Building Better Credit Scores:
Machine Learning and NLP for Optimized Risk Assessment

Aman Kar
akar@ucsd.edu

Daniel Mathew
drmathew@ucsd.edu

Tracy Pham
tnp003@ucsd.edu

Brian Duke
brian.duke@prismdata.com

Kyle Nero
kyle.nero@prismdata.com

Berk Ustun
berk@ucsd.edu

Abstract

Credit scores are pivotal in today's financial landscape, influencing everything from rental eligibility to access to health insurance, yet the formula for calculating creditworthiness has long been shrouded in mystery and often overlooks important nuances. Typically, the credit score is determined based on factors such as payment history and credit length, disadvantaging individuals with limited credit histories—especially young adults, recent immigrants, and those without access to traditional credit products. To address these limitations, we propose the Cash Score, a probability-based model that predicts the likelihood of credit delinquency using detailed bank transaction data. By analyzing income patterns, spending behavior, and account activity, the Cash Score provides a more dynamic and inclusive assessment of financial responsibility. This report details our approach, methodology, and findings, highlighting the potential for transaction-based credit evaluation to complement or even improve upon traditional credit scoring methods.

Website: https://danielrmathew.github.io/prism-data-credit
Code: https://github.com/danielrmathew/prism-data-credit

1 Introduction

Typically, the credit score is determined based on five factors: payment history, amount owed, new credit, credit history, and credit mix. This structure can place individuals with limited credit history, especially young adults who are just starting out building their credit, at a compounded disadvantage, restricting their access to loans, credit cards, employment opportunities, and insurance. This report aims to address these limitations by developing a more comprehensive measure of creditworthiness. The Cash Score model leverages detailed account transaction data to predict the probability of defaulting on a loan (also known as loan delinquency). This report details our approach, methodology, and findings, highlighting the potential for transaction-based credit evaluation to more accurately assess financial risk and improve access to credit, offering a fairer alternative to traditional credit scoring methods.

1.1 Literature Reviews

1.1.1 A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset

Text classification plays a critical role in natural language processing (NLP) applications such as sentiment analysis, spam detection, and document categorization. Among various feature extraction methods, Term Frequency-Inverse Document Frequency (TF-IDF) has emerged as a prominent technique for representing text data in a way that highlights relevant features while diminishing the impact of common terms. The paper by **?** contributes to this domain by comparing TF-IDF with N-Gram feature extraction methods, utilizing datasets from both IMDB movie reviews and Amazon Alexa reviews.

The idea behind using TF-IDF began with **?** in 1988 who focused on retrieving information and text mining. We developed more complex concepts such as n-gram models (analyzing sequences of n-items) as a different approach to capturing contextual information. Research has shown that while N-Gram features capture local context effectively, they may lead to a high dimensionality problem, especially in larger datasets (**?**). Das, Selvakumar and Alphonse (2023) build on this foundation by assessing the effectiveness of these two methodologies in classifying sentiment in unstructured datasets. This highlights the gap in comparative studies on feature extraction techniques.

Moving on to the methodology, the authors of this paper conducted experiments using the two distinct feature extraction techniques we discussed: TF-IDF and N-Gram models. The paper employed various classifiers – Support Vector Machine (SVM), Random Forest, and Logistic Regression – to validate the efficacy of the feature extraction methods. The results indicated that TF-IDF significantly outperformed the N-Gram model across all classifiers, achieving a maximum accuracy of 93.81% and an F1-score of 91.99% using the Random Forest classifier. The authors emphasize that the choice of feature extraction method is crucial, as it directly impacts the classification outcome, thus reinforcing the importance of this area in text mining.

Das, Selvakumar and Alphonse (2023) effectively highlight the strengths of TF-IDF in text classification, suggesting it as a preferred method over N-Gram features in the context of sentiment analysis. The study contributes to the broader literature on feature extraction methods, providing empirical evidence that can inform future research directions. Further investigations could explore additional datasets and classification tasks to validate these findings and enhance the understanding of feature selection in NLP.

1.1.2 Attention Is All You Need

Incredible advances in language modeling and machine translation were made thanks to the introduction of the Transformer architecture by Vaswani et al. (2017). While state of the art NLP models at the time were built on recurrent layers with RNNs and LSTMs, the Transformers lets all of that go and focuses entirely on attention mechanisms (**?**). Built on the traditional encoder-decoder structure with self-attention layers, the architecture topped the previously best results in translation tasks while easier to train and more parallelizable. This architecture became the foundation for our current AI models like BERT and GPT, which all build on the ideas introduced in this paper.

The attention mechanism (Bahdanau, Cho, and Bengio 2014) enables models to focus on different parts of an input selectively, thereby assigning varying importance to each element. For the purposes of the Transformer, self-attention (Vaswani et al. (2017)) is utilized to do the same thing in the content of words in a sentence. It is able to do that regardless of the distance between the words, an important improvement over recurrent frameworks that struggle with understanding words that are far from each other given its sequential structure. The self-attention layer also enables the model to be trained in parallel because it isn't computed sequentially, vastly improving train time and efficiency.

Although the Transformer architecture is groundbreaking and widely used in state-of-the-art AI models, it may not yet be the ideal fit for the financial sector. Traditional machine learning models, such as logistic regression and gradient boosting, combined with simpler NLP techniques like bag-of-words and TF-IDF, are significantly faster to train and predict, often by orders of magnitude. Additionally, model interpretability is crucial in areas like credit scoring, where the rationale behind a consumer's credit score must be transparent and easy to understand. Without clear explanations, companies risk significant legal challenges regarding the basis of their decision-making.

Vaswani et al. (2017) introduction of the Transformer architecture revolutionized language processing in machine learning, reshaping how we approach NLP tasks. Their research has made, and continues to make, significant contributions to the AI/ML space, sparking novel techniques and applications across domains. While Transformers and large language models (LLMs) are not yet the standard for NLP in the financial sector, they are likely to be adopted as hardware and software capabilities advance.

