DSC Capstone Q2 Report

Jevan Chahal
j2chahal@ucsd.edu

Hillary Change
hic001@ucsd.edu

Kurumi Kaneko
kskaneko@ucsd.edu

Kevin Wong
kew024@ucsd.edu

Brian Duke
brian.duke@prismdata.com

Kyle Nero
kyle.nero@prismdata.com

Abstract

The process of capturing what makes a creditor trustworthy or not is especially vital within the confines of bank data, due to the guidelines and ethics of what makes this data usable. Although the quantity of the data is massive, there are only a few available features that are explicitly useful in the confines of machine learning, which calls into question how we should measure customer's trustworthiness towards their creditors. Our methodology details the process of refining bank data into categories using Natural Language Processing, assessing individual's income based on bank data alone, and also measuring their credit worthiness both accurately and efficiently.

Code: https://github.com/hillarychang/dsc180b-capstone-q2

1 Introduction

Access to credit is crucial for financial stability, yet traditional credit scoring models often exclude individuals with limited credit history. The "Cash Score" project aims to address this issue by utilizing transaction data to evaluate financial behaviors rather than just historical credit data. Our goal is to provide a more equitable scoring system that benefits both consumers and financial institutions.

1.1 Data Description

We utilized multiple datasets that provide consumer transaction details, account balances, and delinquency indicators:

- **q2-ucsd-consDF.pqt**: Contains consumer attributes like `consumer_id`, `credit_score`, and `DQ_target` (delinquency indicator).
- **q2-ucsd-acctDF.pqt**: Includes account-level data such as `consumer_id`, `account_id`, `balance_date`, and `balance`.
- **q2-ucsd-trxnDF.pqt**: Captures transactional details including `category`, `amount`, `credit_or_debit`, and `posted_date`.
- **categories.csv**: Maps transaction categories like Rent, Groceries, and Entertainment.

	prism_consumer_id	evaluation_date	credit_score	DQ_TARGET
0	0	2021-09-01	726.0	0.0
1	1	2021-07-01	626.0	0.0
2	2	2021-05-01	680.0	0.0
3	3	2021-03-01	734.0	0.0
4	4	2021-10-01	676.0	0.0

Figure 1: First few columns of the consumer dataset, including consumer ID, evaluation date, credit score, and delinquency target.

	prism_consumer_id	prism_transaction_id	category	amount	credit_or_debit	posted_date
0	3023	0	4	0.05	CREDIT	2021-04-16
1	3023	1	12	481.56	CREDIT	2021-04-30
2	3023	2	4	0.05	CREDIT	2021-05-16
3	3023	3	4	0.07	CREDIT	2021-06-16
4	3023	4	4	0.06	CREDIT	2021-07-16

Figure 2: First few columns of the transactions dataset, showing transaction IDs, categories, amounts, and whether they were credit or debit.

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

Figure 3: First few columns of the transaction categories dataset, mapping category IDs to their corresponding descriptions.

	prism_consumer_id	prism_account_id	account_type	balance_date	balance
0	3023	0	SAVINGS	2021-08-31	90.57
1	3023	1	CHECKING	2021-08-31	225.95
2	4416	2	SAVINGS	2022-03-31	15157.17
3	4416	3	CHECKING	2022-03-31	66.42
4	4227	4	CHECKING	2021-07-31	7042.90

Figure 4: First few columns of the accounts dataset, including account IDs, account types, balance dates, and balances.

		prism_consumer_id	prism_transaction_id	category	amount	credit_or_debit	posted_date	calculated_balance
prism_consumer_id								
0	136802	0	136738	14	27.62	DEBIT	2021-03-16	-27.62
	136767	0	136703	11	1400.00	CREDIT	2021-03-17	1372.38
	136803	0	136739	39	25.10	DEBIT	2021-03-17	1347.28
	136804	0	136740	37	500.00	DEBIT	2021-03-17	847.28
	136805	0	136741	14	25.00	DEBIT	2021-03-18	822.28
...	...	...	...	...	...	...	...	...
9999	1524647	9999	1522635	16	66.63	DEBIT	2023-08-08	-274.02
	1524648	9999	1522636	14	16.91	DEBIT	2023-08-08	-290.93
	1524649	9999	1522637	14	3.52	DEBIT	2023-08-08	-294.45
	1524650	9999	1522638	16	7.99	DEBIT	2023-08-08	-302.44
	1524651	9999	1522639	16	16.99	DEBIT	2023-08-08	-319.43

Figure 5: Dataframe showing all user transactions along with their account balances. Credit transactions (inflows) are added while debit transactions (outflows) are subtracted.

