Understanding Consumer Credit Card Transactions and Income

Lakshmikethan Bethamcharla
lbethamcharla@ucsd.edu

Darren Jiang
dzjiang@ucsd.edu

Sheng Zhou
shz040@ucsd.edu

Victor Thai
v3thai@ucsd.edu

Brian Duke, Berk Ustun
brian.duke@prismdata.com, berk@ucsd.edu

Abstract

Throughout this quarter, our team is focusing and analyzing the process behind classifying credit card transactions and estimating consumer income to later determine a final credit score. Given a dataset with text descriptions containing a variety of text data with examples including grocery, general merchandise, and even travel transactions, we aim to classify such categories for each transaction. With the help of our industry partner Prism Data, our team's goal is to maximize our classification accuracy on the transaction data. To achieve a high accuracy and precise measure on outputs, we use text analysis methods, such as the removal of unnecessary characters, phrases, and symbols to clean our data. This data is then taken and put through text classification models with features such as TF-IDF. For estimating the income of the consumers we identified and summed recurring payments that came in quarterly, monthly, or weekly. In this project, we display how simple methods excel when either attempting to classify transaction data or estimating the income of consumers as compared to more complex techniques.

Code: https://github.com/JiaDarren/DSC_180A

1 Introduction 2
2 Methods 6
3 Results 7
4 Discussion 8
5 Conclusion 9

1 Introduction

Our main project's ultimate goal is to build a representative credit score for any given person. This quarter we focused on two parts: classifying the category of credit card transactions and estimating the income of our consumers within the dataset. Working towards that, we started with classifying credit card transaction data in our dataset. Each transaction contains a text field providing a short description/note on what that transaction was intended for. One example for this is shown when a customer purchases an item from Costco Wholesale, a leading retail company. This respective transaction provides a text field that describes where the purchase was made. Using this text description, we want to classify the potential outflow category that the transaction belongs to. However, taking a look into the initial dataset, this text field was very messy and had to be cleaned in order for us to separate our transaction into its true respective category.

To solve this problem, we have applied the use of multiple methods such as the removal of unnecessary white-spaces, special characters, and phrases in the text field. Combining these cleaning methods we created a stratified and downsampled model using BERT that provided a model which resulted in a 88% accuracy. It is important to note that this high accuracy could be due to the stratification of our data from when we initially sampled it. Our team's next steps were to create some of our own models with our personalized hyperparameters to determine if we can come up with a similar accuracy as with stratifying our initial sampled data. We proceeded to create logistic regression and decision tree classifier models with TF-IDF as an additional feature vector. These two models performed extremely well giving us an accuracy of about 97% and 98%, highlighting the impact TF-IDF has on our classification accuracy.

After looking into our classification task we moved onto estimating the income of our consumers. We utilized a simple estimation by looking into recurring inflows of our consumers. So inflow transactions that came in quarterly, monthly, or weekly were all totaled to estimate the income for each individual. From our explorations we identified that a majority of the individuals make less than a □100,000 and the median income of our dataset was around □63,000.

1.1 Discussion of previous work

For our project our industry partner Prism Data has been guiding us through our process of understanding what a credit score is, how it has been typically generated in the past, and steps in order to solve our current problem: credit card transaction classification. Prism Data has provided us with the datasets to work on this specific task. Since we aren't looking into any specific literature works, Prism Data has been our guide for the flow of this project in which they have been sharing their approaches for data cleaning, splitting, and testing. For example, when it comes to cleaning our text description data, Prism has recommended techniques they personally use themselves, such as removing whitespaces, unnecessary characters, and understanding the significance of certain phrases that could be related to a specific outflow category. Apart from cleaning they also said the feature that

helps them best classify their transactions is TF-IDF. Coming to model classification, Prism Data uses accuracy to measure their model's performance. For income estimation Prism Data uses frequency of payments as well. There suggestions have greatly aided us in approaching our tasks at hand. As we move on, Prism Data will be our main guide as they share more techniques and critiques about our models and data analysis.

