Application of Machine Learning for Creditworthiness in the Financial Industry

Michael Chen mic012@ucsd.edu
Yifan Li yil011@ucsd.edu
Aneesh Navanale anavanal@ucsd.edu
William Xia wxia@ucsd.edu
Brian Duke brian.duke@prismdata.com
Berk Ustun berk@ucsd.edu

Abstract

In the complex landscape of financial services, creditworthiness is vital for managing financial risks, ensuring economic stability, and making reasonable lending decisions. The FICO score, the standard measure of creditworthiness, is updated depending on the creditor's reporting schedule, typically ranging from one month to forty-five days. As a result, the FICO score fails to capture the rapid changes in an individual's financial behavior or the emergence of new financial risks. Our technique aims to refine this process thereby significantly enhancing the efficiency and accuracy of creditworthiness assessment. We aim to accurately predict consumers' risk of defaulting by utilizing the best ML model and creating relevant income, balance, and category spend features for the final credit scoring calculation.

Code: https://github.com/willxia/DSC180

1 Introduction

Credit Score Prediction is a crucial function in the financial industry, allowing banks and other financial institutions to gain insight into consumers' spending habits, improve their risk assessment models, and detect fraudulent activity. Our study introduces advanced predictive models that estimate the likelihood of consumer default, introducing a novel metric we term the "Cash Score." This metric is distinct from the conventional FICO Score, offering more frequent updates and delivering a nuanced understanding of consumer creditworthiness.

This prediction lays the groundwork for the development of the "cash-score" model. By integrating variables such as income, account balances, and expenditure across various categories, the model generates a Cash Score that reflects an individual's potential credit risk. Through this refined assessment mechanism, the Cash Score model aims to facilitate more informed decision-making in the realm of credit provision and financial risk management.

1.1 Literature Review

According to Vincenzo et al. (2021), machine learning techniques can be applied to model credit risk assessment as a binary classification problem based on debt repayment. They address the class imbalance by using different sampling techniques, and they use metrics like AUC, Sensitivity, and Specificity, to find the best-performing models. However, the lack of borrowers' credit history poses a significant challenge in evaluating loan applicants' creditworthiness. P2P platforms generate a large amount of unlabeled data, requiring real-time analysis to support lenders' decisions. With that being said, David et al. (2021) suggest AI and machine learning, using alternative data sources, such as public data, satellite images, and data from social media interactions, can address issues like information asymmetry, adverse selection, and moral hazard, enabling financial institutions to assess credit risk and grant credit to underserved populations. Furthermore, to improve the classification algorithm, Pawet et al. (2020) introduced a new Deep Genetic Hierarchical Network of Learners (DGHNL) system with a fusion-based 29-layer structure. Essentially, they combine simple algorithms, and they use an ensemble learning technique with many layers to boost the performance of the CS prediction system. Compared to the prior approaches, We have more complete data and advanced access to private information, so we expect to build a better model by trying out more customized algorithms.

1.2 Dataset Information

The dataset compromises 2,576,829 transaction records (outflows) and each contains five distinct columns: Consumer ID, Account ID, Memo Clean, Amount, and Category Description. The memo clean column contains transaction descriptions. The category description column contains 28 unique categories.

	Category	Count	Percentage
1	GENERAL_MERCHANDISE	516039	20.03%
2	FOOD_AND_BEVERAGES	467667	18.15%
3	EXTERNAL_TRANSFER	323631	12.56%
4	GROCERIES	220227	8.55%
5	AUTOMOTIVE	197638	7.67%
6	UNCATEGORIZED	117390	4.56%
7	ATM_CASH	114602	4.45%
8	LOAN	92650	3.60%
9	ENTERTAINMENT	85238	3.31%
10	CREDIT_CARD_PAYMENT	79785	3.10%

Figure 1: Top 10 most prevalent categories in outflows dataset

The dataset contains approximately 5% of data that is labeled “uncategorized”. Majority of the transactional data, around 50%, falls into general merchandise and food categories. Frequent transactions like mortgage, rent, utilities, insurance, BNPL (Buy now pay later) and loan repayments are helpful in establishing creditworthiness. The distribution of these categories is useful for “cash-scoring” as it provides a comprehensive view of a consumer’s spending habits. To further refine our model, we are currently working on incorporating ‘date’ and ‘amount’ data as it would allow us to create time-sensitive features and assess overall creditworthiness.

