Areeb Abubaker & Jordan Bickelhaupt

Prof Besser

DSC 324 Final Paper

19 March ‘21

I. **Non-technical**:

With the changing tides of the market in mind, the set objective is to derive critical and meaningful analysis which may lead to a determination for which opportunities will profit, and which will not. Due to the enormous size of the data collected, we operated and researched smaller subsets of the data to predict opportune times to decide a particular action for each opportunity.

The Jane Street Market dataset is composed of features that represent various stock market indicators. The dataset is used to predict entries are valid candidates to purchase any given stock opportunity. Along with features, weight and resp are variables that depict the returns of the trade. Also included are the date and ts_id, which is a time signature ordering.

Using the data collected, we created some predictions in which scenarios would be most advantageous to take action on an opportunity that would yield the best returns.

II. **Application, dataset, research question: *Predict optimal buy signals***

The application of the dataset is to predict whether a trading opportunity should be bought or not. For this application, there is no sell or stop-loss trading options. Each entry of the dataset is an opportunity which has various signals (features) which may help or hinder how

desirable a given opportunity may be. The question that arises is whether market data can be effectively divided to ensure profit.

III. Brief background

Many factors can influence and affect the market, from governmental actions, economic situations, employment, social media to technical advancements. Since the market can be easily controlled, it is purely impossible to predict the market 100% of the time since there is always new data or news coming that will disprove a previous action. Many models have been tried on market data. While some are successful, there isn't a straightforward and foolproof method to ensure an opportunity's profitability. With each technique that will be used, the predicted outcome may not be guaranteed; however, if most effects are expected correctly, profitability may result.

IV. Description of methods (modeling)

Our attempt at figuring out the best times to make an action was to make a single model with the factors given by the dataset, but that resulted in a poor performance in predicting [opportune] times. We then tried to cluster the thousands of entries into two in order to determine whether any given opportunity will be profitable or not.

V. Results (laymen terms)

After hours of analysis, we found success in a cluster analysis where we were able to predict accurately (about 52%) of the best times to invest into an opportunity. Even though we weren't

able to increase this accuracy performance, we were still able to produce a positive outcome. 52% correct may not seem entirely accurate, however, provided that this model is used at a high frequency, profits will be made immediately evident, so long as the accuracy is above 50%.

VI. Limitations

The market can be unpredictable. Even today, a full-proof prediction method has not yet been developed. While some models have been successful, there isn't a guaranteed model despite the hype surrounding market prediction. We also understand that with only a subset of the analyzed data, we aren't accounting for some external factors such as economic down turns and economic policies.

Understanding the economic market requires years of experience and other economic factors that weren't considered when conducting our research. Our study promises some reliability in predicting market orders. We stress that there is much to be desired even after finding some particular discoveries and market volatility patterns.

VII. Final Conclusions about research

Our research concludes by revealing that even though the anonymized predictors for our dataset may have hindered our progress in making connections and our modeling techniques, we were successful in finding traceable patterns with our clustering, which demonstrated a 52% accuracy statistic with predicting opportune investment opportunities. The research done can be further used as another modeling study and give ideas on how to invest based on investment proportionality.

ABSTRACT

The market is one of the most unpredictable environments in the world, to this day, there isn't a full-proof modeling technique. For this dataset's application, there is a weighted response variable which is a proportionate ratio for the amount of money returned on an investment opportunity. Linear regression, principal component analysis as well as cluster analysis are a few methods of which have had some success in predicting this value. Linear regression revealed relevant correlations and provided some valuable insights; however, the regression was unsuccessful in predicting opportunities. Some analytical insight that was taken into account from regression was a better understanding of the variables that were being dealt with, as well as which were truly collinear and were rendered useless to include in a final model. Principal component analysis revealed a further insight that there was an underlying bidirectional pattern to the data that regression was unable to capture.

After principal component analysis, there was a greater amount of optimism for discovering new patterns. The results revealed the amount of variance that was captured within only a few principal components. After a proper cluster analysis validation, the bidirectional data was captured and a two-cluster model was fit which predicted a profitable opportunity at a 52% accuracy. For future applications, principal component analysis, k means cluster analysis, as well as a combination of each of these are valid and seemingly profitable approaches to determining a profitable approach to predicting the market.

