Sneaky Machine Learning

Jayme Gerring, Brendan Ok, Pin-Yun Lin

5/9/2022

Abstract

blah blah blah

Introduction

Pulling in $70 billion in 2020, the sneaker market has a powerful influence within American retail. Because of the high demand for these sometimes rare and unique shoes, a powerful resale market has also emerged. The sneaker resale market was worth as much as $2 billion in 2019, a figure that has only increased as more and more players try to get in on the sometimes over 2000% profit margin earned from the rarest of sneakers.

As three certified 'sneakerheads' we were interested in using machine learning methods to accurately predict the premiums that result from reselling popular sneakers.

Price Premium is defined as:

$$Premium(\%) = \frac{ResalePrice(\$) - RetailPrice(\$)}{RetailPrice(\$)}$$

Why is this relevant? Premiums are a quick and simple benchmark to measure the profitability and desirability of a specific sneaker. Many characteristics, such as colorway[1], brand, size, and material can make or break a shoe sale. The physical characteristics of shoes are not the only determining factors for premiums, much like other retail goods, shoe sales have a seasonality component as well. This makes understanding the timing of a sale crucial. Premiums can demonstrate to resellers which characteristics make a shoe more profitable. Premiums can also be useful to buyers: based on characteristics, what price is a good deal and what prices border on irrational?

(Think about adding more)

Methodolgy

The specific shoe data for this project was collected from the popular resale website StockX. The dataset contains the details of 99,956 orders of *Yeezy* and *Nike X Off-White* shoes made on StockX from September 2017 to February 2019. The variables associated with orders are: *Buyer Region* (State), *Order Date*, *Brand*, *Shoe Name*, *Retail Price*, *Sale Price*, *Release Date*, and *Size*.

Premium was created from this initial dataset using the formula described in the previous section.

[1] Colorway is a term used to quickly sum up the colors of the sneakers, in our dataset we have colorway categorized as primary, secondary and tertiary colors.

We collected additional variables regarding characteristics of each shoe including: *Material*, *Lace Type*. *Primary Color*, *Secondary Color*, and *Tertiary Color*.

Because certain buying choices could be reflective of economic conditions, we added the variables *USA Monthly Retail Sales* (Monthly), *State Disposable Income per Capita* (Yearly), and *State Population* (Yearly)

(We should drop variables from the CSV if we're not going to use them)