1.2 Dataset

We were provided with two different datasets in order to calculate a representative credit score: an outflow and an inflow dataset. Ideally, by using both of these datasets to calculate a more representative credit score, we can address the limitations of traditional credit assessments, which often overlook critical financial behaviors, especially for individuals with limited credit histories, who are referred to as 'credit invisible'. An example of this can be people that have not had a rich history in their credit usage, however, they tend to use otehre forms of payment such as paycheck, charges amongst friends, or even in investment returns. These forms of payment aren't shown in a traditional credit check and aren't weighted with much importance. As a result, it's important that we create a new modernized credit score to determine consumer risk.

Both datasets have the same key columns shown in Table 1. The only difference is that the transactions in the outflow dataset are negative transactions (since they are transactions that decrease the amount of money in the consumer's bank account, e.g. purchases) while the transactions in the outflow dataset are positive transactions (since they are transactions that increase the amount of money in the consumer's bank account, e.g. income).

Overall, the outflow dataset holds around 1,306,452 transactions, whereas the inflow dataset has a total of 513115 transactions. With the use of these dataset, our team's goal is to develop a more comprehensive and equitable credit assessment system, utilizing past banking records and transaction data. From our analysis so far, we have discovered that the 'memo' descriptions and provided transaction categories contain a majority of an individuals' financial behaviors and patterns which we plan to use for further analysis.

1.3 Potential Biases/Problems

For this quarter, the main problem we're working on is determining the amount of risk a consumer has for not paying their bills, based on their transaction history. The goal is to create a new score similar to one's current credit score that predicts and communicates that risk to a potential client, but tailored more towards their credit card transaction history instead. This problem can be broken down into several sub-problems, each with their own potential biases and problems:

Problem 1: Memo Categorization for Outflow Transactions
The first step in determining risk is categorizing each transaction a consumer makes. We know from our current dataset that a majority of the outflow transactions are categorized as "FOOD_AND_BEVERAGE" or "GENERAL_MERCHANDISE" as seen in Figure 1.

Table 1: Column Description of Outlfow/Inflow Transaction Dataset

Column Name	**Description**
prism_consumer_id	An identifier for individual consumers.
prism_account_id	An identifier for the associated bank account.
amount	The amount of money associated with the transaction.
memo	Descriptions or memos associated with each transaction, providing insights into spending habits.
category	The category to which each transaction belongs, including the following categories: 'GROCERIES,' 'GENERAL_MERCHANDISE,' 'FOOD_AND_BEVERAGES,' 'TRAVEL,' 'PETS,' 'EDUCATION,' 'OVERDRAFT,' 'RENT,' and 'MORTGAGE.'
posted_date	The date that the transactions was made/recorded.

If not addressed, this could create a categorization model that's unduly biased towards categorizing transactions as those categories, reducing the overall accuracy of the model. Additionally, it's likely that there are many more transactions during holiday months, such as December, which could create additional bias that may affect the model's accuracy. In other words, we would also need to consider seasonality amongst our predictions.

Problem 2: Identifying Income Estimates using Inflow Transactions
Another essential problem we would need to solve would be creating accurate income estimates for each consumer. There's two major issues that need to be addressed when creating these estimates: determining whether a transaction is recurrent or not, and creating a consistent time range for income estimation

The first issue we may encounter is having to determine whether a source of income is regular/recurring or not. It's possible that a recurring income is significantly more common than non-recurring income, which would need to be taken into account when categorizing whether a source of income is recurring, or if money coming in should be considered income at all. To further this, there could be a scenario where someone is receiving deposits as income may not receive amounts in a consistent manner. There can be hiccups or skipped payments in recurring income, however, there should be an overall behaviorial pattern which can be identified that we consider to be regular/recurrent income.

Beyond that, the dataset has a varied range of dates (Figure 2) for the transactions, and different consumers have different ranges of transactions. It's possible that we may have to standardize this date range in order to improve the accuracy of our income estimation model. In our dataset specifically in the histogram below, our oldest transation dates are from 2018 with the most recent transactions from 2023. This 5 year range definitely gives us more insight into a consumer's behavior for credit scoring, but poses as an issue to determine what would be a consistent pattern for a given time frame such as annual income.

