Predicting Financial Transaction Categories: A Comparative Analysis of Machine Learning Based Methods for Optimal Classification

Brandon Dioneda bdioneda@ucsd.edu
Mert Ozer mozer@ucsd.edu
Qianjin Zhou q9zhou@ucsd.edu

Brian Duke (Mentor) brian.duke@prismdata.com
Kyle Nero (Mentor) kyle.nero@prismdata.com
Berk Ustun (Mentor) berk@ucsd.edu

Abstract

Evaluating creditworthiness has been a challenge since the 19th century, with FICO scores becoming the most widely used metric today. However, traditional credit scoring models often overlook consumers who lack conventional credit history, leading to unequal access to credit. With the advent of digital technology, alternative data sources now offer the opportunity for more inclusive credit assessments. This project addresses these gaps by analyzing non-personal financial information, such as categorizing bank transactions and predicting personal income to assess creditworthiness more fairly.

Code: https://github.com/mozer123/credit_score/

1 Introduction 2
2 Methods 4
3 Results 9
4 Discussion 9
5 Conclusion 10

1 Introduction

1.1 Background Information

Creditworthiness assessment has been a longstanding challenge, and while modern credit scores became widely adopted in 1989, these traditional models often fail to account for the financial profiles of individuals who lack conventional credit histories. As a result, millions are excluded from fair access to credit. Moreover, early credit evaluations were frequently marred by discriminatory practices, factoring in age, race, and marital status, while even today, reliance on conventional credit history can reinforce socioeconomic biases.

This project leverages bank data from 2017-2023 to develop a fairer, non-discriminatory model for assessing credit risk. By categorizing individual bank transactions and predicting personal income as alternative indicators, our approach emphasizes unbiased and responsible credit assessments, aiming to expand equal access to credit.

1.2 Literature Review and Discussion of Prior Work

In their review, Markov, Seleznyova and Lapshin (2022) discuss the evolving trends in credit scoring methodologies, highlighting the shift from traditional statistical methods, such as logistic regression, toward more complex machine learning models, including decision trees, neural networks, and ensemble methods. The authors emphasize the growing popularity of explainable AI (XAI) in credit scoring, as financial institutions are required to provide transparency in their decision-making processes, especially in highly regulated environments. They also underscore the importance of considering biases in model training, especially when alternative data sources are involved, as biased data can lead to discriminatory credit outcomes. This work sets the stage for understanding the inherent challenges in balancing model accuracy and fairness, particularly when advanced techniques are employed.

Litty (2024) explores the integration of alternative data sources, such as social media behavior, mobile phone usage patterns, and even psychometric data, in credit risk assessment models. These sources offer the potential to improve credit scoring accuracy, especially for individuals with limited or no traditional credit history, thus addressing the issue of financial inclusion. Litty's work outlines the effectiveness of AI-powered models in capturing complex relationships within these unstructured data sources, which are often ignored in conventional credit scoring models. Furthermore, Litty's research identifies the potential risks of privacy invasion and data security when using personal and behavioral data, suggesting that such risks must be carefully managed to avoid ethical concerns.

Together, these studies highlight the transformative role of AI in credit scoring, offering insights into both the benefits and challenges of these new methods. Unlike traditional models that rely on static financial data, our approach leverages natural language processing (NLP) on transaction memos and additional data sources to develop a more holistic view of consumer behavior. By focusing on these unstructured transaction details, we aim

to extract latent features that may reveal patterns of financial responsibility not captured in traditional scoring systems. This approach not only improves the granularity of credit assessment but also offers a pathway to more personalized credit models, contributing to ongoing efforts in making credit scoring more inclusive and responsive to diverse financial behaviors.