We worked on the categorized inflows (507943 records) and outflows which included the ’posted_date’ column. The inflows consists of 14 unique categories.

	prism_consumer_id	prism_account_id	memo_clean	amount	posted_date	category_description
0	0	acc_0	TRANSFER FROM CHK XXXXXXXXX	25.00	2022-05-04	SELF_TRANSFER
1	0	acc_0	TRANSFER FROM CHK XXXXXXXXX	25.00	2023-01-18	SELF_TRANSFER
2	0	acc_0	TRANSFER FROM CHK XXXXXXXXX	25.00	2023-03-01	SELF_TRANSFER
3	0	acc_0	INTEREST PAYMENT	0.05	2023-02-28	INVESTMENT_INCOME
4	0	acc_0	INTEREST PAYMENT	0.07	2023-01-31	INVESTMENT_INCOME

Figure 2: Inflows dateset

The account dataset contains 4969 records each containing account information. Each consumer can upload multiple types of accounts (Savings, Checking, Credit Card, Money Market, Prepaid, CD, and Cash Management). The consumer dataset contains the consumers’ credit card application evaluation date, approved status, and the FPF_Target. The FPF_Target tells us whether the consumer missed their first payment or not. 18.83% of the consumers missed their first payment.

	prism_consumer_id	prism_account_id	account_type	balance	balance_date
0	0	acc_0	SAVINGS	6182.60	2023-04-13
1	0	acc_1	CHECKING	9907.23	2023-04-13
2	2	acc_12	SAVINGS	17426.83	2022-02-15
3	2	acc_11	CHECKING	8079.43	2022-02-15
4	4	acc_16	SAVINGS	0.00	2021-08-13

Figure 3: Account dataset

	prism_consumer_id	evaluation_date	APPROVED	FPF_TARGET
0	658	2022-06-24	1	0.0
1	539	2023-05-10	1	0.0
2	540	2021-12-21	1	0.0
3	787	2022-06-22	1	0.0
4	1141	2022-07-08	1	0.0

Figure 4: Consumer dataset

1.3 Exploratory Data Analysis - Inflows

Based on the ”Amount” category in both inflows and outflows data, we can analyze the inflows distribution and financial behavior for this group of consumers.

