Capstone Technical Report

Christopher Kim

07/12/2017

Abstract

Sneakers have been rising in their popularity, especially so since 2015. This report aims to find which features make a sneaker rare. The three brands chosen were Nike, Adidas, and the Jordan brand. Using linear regression, the rarity of a shoe could be predicted with an R^2 score of 0.86. Determining if the shoe is rare or not based on the median was modelled using logistic regression. A mean accuracy score of 0.80 was calculated using five principal component.

1. Introduction

The purpose of this project was to determine if one could determine the factors that make up the rarity of a sneaker. Since the mid-1980's, coinciding with Michael Jordan's rise to basketball superiority, athletic shoe manufacturers have been cultivating America's sneaker collecting sub-culture. A rise in the aftermarket for sneakers in the past several years can be attributed to both the marketing and product development efforts of the three big brands: Nike, Adidas, and Under Armour. Jordan Brand, a division of Nike that specializes in Michael Jordan shoes and apparel, paved the way for young athletes to garner their own shoe brand. The advent of social media, online shopping, and intersection of sports and popular culture has been a boon to the aftermarket for shoes. However, the rarity of a sneaker, determined in part by how much demand and little supply it will gather, is difficult to determine based on simple terms like brand, retail price, or year released. This project aims to find a relationship between shoe and rarity. Part 2 discusses the methods used to determine any relationships. Part 3 discusses the data sources and preparation of the data for analysis. Part 4 is the analysis of the data. Part 5 is the results discovered. Finally, part 6 is the conclusion based on the results.

2. Methods Used

Only one programming language was used in this project: Python 2.7. Multiple libraries were used, namely Scikit-Learn, Pandas, Numpy, BeautifulSoup, Seaborn and the Natural Language Toolkit library. All coding was done in Jupyter Notebook and many outside sources were used and/or adapted for the purpose of this project. Sources are listed inside the Jupyter Notebook.

As an informative reach, neural networks were trained using Keras and Tensorflow. This was performed in a virtual machine on Python 3.6. Much of this code was adapted from the website machinelearningmastery.com

3. Data Sources & Preparation - **describe what data looks like**

Shoe data was sourced from the StockX API. StockX is a stock exchange layer on the collectible sneaker aftermarket, amongst other accessories. The website was created by Josh Luber, a former consultant for IBM, who loved data and sneakers and eventually married his two passions together.

The data was prepared by analyzing hidden fields in the json format that was originally downloaded. A scrape of the website's listing was done multiple times a week to account for new sneakers in the market. Each shoe listing had approximately 60 features; the majority which were dropped before analysis was done. In the end, 2016 individual listings were transformed into a usable dataframe.

StockX contains many different brands, but only the top brands with more than 100 shoe listings were considered. Although Under Armour is the third largest brand for sneakers in the United States, less than 15 shoe listings on StockX for Under Armour was found. Secondly, the Jordan Brand was categorized separately from Nike; Jordan shoes are so iconic that they necessitate a alternate view than under the Nike umbrella. Lastly, a metric was made to determine rarity of a shoe, defined as:

Average deadstock Price / Retail Price (1)

The median rarity was used as a cutting off point for whether a shoe was rare or not.

4. Analysis

Figure 1 shows the relationship between retail price and average aftermarket price. The black dashed line represents a 1 to 1 relationship. Adidas seems to be almost parallel, but both Nike and Jordan seem to have a slope of less than 1. The slopes were calculated and Nike has a slope of 0.29 whereas Adidas has a slope of 1.15.

StockX doesn't have any shoes sold by Adidas before 2014, whereas Nike and Jordan both have shoes dating back to the early 1990's. Adidas came into the market largely in 2015 due to the joint venture with Kanye West and the new Boost technology. Since then, Adidas has climbed to

40% of the StockX market, as measured by summation of rarity. A t-test was performed to determine if the average of Nike's rarity (1.58) is statistically different than that of Adidas (1.76).

Next, natural language processing, NLP, was used to determine if word data could predict the rarity of a shoe. An n-gram range of 2-5 was used for word vectorizers along with many classification models: support vector machines, logistic regression, bernoulli, multinomial naive bayes, random forest, and ensemble methods to classify whether a shoe is rare or not. The texts analyzed were colorway, category, title, and name.

As for the modeling, extra features such as colors of the shoe and which general category were extracted to better predict rarity as a continuous measure and if a shoe was above or below the median rarity. Because of the high dimensionality when the top 20 colors were extracted, principal component analysis was performed.

Finally, a neural network was used to try to analyze any importance from the images of each listing. The layers used were convolutional 2D layers, dropout layers, and finally a dense layer.