1. INTRODUCTION

Within any given moment, another source of news may come to influence the reaction of a stock's performance. At the drop of a penny, a reputable source may indicate a change in the environment which may positively impact one stock and negatively another. These changes impact how people will look into a particular market and their investment strategies. These factors can significantly affect the stock market's overall flow and momentum which can range from government intervention, trade wars, economic downturn, and technological advancements. Housing all these factors into a singular dataset would take a large toll on performance as not only would the size of the dataset be enormous, but the number of factors which may contribute are astronomical. As the initial size of the Jane Street Market dataset was two million observations, and data analysis was performed within R 3.6, there wasn't an immediate way to perform analysis of the entire dataset in a timely fashion without crashing the process.

For this reason, a subsetted number of rows was taken from the large dataset for a fair performance analysis. Previous applications of stock market analysis have usually resulted in some success in short-term increments, but long-lasting efforts are fruitless. Analysis techniques including linear regression and non-parametric analysis have been uneventful, but with more robust machine learning algorithms, such as k-means clustering and multidimensional scaling, market prediction's accuracy statistics have improved for more extended periods. By taking into account some external factors, these predictions have increased accuracy, and with further applications and research, more factors can be accounted for to make a seemingly near-perfect model.

The Jane Street Market dataset contains a return value as well as various factors for each stock opportunity. The dataset includes a variable labeled as response which indicates the amount

of money returned on the day that the stock was purchased. This research project's significant predictors came down to the anonymized features variables, which were randomly numbered from 0 to 149. As the dataset was part of a competition, the features may have been randomized and anonymous to prevent interfering with real-world stock features and data, further skewing the competition. These features were interpreted as market signals that created an effect on the market somehow that wasn't distinguished. We went ahead and used these features and combined them with the weight variable to have an understanding of the dollar proportionality invested into a particular opportunity.

To have some idea in proper spending measures, we conducted a linear regression (for baseline purposes) with more efforts conducted with PCA, Multidimensional analysis, Cluster analysis, and concluding our modeling techniques with the K-Means algorithm. The dataset gave us some insight into buying patterns with proportionality to the dollar amount invested, which furthered our spending returns.

2. LITERATURE REVIEW

As mentioned previously, the stock market has various elements that can be used to analyze and predict future investment opportunities. The research team took inspiration from [1], where they were able to do some data mining techniques and cleaning with principal component analysis, and applying a k-means algorithm. Their application was useful because it heavily relied on their labeling procedures and cleaning measures that were needed for rich datasets. Their analysis retrieved actual data from companies that helped them study the stock market in understanding growth factors for a company. [2]'s research created models that helped them know which markets influenced other external markets and their reactions to a specific economic

growth or downturn. These noteworthy techniques were considered and used to further our visualization and analysis of our dataset.

The research team also considered [2]'s methodologies, where their study provided much of our inspiration for modeling techniques and understanding of the market data. Their research consisted of macroeconomic factors such as economic states, inflation, unemployment, and company growth ratios to determine their best investment opportunities.

The limitation within the Jane Street Market project comes down to the dataset where the features were anonymized and thus was harder to distinguish which was which, but nevertheless gave a giant boost in how similar markets are when a certain action is taken. We also considered that because it was impossible to predict the data 1-1, we made sure not to overfit our data by performing a split to ensure proper validations and model analysis for further applications. For this reason, our analysis will constitute a much broader application in determining investment opportunities.

3. METHODS

The analytic techniques that will be applied to analyze and predict the Jane Street market data include: linear regression, principal component analysis, classical multidimensional scaling, as well as cluster analysis. Linear regression seemed to be an appropriate analysis technique for market prediction as features should be correlated to the performance and return of an investment but was not entirely the case and was unsuccessful. Principal component analysis revealed that there was a significant trend of covariance occurring across only a few principal components. Classical multidimensional scaling further supported that there was a clear bidirectional trend

found within the dataset. Cluster analysis was then used to capture this covariance and predict which entries would have positive responses.

4. DISCUSSION & RESULTS

After data cleaning and preparation, a linear regression analysis was fit to the response variable within the dataset. This proved to be unsuccessful even after removing finetuning as the r-squared values would not reach 0.5 and was even difficult to exceed 0.2. [4]'s outputbecame clear that the most significant correlations to a positive response were not prominent enough to relay strong confidence predictions.