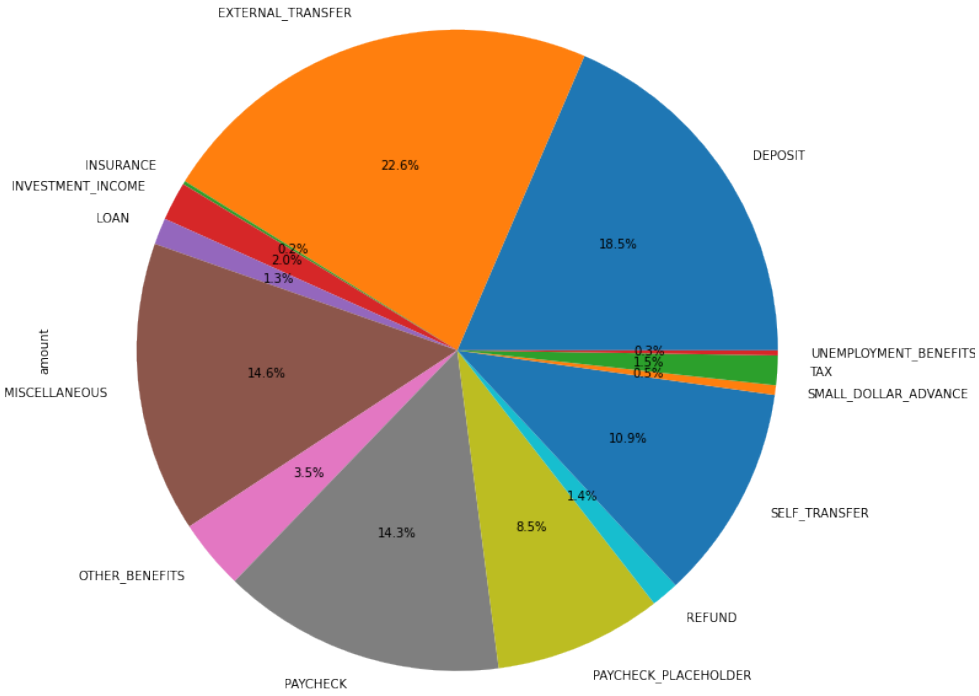

Figure 5: Inflows Source Distribution

Figure 5 shows that the main sources of inflows are External Transfers, Deposits, and Paychecks, indicating a mix of internal earnings and external transfers. The largest inflow source is from external transfers, amounting to USD 81,586,232.14. The high figures in deposits and paychecks can indicate robust earnings, either from a business or professional employment.

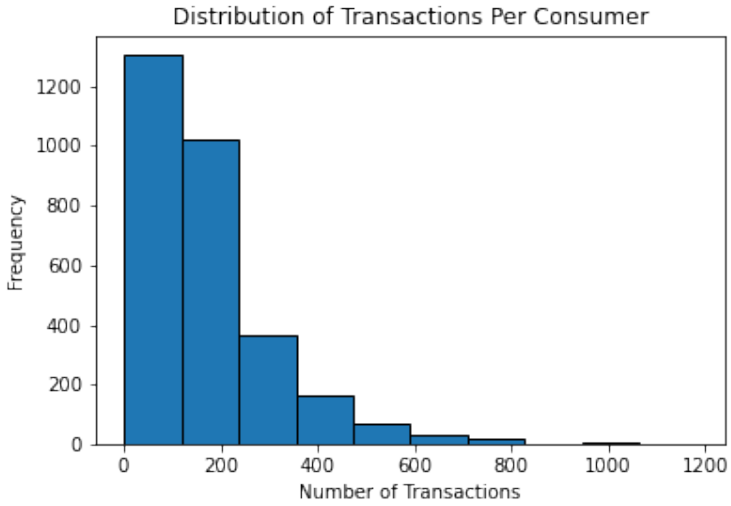

Figure 6: Count of Transaction Distribution

By looking at the count of transactions for customers in Figure 6, we notice that consumers make about 172.53 transactions on average. A standard deviation of 148.45 suggests moderate variability in the number of transactions among different consumers. At the same time, the wide range in the number of transactions indicates varied levels of engagement or financial activity among consumers.

We can retrieve net balance by subtracting total consumption from the total inflows of the individual. From the statistics in Table 1, we can notice that there is a huge standard deviation (39235.0), indicating significant variability in net balance among customers. Also, the median (30.6) being much higher than the mean (-807.9) implies a skewed distribution, likely with a few high-balance individuals. While some customers have a negative net balance, it could be that they consumed more in this period of time compared to what they earned.

Table 1: Consumer Net Income Statistics

Index	**Amount**
count	2978.0
mean	-807.9
std	39235.0
min	-1492984.2
25	-619.2
50	30.6
75	747.7
max	575143.9

1.4 Exploratory Data Analysis - Balance Over Time

On the other hand, we are also interested in examining the consumers' balance over time because we think a stable or growing balance might indicate the ability to repay and a higher absolute balance amount also tells one's strong financial status.

