**What are the best Sneakers for Resellers to Make Profit?**

**A display of shoes on shelves

Description automatically generated**

*Analysis of Sneaker Resale Market Using StockX Data*

Marcia Williams  
DATA 6150  
 Wentworth Institute of Technology   
 Boston, MA USA  
willliamsm25@wit.edu

ABSTRACT

This project focuses on understanding how real-time market data drives buying and selling decisions by analyzing trends and consumer behavior. The primary goal is to predict sneaker retail prices using a combination of categorical and numerical features. Leveraging gradient boosting trees, specifically XGBoost,

I aim to identify factors that influence pricing and develop a model capable of delivering accurate predictions while accounting for variability within the dataset.

KEYWORDS

Sneaker, StockX, Sale Price, Retail Price, Brand, Buyer Region

1 Introduction

StockX operates as a live bid-and-ask marketplace, often described as the “Stock Market” for sneakers, streetwear, electronics, and more. It emphasizes transparency and authenticity, ensuring consumers receive verified products. StockX has revolutionized there sale market for limited-edition sneakers and hype products,

highlighting the economic significance of a secondary marketplace.

For data scientists, it is essential to consider the target audience when analyzing data. StockX is a vital platform for sneaker enthusiasts and resellers, offering insights into profit margins across different sneaker types and seasons. By understanding trends in demand, the profitability of specific sneakers, and them fluctuations over time, we can better comprehend why users turn to StockX for their sneaker needs.

2 Data Description

The dataset used in this analysis comes from the 2019 StockX Sneaker Data Contest and consists of 99,964 rows across eight columns. It includes numerical and categorical data, such as retail price, sale price, brand, sneaker name, buyer region, and order and release dates. The dataset required minimal formatting and cleaning, as no missing values existed. This rich and well-structured dataset offers a comprehensive view of the sneaker market, enabling the analysis of patterns and the development of predictive models with efficiency and precision.

**Numerical Data:** Retail Price, Sale Price. Categorical Data: Brand, Buyer Region, Sneaker Name. Temporal

**Features:** Order and Release Dates.

2.1 Source of dataset

**Data**

**Source:** [**https://www.kaggle.com/datasets/hudsonstuck/stockx-data-contest**](https://www.kaggle.com/datasets/hudsonstuck/stockx-data-contest)

* **Primary Data Source:** StockX sneaker sales data. This dataset contains detailed information on sneaker sales, including model names, sales dates, sale prices, and profit margins.

2.2 Characters of the datasets & Preparing Data

1. Import the data and inspect its structure.

2. Convert Retail Price and Sale Price to numeric values.

3. Encode categorical variables using appropriate methods (e.g., one-hot encoding).

4. Handle missing values or inconsistent data if present.

**Characteristics:** Columns, Bar graphs,

3 Methodology

For this project, I utilized Gradient Boosting Trees (GBTs) as the predictive modeling approach, chosen for their ability to effectively handle numerical and categorical features. GBTs, with their unique ability to handle large datasets like ours, offer valuable insights into feature importance and allowed me to focus on the most impactful variables, providing reassurance about the scalability of my approach. We implemented the XGBoost algorithm, a renowned choice for its speed, efficiency, and scalability. This makes it an excellent choice for analyzing sneaker retail prices. Our approach with XGBoost not only enabled us to achieve highly accurate results but also maintained computational efficiency, demonstrating our appreciation for optimization in our approach.

3.1. Feature Engineering

New features were added for order and release dates, separating days, months, and years to better capture temporal importance. These features allowed us to capture the influence of specific days, months, and years on retail prices, avoiding skewed feature importance.

Gradient Boosting Trees were used to identify the most important features. Brand, Release.Month, Release.Day, and Sale.Price were the most significant.

A blue bar graph with white text

Description automatically generated

A table with numbers and letters

Description automatically generated with medium confidence

4 Results

The results from tuning our hyperparameters for our XGBoost model shows strong prediction accuracy for predicting the retail price of sneakers. For starters, our low RMSE value of ~6.77 shows that our XGBoost model deviates only slightly away from the actual retail prices by an average of $6.77. This means that the tuned XGBoost model can reasonbaly give good predictions of retail prices. Alongside this, our Rˆ2 value being ~0.93 means that our model explains 93% of the variance from retail price (Retail.Price) with the four features that we have selected: Brand, Release Month, Release Day, and Sale Price. Finally, the MAE being at ~1.42 means that the absolute difference between the predicted and actual price is small. These metrics show that our model can give accurate predictions of retail prices of sneakers within the data set.