1.1.3 Credit scoring methods: Latest trends and points to consider

Recent studies (2016-2021) show a growing use of machine learning techniques like support vector machines, ensemble methods (e.g., random forests, boosting), and neural networks in credit scoring due to their ability to capture complex patterns in large datasets. However, traditional models like logistic regression and decision trees remain popular as baseline models, valued for their interpretability, which is often required by regulators. Logistic regression, still widely used, appeared in around 70 studies but generally shows average performance compared to advanced models, with 16 studies rating it ”best” and 19 as ”worst.”

In terms of efficiency, Markov, Seleznyova and Lapshin (2022) showed ensemble methods—including bagging, boosting, and stacking—are known as the alternative, more challenging models that frequently outperform traditional models, delivering higher accuracy. Interestingly, support vector machines (SVMs) are used as both baseline and advanced alternatives, highlighting their versatility in various studies. Additionally, the ”Other” category of models—comprising techniques like k-nearest neighbors (kNN), deep genetic hierarchical networks of learners (DGNHL), and generalized additive models (GAM)—achieved the highest ”best” to ”worst” model ratio, highlighting the potential of these diverse methods in credit scoring.

Data preprocessing practices, essential for ensuring data quality, have remained largely consistent over the last four years. These typically involve imputation of missing values, feature selection, transformation, and rebalancing (resampling) of datasets, though all stages are usually applied simultaneously. The review underscores how each preprocessing step contributes to model reliability and performance. Notably, between 2016 and 2020, researchers have seen an increasing interest in missing value imputation rather than dropping data with missing values, which has been the most popular approach.

Performance evaluation metrics play a critical role in determining model reliability. Confusion matrix-based measures were traditionally used as they are easy to interpret, however, they are dependent on misclassification cost functions and thresholds as well as being sensitive to good/bad loan imbalances in data. Thus these measures are no longer used as a standalone misclassification criterion but rather used alongside other metrics such as accuracy, F1 score, Brier score, and AUC-ROC to assess credit scoring models. These measures help financial institutions ensure that predictions are both accurate and actionable. Overall, the article provides a comprehensive overview, offering insights into new innovative machine learning approaches while reaffirming the relevance of traditional statistical methods in credit scoring.

1.2 Data Description

Our analysis leverages four key datasets that provide insights into consumer accounts, transaction histories, and credit scores. As the datasets were prepared and preprocessed by Prism Data, this minimized the need for extensive data cleaning. Our primary focus in terms of data cleaning was reviewing the data for consistency, addressing any remaining missing

values, standardizing variables, and structuring time series data to optimize it for modeling.

1.2.1 Account Data

Table 1: Head of the `acctDF.csv`

prism_consumer_id	prism_account_id	account_type	balance_date	balance
3,023	0	SAVINGS	2021-08-31	90.57
3,023	1	CHECKING	2021-08-31	225.95
4,416	2	SAVINGS	2022-03-31	15,157.17
4,416	3	CHECKING	2022-03-31	66.42
4,227	4	CHECKING	2021-07-31	7,042.90

The `acctDF.csv` dataset provides detailed information about consumer financial accounts, such as account types, balances, and balance dates. A preview of this data is shown in Table 1, where you can see each consumer's balance on their accounts on specific dates.

1.2.2 Consumer Data

Table 2: Head of `consDF.csv`

prism_consumer_id	evaluation_date	credit_score	DQ_TARGET
0	2021-09-01	726	0
1	2021-07-01	626	0
2	2021-05-01	680	0
3	2021-03-01	734	0
4	2021-10-01	676	0

Next, Table 2 displays the `consDF.csv` dataset which provides credit scores, evaluation dates, and delinquency targets for each consumer. This dataset is essential for building a model of credit risk, as it contains direct indicators of a consumer's creditworthiness. The delinquency targets serves as the dependent variable, enabling us to assess our model's performance in predicting credit risk. If `DQ_TARGET` $= 0$, the consumer has not gone delinquent before, and if `DQ_TARGET` $= 1$, the consumer has gone delinquent before.

1.2.3 Transaction Data

Table 3: Head of `trxnDF.csv`

prism_consumer_id	prism_transaction_id	category	amount	credit_or_debit	posted_date
3,023	0	4	0.05	CREDIT	2021-04-16
3,023	1	12	481.56	CREDIT	2021-04-30
3,023	2	4	0.05	CREDIT	2021-05-16
3,023	3	4	0.07	CREDIT	2021-06-16
3,023	4	4	0.06	CREDIT	2021-07-16

The `trxnDF.csv` dataset, as shown in Table 3, records individual transactions, including transaction category IDs, amounts, and whether the transaction was a credit or debit. These transactional data are vital for modeling consumer behavior, such as income sources, spending habits, and cash flow.

1.2.4 Category Codes Mappings

Table 4: Sample of `cat_map.csv`

category_id	category
0	SELF_TRANSFER
1	EXTERNAL_TRANSFER
2	DEPOSIT
3	PAYCHECK
4	MISCELLANEOUS

Finally, the `cat_map.csv` dataset maps transaction categories to their corresponding category IDs, allowing us to classify and interpret the transactions effectively.Table 4 shows an excerpt of this category mapping.

It is important to note that, in compliance with the Equal Credit Opportunity Act (ECOA), we excluded specific transaction categories that could introduce bias in credit decision-making. These excluded categories include CHILD_DEPENDENTS, HEALTHCARE_MEDICAL, UNEMPLOYMENT_BENEFITS, EDUCATION, and PENSION. Removing these categories ensures that our model does not inadvertently discriminate against individuals based on protected attributes.

