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

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.

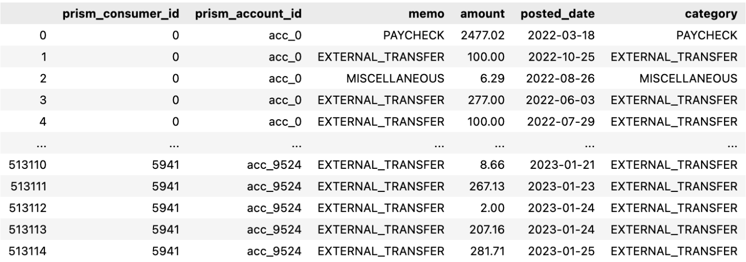
Litty (2024) explores the integration of alternative data sources, such as social media behavior, mobile phone usage patterns, and even psychometric data, in credit risk assessment models. These sources offer the potential to improve credit scoring accuracy, especially for individuals with limited or 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

Data provided by PrismData which are bank transactions that are from customers that use PrismData credit products from years 2017-2023:

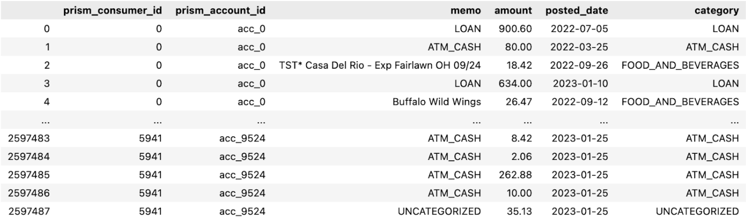
- Inflows.pqt: Contains transaction-level information on inflowing transactions (such as: paychecks, refunds, etc.)



	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:**

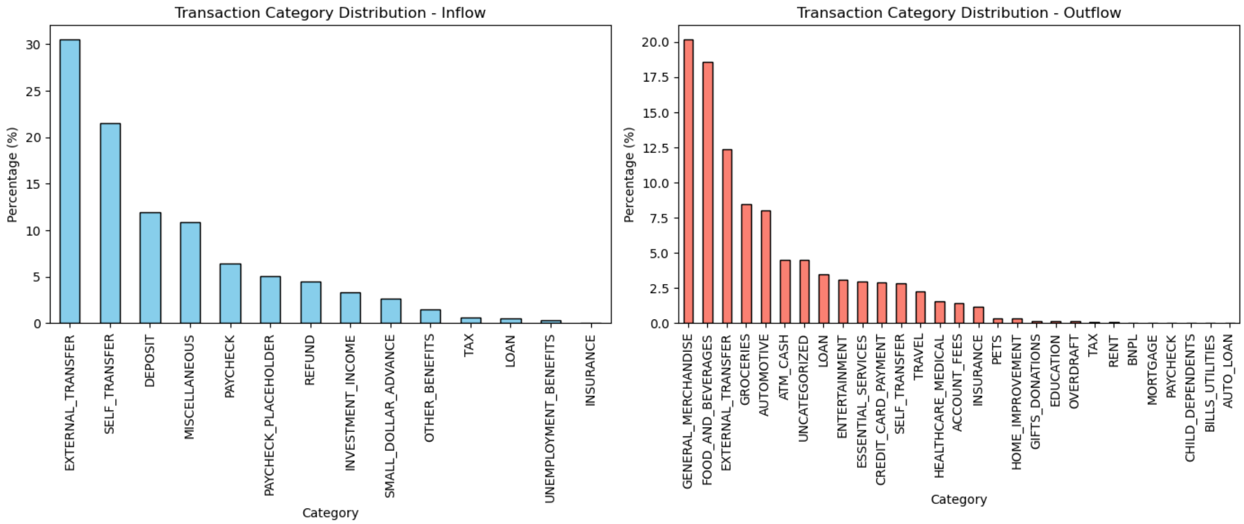
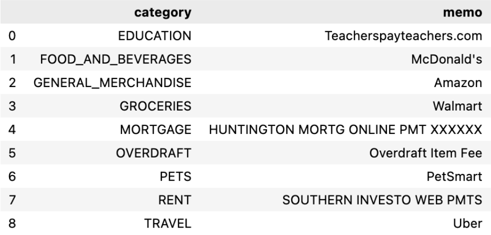


Figure 3: Category Distribution

   - 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 both outflow and inflow 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

