learning to use decentralized data, there is almost no performance dropoff and is comparable to using centralized data in classical machine learning. This also preserves localization and the privacy of the credit card holders. We have recognized the potential of federated learning in the classification of credit score and its advantages over using centralized data and thus, we will try to implement this approach in our project.

Secondly, We will also be implementing Boosting models like XGBoost and AdaBoost in the process of model comparison as these models have not been explored in previous studies. Additionally we will also be performing feature engineering techniques that include but are not limited to oversampling minority classes and undersampling majority classes to ensure that classes have near equal representation.

Thirdly, we hope to explore the usage of Graph neural networks (GNN) for this problem. GNN's have been known to be excellent in capturing the significance of features by representing features as nodes and relationships as edges. Node feature embeddings and edge dependencies can improve predictive accuracy.

Description of the dataset

The following database includes a tabular dataset with a description of banks' clients and focuses on their credit status. The dataset includes around 100,000 samples totally with 27 distinct features. A description of the 27 features are as the following:

1. **ID:** An identifier for each record in the dataset, typically a unique number or code.
2. **Customer ID:** An identifier for each customer, which can be used to link multiple records for the same customer.
3. **Month:** The month or time period to which the data corresponds. It may indicate when the data was collected or when the credit scoring assessment was made.
4. **Name:** The name of the customer or borrower.
5. **Age:** The age of the customer, which can be a factor in credit scoring.
6. **SSN:** The Social Security Number or a unique identification number for the customer, often used for verification purposes.
7. **Occupation:** The customer's occupation or employment status, which can provide insight into their financial stability.
8. **Annual Income:** The customer's total annual income, a key factor in determining creditworthiness.
9. **Monthly Inhand Salary:** The customer's monthly take-home salary after deductions.
10. **Num Bank Accounts:** The number of bank accounts held by the customer, which may indicate financial stability.
11. **Num Credit Card:** The number of credit cards owned by the customer.
12. **Interest Rate:** The interest rate associated with a loan or credit, if applicable.
13. **Num of Loan:** The number of loans currently held by the customer.

14. **Type of Loan:** The type or category of the loan(s), such as personal loan, mortgage, or car loan.
15. **Delay from due date:** The delay in making payments from the due date, which may indicate a history of late payments.
16. **Num of Delayed Payment:** The number of delayed or missed payments.
17. **Changed Credit Limit:** Whether there has been a change in the customer's credit limit.
18. **Num Credit Inquiries:** The number of times the customer's credit report has been accessed by creditors or lenders.
19. **Credit Mix:** The variety of credit types used by the customer, such as credit cards, loans, and mortgages.
20. **Outstanding Debt:** The total amount of outstanding debt owed by the customer.
21. **Credit Utilization Ratio:** The ratio of credit used to credit available, often associated with credit cards.
22. **Credit History Age:** The length of the customer's credit history, which can impact credit scores.
23. **Payment of Min Amount:** Whether the customer has consistently made at least the minimum required payments on loans or credit cards.
24. **Total EMI per_month:** The total Equated Monthly Installments (EMI) paid by the customer.
25. **Amount invested monthly:** The amount the customer invests or saves monthly.
26. **Payment Behavior:** An indicator of the customer's payment behavior, such as "good," "fair," or "poor."
27. **Monthly Balance:** The customer's monthly account balance or financial position.

The dataset contains 3 classes pertaining to the credit score which is given to the customer according to several calculations and factors to prove their creditworthiness. The following classes are poor (28998 occurrences), standard (53174 occurrences), and finally good (17828 occurrences) respectively.

Classes:

1) Good
2) Standard
3) Poor

Number of samples in each class:

	Class	Count
0	Standard	53174
1	Poor	28998
2	Good	17828

References

Link to the dataset:
https://www.kaggle.com/datasets/mohammedobeidat/credit-score-classification/data

McMahan, B., & Ramage, D. (2017). Federated learning: Collaborative machine learning without centralized training data. Google AI Blog.