2 Methods

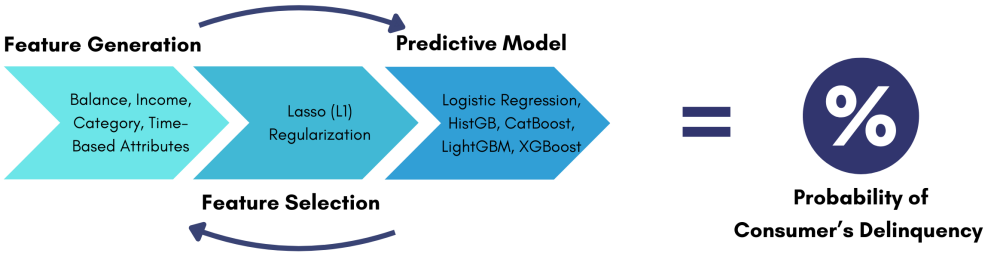


Figure 1: Cash Score Model Overview

We adopted an iterative approach to model development, emphasizing continuous refinement and enhancement of features alongside model selection and performance evaluation. We began with logistic regression to establish a baseline and identify key features. As the process evolved, we integrated more advanced algorithms like HistGB, CatBoost, LightGBM, and XGBoost, chosen for their ability to handle complex data patterns. Throughout the iterations, we focused on refining and enhancing feature generation, selecting the most relevant ones to improve the model's performance. This iterative process allows us to optimize the model's predictive power.

2.1 Exploratory Data Analysis

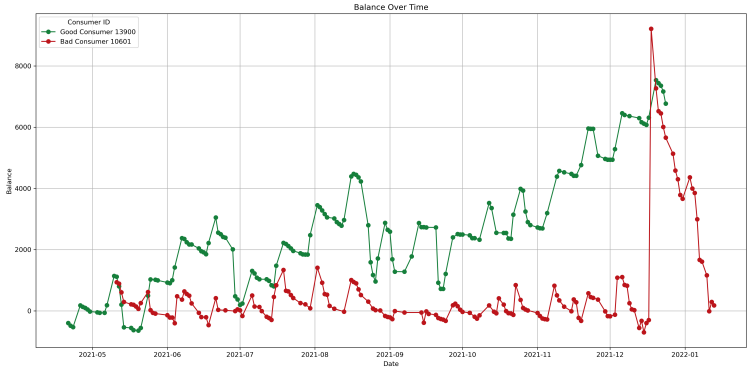


Figure 2: Balance Over Time of Delinquent vs. Non-Delinquent Consumer

Comparing bank balances over time between a randomly selected delinquent and non-delinquent consumer reveals distinct financial patterns. The delinquent consumer's balance remains mostly stagnant below 0, with a single large spike that quickly drops. In contrast, the non-delinquent consumer maintains a steady, positive balance with gradual growth, indicating stable income, controlled spending, and savings. This suggests that bank balance trends can serve as a strong predictor of creditworthiness.

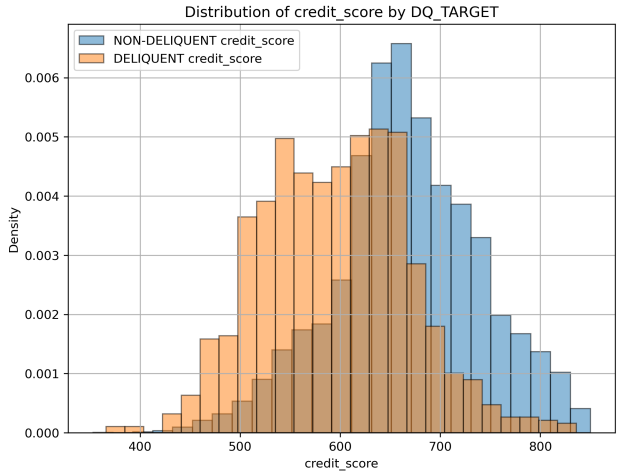
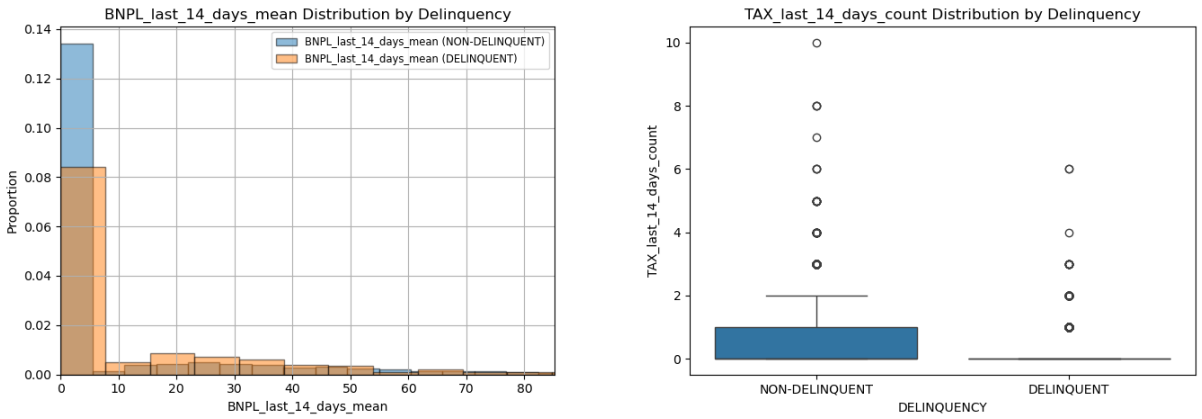


Figure 3: Distribution of Credit Score by Delinquency Status

The normal distribution of delinquent credit scores, compared to the left-skewed distribution of non-delinquent credit scores, shows that non-delinquent individuals typically have higher credit scores, while most delinquent individuals fall within the lower middle of the credit score range. This reinforces that credit scores are already a strong indicator of delinquency. This provides a solid foundation for our model, allowing us to build upon the credit score feature to potentially outperform traditional credit scoring models at predicting delinquency.



(a) Distribution of BNPL_last_14_days_mean

(b) Distribution of TAX_last_14_days_count

Figure 4: Distribution of Potential Transaction Features by Delinquency Status

Identifying ”Buy Now, Pay Later” (BNPL) as a risky category, we analyzed this category further. Figure 4a reveals that a significantly higher proportion of non-delinquent consumers fall into the lowest bin for mean BNPL transactions. However, delinquent consumers tend to have higher proportions in the upper bins, indicating that they engage in larger BNPL transactions compared to non-delinquent consumers.

Figure 4b highlights the distribution of tax transactions for delinquent versus non-delinquent consumers using a box plot. The plot reveals a wider range of tax transactions over the last two weeks for non-delinquent consumers, while the number of tax transactions for delinquent consumers centers around 0. This suggests that non-delinquent consumers are more active and consistent in handling their tax-related transactions, which could indicate better financial management and stability compared to delinquent consumers.