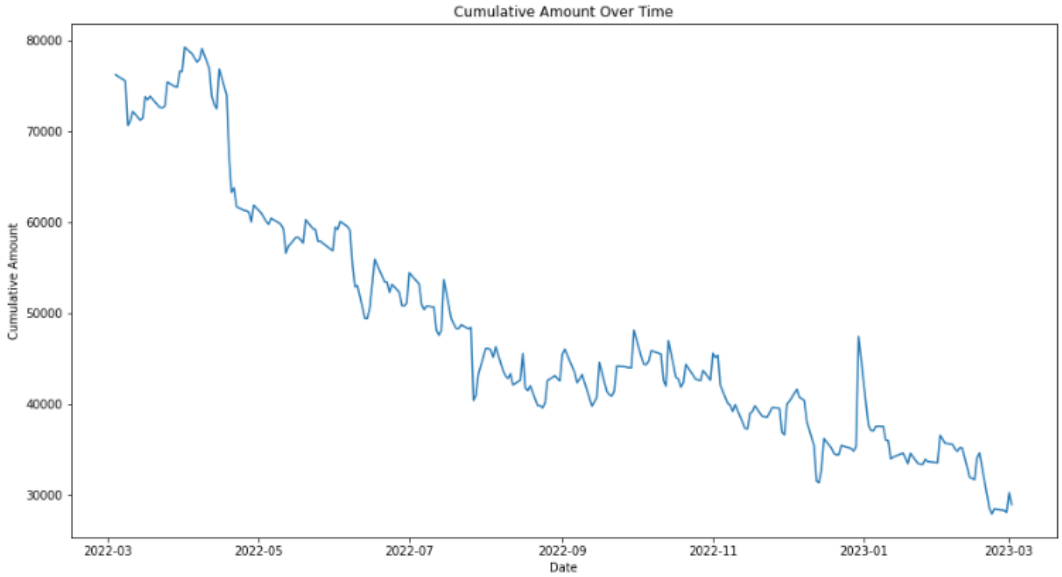

Figure 7: A consumer's balance over time

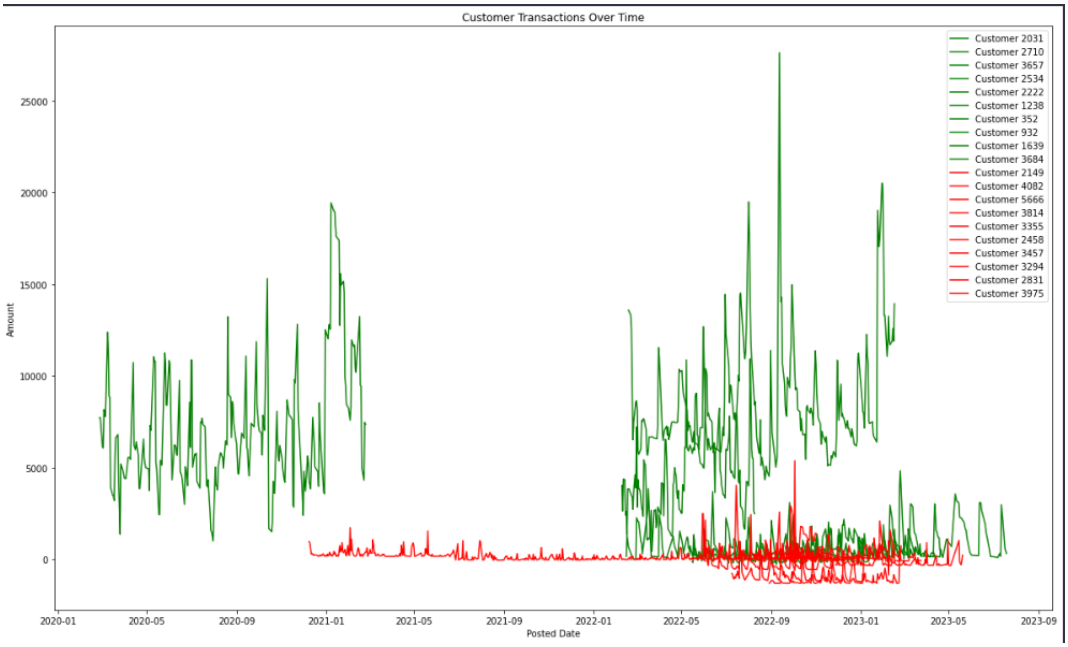

Figure 8: Comparison of consumers' balances

To build features based on the balance curve, we randomly selected 10 consumers from each

target class and plotted it in the same figure. The green curves represent the balance over time for the random consumers who will pay the loan, and the red ones are the curve for those who won't. As shown in the graph, we discovered that green consumers usually have higher absolute amounts and more volatile changes of balance over time. This encouraged us to build more features based on one's balance.

2 Methods

2.1 Overview

Our group would utilize the categories and income estimates to build a score to predict the risk of a consumer not paying his/her bills. To predict the probability of someone defaulting or not, we need to train a model using information such as an individual's income, balance, and categories as the features to make the prediction.