1.3 Description of Relevant Data

Consumers' bank transactions provided by PrismData which are from years 2017-2023:

- Inflows.pqt: Contains transaction-level information on inflowing transactions (such as: paychecks, refunds, etc.)

	prism_consumer_id	prism_account_id	memo	amount	posted_date	category
0	0	acc_0	PAYCHECK	2477.02	2022-03-18	PAYCHECK
1	0	acc_0	EXTERNAL_TRANSFER	100.00	2022-10-25	EXTERNAL_TRANSFER
2	0	acc_0	MISCELLANEOUS	6.29	2022-08-26	MISCELLANEOUS
3	0	acc_0	EXTERNAL_TRANSFER	277.00	2022-06-03	EXTERNAL_TRANSFER
4	0	acc_0	EXTERNAL_TRANSFER	100.00	2022-07-29	EXTERNAL_TRANSFER
...	...	...	...	...	...	...
513110	5941	acc_9524	EXTERNAL_TRANSFER	8.66	2023-01-21	EXTERNAL_TRANSFER
513111	5941	acc_9524	EXTERNAL_TRANSFER	267.13	2023-01-23	EXTERNAL_TRANSFER
513112	5941	acc_9524	EXTERNAL_TRANSFER	2.00	2023-01-24	EXTERNAL_TRANSFER
513113	5941	acc_9524	EXTERNAL_TRANSFER	207.16	2023-01-24	EXTERNAL_TRANSFER
513114	5941	acc_9524	EXTERNAL_TRANSFER	281.71	2023-01-25	EXTERNAL_TRANSFER

Figure 1: Inflows.pqt

- Outflows.pqt: Contains transaction-level information on outflowing transactions (such as: groceries, rent, etc.)

	prism_consumer_id	prism_account_id	memo	amount	posted_date	category
0	0	acc_0	LOAN	900.60	2022-07-05	LOAN
1	0	acc_0	ATM_CASH	80.00	2022-03-25	ATM_CASH
2	0	acc_0	TST* Casa Del Rio - Exp Fairlawn OH 09/24	18.42	2022-09-26	FOOD_AND_BEVERAGES
3	0	acc_0	LOAN	634.00	2023-01-10	LOAN
4	0	acc_0	Buffalo Wild Wings	26.47	2022-09-12	FOOD_AND_BEVERAGES
...	...	...	...	...	...	...
2597483	5941	acc_9524	ATM_CASH	8.42	2023-01-25	ATM_CASH
2597484	5941	acc_9524	ATM_CASH	2.06	2023-01-25	ATM_CASH
2597485	5941	acc_9524	ATM_CASH	262.88	2023-01-25	ATM_CASH
2597486	5941	acc_9524	ATM_CASH	10.00	2023-01-25	ATM_CASH
2597487	5941	acc_9524	UNCATEGORIZED	35.13	2023-01-25	UNCATEGORIZED

Figure 2: Outflows.pqt

Columns:

- prism_consumer_id: ID associated with the consumer
- prism_account_id: ID associated with the account
- memo: Descriptive text for the transaction
- amount: Transaction amount
- posted_date: Date when the transaction was posted
- category: Type/category of the transaction

2 Methods

2.1 Exploratory Data Analysis and Bias Checking

1. **Summary Statistics of Transaction Amounts:**
   The variable (transaction) ”amount” is important information that could potentially help us build our models, hence we calculated the summary statistics of “amount” for both inflow and outflow datasets.
   - Inflow Data:
     - Count: 513,115 transactions
     - Mean amount: $734.70
     - Median amount: $100.00
     - Standard deviation: $5,296.57
     - Range: $0.01 to $1,154,966.00
   - Outflow Data:
     - Count: 2,597,488 transactions
     - Mean amount: $145.13
     - Median amount: $24.26
     - Standard deviation: $1,697.88
     - Range: $0.00 to $654,853.20

   As shown in the summary statistics above, inflows tend to have higher transaction amounts compared to outflows, with a significant difference in mean and maximum values.
2. **Category Distribution:**

