Preventing Credit Risk: Categorizing Transactions Using NLP

Jerry Luo
Halıcıoğlu Data Science Institute
University of California, San Diego
jcl009@ucsd.edu
Sammer Alomair
Halıcıoğlu Data Science Institute
University of California, San Diego
salomair@ucsd.edu
Kevin Wong
Halıcıoğlu Data Science Institute
University of California, San Diego
kkw002@ucsd.edu
Sruti Bandaru
Halıcıoğlu Data Science Institute
University of California, San Diego
srbandar@ucsd.edu

March 15, 2023

Abstract

Categorizing credit transactions is critical in the finance industry for understanding spending behavior, detecting fraud, and preventing losses. Natural language processing (NLP) can automate the categorization process by analyzing transaction memos. However, text cleaning poses challenges, such as handling misspellings, abbreviations, and non-standard language. We explore NLP techniques for categorizing credit transactions and evaluate several machine learning algorithms. Our study identifies the best-performing algorithm for transaction categorization, and we discuss its potential applications in fraud detection and customer segmentation. We also highlight limitations and future research directions. Our research demonstrates the effectiveness of NLP methods for automating the categorization of credit transactions, improving efficiency and accuracy.

1 Introduction

The categorization of credit transactions is an essential task in the finance industry. It involves grouping transactions into meaningful categories based on their purpose, which enables financial institutions to better understand their customers' spending behavior, detect suspicious activities, and prevent fraud. However, manual categorization can be a tedious and time-consuming process, especially when dealing with large volumes of transactions. To overcome this challenge, natural language processing (NLP) techniques can be used to automatically categorize credit transactions based on their memos.

Additionally, Credit scoring is a system creditors use to predict whether someone is worthy of receiving a loan. Given the sheer amount of credit seekers, it is impossible to evaluate each case on a personalized basis. Instead, there are automated methods to predict someone's creditworthiness. Linear regression was used for these purposes starting in the 1940s. Today, predictive classification methods are generally used by making use of more advanced personalized models that can learn from vast amounts of data.

In recent years, NLP has shown significant potential for automating transaction categorization. With the increasing volume of credit transactions, the use of NLP techniques can significantly improve the efficiency and accuracy of the categorization process. However, categorizing transactions based on their memos presents several challenges. For instance, transaction memos often contain non-standard language, abbreviations, and misspellings, making text cleaning a crucial step in the process. Furthermore, a transaction's memo might be insufficient to categorize it accurately, requiring additional data sources to achieve better results.

As for credit scoring, statistical methods such as regression have been widely used in credit scoring techniques and applications. More advanced techniques, such as neural networks, have also proved useful in building scoring models. Thus far, the literature has shown that both techniques, amongst other statistical methods, lead to very similar accuracies in predicting creditworthiness. Additionally, numerous evaluation criteria, such as Gini coefficients and ROC curves are used to evaluate the limitations and effectiveness of scoring models.

For these cases, it is useful to have historical data that consists of credit seekers, whether or not they're delinquent, their annual income, and their demographics, such as age, gender, city/state of residence, etc. In our case, the data consisted of a subset of about 50,000 datapoints, and 832 features, sourced from our industry partner, Petal. We have the aforementioned data of each time someone applied for a Petal card. In most cases, and in our case, the vast majority of the datapoints will represent credit seekers who are not delinquent. Thus, to avoid naive predictions due to imbalanced data (i.e., a model that just predicts everyone as creditworthy), cost-sensitive methods ought to be used, and/or data

manipulation techniques that accentuate the correlations between features and the rare class (not creditworthy) so that the model may represent the ground truth better.

2 Methods

When analyzing a credit transaction dataset, it is essential to preprocess the data before applying machine learning algorithms. In this section, we describe the methods we used to clean and preprocess the transaction memos and extract relevant features for categorization.

Firstly, we converted the memo column of the dataset to lowercase and used regular expressions to remove any punctuation marks, such as commas, backslashes, and periods, that may hinder the accuracy of the categorization model. We also removed any digits that were present in the memo column as they would not be relevant to the categorization process. Although some transaction memos may contain relevant numerical information, removing them resulted in better performance of the model overall.

Moreover, we enriched the dataset by adding features from the transaction date column, such as the day of the week and month, which could provide valuable information about the spending behavior of credit users. These features were extracted and added to the dataset to supplement the information present in the memo column.

We then used a term frequency-inverse document frequency (tf-idf) object to vectorize the transaction memos, with the stop words argument set to English and the n-gram range set to (1,3). The tf-idf object is commonly used in NLP to quantify the relevance of a term in a document based on its frequency and importance across all documents in the corpus. This allowed us to identify and extract the most relevant features for categorizing credit transactions.

To train and validate our model, we split the dataset into training and validation sets. We fit and transformed the training set using the tf-idf object and then used logistic regression to categorize the transaction memos. We chose logistic regression as it is a simple yet effective classification algorithm that can be easily implemented and interpreted. We evaluated the performance of our model using the accuracy metric provided by the scikit-learn library.

In summary, we used various techniques to preprocess and extract relevant features from the transaction dataset, including text cleaning, feature engineering, and vectorization. We then applied logistic regression to categorize the transactions and evaluated the performance of our model using accuracy metrics. Our results demonstrate the effectiveness of these methods in accurately categorizing credit transactions and highlight their potential for automating the

categorization process in the finance industry.