2 Methodology

2.1 Exploratory Data Analysis

- Identified differences in transaction patterns between delinquent and non-delinquent consumers.

- Examined seasonal trends, payday effects, and spending fluctuations.
- Estimated income using reoccuring transactions
- Analyzed the impact of account fees, buy-now-pay-later (BNPL) transactions, and overdrafts.

2.2 Balance Trends for Delinquent vs. Non-Delinquent Consumers

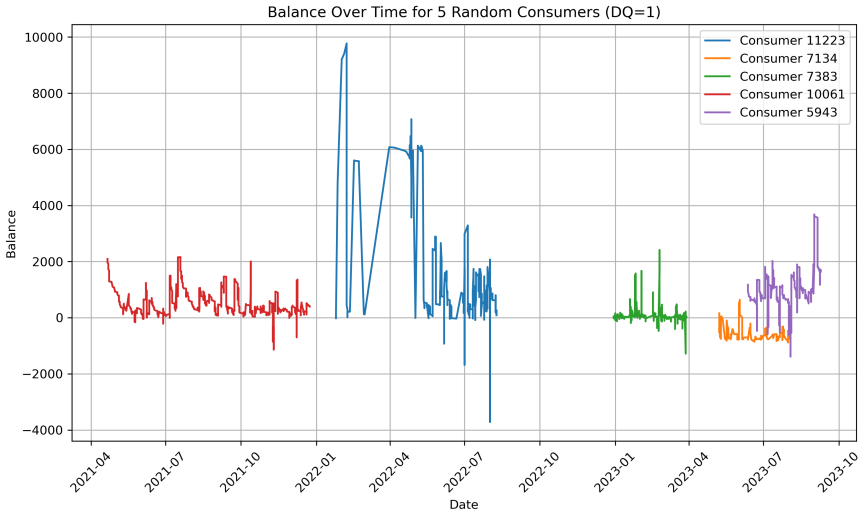

Figure 6: Balance trends over time for five randomly selected delinquent consumers. The plot illustrates fluctuations and frequent occurrences of negative balances, highlighting financial instability.

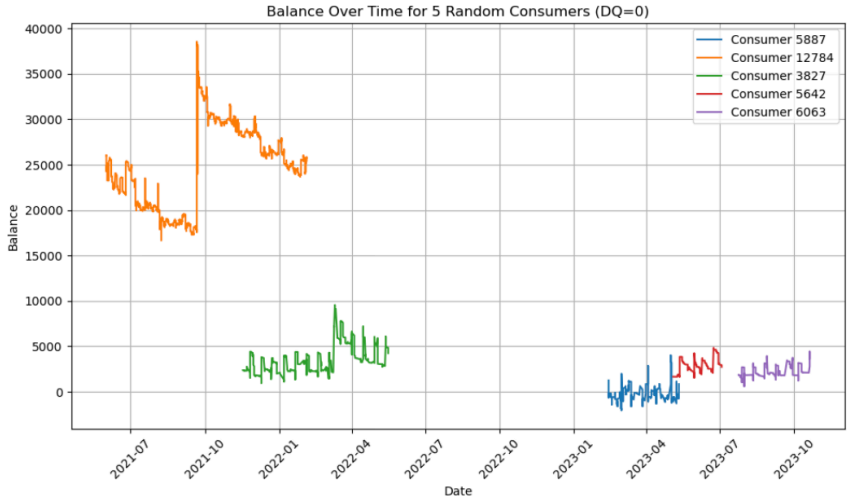

Figure 7: Balance trends over time for five randomly selected non-delinquent consumers. Compared to delinquent consumers, these users maintain more stable balances with fewer instances of overdrafts.