2.2 Feature Generation

Time Window Analysis: We analyzed transaction behaviors over multiple time frames, including 14-day, 30-day, 3-month, 6-month, 1-year periods, and all history. By incorporating different time windows, we ensured that our features reflected both immediate financial activity and historical financial stability.

Aggregated Statistics: To gain deeper insights into transaction patterns, we calculated various summary statistics for both categorical spending patterns and balance trends (inflows and outflows). For each transaction category and balance trend, and for each time window, key summary statistics were computed, including mean, median, standard deviation, minimum, maximum, sum, count, and percentage of transactions.

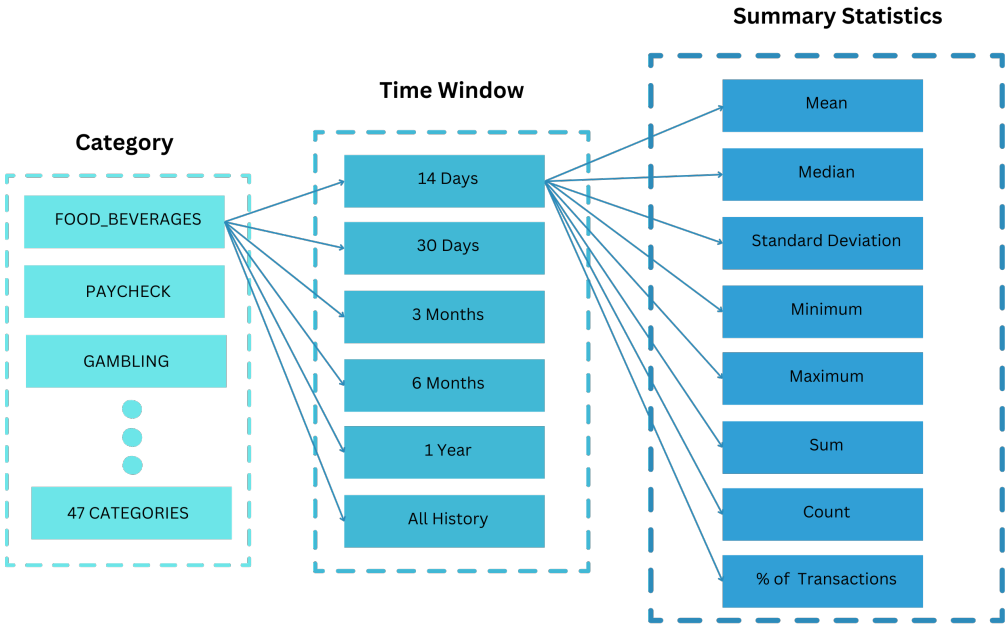


Figure 5: Category-Based Feature Generation Process

Figure 5 showcases our process for generating category-based features. For example, one of the features created through this process is `FOOD_BEVERAGES_last_14_days_mean`, which represents the average transaction amount within the "Food & Beverages" category over the past 14 days. By analyzing these features, we aim to capture spending habits.

Balance Features: A running balance for each consumer was calculated after every transaction, capturing changes in balance over time. This allowed us to create features that reflect balance fluctuations. Features such as balance deltas, rolling averages, and recent trends.

Income Features: Income-based features such as the number of income sources and income standard deviation were calculated to assess the diversity and variability of a consumer's income.

Risk Indicators: We created binary features to flag high-risk transaction categories linked to financial instability, such as gambling, overdraft fees, and "buy now, pay later" services. By incorporating these indicators, we aim to enhance the model's ability to detect potential risks in consumer behavior.

Standardization: To ensure consistent scaling and comparability across different features, we applied standardization techniques to all non-categorical variables. This step is crucial in preventing features with larger magnitudes from dominating the model.

Resampling: To address class imbalance within our dataset, we implemented a combination of Synthetic Minority Over-sampling Technique (SMOTE) and undersampling strategies. SMOTE was used to synthetically generate new samples for the minority class, increasing its representation, while undersampling was applied to the majority class to prevent overfitting. This resampling approach helped create a more balanced training dataset, allowing the model to learn meaningful patterns from both classes effectively.

2.3 Notable Features Observations:

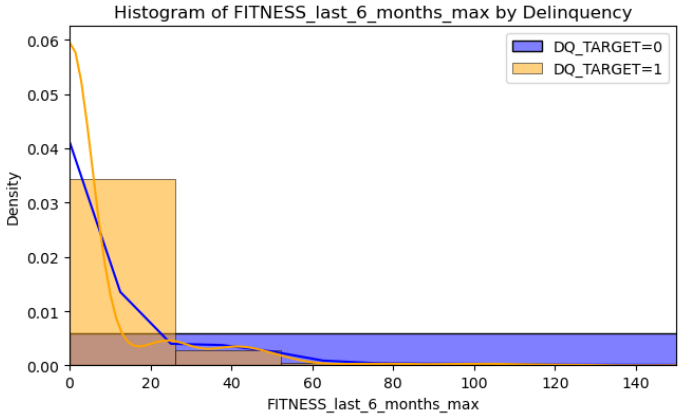


Figure 6: Distribution of FITNESS_last_6_months_max by Delinquency Status

The distribution of maximum fitness-related expenses in the last 6 months, one of the time window filtering category features we created, shows the spending differences between delinquent (`DQ_TARGET=1`) and non-delinquent (`DQ_TARGET=0`) consumers. Delinquent consumers are mostly concentrated in the lowest spending bin, whereas non-delinquent consumers are more evenly distributed across the range.

