Swipe Right on Returns: Forecasting Stocks with Consumer Transaction Trends

Al Shodiev

Contents

1 Executive Summary

This report examines the complex relationship between total spending across different credit card user cohorts and daily stock returns for 100 unique stocks. By utilizing a linear model with lagged regressors and additional models such as Random Forest and Lasso, I uncover nuanced insights into how spending can be used to forecast daily returns.

My findings reveal a significant dependence on immediate and lagged total spend metrics as key predictors of daily stock returns. Additionally, features like cohort spend momentum and sector momentum provided complementary information.

2 Introduction

2.1 Background

Recall a common scene in high-frequency trading: algorithms sifting through billions of credit card transaction data in real-time to buy, sell or hold. A sudden spike in spending data of a certain sector of stocks triggers a bull run in the market, well before financial reports are out. Already, financial giants like Citi and BAML have began to track weekly credit card spending data that allows traders react faster to shifts in consumer demand. This has been the evolving landscape of markets where going towards alternative data territory is the new frontier for generating alpha. But why would credit card transactions be predictive of future stock returns? This is because there is a direct relationship between company performance and consumer spending. Credit card spending data offers nuanced insights into company performance such as spending trends across different groups from various demographics that have direct influence on company revenue and growth, providing information well before the release of official earnings' reports.

2.2 Research Objective

To forecast the stock return over period $(t, t+1)$ using data up to and including day t.

2.3 Datasets

There were 3 datasets - card spending, stock metadata, stock returns. Card spending data consisted of credit card transactions for 10 unique user cohorts for 100 unique stocks, spanning from Jan 1, 2015 to Jan 1, 2023. Stock metadata contained a mapping of stocks to their respective sectors. Stock returns consisted of returns for 100 unique stocks, spanning Jan 1,2015 to Dec 30, 2022. To clean and consolidate our predictor variable, I performed these steps:

1. For Card spending data, I dropped all the negative values and strings ('a298a') in spend column.
2. To deal with 14,820 missing values in card spending data, I checked missingness type and determined that data was missing at random so I dropped them.
3. I indexed each entry by the 'date' in card spending and returns datasets, converting it to a 'datetime' object for ease of manipulation. Additionally, I aligned the datasets by dropping holidays and weekends from returns and card spending data.
4. For Card spending data, I applied Box-Cox transformation to right-skewed spend column, which had extreme values up to $1e12$ to stabilize the variance and make the data more normally distributed since I will fit a linear model later which requires normality
5. I aggregated the spend across cohorts and regrouped by columns for cross-sectional analysis. For missing cohort-stock-day combinations, I filled with zeroes since absence of those combinations suggests that the cohorts did not spend on the stock on that day.
6. I regrouped stock returns into columns of time series for each stock, hence creating 2923 x 100 response variable dataset

2.4 Exploratory Data Analysis

To rigorously define the relationship between spend and stock returns, there are three specific questions I wish to answer (See Notebook for detailed plots):

1. How does the spending distribution vary across different cohorts and sectors, and are there any distinct patterns or outliers observed in specific sectors?
2. What are the key trends and seasonal variations in spending over months, quarters, or years? Are there any noticeable cyclical patterns or anomalies?
3. What relationships exist between the spending behaviors of different cohorts? Which stocks get the highest and lowest spending, and how might these relationships give us nuanced insights for forecasting?

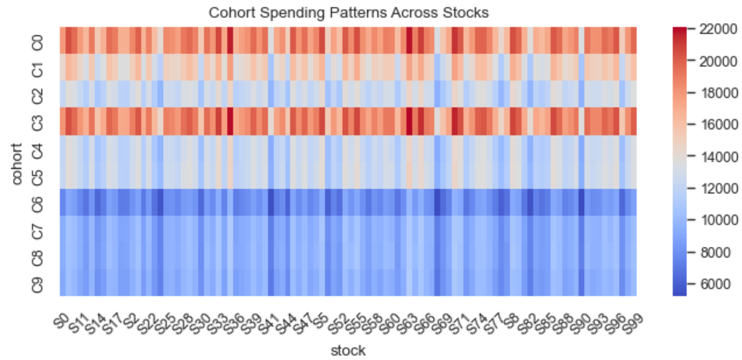

Figure 1: Heatmap between Cohort Spending and Stocks

After aggregating cohort spending into 'total spend', I found that the total spend per stock is strongly related to the number of active cohorts - 0.95 - which suggests as the number of active cohorts increases, the total spend per stock increases over time. Additionally, I calculated the correlations between Cohort Spending and Sector and performed t-test to test their significance. I found them to be weak which could indicate the impact of cohort spending on stock returns may not be immediate. **By introducing lagged spending data (e.g., spending from the previous week or month), we might uncover stronger relationships.**