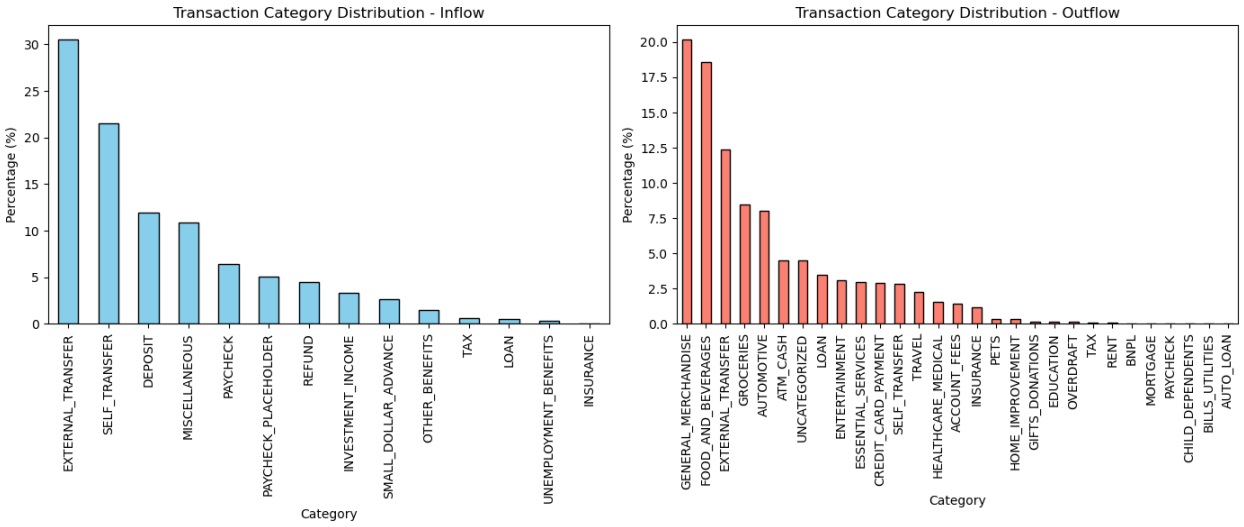

Figure 3: Category Distribution

- Inflow Data: Most common categories are EXTERNAL_TRANSFER (30.5%) and SELF_TRANSFER (21.5%).
- Outflow Data: Most common categories are GENERAL_MERCHANDISE (20.2%),

FOOD_AND_BEVERAGES (18.6%), and EXTERNAL_TRANSFER (12.4%).
We also computed the most common merchants("memo") per category for the outflow dataset. The results for the inflow dataset are not displayed because all memos exactly match their categories.

	category	memo
0	EDUCATION	Teacherspayteachers.com
1	FOOD_AND_BEVERAGES	McDonald's
2	GENERAL_MERCHANDISE	Amazon
3	GROCERIES	Walmart
4	MORTGAGE	HUNTINGTON MORTG ONLINE PMT XXXXXX
5	OVERDRAFT	Overdraft Item Fee
6	PETS	PetSmart
7	RENT	SOUTHERN INVESTO WEB PMTS
8	TRAVEL	Uber

Figure 4: Most Common Merchant per Category - Outflow Dataset

We can see that McDonald's, Amazon, Walmart, Uber, etc. are the most common merchants for each category they belong to (see Figure 4). It is worth noting that the process of finding the most common merchant is based on the uncleaned "memo" column.

2.2 Data Cleaning

To prepare our dataset for further analysis and modeling, we cleaned the memo column. Here is how we cleaned the memo with considerations we made along the way:

- Filtered rows where the memo field differs from the category field, which would allow our future models to look for more meaningful features
- The resulting unique categories are: ['FOOD_AND_BEVERAGES', 'GENERAL_MERCHANDISE', 'GROCERIES', 'PETS', 'TRAVEL', 'MORTGAGE', 'OVERDRAFT', 'EDUCATION', 'RENT']
- Converted all bank memos to lowercase
- Removed special characters and numbers to focus on text-based features
- Removed placeholders (e.g., sequences like "xxx")
- Trimmed extra spaces to clean up the text field further