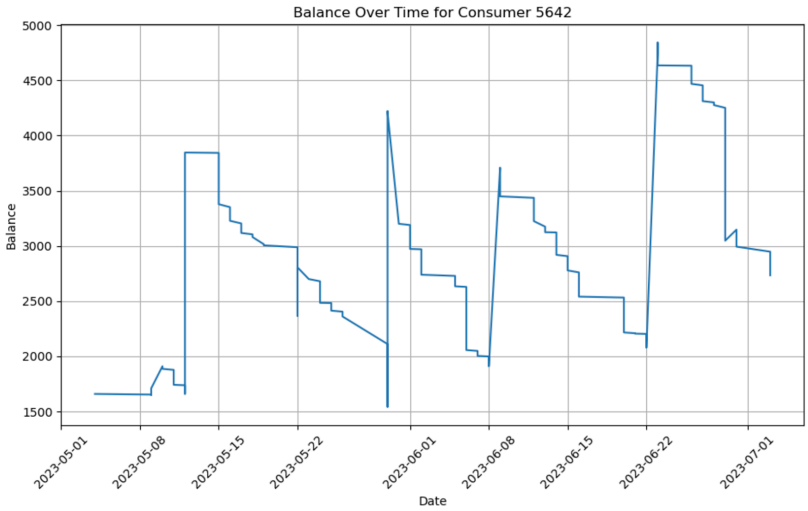

Figure 8: Balance over time for a single non-delinquent consumer. The balance exhibits periodic fluctuations, potentially due to income deposits and spending patterns, but remains above zero during the observed period.

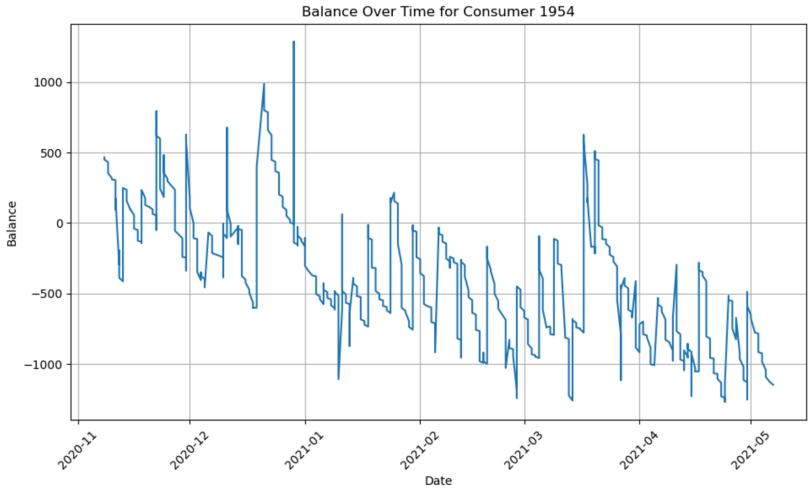

Figure 9: Balance over time for a single delinquent consumer. This consumer frequently experiences negative balances, indicating financial distress and an increased risk of missing payments.

2.3 Feature Engineering

We engineered multiple features relevant to the prediction of delinquency:

- **Balance Features**: Negative balance ratio, balance trends, payday effects.
- **Transaction-Based Features**: Credit vs. debit transaction volume, category-based spending breakdown.
- **Temporal Features**: Spending frequency over time, account for longevity effects.

- **Account Types**: Features based on the types of accounts a consumer has

	prism_consumer_id	amount	balance	spending_balance_ratio	evaluation_date	credit_score	DQ_TARGET
0	0	29295.23	320.37	91.157326	2021-09-01	726.0	0.0
1	1	48002.17	3302.42	14.531053	2021-07-01	626.0	0.0
2	10	42343.16	824.24	51.310116	2022-02-01	654.0	0.0
3	100	74979.45	2655.47	28.225220	2021-12-01	750.0	0.0
4	1000	156268.06	95.25	1623.564260	2021-03-01	756.0	0.0