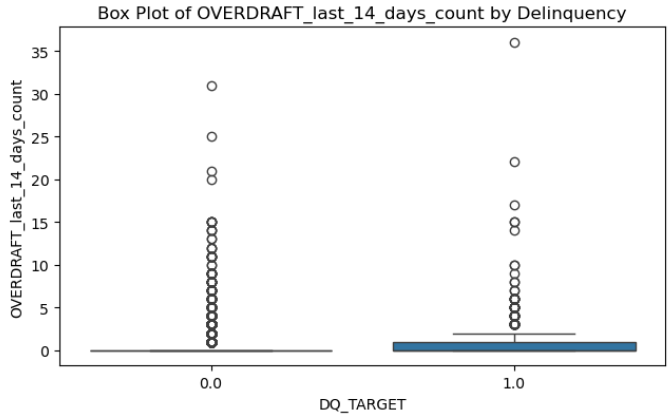


Figure 7: Distribution of OVERDRAFT_last_14_days_count by Delinquency Status

This box plot highlights the frequency of overdrafts in the past two weeks, emphasizing that while most consumers have low overdraft counts, the delinquent group exhibits a higher median and a wider range, indicating a greater frequency of overdrafts overall.

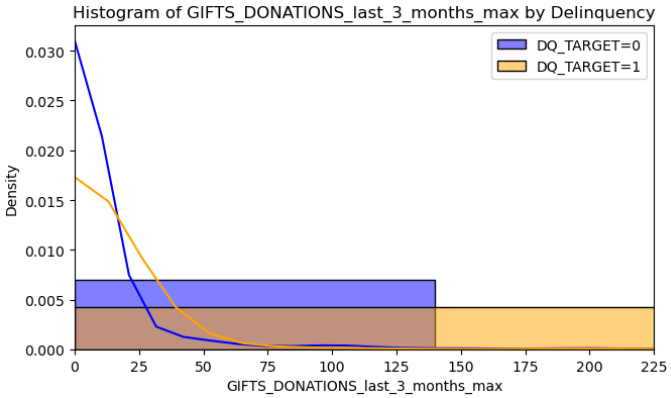


Figure 8: Distribution of GIFTS_DONATIONS_last_3_months_max by Delinquency Status

Figure 8 illustrates the distribution of maximum gift donations transactions over the last 3 months. While both delinquent and non-delinquent individuals tend to have lower donation amounts, the delinquent group exhibits a slightly broader distribution, suggesting a higher frequency of larger donation amounts. This is particularly intriguing, as one might assume that financially distressed individuals would reduce discretionary spending, yet the data reveals otherwise thus is important that we create diverse set of features to capture these seemingly counterintuitive financial behaviors, as they may provide valuable predictive power in assessing risk and understanding consumer decision-making patterns.

2.4 Feature Selection

Correlation Analysis: We conducted a correlation analysis to remove highly correlated features. Features with a correlation coefficient above the threshold of 0.9 were examined,

and only one representative feature from each correlated pair was retained to prevent multicollinearity and redundancy in the model.

Lasso (L1) Regularization: To further refine feature selection, we applied Lasso (L1) regularization, which penalizes less important features by driving their coefficients toward zero. This method allowed us to retain only the most influential features.

Embedded Method: We utilized a Random Forest model to rank feature importance based on impurity-based criteria. Features with higher importance scores were selected, ensuring that the model prioritized variables that contributed the most to prediction accuracy. This method provided an additional layer of validation for selecting meaningful features beyond correlation and regularization techniques.

By combining these selection strategies, we refined our feature set to improve model performance while reducing noise and redundancy. This approach enabled us to maintain a balance between interpretability and predictive accuracy.

2.5 Final Feature Set

The final dataset contained more than 2,000 features, with the dataframe shape being 15000 rows $\times$ 2430 columns. Since displaying all the columns in this report would not be feasible, the table below shows only a subset of the features, including some of the key metrics such as `balance_mean`, `credit_minus_debit`, and binary features like `HAS_SAVINGS_ACCT`.

Table 5: Head of final features subset, with selected columns.

prism_consumer_id	balance_mean	credit_minus_debit	INFLOWS_amt_last_14_days_max	BNPL	HAS_SAVINGS_*ACCT*
0	1,256.68	−521.59	1,250.30	0	1
1	−1,217.06	1,805.43	958.47	1	1
2	1,675.20	430.13	12.00	0	1
3	−1,439.71	2,795.24	1,400.00	0	1
4	1,736.36	−2,543.60	760.00	0	1

2.6 Cash Score Models

Baseline Model: We started with a logistic regression model as the baseline for predicting cash scores.

Advanced Modeling Approaches: To improve upon the baseline, we explored and evaluated several advanced machine learning models, including:

- Histogram-based Gradient Boosting (HistGB)
- Categorical Boosting (CatBoost)
- Light Gradient-Boosting Machine (LightGBM)
- Extreme Gradient Boosting (XGBoost)

2.7 Model Evaluation

We used the following metrics to assess model performance:

- **ROC AUC:** Measures class distinction, with higher values indicating better performance and discriminative power between positives and negatives.
- **Accuracy:** Proportion of correct predictions.
- **Precision:** Ratio of true positives to predicted positives.
- **Recall:** Ratio of true positives to actual positives.
- **Confusion Matrix:** Displays true/false positives and negatives to assess errors.

3 Results

3.1 Feature Importance

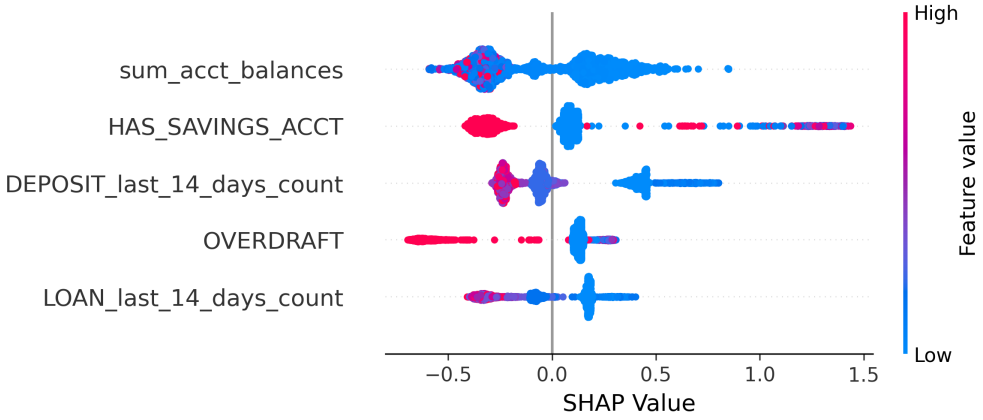


Figure 9: Top SHAP Values

SHAP (SHapley Additive exPlanations) is a method used to explain model predictions by attributing each feature's contribution to the final prediction.