3 Results

First, we investigated the results of our tf-idf objects. When looking at the most frequent multi-word phrases (capped at 3), we saw a variety of service platforms, such as Uber and Netflix as well as physical stores, such as Walmart and Starbucks. Large city names such as Los Angeles and Las Vegas also appeared frequently.

Our Sci-kit learn accuracy score gave our model an accuracy of 0.86, which indicates that it can accurately categorize a vast majority of the credit transactions. This score was our subset accuracy in a multi-label classification, as it returns the fraction of correctly classified samples out of our overall data set.

With our baseline accuracy of 0.86, we set out to improve our accuracy score by tuning the model. While we were able to achieve a 0.92, this score introduced the risk of our model over-fitting to the training data-set. Thus, we decided to finalize the initial results of 0.86, which satisfied industry standards in terms of classification accuracy.

4 Discussion

The baseline model without any transformations of any of the columns had a relatively low accuracy, which means that even when preprocessing the data, the Softmax Regression model wasn't a good fit for classification in this case. We also saw that the performance of the XGBoost model and Random Forest models were relatively similar, even after optimizing their hyper-parameters. Possible approaches to increasing the accuracy of the predictions made by the models would be perhaps upsampling the training set so that the classification labels would be more balanced. Furthermore, the creation of self made features based on already existing features could also possibly increase the accuracy of label predictions.

Regarding prior work in the field, credit scoring models have been in use for many decades. Linear regression, in particular, was popular in the 1940s. Since then, however, better performing statistical techniques have come to use, such as the ones we have been using. Today, many creditors still use variations of logistic regression, XGBoost, random forest, KNN, and others. The real differentiators between the good models and the exceptional models are moreso the quality/quantity of the data they're trained on (through acquisition and preprocessing) than the type of model itself. There doesn't appear to be a consensus

in the field as to what the single best model architecture looks like, as that may very depending on the data.

When benchmarking our approach against prior work in the field, we found that our methodology follows a similar path to industry-standard practices. However, the models deployed in industry tend to have higher performance, typically exceeding an accuracy of 90%. Upon closer examination, we identified several key factors that contribute to the performance gap between our model and industry-standard models.

One significant factor is data pre-processing, which plays a critical role in the performance of any NLP model. Industry models tend to have access to larger and more diverse datasets, enabling them to perform more extensive pre-processing on the data. In contrast, our dataset may not be as comprehensive, resulting in reduced performance.

Another factor is feature engineering and selection, where industry models tend to employ more sophisticated techniques to extract and select relevant features. For instance, industry models may use more advanced NLP techniques, such as part-of-speech tagging, named entity recognition, or sentiment analysis, to enrich the features and improve the model's performance. Our model used a simpler approach to feature engineering, which may have led to reduced performance.

Moreover, industry models typically have more access to data and resources, including more powerful computing infrastructure capable of processing large volumes of data quickly. This enables them to train more complex models with greater precision and accuracy. In contrast, our model was trained on a smaller dataset and used less computing power, potentially contributing to the lower performance.

Despite the performance gap, our model provides a solid foundation for automating credit transaction categorization and demonstrating the potential of NLP in the finance industry. By identifying areas for improvement and incorporating more advanced techniques, such as those used in industry-standard models, we can work towards closing the performance gap and creating models that are more accurate and effective.

As we continue to refine our model, we recognize the importance of gaining a deeper understanding of credit scoring systems and financial transaction categorization. By improving our knowledge in these areas, we can better process the data and select/engineer more effective features for the model.

To achieve this, we plan to conduct further research and analysis on credit scoring systems, including exploring how different credit factors affect credit scores, such as payment history, credit utilization, and length of credit history. Addi-

tionally, we will investigate the factors that contribute to financial transaction categorization and develop more sophisticated techniques for feature extraction and selection.

Furthermore, we acknowledge that larger and more diverse datasets are crucial for improving the performance of our model. Therefore, we plan to acquire additional data sources and leverage more advanced computing resources to train larger and more complex models.

We believe that these efforts will ultimately lead to significant improvements in our model's performance, enabling us to make more accurate predictions about credit users and detect suspicious financial activity more effectively. Moreover, these advancements in NLP-based credit scoring and financial transaction categorization could have broader implications for the finance industry, improving the accuracy and efficiency of financial decision-making and enhancing overall financial stability.

5 References

Bakar, Engku. (2017). Credit scoring models: techniques and issues. 7.

N. T. Cao, L. H. Tran and A. H. Ton-That, "Using machine learning to create a credit scoring model in banking and finance," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2021, pp. 1-5, doi: 10.1109/CSDE53843.2021.9718414.

Knutson, Melissa L. "CREDIT SCORING APPROACHES GUIDELINES", The World Bank, 2019.

Sri charan, Manas, "Using NLP for Banking Transaction Categorization and KPI Augmentation", Saksoft Blog, 2019. https://www.saksoft.com/blog/using-nlp-for-banking-transaction-categorization-and-kpi-augmentation/

Cui, Jin. "Categorize Free-Text Bank Transaction Descriptions Using BERT", Towards Data Science, 2023. https://towardsdatascience.com/categorize-free-text-bank-transaction-descriptions-using-bert-44c9cc87735b

Chow, Ern. "Categorising Short Text Descriptions: Machine Learning or Not? ", Towards Data Science. 2022. https://towardsdatascience.com/categorising-short-text-descriptions-machine-learning-or-not-d3ec8de8c40