	prism_consumer_id	prism_account_id	memo_default	memo	amount	posted_date	category
2	0	acc_0	TST* Casa Del Rio - Exp Fairlawn OH 09/24	tst casa del rio exp fairlawn oh	18.42	2022-09-26	FOOD_AND_BEVERAGES
4	0	acc_0	Buffalo Wild Wings	buffalo wild wings	26.47	2022-09-12	FOOD_AND_BEVERAGES
6	0	acc_0	Oculus CA 04/16	oculus ca	11.73	2022-04-18	GENERAL_MERCHANDISE
7	0	acc_0	LOS GIRASOLES STOW OH 03/08	los girasoles stow oh	30.04	2022-03-09	FOOD_AND_BEVERAGES
8	0	acc_0	BUZZIS LAUNDRY 1 OH 03/28	buzzis laundry oh	4.16	2022-03-29	GENERAL_MERCHANDISE
...	...	...	...	...	...	...	...
2597457	5941	acc_9524	DEBIT CARD WITHDRAWAL PURCHASEAmazon Prime*TI4...	debit card withdrawal purchaseamazon prime ti ...	15.93	2023-01-16	GENERAL_MERCHANDISE
2597462	5941	acc_9524	POS WITHDRAWALAZ LOT QUIKTRIP XXXX XXXX E INDI...	pos withdrawalaz lot quiktrip e indian school ...	25.00	2023-01-18	EDUCATION
2597465	5941	acc_9524	POS WITHDRAWALWAL-MART #XXXX XXXX E MCKELLIPS ...	pos withdrawalwal mart e mckellips rd mesa az ...	3.68	2023-01-18	FOOD_AND_BEVERAGES
2597468	5941	acc_9524	WITHDRAWAL Salt River ProjeTYPE: ONLINE PMT CO...	withdrawal salt river projetype online pmt co ...	90.00	2023-01-20	FOOD_AND_BEVERAGES
2597476	5941	acc_9524	POS WITHDRAWALFRYS-FOOD-DRG #1 435 S. E MESA A...	pos withdrawalfrys food drg s e mesa az card mcc	7.74	2023-01-21	FOOD_AND_BEVERAGES

Figure 5: Original Memos vs Cleaned Memos

2.3 Feature Engineering

- Features based on Memo:
  - TF-IDF: Process that helps pick out the most distinctive and meaningful words from the bank memos. Some words are present in certain transaction types. For example, the term ”nsf” (which stands for ”non-sufficient funds”) is only present in OVERDRAFT transactions
- Features based on Date:
  - `date_month`: Indicates the month that a transaction was processed.
  - `date_day`: Indicates the weekday that a transaction was processed.
  - `date_weekend`: Indicates if a transaction was processed on a weekend. Certain transactions are never processed over Saturday/Sunday.
- Features based on Amount:
  - `is_whole_num::` Indicates whether a transaction amount is a whole number. This can provide insight into the nature of the transaction—for example, most food and beverage transactions typically have fractional amounts, whereas rent payments are often whole numbers.

2.4 Models for Categorization

These are the various methods we tried using as a way to categorize an individual's bank transactions. We believe that being able to accurately categorize one's bank transactions helps us gain insight into people's financial behavior. How do they spend their money? What do they spend it on? By answering these questions, we aim to build a clearer picture of an individual's financial stability and habits, ultimately enabling a more reliable assessment of their creditworthiness.

Traditional Classifiers:

1. Logistic Regression as a Baseline
   - Logistic Regression was used as a starting point for comparison, since it is easily to implement and prediction time is relatively fast
   - Simply used default parameters from sklearn Logistic Regression that only used TF-IDF on the memo column for features
2. FastText
   - FastText is a lightweight and efficient text classification library by Facebook. It treats text data as a bag of words or bag of n-grams, providing a simple but effective approach for capturing word relationships and context
   - We processed the data to align with the format required by the FastText model and trained the model with hyperparameters (epoch=30, lr=0.1, wordNgrams=2) optimized through grid-search and cross-validation
3. XGBoost
   - We also attempted our predictive task with the gradient-boosted decision tree model from the *xgboost* library. XGBoost excels in structured data tasks. It

captures complex feature interactions and is optimized for speed and accuracy. The XGBoost classifier was trained using TF-IDF features derived from the memo text data.