Figure 10: Spending balance ratio feature created to measure how much consumers spend relative to their balance. This helps assess financial stability and risk of delinquency.

	prism_consumer_id	evaluation_date	credit_score	DQ_TARGET	balance	std_credit	std_balance
0	0	2021-09-01	726.0	0.0	320.37	0.846851	-0.146222
1	1	2021-07-01	626.0	0.0	3302.42	-0.459894	-0.090027
2	2	2021-05-01	680.0	0.0	2805.36	0.245748	-0.099394
3	3	2021-03-01	734.0	0.0	7667.01	0.951391	-0.007780
4	4	2021-10-01	676.0	0.0	394.55	0.193478	-0.144824
...	...	...	...	...	...	...	...
14995	14995	2022-03-08	655.0	NaN	NaN	-0.080938	NaN
14996	14996	2022-01-15	625.0	NaN	6821.92	-0.472962	-0.023705
14997	14997	2022-01-31	688.0	NaN	NaN	0.350288	NaN
14998	14998	2022-03-08	722.0	NaN	NaN	0.794581	NaN
14999	14999	2022-02-12	751.0	NaN	2000.94	1.173537	-0.114553

Figure 11: Feature engineering step where credit and balance were standardized to allow for easier model interpretability and comparisons across different financial profiles.

2.4 Model Training

We trained the following machine learning models:

- **Logistic Regression**: Simple, interpretable baseline model.
- **Random Forest**: Captures non-linear financial relationships.
- **XGBoost**: Optimized for structured financial data.
- **Neural Networks**: Captures complex spending patterns.
- **LightGBM**: Efficient with categorical features and handling imbalanced data.
- **Balanced RF**: Handles class imbalance by weighting classes or resampling.
- **CatBoost**: Handles categorical features automatically, robust to overfitting.
- **HistGB**: A histogram-based gradient boosting model
- **RUSBoost**: Combines Random Under-Sampling with boosting to address class imbalance while maintaining predictive performance.

2.5 Model Evaluation

Key metrics for model evaluation include:

- **Accuracy and F1-Score**: Measures classification performance.
- **ROC-AUC**: Evaluates the model's ability to differentiate delinquent users.
- **Precision and Recall**: Precision measures correct positives; recall measures detected positives.
- **Training Time**: Time required to train the model.
- **Prediction Time**: Time taken to make predictions.
- **Feature Importance**: Highlights predictive variables.

To mitigate the class imbalance (delinquents only 8.4% of dataset), we used:

- **SMOTE & SMOTEENN**: Oversampling techniques.
- **Feature Normalization**: Standardization of key variables.

3 Results

3.1 Feature Performance

To identify key predictors of delinquency, we utilized models like XGBClassifier, which highlighted several influential financial indicators. Table 1 presents the most important features based on their contribution to the model.

The top predictors include:

- **Account Type - Savings**: If a consumer has a savings account
- **Account fees**: Sum of all account fees
- **Credit Score**: Credit score of a consumer
- **Overdraft Median**: Median amount of Overdraft
- **Account fees**: Median amount of Account fees
- **BNPL Std**: Standard Deviation of BNPL transactions
- **Overdraft count**: Number of overdraft transactions
- **Investment Income Median**: Median amount of Investment income
- **Investment Income Count**: Number of Investment income transactions
- **Banking Catch All Std**: Standard deviation of a consumer's transactions within this category

Table 1: Top Features of XGBClassifier

Feature	Importance	Correlation
account_types_savings	0.044091	-0.099071
account_fees_count	0.037083	0.020680
credit_score	0.030284	-0.249976
overdraft_median	0.026024	0.000407
account_fees_median	0.021387	0.001497
BNPL_std	0.021323	0.034083
overdraft_count	0.021319	0.066101
investment_income_median	0.018767	0.004675
investment_income_count	0.017630	-0.026354
banking_catch_all_std	0.014733	-0.010585

The credit score has the strongest negative correlation with delinquency (-0.25), reaffirming its importance in assessing financial risk. Overdraft-related features and account fees also exhibit relatively strong relationships with delinquency, as consumers with high overdraft usage and excessive fees are more likely to struggle with repayments.