In our model, SHAP identified the following key features:

- **sum_acct_balances:** Higher account balances suggest lower delinquency risk.
- **HAS_SAVINGS_ACCT:** Having a savings account reduces delinquency risk.
- **DEPOSIT_last_14_days_count:** Recent deposits indicate financial stability.
- **OVERDRAFT:** Frequent overdrafts increase delinquency risk.
- **LOAN_last_14_days_count:** Recent loans may signal financial stress.

These features were found to be the most important in predicting credit delinquency.

3.2 Reason Codes

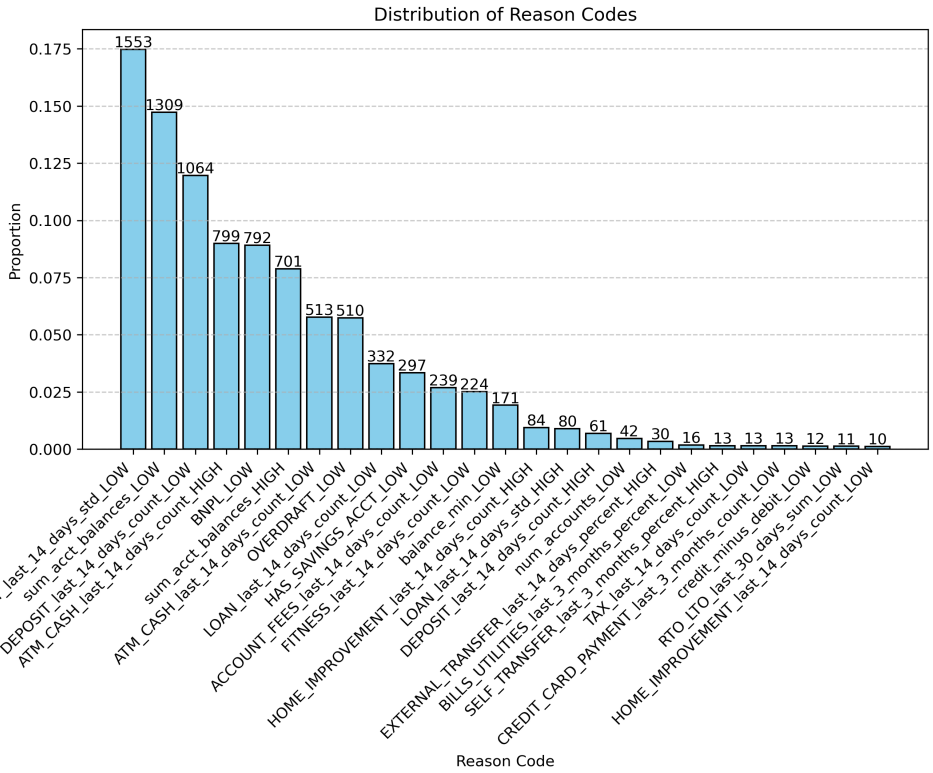


Figure 10: Distribtion of Top Reason Codes

This bar chart displays the distribution of reason codes used in our model to explain cash score predictions. When a consumer's cash score is generated, the top three contributing factors are identified as reason codes, based on model features.

For example, "DEPOSIT_last_14_days_std_LOW" and "sum_acct_balances_LOW" are among the most frequently used reason codes, indicating that low deposit activity and account balances are common drivers of lower scores.

3.3 Model Performance

In order to access its performance, we evaluated our Cash Score model both independently and with the inclusion of credit score, which is already historically recognized as a strong indicator of credit risk. This comparison allows us to understand the added value of incorporating credit score into the model and determine if our Cash Score model can effectively predict credit risk on its own.

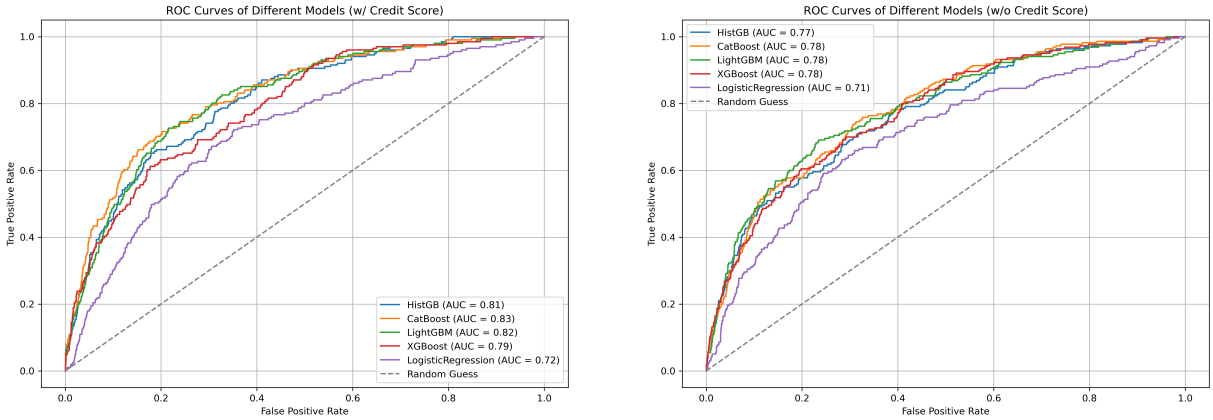


Figure 11: Comparison of ROC curves (**Left**: w/ Credit Score, **Right**: w/o Credit Score)

Table 6: Cash Score Model (w/o Credit Score vs. w/ Credit Score) Performance Evaluation

