A DEEP DIVE INTO THE SECONDARY SHOE MARKET EXCHANGE STOCKX

A PREPRINT

Anton Njavro
Boston University
njavro@bu.edu

David Neary
Boston University
dneary@bu.edu

December 17, 2019

ABSTRACT

This paper performs a detailed analysis of peer-to-peer secondary shoe market platform StockX with special emphasis on price prediction modeling. We focused on combining the data provided by StockX with external trend data such as Google Trends in order to analyze the dynamics of the marketplace. Using basic data analysis we discovered a secondary marketplace of great magnitude whose value does not necessarily track the underlying commodity (sneaker). We tried to perform predictive analysis ranging from simple linear regression to more advanced machine learning algorithms such as Light Gradient Boosting Model (LGBM) pioneered by Microsoft. Then, we focused on using classical financial modeling techniques which surprisingly gave a good performance, on par with LGBM, which goes to show that this marketplace could be further analyzed similarly to classical financial markets.

***K*eywords** StockX · Secondary Market

1 Introduction

The global market for sneakers has been valued at around 58 billion dollars in 2018. It is projected to have an impressive CAGR of 7% in the following six years. With such growing trends followed by increasing influence of social media on the world of fashion, there has been a significant increase in the domain of "limited edition" shoes, whose limited supply paired with great demand results in a sprawling network of secondary markets. Many companies in recent years have come to serve as facilitators of those secondary markets, offering a centralized platform that helps to pair sellers and buyers while providing authenticity check for the shoes themselves. One of the largest companies in this market, StockX, recently reached a 1 billion dollar valuation signaling great potential that such marketplace platforms might have. With the marketplaces essentially acting as a platforms/brokers between buyers and sellers there are viable questions as to what are the market dynamics of such platforms, and could the goods traded on them (sneakers) perhaps be modeled as some financial instruments.

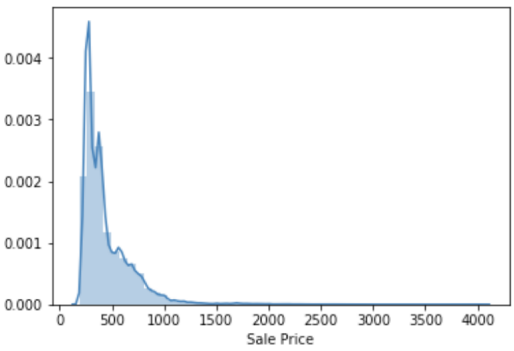
2 Data

The main dataset used throughout this paper was provided by StockX in February of 2019. Data consists of randomly sampled sales (U.S only) of two major sneaker brands: **Yeezy 350** and **Off-White x Nike**. Data spans from the beginning of September 2017 to February 2019. There are 99,956 sales recorded with 72,162 being Yeezy shoes and 27,794 being Off-White shoes. There are also 8 included variables: Order Date, Brand, Sneaker Name, Sale Price ($), Retail Price ($), Release Date, Shoe Size, and Buyer State (the U.S. state the buyer shipped to). In addition to the dataset provided by StockX, we have also collected the Google Trends data in order to have an external data source for trend following. That enabled us to observe the sales/prices in relation to social interest related to these products. The main three terms on which we collected the Google Trends data are: "Kanye West", "Yeezy" and "Off-White". Google

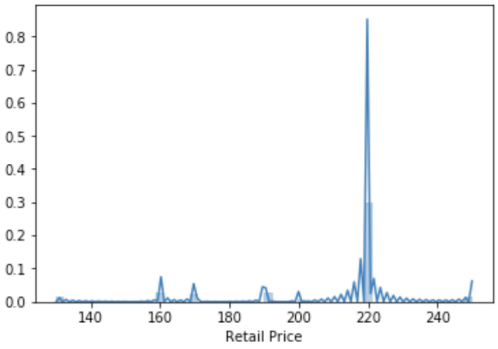
Trends data tracked the interest of our search terms on a weekly basis from September 2017 to February 2019, with interest being relative to the highest point on for the given region and time.