3.2 Model Performance

Table 2: Comparison of model performance

Model	ROC-AUC	Accuracy	Precision	Recall	F1-Score
HistGB	**0.842428**	0.913528	0.889776	0.913528	0.899533
XGBoost	0.839279	0.910069	0.889033	0.910069	0.898105
LightGBM	0.828857	0.913912	0.889361	0.913912	0.899373
CatBoost	0.823153	0.915834	0.887905	0.915834	0.898882
RUSBoost	0.805893	0.826287	0.905010	0.826287	0.857906
Random Forest	0.794283	0.915450	0.885290	0.915450	0.897285
Balanced RF	0.791598	**0.919677**	0.892594	**0.919677**	**0.902200**
Logistic Regression	0.761125	0.759416	**0.906664**	0.759416	0.814041

- HistGB achieved the highest ROC-AUC score (0.8424), indicating the best overall ability to distinguish between classes.
- Balanced RF attained the highest accuracy (0.9197), recall (0.9197), and F1-score (0.9022). This suggests that the model effectively classifies both delinquent and non-delinquent cases, particularly excelling in identifying delinquent cases

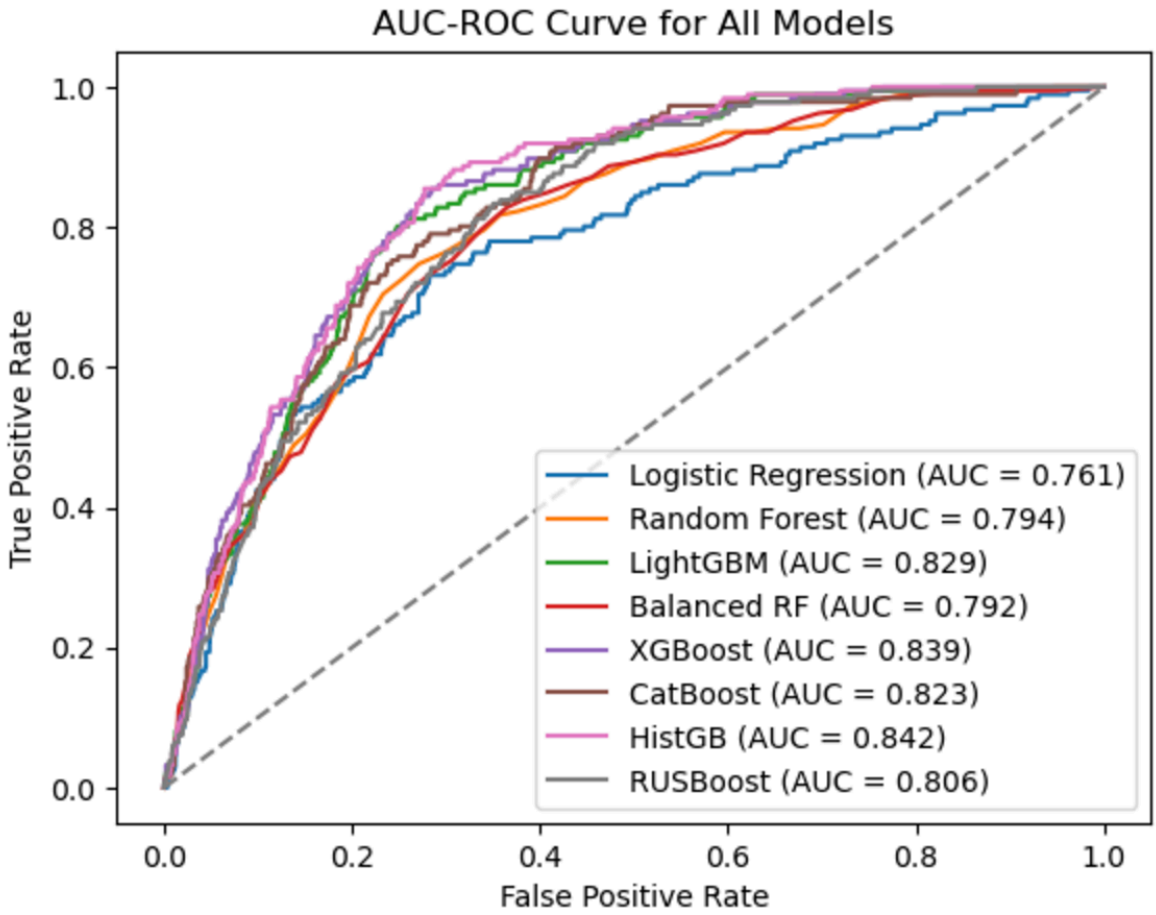

Figure 12: Comparison of all AUC_ROC scores of the different models we trained, further demonstrating the relative strengths of each approach

Apart from standard classification metrics, training and prediction time are crucial considerations. Table 3 compares the models based on:

- **Training Time**: The time required to train the model.
- **Prediction Time**: The time taken for the model to make a single prediction.

Table 3: Comparison of time-based model performance

Model	Training-Time	Prediction-Time
HistGB	8.755626	0.000015
XGBoost	5.682351	0.000009
LightGBM	4.397599	0.000012
CatBoost	42.606373	0.000009
RUSBoost	23.319771	0.000042
Random Forest	22.369477	0.000036
Balanced RF	27.226752	0.000035
Logistic Regression	**2.352007**	**0.000008**

3.3 Interpretability & Reason Codes

Top reason codes contributing to a high delinquency risk score:

- **Frequent Negative Balances**: A high negative balance ratio and frequent occurrences of negative balances per day were strong predictors of delinquency.
- **Overdraft Usage**: The total number and magnitude of overdraft transactions contributed highly to delinquency risk.
- **Buy-Now-Pay-Later (BNPL) Transactions**: Consumers with excessive BNPL spending, specifically those with a high standard deviation in BNPL transactions, displayed greater financial risk.
- **High Account Fees**: Frequent account fees, including overdraft penalties and maintenance charges, were associated to a higher likelihood of delinquency.