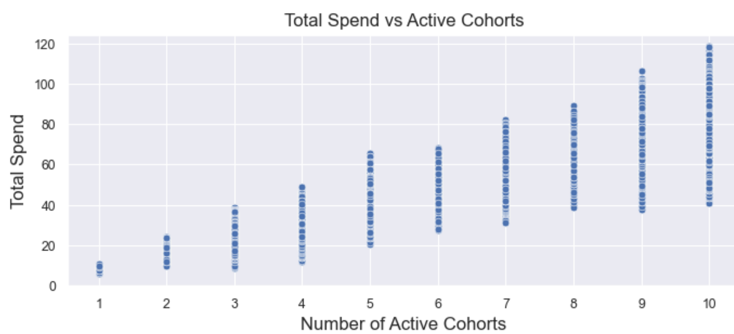

Figure 2: Total Spend vs. Active Cohorts

3 Methodology

3.1 Constructing the Spend Time Series

I constructed the Spend Time series but since we are dealing with 100 stocks, I only plotted the top 3 highest spend and top 3 lowest spend stocks.

Figure 3: Top 3 Highest Total Spend and Lowest Total Spend

3.2 Seasonality Analysis and Correction

Evidently, there is seasonality in 4b. All 6 stocks exhibit a gradual trend up in total spend over time, with noticeable dips around 2018 and 2020 which coincide with geopolitical events and pandemic. There are noticeable increases in spending around 2022 and 2023. See Notebook for more analysis.
After implementing a seasonality correction to the total spend estimate per stock and plotting the resulting predictor against the future stock return (I shifted response back by 1 in order to align with our research objective), Forecast power was 0.8, suggesting that predictor has a linear relationship with future returns which we could use for forecasting future returns

3.3 Feature Engineering

With the goal to extract more information from our data, I lagged 'total spend' by 2,5,7 days because I observed 2-6 significant spikes in Partial Autocorrelation Test plot, showing strong correlations for those lags, enough to predict future value for question 5c and question 6. I also created the following features to test additionally:

Cohort-based featuress	**Sector-based features**
Average Spend per Cohort	Average Spend per Sector
Cohort Spend Growth Rate	Sector Spend Relative
Cohort Spend Momentum	Sector Spend Momentum
Active Cohort Ratio	

The goal of smoothing is to remove short-term fluctuations and reduce noise to improve our predictive signal. I applied smoothing with a window size of 2 to 'total spend' and calculated Forecast Power which was -0.0031. We see an increase in magnitude but the value is still negative. Since stock returns are notoriously hard to predict with any single indicator, low forecast power is expected in these cases.

3.4 Modeling

I initially used a linear regression model to forecast returns using lagged regressors of 'total spend'. Additionally, I tested Random Forest and Lasso to reduce any multicollinearity between predictors and capture non-linear relationships.

4 Results and Discussion

4.1 Model Performance

With a focus on achieving a strong predictive signal, I applied smoothing with a window size of 2 to "total spend" and calculated the Forecast Power, which was 0.8. This result indicates a substantial positive relationship, suggesting that our smoothed total spend data is effective at forecasting stock returns. Given the notoriously difficult nature of predicting stock returns, this high forecast power demonstrates the strength of our approach.

Overall, the results offer valuable insights into the model's effectiveness and potential for refinement. The models exhibited very low RMSE and MAE values (around 0.5 and 0.4, respectively, for most stocks), indicating that they explain a considerable amount of variation in daily stock returns. This provides a strong foundation for further refinement. The current model achieves significance at the 0.05 level, demonstrating its robustness, but exploring additional predictors, adjusting lag structures, or experimenting with other techniques (e.g., non-linear models) may still further enhance forecast power. I also calculated Durbin-Watson (DW) statistics and Jarque-Bera (JB) p-values to assess regression assumptions. Details are available in the notebook.