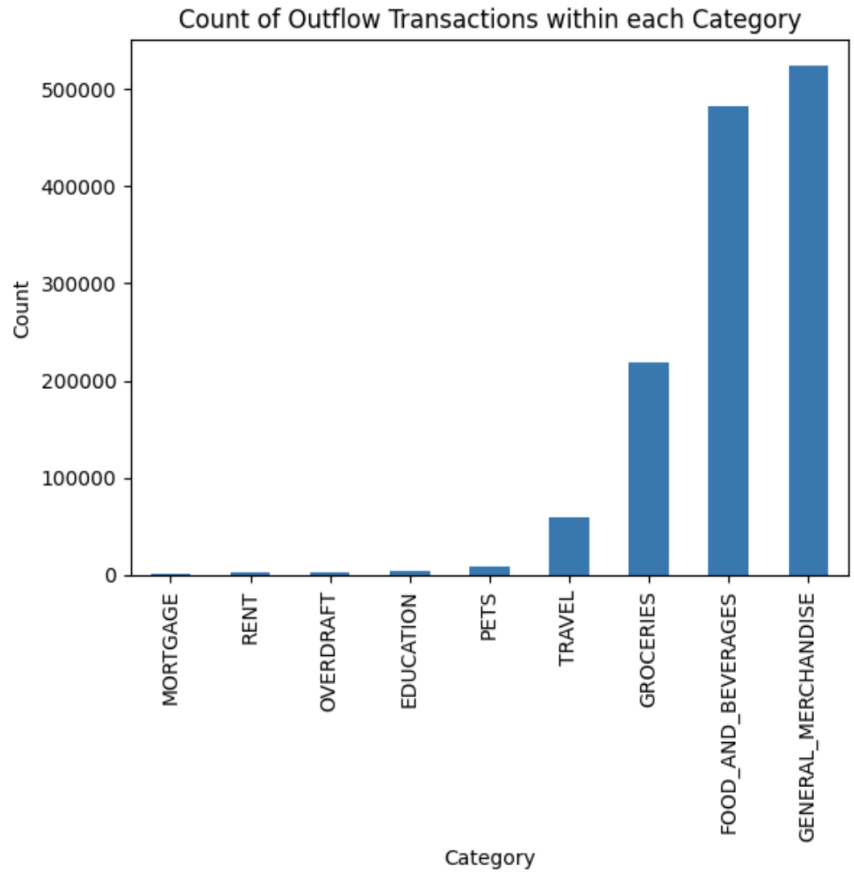

Figure 1: Histogram of Income Distribution for Consumers

Problem 3: Creating the risk assessment score
Even if all the categorizations that we did earlier are correct, we'd still have to create a score that accurately measures a consumer's risk. While it's difficult to know exactly what problems we may encounter, some potential issues that could come up is if the majority of the data consists of consumers with a certain range of income, predictions of risk for consumers that fall above or below that range may be inaccurate. Additionally, it's possible that small inaccuracies in earlier problems such as memo categorization or income identification could compound to become much larger problems, causing the overall predicted risk score to be much less accurate.

1.4 Model Assessment Metrics

Our group is currently using accuracy to assess how well our categorization model is doing at categorizing memo statements. This accuracy metric is simply the number of correct categorizations compared to the actual category of that transaction. However, during model training for BERT we used F1-Score to optimize the model. Ideally, this would allow the model to be less affected by outliers or biases within the data during training, and after some testing, we determined that this metric led to better results compared to just using