- **Spending in High-Risk Categories**: High expenditures in categories such as payday loans, fast cash lending, and gambling suggested increased financial instability.
- **Unstable or Low Income Deposits**: Consumers with volatile paycheck deposits or irregular income streams exhibited increased delinquency risk.
- **Late Payment History**: Features related to recurring payments and whether they were missed or delayed indicated high delinquency.

4 Conclusion

4.1 Key Findings

Through feature engineering and model evaluation, we identified several key findings relating to delinquency prediction:

- **Balance Trends Matter**: Consumers with frequent negative balances or high overdraft usage are significantly more likely to default.
- **Spending Patterns are Predictive**: Categories such as Buy Now Pay Later (BNPL), account fees, and high-risk financial services are disproportionately represented among delinquent consumers.
- **Income Stability is Crucial**: Variability in paycheck deposits and reliance on short-term lending options are strong indicators of delinquency.

From a modeling perspective, we evaluated multiple approaches, with HistGB and Balanced RF emerging as the top performers. HistGB achieved the highest ROC-AUC score (0.8424), while Balanced RF demonstrated the best overall recall and precision, making it a solid model as well.

4.2 Next Steps: Developing the Cash Score

Our next steps will be on implementing a Cash Score, a scoring system based on transactional data. We will focus on enhancing reason codes to provide clearer insights into

the factors contributing to delinquency risk. We will also work on finalizing our model predictions, ensuring that they align with both accuracy and fairness in assessing credit risk. Additionally, we will prepare stakeholder presentations to communicate our findings, methodology, and the potential impact of integrating the Cash Score into existing risk assessment frameworks.

5 Literature Review

5.1 Attention is All You Need (2017)

One of the key papers in machine learning, *"Attention is All You Need"* by Vaswani et al. (2017), introduced the Transformer architecture along with a cohort of innovative ideas for natural language processing, becoming the foundation for modern-day language models. Specifically, the architecture addressed limitations of traditional sequence models and outdated bigram models by introducing an attention mechanism that does not rely on recurrence or convolution operations, instead focusing on weights that allow models to weigh the importance of different parts of the input sequence during processing.