Model	ROC-AUC	Accuracy	Precision	Recall	F1-Score	Training	Prediction
Logistic Regression (w/o Credit Score)	0.7079	0.8445	0.2383	0.2785	0.2568	1.3368	0.4016
Logistic Regression (w/ Credit Score)	0.7241	0.8571	0.2674	0.3548	0.3050	1.7175	0.3315
LightGBM (w/o Credit Score)	0.7796	0.8991	0.3878	0.0802	0.1329	4.1249	0.0931
LightGBM (w/ Credit Score)	0.8162	0.9068	0.4167	0.1382	0.2076	3.9720	0.0859
CatBoost (w/o Credit Score)	**0.7704**	**0.9019**	**0.4474**	**0.0717**	**0.1236**	**38.6703**	**0.0788**
CatBoost (w/ Credit Score)	**0.8260**	**0.9170**	**0.4681**	**0.1095**	**0.1774**	**40.9512**	**0.0960**

Key Insights:

- **Impact of Credit Score:** Incorporating credit scores consistently improves all performance metrics across models. For example, in Logistic Regression, ROC-AUC increases from 0.7079 to 0.7241, and F1-score improves from 0.2568 to 0.3050.
- **Best Performing Model: CatBoost (w/ Credit Score)** achieves the highest ROC-AUC (0.8260) and accuracy (0.9170), making it the most effective model. However, despite its strong overall performance, it has a relatively low recall (0.1095) and F1-score (0.1774), indicating that it struggles to correctly identify positive cases. Additionally, it has the highest training time (40.9512), making it computationally expensive.
- **LightGBM vs. CatBoost:** LightGBM (w/ Credit Score) also performs well, achieving a ROC-AUC of 0.8162 and accuracy of 0.9068. While slightly behind CatBoost in these metrics, it has a better balance of recall (0.1382) and F1-score (0.2076), making it a more effective choice in scenarios where capturing positive cases is important. Additionally, its lower training time (3.9720) makes it a more computationally efficient alternative.
- **Best Model Without Credit Score:** Among models trained without credit score data, **LightGBM (w/o Credit Score)** is the best performer with a ROC-AUC of 0.7796 and accuracy of 0.8991. While it lags behind its credit score-enhanced counterpart, it slightly outperforms CatBoost (w/o Credit Score) in ROC-AUC.
- **Precision-Recall Tradeoff:** Logistic Regression (w/ Credit Score) achieves the highest recall (0.3548), meaning it captures more positive cases. However, its lower

precision (0.2674) results in a higher number of false positives. In contrast, both CatBoost and LightGBM have higher precision but lower recall, which means they are more conservative in predicting positive cases but also risk missing some.

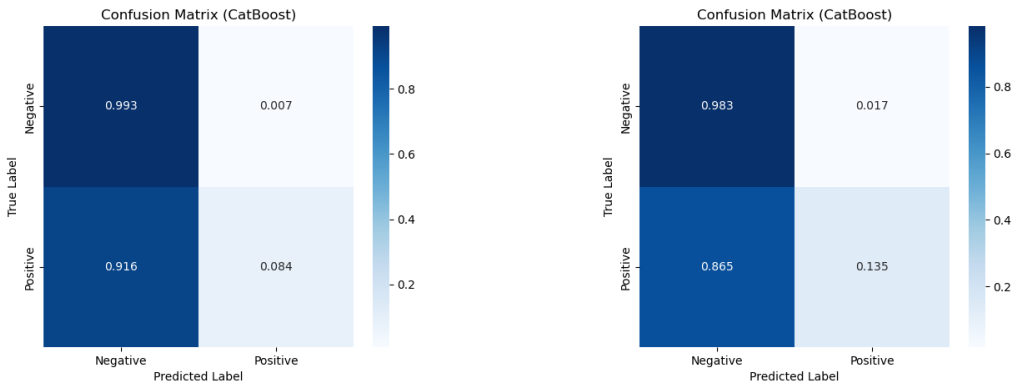


Figure 12: Confusion Matrix of Catboost (Left: w/o Credit Score, Right: w/ Credit Score

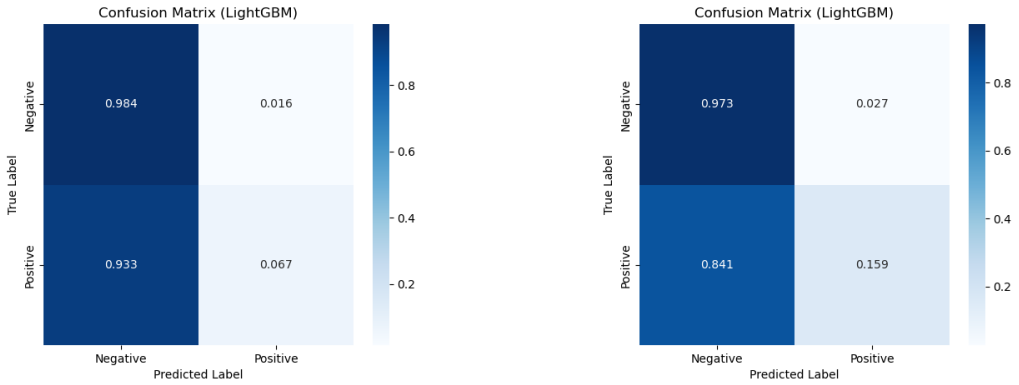


Figure 13: Confusion Matrix of LightGBM (Left: w/o Credit Score, Right: w/ Credit Score

From the confusion matrices (Figures 12 and 13), we observe that both CatBoost and LightGBM improve slightly with credit score inclusion. And they both remains highly conservative, predicting very few positive cases, which results in high precision but low recall.