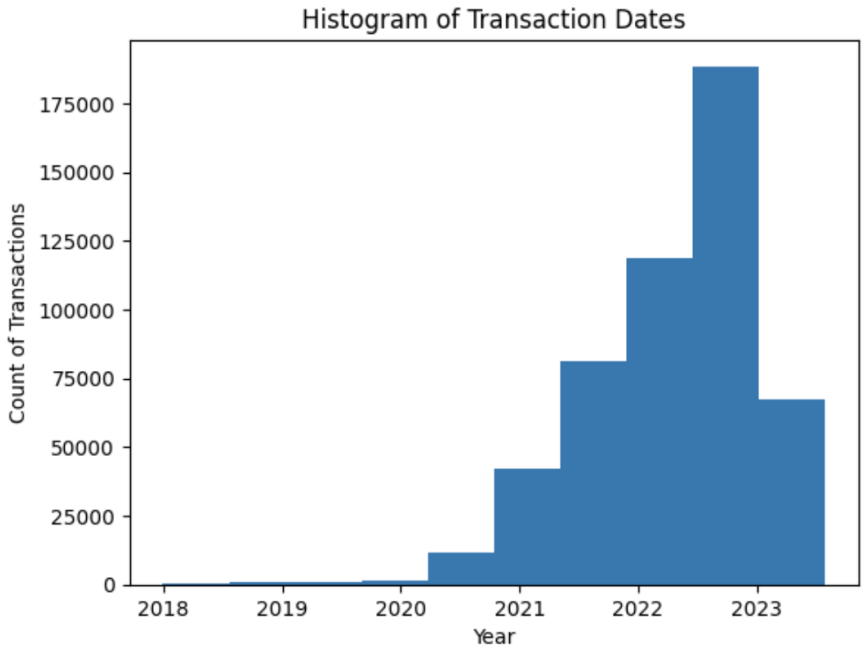

Figure 2: Histogram of the date range for transactions

accuracy for the training process. Besides BERT, the rest of the models were built and optimized upon their accuracies.

2 Methods

2.1 Categorizing outflow memos

One of the first steps towards modeling was to present a new schema for the way that the outflow memos were categorized. Initially our dataset had given us predefined categories which had a total of 9 categories for the outflow data. For our model to raise significantly more accurate results rather than just using 9 general category types, we decided to come up with a great amount of categorizations that included more tailored categories to our memo data. For this step, we decided to use TF-IDF Vectorization, also recognized as Term Frequency - Inverse Document Frequency, which measures the importance of terms within a document in relation to the collection of words per document. This was a pivotal step in our data preprocessing especially since we were able to generate 96,000 additional features for our model to use in its predictions. With the addition of 96,000 features in our vectorization process allows the model to use more clues in its analysis. Looking at the effects of this on our models after vectorization, it generated a 97% accuracy for Logistic Regression and a 98% accuracy for the decision tree classifier. This is an increase of 10% compared to our initial accuracy of 88% from BERT.

2.2 Estimating Income

Determining a user's income is arguably the most complex method in our project. There are many factors that would need to be considered for this specific task. Some questions to consider include whether the transaction is recurring, if they are of an amount to be considered income, are there other activities which aren't listed that are lacking to show a representative income, etc. Credit card transactions do have a limit as to how representative they are since there may be income that isn't exactly shown on paper all the time such as cash transactions or payment that is other than credit. However, for transactions shown in our dataset, we used a simple estimation that looked into the recurring payments offered in the dataset on whether a transaction happened more than 4 times or if it was incoming on a specific day for more than 3 times. This would consider any inflow transactions that came in quarterly, monthly, or weekly.

2.3 Primitive Assessment of Maximum Amount of Disposable Income Available

To further the income estimation, we decided to also explore the maximum amount of income consumers have according to their credit card transactions. Following with the identification of whether transactions were recurring or not, we continued with the subset of data that we obtained from the income recognition step and totaled up all the income of a respective consumer based on their consumer_id. Since our dataset was also dating back to 2018, we took the 2023 subset of the data which would show annual income for a consumer and got a right skewed histogram on all consumers. The range of consumer incomes concluded being from □1 - □1,000,000 per year. Consumers were heavily weighted towards the lower end of the spectrum below □100,000 per year.

3 Results

For the first part of our Quarter One Project we worked on classifying memos by category of credit card transactions. We utilized our cleaned data with three different models to see which could perform the best. Our accuracy results on the classification task are shown in Table 2.