2.1 Exploratory Data Analysis

After obtaining our data, exploratory data analysis (EDA) was needed in order to get a general sense of data. After analyzing "Sale Price" and "Retail Price" columns, we realized there were significant differences between these two. The mean price of shoes sold on StockX was around $446, while the mean retail price was $208. That is a significant markup of almost 114%, showing us that there exists a significant miss-pricing on the side of shoe brands. It also goes to show that there exists an unexpectedly large secondary market whose existence can best be explained due to shallow supply. From graphs, it is also seen that resale prices on StockX have a significantly larger standard deviation ($255) when compared to retail prices ($25). That can be explained due to the fix-price nature of shoes sold in retail, unlike in StockX where exactly the same models can be sold for significantly different prices. In our EDA we also

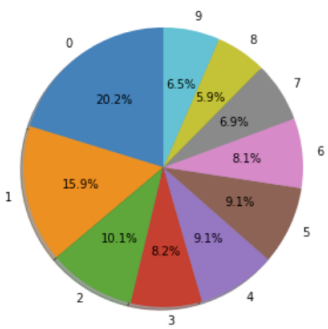


Resale prices

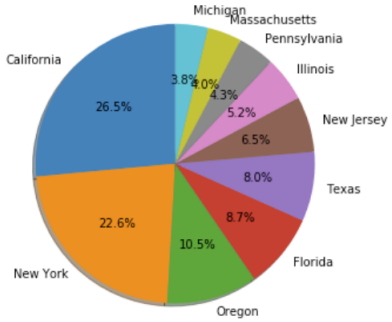


Original Price

created our own features that were focused on date distances between a sale and certain events. The first feature we added was "Date Difference" which calculated the absolute value between release date and sale date on StockX. We hypothesized that meaningful correlations could be extracted from such feature since prices might alter depending on the novelty of the shoe. We then analyzed percentages of top 10 date differences among StockX sales, since all the other date differences were represented in insignificantly smaller percentages following pie-charts depict distribution among the top 10 dates. As we can see, significant portion of sales on StockX occur within the same day of release which indicated a large number of speculators that instantly re-sell for higher price. That phenomenon can be attributed to "sniping", where high-speed bots purchase shoes off retailers websites only to instantly re-sell them elsewhere. We also focused on the analysis of buyers region where we hypothesized that states such as New York and California would have larger representation due to major metropolitan areas that serve as large fashion hubs.



Percentage of date differences



Buyers sate percentage

2.1.1 Google Trends data

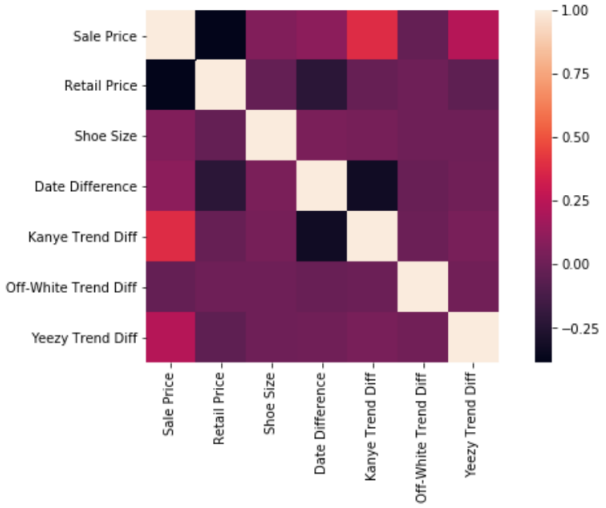
In addition to StockX data, we also utilized Google Trends in order to observe movements in interests related to shoes. Since we were only interested in sudden and meaningful increases in Google Trends we have primarily decided to define interest points to be only those that are 1 standard deviation above mean, however throughout later analysis, in particular, the prediction modeling, such cut-off was not the optimal thus we decided to impose a looser constraint of observing all the points above mean. Imposing a looser definition on interest points in Google Trends and then calculating the difference between sales date on StockX and closest prior interest point in Google Trends data resulted in relatively correlated data. That showed us that spikes in Google Trend data do have a certain effect on sales price as well.

3 Prediction

3.1 Linear Regression

After performing a detailed analysis of our datasets we proceeded to tackle the problem of predicting the sale prices on StockX. Being able to predict the price of the shoe on such a secondary market platform would have a great benefit for both buyers/sellers and potential individuals who might speculate on future prices in order to obtain financial gains. Buyers and sellers would benefit greatly from the recommended potential price for a certain shoe since the market itself do not have clear pricing guidelines yet, leaving buyers/sellers to feel lost. Relating to price prediction for "trading", one would also need to have a much better understanding of market microstructure, including things such as depth of an order book. Market microstructure and dynamics are out of the scope for this paper and should be studied in future studies.

Our prediction analysis started by studying our data through various correlation matrices. By plotting correlation matrices we noticed that, as relatively expected, our data were weakly correlated. We then analyzed correlation matrices for each brand individually and found out that our data had stronger correlations on the Yeezy brand than on the Off-White brand. We attribute that to Off-White being a much smaller and niche brand in comparison to Yeezy. Also with the Yeezy brand is highly influenced by its creator, and famous rapper Kanye West, Yeezy branded shoes had a closer correlation with its related Google Trends terms than Off-White did.