Large Language Models (LLMs):

1. DistilBERT
    - DistilBERT is a lightweight transformer model derived from BERT, designed to retain most of BERT's accuracy while being faster and more efficient. It utilizes deep contextual embeddings to understand semantic relationships in text, making it effective for text classification tasks.
    - The dataset was reduced to 12,500 rows due to computational resource constraints (see Table 1). DistilBERT was fine-tuned for the task using the Hugging Face library, with hyperparameters including 3 epochs of training. The model leveraged pre-trained weights and adapted them to the transaction classification task.
2. Nemotron 70Bl
    - We utilized the open-source model Nemotron 70B from NVIDIA. Each prompt included five examples demonstrating the desired response format and predefined category options. Inference was conducted via an API service rather than being hosted locally.
3. Llama-3.2-1B
    - We employed an instructable version of the open-source model Llama-3.2-1B from Meta for this task. To optimize the fine-tuning process, we leveraged Unsloth, a library designed to enhance processing speed and reduce memory consumption. The fine-tuning was conducted in the Google Colab environment.

2.5 Income

We analyzed the inflows dataset further to investigate consumer's income. Doing this would help us answer questions like: Does this person have a steady inflow of money to be able to pay their bills? Do they have an income in the first place?

Major sources of income that we want to utilize later:

- DEPOSIT: 20.2%
- EXTERNAL_TRANSFERS: 51.5%
- INVESTMENT_INCOME: 5.7%
- OTHER_BENEFITS: 2.5%
- PAYCHECK: 19.5%
- UNEMPLOYMENT_BENEFITS: 0.6%

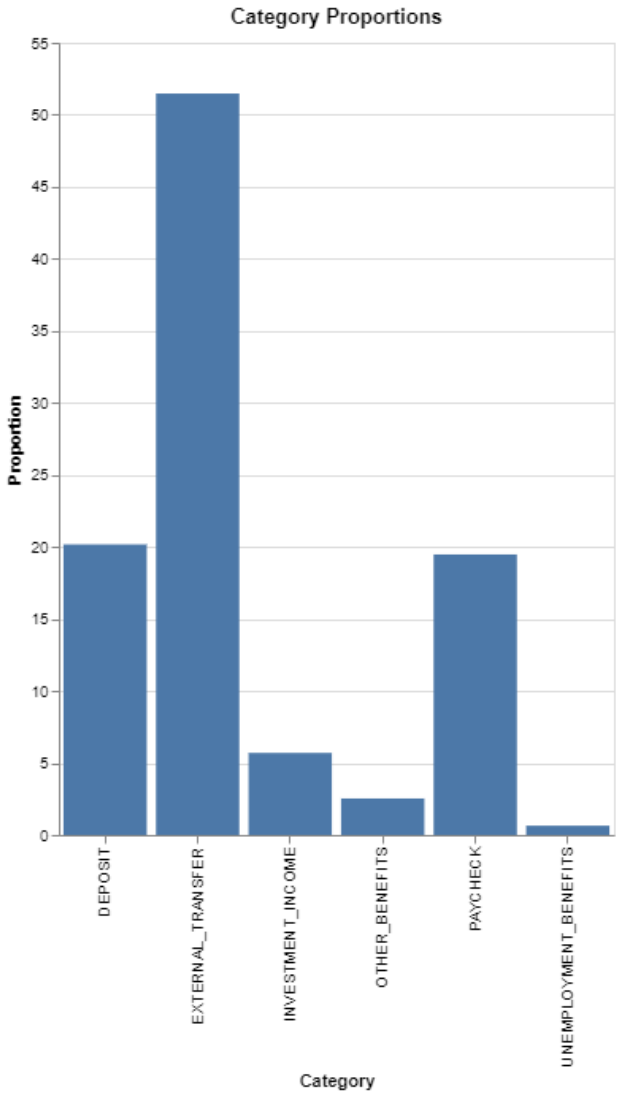

Figure 6: Income Categories Distribution

3 Results

Table 1: Performance comparison of models on transaction categorization.