Tripathi, D., Edla, D. R., Bablani, A., et al. (2021). Experimental analysis of machine learning methods for credit score classification. Progress in Artificial Intelligence, 10(2), 217–243. https://doi.org/10.1007/s13748-021-00238-2

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. Technology in Society, 63, 101413. https://doi.org/10.1016/j.techsoc.2020.101413

Appendix:

Code used:

```
import os
import pandas as pd
dataset = "C:/Users/koush/OneDrive/Documents/Machine learning project/train.csv"
credit = pd.read_csv(dataset)
credit
```

	ID	Customer_ID	Month	Name	Age	SSN	Occupation	Annual_Income	Monthly_Inhand_Salary	Num_Bank_Accounts	...	Credit_Mix	Ou
0	5634	3392	1	Aaron Maashoh	23.0	821000265.0	Scientist	19114.12	1824.843333	3.0	...	Good	
1	5635	3392	2	Aaron Maashoh	23.0	821000265.0	Scientist	19114.12	1824.843333	3.0	...	Good	
2	5636	3392	3	Aaron Maashoh	23.0	821000265.0	Scientist	19114.12	1824.843333	3.0	...	Good	
3	5637	3392	4	Aaron Maashoh	23.0	821000265.0	Scientist	19114.12	1824.843333	3.0	...	Good	
4	5638	3392	5	Aaron Maashoh	23.0	821000265.0	Scientist	19114.12	1824.843333	3.0	...	Good	
...	...	...	...	...	...	...	...	...	...	...	...	...	
99995	155625	37932	4	Nicks	25.0	78735990.0	Mechanic	39628.99	3359.415833	4.0	...	Good	
99996	155626	37932	5	Nicks	25.0	78735990.0	Mechanic	39628.99	3359.415833	4.0	...	Good	
99997	155627	37932	6	Nicks	25.0	78735990.0	Mechanic	39628.99	3359.415833	4.0	...	Good	
99998	155628	37932	7	Nicks	25.0	78735990.0	Mechanic	39628.99	3359.415833	4.0	...	Good	
99999	155629	37932	8	Nicks	25.0	78735990.0	Mechanic	39628.99	3359.415833	4.0	...	Good	

100000 rows × 28 columns

```
import pandas as pd
print(credit.head(20))
```

```
      ID  Customer_ID  Month            Name   Age          SSN Occupation  \
0   5634         3392      1    Aaron Maashoh  23.0  821000265.0  Scientist
1   5635         3392      2    Aaron Maashoh  23.0  821000265.0  Scientist
2   5636         3392      3    Aaron Maashoh  23.0  821000265.0  Scientist
3   5637         3392      4    Aaron Maashoh  23.0  821000265.0  Scientist
4   5638         3392      5    Aaron Maashoh  23.0  821000265.0  Scientist
5   5639         3392      6    Aaron Maashoh  23.0  821000265.0  Scientist
6   5640         3392      7    Aaron Maashoh  23.0  821000265.0  Scientist
7   5641         3392      8    Aaron Maashoh  23.0  821000265.0  Scientist
8   5646         8625      1  Rick Rothackerj  28.0    4075839.0    Teacher
9   5647         8625      2  Rick Rothackerj  28.0    4075839.0    Teacher
10  5648         8625      3  Rick Rothackerj  28.0    4075839.0    Teacher
11  5649         8625      4  Rick Rothackerj  28.0    4075839.0    Teacher
12  5650         8625      5  Rick Rothackerj  28.0    4075839.0    Teacher
13  5651         8625      6  Rick Rothackerj  28.0    4075839.0    Teacher
14  5652         8625      7  Rick Rothackerj  28.0    4075839.0    Teacher
15  5653         8625      8  Rick Rothackerj  28.0    4075839.0    Teacher
16  5658        11708      1          Langep  34.0  486853974.0   Engineer
17  5659        11708      2          Langep  34.0  486853974.0   Engineer
```

```
credit.columns
```

```
Index(['ID', 'Customer_ID', 'Month', 'Name', 'Age', 'SSN', 'Occupation',
       'Annual_Income', 'Monthly_Inhand_Salary', 'Num_Bank_Accounts',
       'Num_Credit_Card', 'Interest_Rate', 'Num_of_Loan', 'Type_of_Loan',
       'Delay_from_due_date', 'Num_of_Delayed_Payment', 'Changed_Credit_Limit',
       'Num_Credit_Inquiries', 'Credit_Mix', 'Outstanding_Debt',
       'Credit_Utilization_Ratio', 'Credit_History_Age',
       'Payment_of_Min_Amount', 'Total_EMI_per_month',
       'Amount_invested_monthly', 'Payment_Behaviour', 'Monthly_Balance',
       'Credit_Score'],
      dtype='object')
```

```
import pandas as pd

class_counts_df = pd.DataFrame({'Class': class_counts.index, 'Count': class_counts.values})
class_counts_df
```

	Class	Count
0	Standard	53174
1	Poor	28998
2	Good	17828