- Based on the DW test, most stocks have DW values around 1.8 to 2.2, suggesting little to no autocorrelation in the residuals. Some stocks, such as S1 and S32, exhibit mild autocorrelations.
- Based on the JB test, most p-values are greater than 0.05, indicating that the residuals appear to be normally distributed. Exceptions include stocks S0 and S32, which suggest potential issues with model specification.

4.2 Sector Performance Analysis

I evaluated each linear model from question 6 by sector by aggregating their RMSE and MAE scores.

Although in figure 4, performances look neglible, I sorted the values in ascending order, showing best to worst performances. Despite there being discrepancies in the middle bars, both metrics display an agreement that Sector 10 is the best performing and sector 20 is the worst performing.

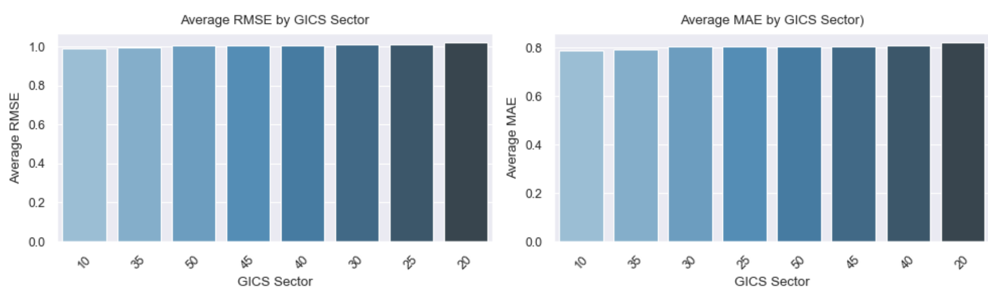

Figure 4: Sector Results

Regarding the inclusion-exclusion of worst performing sectors, it's important to weigh the pros and cons. Removing poorly performing sectors can make the model less generalizable, especially if future data includes those sectors. Additionally, removing sectors may reduce diversification and increase the risk of sector-specific idiosyncrasies dominating the portfolio of stocks. However, removing the worst-performing sectors can simplify the analysis and reduce error metrics. Given additional time, I would first analyze why some sectors perform worse and explore ways to improve the model for those sectors before considering discarding them from the model. I would hypothesize that the well-performing sectors are probably sectors heavily dependent on consumer spending such retail and worse-performing sectors are those that are not directly dependent on consumer spending.

4.3 Trading Strategy

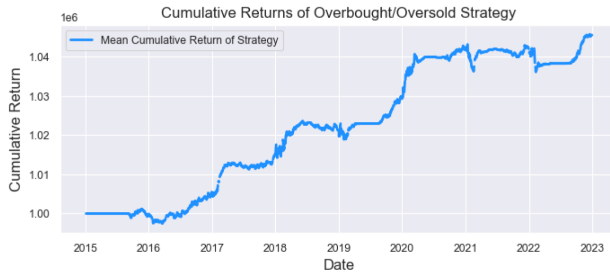

Figure 5: Mean-Reversion Strategy: Cumulative PnL

5 Conclusion

The integration of alternative data in forecasting stock returns offers a fresh lens through which to view market movements. This report explored the relationship between credit card spend data and daily stock returns. My findings show that spend data has a notable impact on stock return movements.

I posit two avenues for further research: Firstly, since we are predicting company performance, additional data on company's financial health such as growth rates and ratios as these provide a more accurate comparison across firms of different sizes and sectors. In addition, we could also extract discretionary spending trends from credit card data and align it with Personal Consumption Expenditure (PCE), we can estimate consumer confidence. Secondly, I would reframe the problem into a classification task where instead we consider a portfolio with n stocks and we attempt to capture directional changes in returns for the next-day portfolio return. This would reduce the complexity of the problem where noise and idiosyncratic risks associated with each stock is diversified.