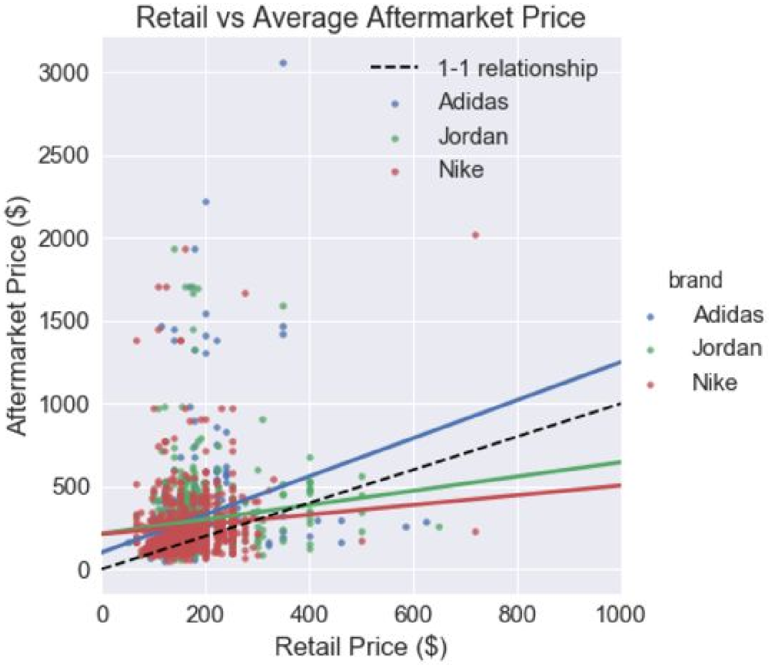

Figure 1. Retail vs. Average Aftermarket Price for the three brands: Adidas, Jordan, and Nike. The dashed black line indicates a linear 1:1 relationship.

5. Results

An alpha value of 0.05 was chosen for the t-test. The test resulted in a p-value of 0.035. As a result, statistically, Adidas has a higher average rarity than Nike. One has to be wary of this result, since Nike has 748 listings on StockX while Adidas has 465. Another test performed was the Mood's median test. With a p-value of 0.13, the null hypothesis, equal medians, cannot be rejected. So although the averages of rarity are statistically different, their medians are not statistically different.**Talk more about the difference in graphs and show in paper**

With a baseline of 0.50, any model that was at least a 15% improvement was examined. Only six models fulfilled this requirement, with the most accurate model consisting of a total frequency-inverse document frequency (tfidf) and logistic regression on the title feature. The accuracy score was 0.598, only slightly better than the baseline. However, all these scores came from no input of any original numerical data and used only the texts describing the shoes.

Next, the original numerical features besides average deadstock and retail price were used to try to predict rarity. The model resulted in a R^2 score of 0.86 on the testing set using 3 principal components. The logistic regression model performed slightly worse with an R^2 score of 0.80 using the same training and testing set and only 1 principal component.

The features were ranked from most important to least important for the logistic regression model. Last sale, or price shoe was last sold at in aftermarket, was the highest. Second was the lowest aftermarket price in the previous 12 months. Thirdly was the lowest price someone is currently selling the product for on StockX. All were positive coefficients, so the more positive or higher, the more likely the shoe is to be rare.

Lastly, the neural network performed at near baseline accuracy. Only 25 epochs were used and images had to be scaled down so that a local machine could run the neural network.

6. Conclusion

Although the title could possibly determine if a shoe was rare more accurately than the baseline, the original data of most recent sale price, annual low price and lowest price someone is currently selling are better predictors for rarity. With this information, one could purchase shoes knowing if the shoe is rare or not. Alternatively, one could determine how rare the shoe is using the linear regression model.

Since NLP was done on very short phrases or terms, it was not as effective as a model based on the sales data. However, if reviews from popular aftermarket retailers like Amazon or Zappos were used, one could surmise a higher accuracy score.

StockX provided data of the aftermarket, but no actual sales data on retail was used. This was partly due to the lack of available and free information online. Like the reviews from aftermarket retailers, actual sales data might be able to predict a more accurate model than one based on the features given solely from StockX.

Finally, the neural network could have been trained better using the full non-scaled image, along with image transformations. When training the neural network, only one side profile image of each shoe was used. However, StockX collects 360° of images of each shoe. Due to machine capabilities, only one picture was chosen. However, the purpose of using the neural networks was for the experience. In the future, a service such as Amazon Web Services could be used to train using the full breadth of data unused in this project's scope.