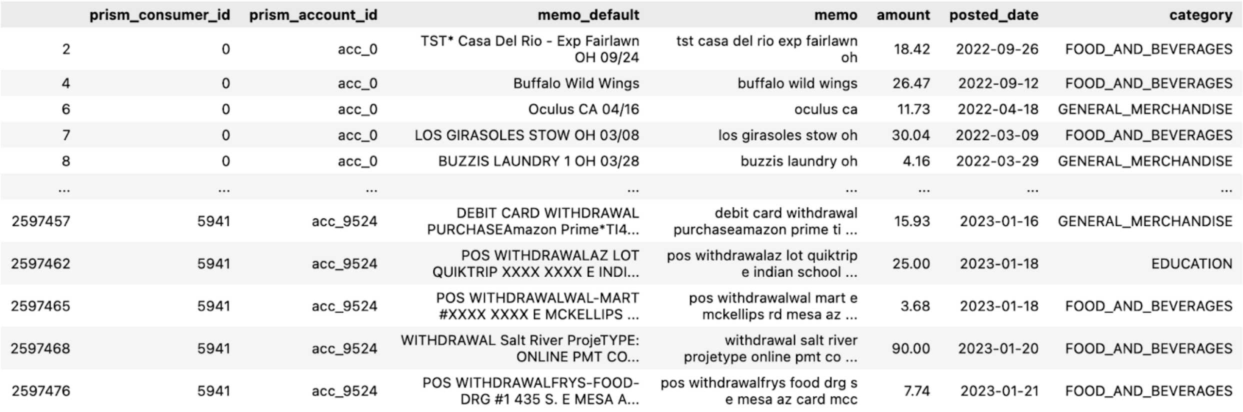
Figure 4: Most Common Merchant per Category - Outflow Dataset

From Figure 4 we can see that McDonald’s, Amazon, Walmart, Uber, etc. are the most common merchants for each category they belong to. It is worth noting that the process of finding the most common merchant is based on the uncleaned “memo” column.

2.2 Cleaning Memo

Here is how we cleaned the memo and various considerations.

- Filtered rows where the memo field differs from the category field
- The resulting unique categories are: [’FOOD_AND_BEVERAGES’, ’GENERAL_MERCHANDISE’, ’GROCERIES’, ’PETS’, ’TRAVEL’, ’MORTGAGE’, ’OVERDRAFT’, ’EDUCATION’, ’RENT’]
- Converted memo 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 Categorization Approaches

These are the various methods we tried.

2.4 Feature Engineering

2.5 LLMs

2.6 Income

3 Results

3.1 tfidf+svm

Confusion matrix and accuracy of each method. Testing conditions and environments. Inference and training times.

3.2 other method

4 Discussion

4.1 Summary of Methods

Table summarizing all factors in choosing the best method.

Table 1 presents some summary of the data.

Table 1: Some Table Caption

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Table 2 presents some summaries of the performance of our model.

4.2

This is the best model overall but these are also useful in these constraints.

Table 2: Some Other Table Caption

Method	Modality	Edit distance	BLEU	METEOR	Precision	Recall	F1
PDF	All	0.255	65.8	82.1	77.1	81.4	79.2
GROBID	All	0.312	55.6	71.9	74.0	72.1	73.0
+ LaTeX OCR	Tables	0.626	25.1	64.5	61.4	80.7	69.7
	Plain text	0.363	57.4	69.2	82.1	70.5	75.9
	Math	0.727	0.3	5.0	11.0	8.6	9.7
Nougat base (350M*)	All	**0.071**	**89.1**	**93.0**	93.5	**92.8**	**93.1**
	Tables	0.211	69.7	79.1	75.4	80.7	78.0
	Plain text	0.058	91.2	94.6	96.2	95.3	95.7
	Math	0.128	56.9	75.4	76.5	76.6	76.5

5 Conclusion

5.1 Inline Citation Examples

Citation in text (no parentheses): use `\cite{citekey}`. For example, Breiman (2001), Devlin et al. (2019).

Citation in parentheses: use `\citep{citekey}`. For example: (Vaswani et al. 2017), (Karras, Laine and Aila 2019).

To edit the contents of the "References" section, edit `reference.bib`. Many conference websites format citations in BibTeX that you can copy into `reference.bib` directly; you can also search for the paper on Google Scholar, click "Cite", and then click "BibTeX" (here's an example).

References

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–215. [Link]

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. [Link]

Karras, Tero, Samuli Laine, and Timo Aila. 2019. "A Style-Based Generator Architecture for Generative Adversarial Networks." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [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 Training Details

A.2 Additional Figures

A.3 Additional Tables