This key innovation enabled Transformer models to capture long-range dependencies more effectively and efficiently compared to prior models such as RNNs and LSTMs. Additionally, Vaswani et al. demonstrated that Transformer models could achieve faster parallelism, requiring less training time.

For our application in banking transaction categorization and credit default prediction, Transformer models are particularly relevant. They offer a promising approach for future banking categorization methods due to their strong capability in self-attention, which allows precise categorization of transaction memos. This specificity enhances the richness of the data available for analysis. Furthermore, the Transformer model's ability to learn patterns in spending behavior and recognize relationships between transactions over time offers potential in detecting financial distress or increased credit risk, as it learns from sequential transaction data associated with individual consumers.

5.2 Language Models are Unsupervised Multitask Learners (2019)

The paper *"Language Models are Unsupervised Multitask Learners"* by Radford et al. (2019) marks a shift in natural language processing by introducing GPT-2, a model that emphasizes unsupervised multitask learning. Traditionally, NLP models rely on supervised learning, where they are trained on labeled data. In contrast, GPT-2 was trained on a massive dataset without any task-specific labels, allowing it to perform multiple tasks in a *zero-shot setting*, meaning it can handle tasks without additional data or fine-tuning for each one.

GPT-2's ability to generalize across a range of tasks is primarily attributed to its use of the Transformer architecture, which employs a self-attention mechanism to capture long-range dependencies in text more effectively than previous models such as RNNs or LSTMs. The

Transformer's design also supports parallel processing, making it both faster to train and more efficient in application. For training, the authors created WebText, a massive dataset comprising data from Reddit web pages.

GPT-2's zero-shot learning capability makes it highly adaptable for categorizing transaction memos without requiring labeled data for each specific task. By identifying patterns in transactions, GPT-2 could potentially discover trends across transaction types as well. However, despite these strengths, GPT-2 has limitations, especially in complex reasoning, which suggests room for further refinement and improvement.

5.3 DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring (2020)

In financial risk management, accurate credit scoring and transaction categorization models play a crucial role in understanding borrower behavior and assessing creditworthiness. Traditional models like logistic regression offer simplicity but often fail to capture the complex, nonlinear patterns found in financial data. More advanced machine learning techniques, such as neural networks and ensemble methods, have demonstrated improved predictive accuracy by modeling these complexities. However, these approaches can be challenging to interpret, an essential factor for regulatory requirements in financial applications. The Deep Genetic Hierarchical Network of Learners (DGHNL) offers a solution by using a combination of genetic algorithms (GAs) and hierarchical neural networks to optimize feature selection and learning in a structured, layered way. This hybrid approach is relevant to both credit scoring and transaction categorization models, as it can capture intricate patterns in bank transaction data—vital for understanding spending behaviors and identifying risky patterns. By categorizing transactions more accurately, DGHNL could support better credit assessments by providing a clearer view of a borrower's financial habits. DGHNL's hierarchical architecture allows the model to focus on various data levels, potentially making it highly suitable for categorizing diverse transaction types. This structured learning enables DGHNL to balance detailed data analysis with high-level abstractions, enhancing both accuracy and interpretability. Consequently, the model not only improves risk predictions but also supports nuanced transaction categorization, helping to bridge gaps in credit scoring and financial behavior analysis.

References

Plawiak, Pawel, Moloud Abdar, Joanna Plawiak, Vladimir Makarenkov, and U. Rajendra Acharya. 2020. "DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring." *Information Sciences* 508 : 394–409. [Link]

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." [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]

Plawiak et al. (2020) Radford et al. (2019) Vaswani et al. (2017)

6 Appendix

Project Proposal:
https://drive.google.com/file/d/1G-DzwBNvGlgd32JJwjMDrFrnr5IMGV8t/view?usp=sharing

7 Contributions

- **Hillary**: EDA, creating & testing features, report, website
- **Kevin**: EDA, creating & testing features, creating graphs, wrote the main code that tests all of our features
- **Kurumi**: EDA, creating & testing features, creating graphs, report
- **Jevan**: EDA, creating & testing features

Everyone has met all deadlines and contributed equally.