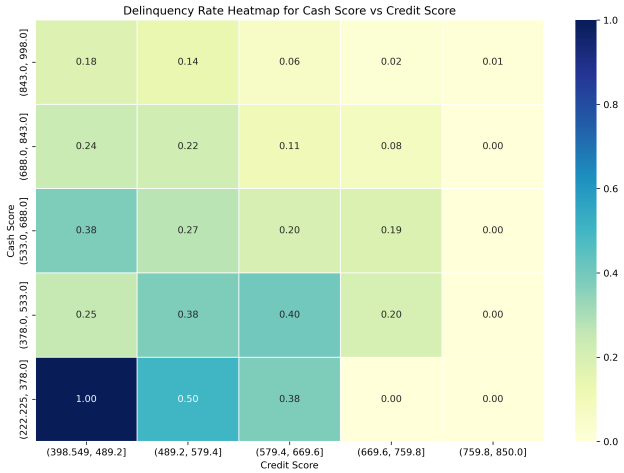


Figure 14: Delinquency Rate Heatmap for Cash Score vs. Credit Score

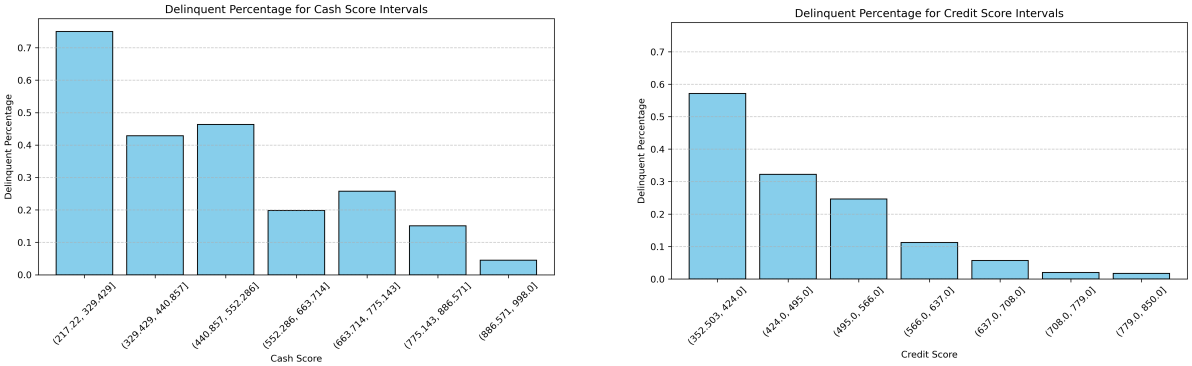


Figure 15: Delinquency Percentage by Cash Score(right) vs. Credit Score (left)

Figure 14 and Figure 15 collectively illustrate the strong inverse relationship between delinquency rates, cash scores, and credit scores. In the heatmap (Figure 14), delinquency rates are represented by both color intensity and numerical values, with darker regions indicating higher delinquency. The bottom-left region, where scores are lowest, shows delinquency reaching 100%, while the top-right region, representing higher scores, exhibits near-zero delinquency. This pattern is reinforced by the bar charts in Figure 15, which show a similar negative correlation. For instance, in the cash score chart, delinquency exceeds 70% for the lowest score range (217-329) but approaches zero for the highest (887-998). Likewise, in the credit score chart, delinquency surpasses 60% for the lowest range (353-424) and drops significantly at higher scores (779-850). Together, these figures highlight cash and credit scores as strong indicators of financial risk, with higher scores consistently associated with lower delinquency rates.

However, unlike the credit score model, where delinquency declines more smoothly as scores increase, the cash score model exhibits sharper drops and fluctuations across score ranges. This uneven distribution highlights room for improvement in refining cash score

thresholds to ensure a more consistent relationship with delinquency risk, potentially enhancing their predictive power and usability in financial assessments.

4 Conclusion

Our research demonstrates that integrating detailed bank transaction data into credit scoring models performs at a level similar to that of traditional models without the need for credit history. By analyzing transactional histories, we improve accuracy, fairness, and transparency in evaluating an individual's credit risk, increasing financial inclusion, expanding credit opportunities for individuals who have been previously overlooked or excluded.

Next Steps

- **Feature Engineering:** We aim to optimize aggregated feature metrics based on transaction categories and time windows. Additionally, we plan to implement clustering algorithms to identify and select the most relevant features for improved model performance.
- **Model Refinement:** We intend to explore deep learning models, incorporating extended hyperparameter tuning sessions to uncover more complex patterns in the data and improve predictive accuracy.
- **Bias & Fairness:** To ensure equitable credit assessments, we will evaluate the potential for biases in predictions across different demographic groups and implement fairness constraints to mitigate any identified disparities.

References

Das, M., K. Selvakumar, and P. J. A. Alphonse. 2023. "A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset." *arXiv*. [Link]

Markov, Anton, Zinaida Seleznyova, and Victor Lapshin. 2022. "Credit scoring methods: Latest trends and points to consider." *The Journal of Finance and Data Science* 8: 180–201. [Link]

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*. [Link]

Appendices

A.1 Latest Project Proposal

For reference, the link below provides access to our latest project proposal, originally created last quarter before we received our current datasets. The core ideas and general plan remain the same, with minor adjustments to the timeline and progress to align with the datasets: https://tinyurl.com/up2e5ft4.

A.2 Contributions

Aman Kar

- Developed balance-related features across multiple time periods.
- Computed descriptive statistics and performed statistical tests to assess feature relevance.
- Generated running balance features for all consumers using transaction data.
- Created plots to visualize feature distributions and balance trends for delinquent and non-delinquent consumers.
- Assisted in feature selection to refine 2000+ features to a final subset.
- Standardized the final dataset and trained models.
- Contributed to writing the Quarter 2 Project Report.
- Created build scripts for code reproducibility.
- Generated visualizations for final deliverables.
- Created application to showcase Cash Score model.

Daniel Mathew

- Developed outflow-related features across multiple time periods.
- Computed descriptive statistics and performed statistical tests to assess feature relevance.
- Created binary flags for risky transaction categories.
- Standardized the final dataset and trained models.
- Applied feature selection techniques to reduce 2000+ features to a final subset.
- Worked on model performance through enhancing feature engineering and model optimizations.

- Fit SHAP explainer to model and compiled reason codes for predictions.
- Contributed to writing the Quarter 2 Project Report.
- Created build scripts for code reproducibility.
- Generated visualizations for final deliverables.

Tracy Pham

- Developed category-related features across multiple time periods.
- Computed descriptive statistics and performed statistical tests to assess feature relevance.
- Assisted in feature selection to refine 2000+ features to a final subset.
- Reviewed and standardized code for consistency across feature engineering.
- Helped to create build scripts for code reproducibility.
- Wrote and structured Quarter 2 Project Report.
- Compiled Project Poster.
- Designed Project Website.