A principal component analysis revealed that a high degree of variance (upwards of 80%) was able to be captured within only a few principal components. The scree plot [5] showed that three components were most suited to capture the covariance, however, two components was another valid choice.

Classical multidimensional scaling [6] revealed enough information to support a few assumptions that were made during principal component analysis. After graphically analyzing multidimensional scaling results via the euclidean distance matrix of the data, not only was the data shown to have a cluster behavior, but the bidirectional behavior assumption was verified as well [6].

These graphical commonalities between principal component analysis and multidimensional scaling reassured the assumptions and was finally verified within cluster analysis. The trend that was graphically identified was clearly bidirectional and could seemingly be captured via k-means clustering.

The final analysis technique that the research team conducted was a K-means cluster analysis. Our initial understanding and implications of our study were moving towards a three-cluster[6] analysis that performed better, but we could not intercept the results. The three clusters were able to have distinct patterns and identify when an opportunity should have been bought or passed upon with an unknown cluster. Though the performance gain was needed, the research team went ahead with a two-cluster approach for better interpretation purposes and future implications.

The team decided to discretize the variables of rw to ensure the training and testing splits would be better understood and trained. The split was done in a 50/50 standard in which half the sample was trained and then tested upon. When first trained the data, we used the threshold of any returns less than 0, constituted a pass[7], while anything more significant constituted a bought opportunity. With this threshold[7], our split performed with 45% accuracy statistics. To increase this accuracy, we were able to increase the threshold of our discretization by having the intermediate increase to 2.75 as that was the median value with the best return on investment. The increase in our cutoff helped the accuracy metric up to 52%. Even though 52% seems low in terms of the data science world, taking into account that overfitting didn't occur within our modeling technique[7]. The dataset was only a smaller sample of the entire dataset that was analyzed. This ensures that some patterns can be recognized even with the limited selection that was analyzed upon.

As we've mentioned previously, the research had many fallbacks that hindered our progress in finding distinct patterns for the market exchange. Most of it came down to the dataset and its anonymized features which were difficult to interest and place labels when conducting the principal component analysis. Without this initial labeling, we weren't able to generalize and

come up with some components to create a model prediction model for our study. Our research also does not consider the NA values dismissed as when the data was scaled; they had to be removed for further analysis. These data entries could have been replaced with median values to increase our sample size and have interpretations of the population size or implications of the real-world stock market data. As well as that, the multidimensional scaling performance also hindered our progress in having larger datasets be cluttered and grouped, which was then lowered to get some sort of performance metric for our data sample. With better computational power, this dataset could've been analyzed in a larger capacity and could have helped the research team make better conclusions of the opportunities for buying on an investment.

CONCLUSION

Despite not revealing the perfect market predictor, we have determined that it is not entirely impossible to create approximate predictions about the stock market. Successful results within our cluster analysis, alongside the 52% prediction accuracy, this study will be invaluable to people who will be having further research on predicting stock market investment opportunities and its real-world implications.

APPENDIX

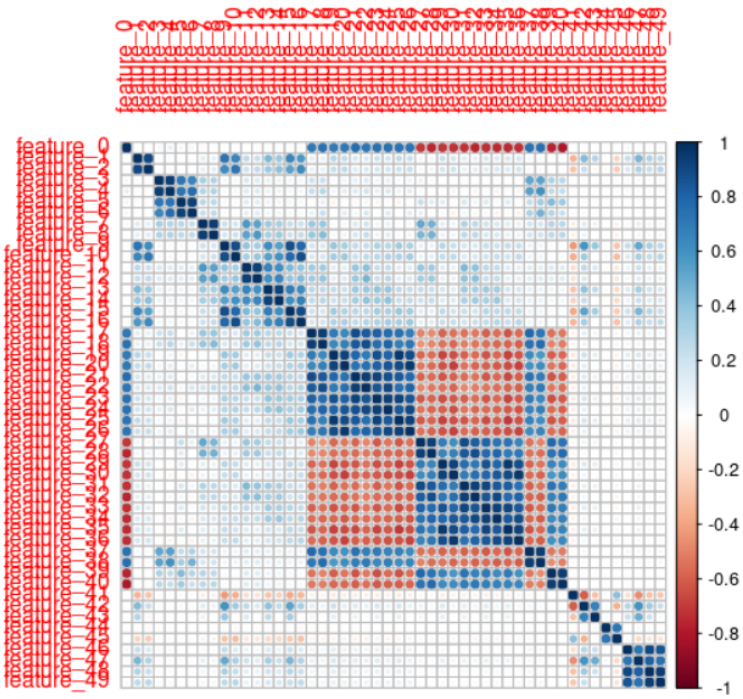
Literature Reviews:

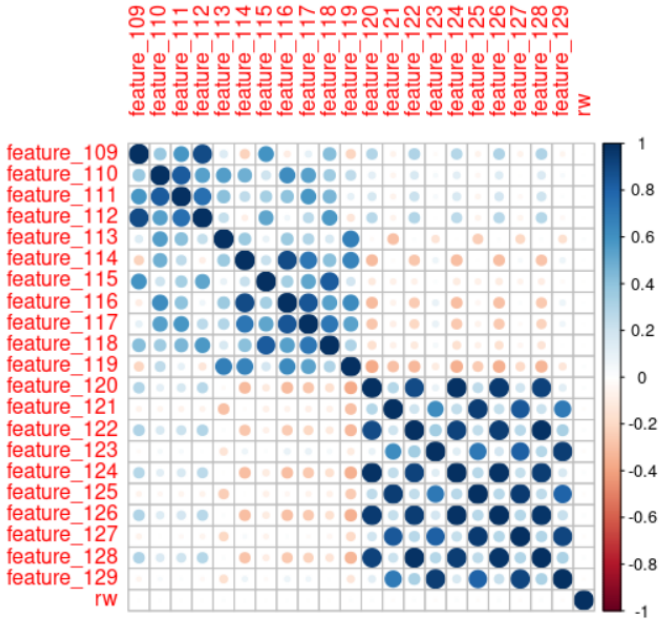
[1] Nguyen, T., & Huynh, N. (2019). Clustering Stock Market Data Using K-Means Clustering

Algorithm. Retrieved from https://athena.ecs.csus.edu/~nguyethi/TN_NH_CSC177_Report.pdf used principal component analysis to identify kmeans clustering model techniques.

[2] Sun, C. (2017, April 25). Application of K-Means Clustering and NeuralNetwork to Stock Return Prediction. Retrieved 2021, from https://cpb-us-w2.wpmucdn.com/blogs.baylor.edu/dist/d/4574/files/2018/01/project_presentation-1kol1u6.pdf utilized kmeans clustering with neural networks to generate predictions about stock performance.

[3] Correlations:





[4] Linear Regression Performance Validation:

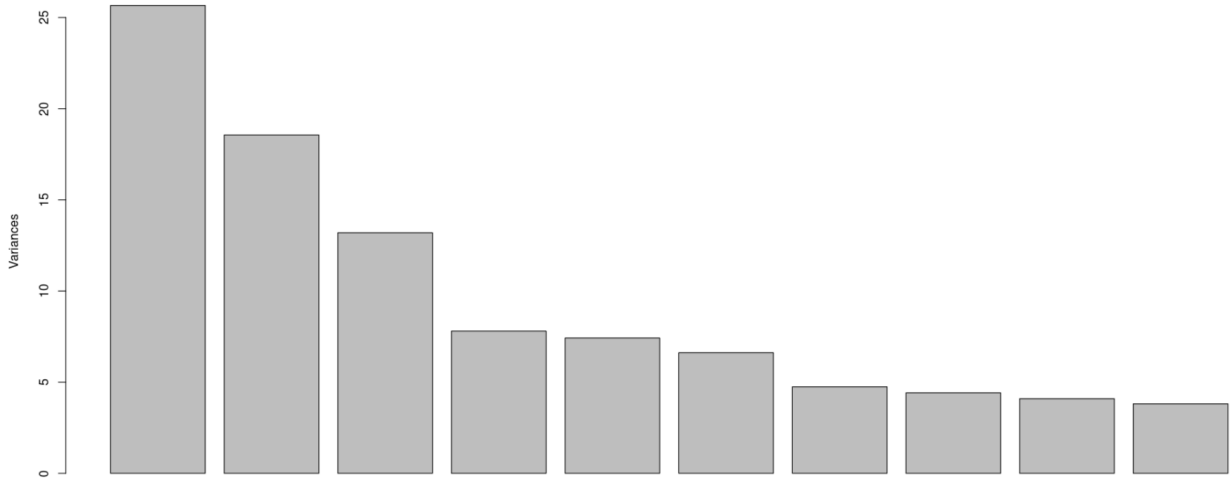
All vars

```
Residual standard error: 0.9732 on 2306 degrees of freedom
Multiple R-squared:  0.1035,    Adjusted R-squared:  0.05296
F-statistic: 2.048 on 130 and 2306 DF,  p-value: 1.406e-10
```