Correlation matrix for "Yeezy" brand data

Prior to starting our prediction analysis, we used SKLearn's library methods to randomly split our data with 80% of data being dedicated to training and 20% of data reserved for testing. Seeing that only meaningful correlations could be seen on a subset of Yeezy sales, we proceeded to perform the predictive analysis only on that subset of sales. We started with the simplest form of linear regression (LR) in order to set a basic threshold for prediction. In our LR model dependent variable was "Sale Price" while independent variables were: "Date Difference", "Kanye West Google Trend",

"Sale Price" and "Yeezy Google Trend". Such a model performed relatively poorly with the explained variance being only 0.37 and with RMSE of $115.

3.2 LGBM

Since we observed relatively weak correlations from our correlation matrix such result from LR was expected. We proceeded to research optimal algorithms that would tackle our issue of weakly correlated data and we decided on LGBM. This relatively new algorithm introduced by Microsoft has become a popular tool in many data science competition with impressive results in prediction. LGBM is a gradient boosting framework that utilizes a tree-based approach. It builds upon its tree vertically as opposed to horizontal build-up. LGBM can attribute popularity to its lightness, speed and focus on the accuracy of results. Our data also proved to be suitable for LGBM since it was relatively large (around 70,000 entries) since LGBM requires a relatively larger dataset (usually greater than 10,000) in order to avoid overfitting. After performing some model-tuning we managed to get a significant increase in our results. We have increased explained variance to 0.81, and have significantly reduced RMSE to $61. Such improvement was better than expected, and it showed that LGBM could very well be the optimal tool for this problem.

3.3 Analysis via financial modeling

StockX exchange shares some characteristics with typical soft commodity exchanges such as the Chicago Mercantile Exchange. A bid-ask spread exists within units, the market volume is variable, and assets can exhibit both seasonal and macro-derived changes. However, a lack of well-defined commodity units, comparably low volume of sales, high transaction costs, and less macro-predictable demand make it in practice, unlike a typical exchange. Despite StockX promoting an image of a quantitative stock-like exchange, the underlying market dynamics do not appear to exist at first glance.

Despite this relative illiquid and only pseudo-commodifiable nature of shoes, we did a test to see if financial models could accurately predict future shoe price. We implemented a basic runs analysis that looked at time-weighted momentum characteristics of the individual shoe, the body of comparable shoes, and the market as a whole. This is a common financial model used that requires minimal outside data or factors and relies only on past sales and trade data. It is also robust in that it makes few assumptions about the market, allowing it to be applied in less-regulated exchanges. Eight new computed attributes were created for each sale to account for these factors. Because this analysis was purely looking at runs, it did not take into account features such as release date, selling location, google trends, or even any other non-market factors. Using this we were able to lower the RMSE to 89 and bring the explained variance to .88. This result is particularly prominent as it does not utilize the second dataset and relies on calculations solely achievable with StockX market data.

4 Conclusion and future steps

Interestingly the classical financial prediction models had similar predictive power as our advanced machine learning model. This suggests that the StockX market may be suited to be treated as a normal exchange, rather than an illiquid one. Implementation of more advanced asset price prediction models may be justified as a result. We also like to implement some linear factors into the analysis to account for categorical information about the shoe as well as more advanced financial modeling tools. For example, in future analysis, we would like to implement concurrent market data from other exchanges such as GOAT. As shoes are not traded on multiple exchanges and in OTC deals, data from multiple exchanges would help with the momentum analysis, especially in cases of cross-market arbitrage.

Both more complete data from StockX as well as better secondary datasets would have helped our analysis. StockX only provided us with a randomized sample of 100,000 sales of two lines of shoes. This represents a minuscule fraction of both sales and shoe lines. As a result, we cannot necessarily expand our same models to the secondary shoe market in general. We also had no specific information on the bid-ask spread over time. This information could also be imperative to recognizing potential future pricing also more pertinent secondary data such as release dates for similar (market-taking) shoe models may have helped our LGBM analysis recognize more pertinent inflection points.

Our analysis can be used in two ways. First, it can be implemented as a suggested price feature on StockX to facilitate sales volume and lower the bid-ask spread. By providing potential sellers with an idea of the range their shoes could sell for it allows a more informed decision by the seller. This could prevent a seller from setting unrealistic expectations and never having their ask met. This information could also be used by those hoping to use shoes as speculative investments in the short run.