Table 2: Classification Accuracy of Utilized Models

Model	**Accuracy**
BERT	88
Logistic Regression with TF-IDF	97
Decision Tree Classifier with TF-IDF	98

For the second part of our Quarter One Project we started working on estimating and understanding the income range of our consumers in the dataset. Table 3 contains the mean and median consumer incomes of our inflow dataset. Overall, the income of our consumers was heavily skewed towards below □100,000 per year as seen in Figure 3.

Table 3: Classification Accuracy of Utilized Models

Income Statistic	**Value**
Mean	□62917.45
Median	□109579.69

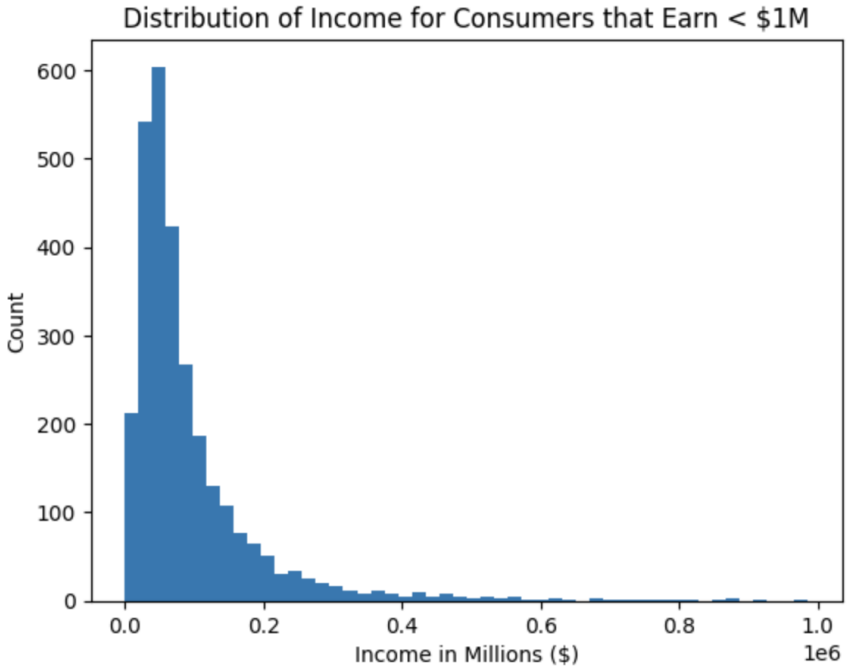

Figure 3: Histogram of Income Distribution for Consumers

4 Discussion

From the beginning of our project, it has been clear from our industry partner, Prism Data, that the ultimate goal for our project was to build a new representative credit score for any given person given the respective transaction dataset we obtained. Working towards that, the tasks for Quarter 1 were to classify the category of credit card transactions and estimate the income of our consumers. The project is an industry application compared to a research initiative so there weren't any specific literature works that we used as support. Our mentors from Prism Data were our main guides as they explained how to approach the task at hand. Therefore when it comes to understanding our results, we compare and share

all updates with our mentors from Prism Data. All of our results are similar to Prism Data as they have suggested many key approaches to solving our tasks, yet we aim to recommend and come up with new insight that Prism Data may not have explored before.

For our first task of classification we use multiple models, but the model that worked best implemented TF-IDF as a feature. These models had a super high accuracy of 97%-98%. This is an accuracy higher than what Prism Data currently uses in their product, standing around 94%. It may seem like our model works better than what Prism Data has in production, but it should be noted that this is specifically only since Prism Data deals with a much larger dataset than the subset we obtained from them. We were given a much smaller sub-sample, explaining the higher accuracy results. For our second task of consumer income estimation we looked into totaling quarterly, monthly, and weekly recurring inflows. Overall, our results are inline with the expectations of Prism Data as we used the text-classification and cleaning methods that they suggested.

5 Conclusion

The majority of the analysis and modeling that we did ourselves were in line with what Prism Data has also developed. Our analysis shows that what Prism Data uses in production is what is most accurate and reliable without much else that can be adapted in their work.

Our next steps for this overall project is to come up with a concrete score ourselves and determine whether or not a consumer is risky or not.