In this section, we will introduce the methodologies employed in our cash score prediction model. To prepare for the cash score calculation, we need to create relevant features to better understand consumers' financial behavior, which allows us to predict the probability of default. We tested 3 different types of models for this prediction task:

- Logistic Regression
- XGBClassifier
- SGDClassifier

We decide to use the machine learning models above due to their advantages in predicting binary outcome variables. Logistic Regression, XGB, and SGD were trained on features from balance, income, and consumption, which are derived from transaction and balance datasets.

2.2 Feature Selection

After receiving an additional dataset that contains information about consumers' account balances, and account type, we are able to generate more features that are representative of consumers' financial behavior.

There are three most important aspects we focus on, which are income, consumption, and account balance.

- **Income**
- **Consumption**
- **Account Balance**

By accessing the changes in consumers' income throughout time, we can capture the fluctuation and measure the stability in their inflow of money, which may reflect their financial well-being. For instance, a regular income from a reliable source would make the consumer less risky compared to someone with an irregular income. Income is also a reliable indicator

that measures an individual's financial health. In general, a higher monthly income means a greater capacity to meet financial commitments, reducing the likelihood of defaulting.

Consumption also helps us to identify if there are any suspicious activities or spending behavior of the individual. By looking at the corresponding category for each outflow transaction, we can distinguish between people's discretionary spending and essential expenses. For instance, if a consumer constantly has a high proportion of discretionary spending compared to her or his income, it may be a red flag. Consumption also reveals the consumer's financial priorities, which helps us to gain a deeper understanding of her or his financial behavior.

Account balance is another important indicator of a consumer's financial health, and it also represents the liquidity of her or his current asset. For example, a higher savings and checking account balance may indicate financial stability and a lower risk of default, whereas lower or negative balances may signal financial distress.

In the following paragraphs, we will elaborate on important features that we found that help the model to learn an individual's financial pattern and spending behavior.

2.2.1 Consumer's Net Income

This feature is based on the consumer's regular payment from inflow through a billing cycle. It's a critical predictor of bill payment capacity, as higher net income generally suggests a higher ability to meet financial obligations.

2.2.2 Individual's Total Transaction Count

This feature is calculated using the count of transactions recorded in both the inflow and outflow datasets, which also reflects the consumer's financial activity level. A higher transaction count might indicate more active financial management or higher consumption patterns, which could influence their payment behavior.

2.2.3 Net Balance and Account Type

This is a direct indicator of a consumer's financial health. A positive net balance suggests that a consumer has surplus funds, potentially lowering the risk of non-payment. Conversely, a negative balance might indicate financial distress.

2.2.4 Consumption on Single Category (e.g., Entertainment)

Spending patterns in specific categories can provide insights into a consumer's financial priorities and discretionary spending habits. High spending on non-essential categories, relative to overall income, might signal a higher risk of financial mismanagement.

2.2.5 Monthly Consumption

Analyzing short-term spending patterns can help to identify consistent behaviors that predict payment risk. Stability in consumption, or significant increases/decreases, might indicate changes in financial status or behavior over time.

2.2.6 Monthly Outflow/Inflow

This represents the monthly financial movement in and out of the consumer's account. A positive net flow suggests financial stability, while a negative net flow might indicate a higher risk of running into payment issues.

2.2.7 Average Monthly Inflow/Outflow Per Category

This represents the mean monthly financial movement in and out of the consumer's account per category.

As we experimented with using the single feature for the model in the beginning and failed to receive a satisfactory result, we decided to incorporate more features as they are able to demonstrate a full picture of consumer financial behavior.

2.2.8 Average Income Per Month

We used a statistical test to classify the inflows with regularity, so they represent one's income source, and we can build features based on that. Then we calculated one's total income per month and took the average so that we came up with this feature, which we think a higher absolute amount of it is positively correlated with one's ability to pay back the loan on time.

2.2.9 Average Percentage Change of Income Over Months

Another feature we built based on one's income is its percentage changes over months since a positive average change means either one got a better job, or one obtained more ways to earn more money, which are both good indicators of one's capability to make on-time repays.