Model	Training Dataset Size	Training Time (s)	Inference (predictions per second)	Testing Dataset Size	Accuracy (%)
Logistic Regression	979,839	32.8	326,613	326,613	96.45
FastText	969,666	19.1	336,786	336,786	96.88
XGBoost	969,666	30.9	300,000	336,786	91.98
DistilBERT	12,500	600	54.67	12,500	89.56
Nemotron 70B	5	N/A	0.44	100	78.00
Llama-3.2-1 B	1,000	72	3.04	100	75.00

4 Discussion

4.1 Traditional Classifiers

The performance comparison between Logistic Regression, FastText, and XGBoost reveals interesting trade-offs among accuracy, training time, and inference speed. As seen in Table 1, FastText achieved the highest accuracy at 96.88%, outperforming both Logistic Regression (96.45%) and XGBoost (91.98%).

FastText had the fastest training time at 19.1 seconds, making it highly efficient for large-scale datasets like this one. Its inference speed was also the fastest, processing approximately 336,786 predictions per second.

XGBoost lagged behind in accuracy at 91.98% and had a slower inference speed of 300,000 predictions per second, despite a reasonable training time of 30.9 seconds.

Overall, FastText stands out as the best-performing model in terms of accuracy, speed, and scalability, making it an excellent choice for text-based transaction classification tasks.

4.2 Thoughts on LLMs

Ideally, we aim for a local solution that offers high accuracy and fast inference while minimizing computational resource requirements. Although longer training times can be a drawback during testing, this is more acceptable compared to other factors, as the few-shot training premodel will only need to be trained once in the production phase.

While inference speed may not be optimal in our scenario, hosting the model locally on a supercomputer can significantly improve processing capabilities. However, accuracy remains a concern, as the LLM often generates generic responses based on its pre-trained knowledge. For instance, we observed it predicting ”GENERAL_MERCHANDISE” more frequently than appropriate.

Fine-tuning the LLMs could address the accuracy issues associated with few-shot training. However, achieving higher accuracy has been difficult due to the substantial computational resources required for fine-tuning. Despite this, the approach demonstrates promise: even when we utilized a much smaller model, it performed comparably to a larger model without fine-tuning.

5 Conclusion

Our study highlights FastText as the most effective model for categorizing financial transactions, achieving superior accuracy (96.88%), efficient training, and rapid inference compared to Logistic Regression and XGBoost. While Logistic Regression provided a reliable baseline, XGBoost, despite its utility in structured data tasks, underperformed in this context. Large Language Models (LLMs) demonstrated potential, but their practical application was limited by high computational costs and slower inference speeds.

Compared to prior works, our results align with Markov, Seleznyova and Lapshin (2022), who noted the growing efficiency of advanced machine learning models over traditional approaches like Logistic Regression in financial applications. However, unlike their findings on decision tree-based models, XGBoost underperformed in our task, likely due to its inability to fully leverage the unstructured text features in transaction memos. Similarly, while Litty (2024) emphasized the transformative role of AI-powered models incorporating alternative data sources, our work demonstrates that simpler models like FastText can outperform resource-intensive methods such as LLMs. Additionally, our focus on categorizing financial transactions based on bank memos complements Litty's findings by providing a practical and interpretable method for assessing financial behavior.

However, our approach is not without limitations. Our reliance on pre-defined transaction categories may miss subtle patterns in unstructured financial data. Furthermore, the computational constraints prevented extensive exploration of LLM fine-tuning, which may have improved their accuracy and generalizability. Finally, while FastText offers excellent scalability, its simplistic representations may not capture nuanced semantic relationships as effectively as LLMs.

Future work should focus on integrating alternative features, such as behavioral patterns or temporal trends, to enhance model performance. For instance, expanding the analysis of the income section could provide valuable insights. By measuring the regularity of inflows—such as the consistency, frequency, and source of transactions—we can better determine whether a particular transaction qualifies as actual income. This refinement could improve the reliability of creditworthiness assessments by distinguishing between steady income streams and irregular inflows.

Researchers building on our work could explore hybrid models that merge FastText's speed and interpretability with the contextual depth of LLMs, effectively bridging the gap between efficiency and complexity for more robust transaction classification. Incorporating income regularity analysis into these hybrid models could further enhance their predictive capabilities, offering deeper and more nuanced financial insights.

References

Litty, Abi. 2024. "Beyond Traditional Credit Scoring: Developing AI-Powered Credit Risk Assessment Models Incorporating Alternative Data Sources." *EasyPrint Preprint*

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