sig vars

```
Residual standard error: 0.9831 on 2427 degrees of freedom
Multiple R-squared:  0.03699,   Adjusted R-squared:  0.03342
F-statistic: 10.36 on 9 and 2427 DF,  p-value: 7.763e-16
```

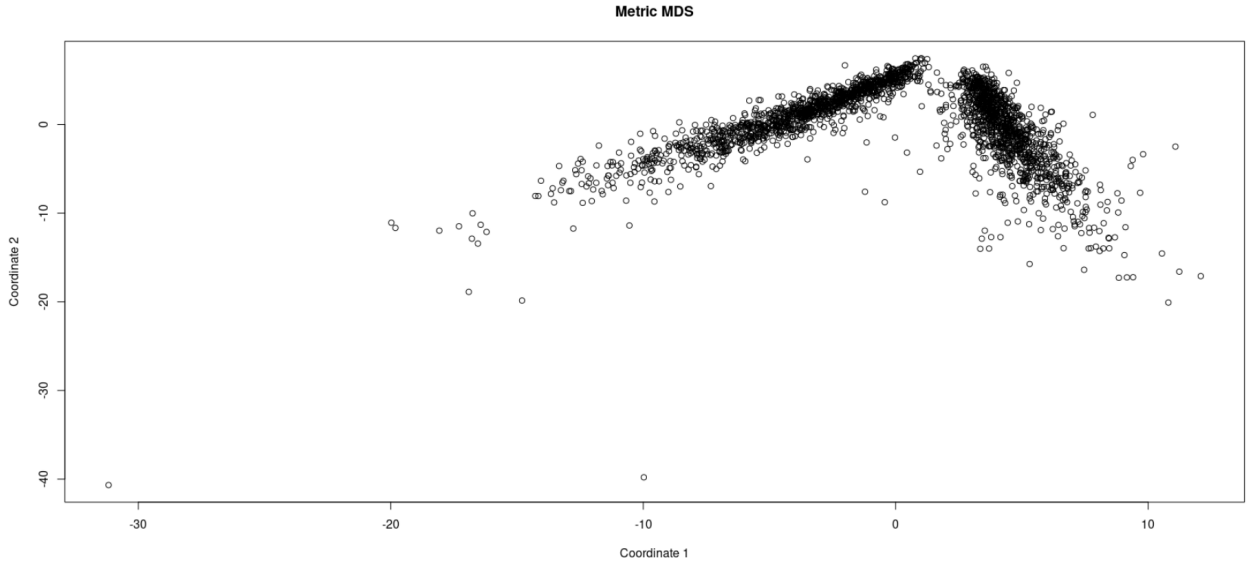
[5] PCA Scree Plot:

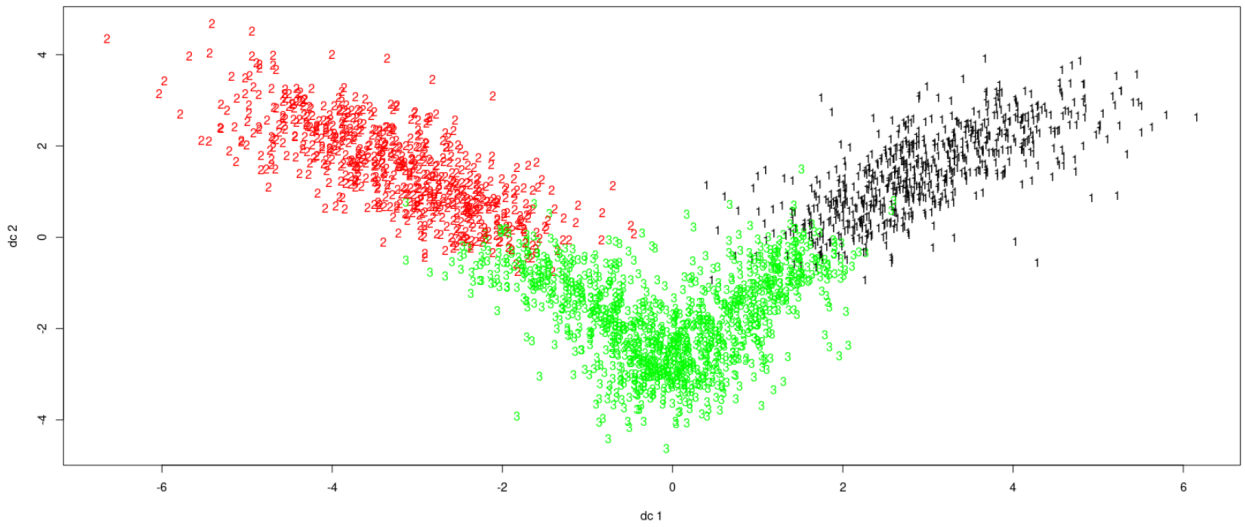


[6]

Cluster and Biplot:

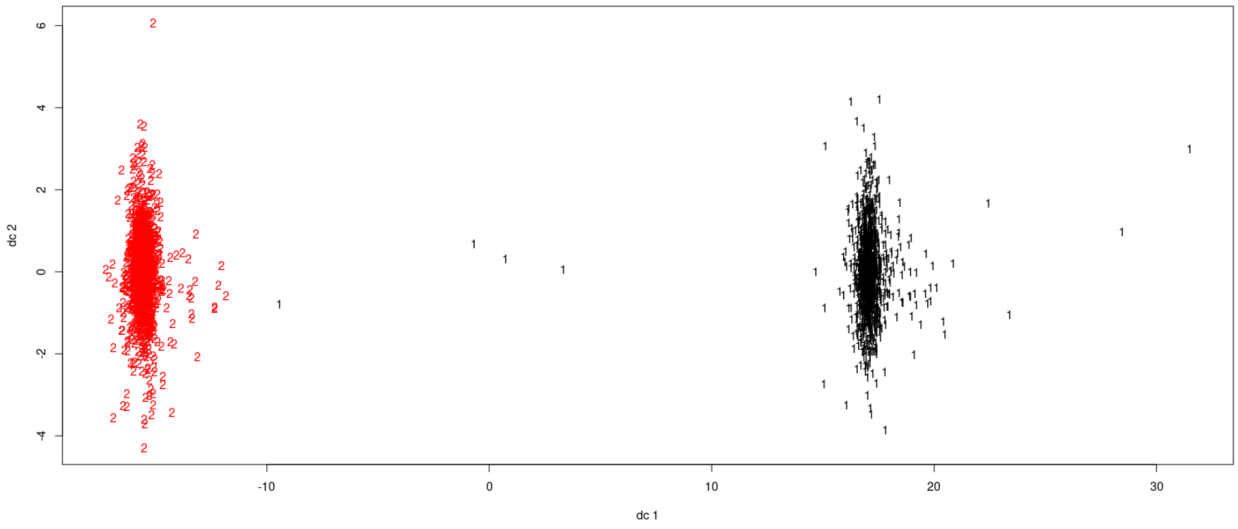
MDS





3 clusters

2 clusters:



[7] K means Performance:

```
>   cluster.stats(d, km1$cluster, km2$cluster)   #compare cluster fit 1 & 2
$n
[1] 2437

$cluster.number
[1] 2

$cluster.size
[1] 1161 1276

$min.cluster.size
[1] 1161

$noisen
[1] 0

$diameter
[1] 81.77554 82.87380

$average.distance
[1] 13.88533 13.92038

$median.distance
[1] 12.96156 13.10854

$separation
[1] 4.616035 4.616035
```