2.3 Models

2.3.1 Logistic Regression Model

We also focus on the logistic regression model for predicting individuals' cash scores, which captures the probability of someone not paying back their bills. It stands out because Lo-

gistic regression is designed specifically for binary outcome variables, making it ideal for predicting whether a consumer will pay (1) or not pay (0) their bills. It estimates probabilities that are bounded between 0 and 1, aligning with the need to assess the risk of a binary event.

We tried to utilize multiple features mentioned above to feed the model and did a training test split around a 6:4 proportion.

2.3.2 XGB Classifier Model

XGBClassifier refers to the eXtreme Gradient Boosting Classifier. It is based on gradient boosting, an ensemble technique where new models are created to correct the errors made by existing models. XGBoost uses decision trees as its base learners. There are several strengths of XGBClassifier that meets our need for the credit assessment task. It includes built-in regularization which helps to prevent overfitting. At the same time, we can adjust a variety of tuning parameters that can be optimized for better model performance.

2.3.3 SGD Classifier Model

As we continued to explore different models, we implemented a Stochastic Gradient Descent (SGD) classifier with hinge loss, a variant that essentially implements a linear Support Vector Machine (SVM). Much like Logistic Regression, this model is well-suited for linearly separable data and has proven resilience against overfitting. The decision to employ the SGD classifier with hinge loss was driven by the nature of our dataset, where discerning linear relationships between features and class labels was crucial. The model's capability to effectively capture these linear boundaries, akin to Logistic Regression, contributed to an enhanced accuracy rate compared to our baseline model. Furthermore, the SGD classifier with hinge loss demonstrated robustness against class imbalance, a prevalent characteristic in our dataset, bolstering its overall performance. This strategic choice in model selection showcases our commitment to leveraging algorithms that align with the dataset's inherent characteristics and enhance predictive accuracy.

3 Results

After feature selection and hyper-parameter tuning, our analysis identified the XGBClassifier as the best performing model among the evaluated algorithms, including Logistic Regression and SGDClassifier. The performance metrics, including accuracy and ROC-AUC score, were used to assess the efficacy of each model.

We hypothesize that XGBoost outperformed the other models because it is capable of capturing non-linear relationships between features and the target variable, which might be present in the dataset but not adequately captured by linear models like Logistic Regression.

Table 2: Performance Comparison of Different Machine Learning Models

Model	ROC-AUC Score	Accuracy
Logistic Regression	0.83	0.83
XGBClassifier	0.86	0.84
SGDClassifier	0.79	0.79

The XGBClassifier's superior accuracy and ROC-AUC score indicate its effectiveness in accurately predicting consumers' risk of defaulting. The model's robust performance underscores its potential utility in refining credit scoring calculations and enhancing the efficiency of creditworthiness assessments in the financial services landscape. These findings highlight the significance of employing advanced machine learning techniques to improve the accuracy and timeliness of credit risk assessments, ultimately contributing to better-informed lending decisions and mitigating financial risks.

Using the XGBClassifier alongside the SHAP Python package, our analysis revealed key factors driving predictive accuracy. We found that features like the average credit card payment amount and consumer balances were highly influential in predicting default risk. Looking at Figure 9, we can understand each feature's impact on the prediction by looking at its SHAP value and its color. For example, with the `balance` feature, we can see that when a consumer's balance plays a role, it tends to push our prediction towards 0.

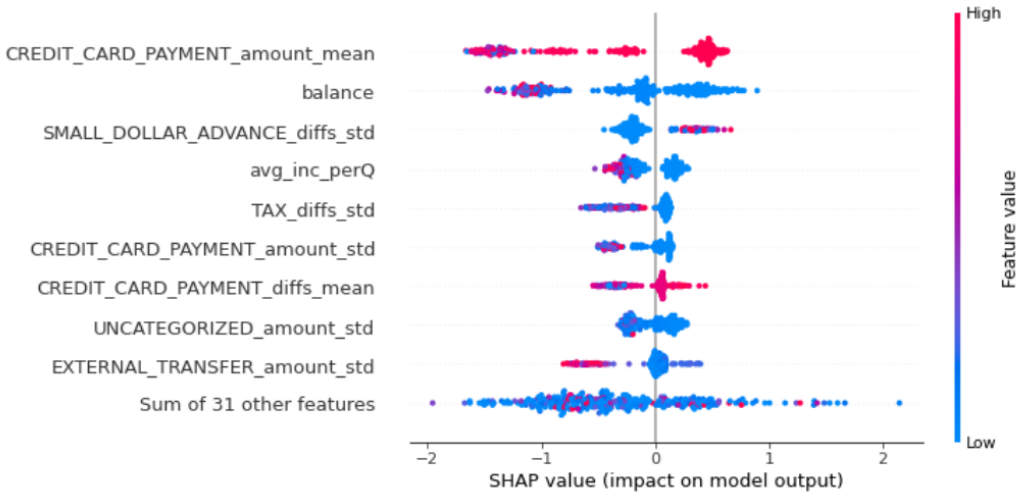

Figure 9: Most Important Features Using SHAP Values

These insights emphasize the importance of understanding consumer credit behavior for better lending decisions. By pinpointing such critical factors, our study offers a clearer path to more accurate credit scoring, enabling financial institutions to make smarter lending choices and mitigate risks effectively.

4 Conclusion

This project aimed to explore various machine learning models to ascertain their efficacy in predicting consumer credit risk, which is crucial in the financial domain. As we can notice from the result section above, XGBoost demonstrated a superior performance over both Logistic Regression and SGDClassifier based on our evaluation of the ROC-AUC Score and accuracy rate. Notably, the XGBClassifier attained a ROC-AUC score of 0.86 and an accuracy of 0.84, outshining its counterparts.

The obtained result resonated with existing literature and XGBoost is capable of handling diverse predictive tasks in the financial domain, particularly where complex, non-linear interdependencies are intrinsic to the data. We also created reason codes that allow consumers to know which features/factors are having the greatest impact on their credit scores. They would be able to see which category makes the final prediction toward the negative direction, which makes them more risky compared to average consumers. Therefore, this can be a potent tool for credit risk assessment as it makes the process more accurate, dynamic, and timely.

There are four main impacts and capabilities of our approach:

- **Reduce the probability of default risk:** Reduction in financial losses associated with bad debts and enhance the overall financial health of lending institutions.
- **Various applications :** Can be readily applied in various decision-making processes within financial institutions.
- **Greater financial inclusion:** providing more individuals and businesses with access to essential financial services.
- **Operational Efficiency:** Integrating advanced machine learning models can streamline the credit assessment process, reducing the time and resources required to evaluate loan applications.

However, despite the positive feedback from our model, there are some limitations of this approach. The efficacy of our model has highly relied on the quality and representativeness of the dataset employed. Without a stable source of dataset, this approach may not generate accurate results that predict the probability of defaulting. At the same time, the interpretability of XGBoost is not completely transparent due to the complexity and non-linear nature of how it makes decisions. XGBoost is an ensemble learning method, specifically a gradient-boosting algorithm that builds many decision trees sequentially, with each new tree aiming to correct the errors of its predecessors. The final prediction is made based on the aggregate predictions of all trees, which makes it difficult to trace how each input feature directly influences the prediction.

Future research can build upon our findings in several meaningful ways. Investigating the integration of additional, potentially unexplored features could further enhance predictive accuracy. If we could collect more personal information that reflects consumer behavior, the model performance could be further improved. Moreover, advancing model interpretability, particularly for sophisticated algorithms like XGBoost, stands as an imperative avenue to engender broader acceptance and application.

In conclusion, our results offer a potential implication for the financial industry, presenting opportunities to advance credit risk assessment practices. By leveraging the predictive power of the XGBClassifier, financial institutions can achieve more accurate, efficient, and equitable lending processes, ultimately contributing to their operational success and the broader economic well-being.

References

Mhlanga, David. 2021. "Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment." *International Journal of Financial Studies* 9(3). [Link]

Moscato, Vincenzo, Antonio Picariello, and Giancarlo Sperlí. 2021. "A benchmark of machine learning approaches for credit score prediction." [Link]

Pławiak, Paweł, Moloud Abdar, Joanna Pławiak, Vladimir Makarenkov, and U Rajendra Acharya. 2020. "DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring." [Link]