Although the model has good accuracy for predicting the retail price of sneakers, the high MSE value does indicate that there are residuals or outliers that are affecting the model. From our data, we saw that there were sneakers that sometimes doubled or even tripled the retail price. This is common within the sneaker community as the rarity of the shoe can influence the sale price. Factors such as marketing with celebrities as well as one-time-drops can influence the bidding dynamics, causing shoes to go higher than retail price. The highest seen in the data set was a pair of Air Jordan 1’s that sold for $4,050 while being retailed for $190. In these specific cases, our model fails to capture this due to the MSE being ~45.87. Even though MSE is a squared value, it indicates that these instances cause the model to deviate further from the actual prices. While the model itself performs well under normal circumstances, there is room for further improvement to address the outliers which could reduce the high MSE value.

# Print evaluation metrics

cat("Evaluation Metrics:\n")

## Evaluation Metrics:

cat("RMSE: ", rmse, "\n")

## RMSE: 6.784757

cat("Rˆ2: ", rSquared, "\n")

## R^2: 0.9281183

cat("MAE: ", mae, "\n")

## MAE: 1.426962

cat("MSE: ", mse, "\n")

## MSE: 46.03293

**4.1** Enhanced Feature Importance

A graph with blue and purple squares

Description automatically generated

This bar chart effectively communicates the relative importance of features in our model in predicting sneaker retail prices. The brand name ‘Off-White’ emerges as the most influential feature, with a gain value of 0.523, highlighting its significant role in determining retail price. Equally critical is the timing of a sneaker’s release, including the release month and release day, which is a key factor in pricing. In contrast, the feature’ sale price’ has a minimal gain value of 0.021, indicating its little impact on predicting retail prices within this model. This visualization reassures the audience about the validity of the model’s predictions, emphasizing the key role of brand and release timing in driving retail price.

**4.2** Predicted vs Actual Retail Prices

A graph showing a price increase

Description automatically generated with medium confidence

This scatterplot vividly illustrates the model’s success in predicting retail prices for sneakers. The dashed line, symbolizing perfect predictions, is a testament to the model’s accuracy. The clustering of points along this line is a visual confirmation of the model’s effectiveness, with most predictions aligning closely with the actual prices. This should reassure you about the model’s reliability in predicting sneaker prices. While there are some slight deviations from the line, particularly at lower and higher price ranges, the model’s overall alignment is substantial. These deviations, which may be due to dataset variability or unaccounted factors, do not diminish the model’s effectiveness. Most points are in close proximity to the dashed line, confirming the model’s high accuracy in predicting sneaker retail prices across the dataset.

**4.3** Relationship Between Sale Price and Retail Price

A graph with blue lines

Description automatically generated

This graph effectively illustrates the relationship between sneaker retail prices and their sale prices. The red dashed line, which represents a perfect match, allows to visualize deviations. Many sneakers, as shown by points well above the red dashed line, have significantly higher sales prices than retail prices. This graph serves as a valuable tool in understanding the pricing dynamics in the sneaker resale market. Notably, extreme outliers are visible at higher sale prices, particularly around a $200 retail price but exceeding $3,000 in sale price. These outliers likely correspond to rare or highly sought-after sneakers that command premium prices in the resale market. The graph underscores the dynamics of the sneaker resale market, where factors such as rarity and brand value, but most importantly, demand, can drive substantial price increases beyond the original retail value.

5 Discussion

StockX dataset analysis reveals several key metrics that demonstrate the model’s performance. The Root Mean Square Error (RMSE) indicates that the model predicts retail prices with an average error of $6.77. This measure highlights the typical deviation between the observed and predicted prices, providing insight into the model’s accuracy. The R-squared (R2) value shows that the model explains 93% of the variance in retail prices, signifying a strong correlation between the predicted and actual values and indicating that the model effectively captures the underlying trends in the data. The Mean Absolute Error (MAE) is approximately 1.42, representing the average absolute difference between the predicted and actual prices. This metric offers a straightforward interpretation of the model’s prediction accuracy. Lastly, the Mean Squared Error (MSE) is around 45.87, influenced by the presence of outliers in the dataset. These outliers contribute to the higher error rate, suggesting areas where the model’s predictions are less reliable. Together, these metrics provide a comprehensive overview of the model’s predictive capabilities and areas for improvement.

With the MSE value being ~45.87, this indicates that outliers are present which is throwing off our model’s accuracy. We noted that several sneakers sold for significantly higher sale prices versus retail prices due to factors like limited-time releases or influencer-driven hype. Because of this, the model struggles to predict

shoes that fall into this extreme circumstance, deviating away by approximately $45.87.

6 Conclusion

To sum it up, after hyper parameter tuning our XGBoost model became slightly more accurate than before, seeing a healthy decrease in both RMSE and MSE. Although we had a slight increase in the result of MAE being (1.418558), this slight increase is a trade-off to decrease our high MSE value originally being (46.34753) before being reduced to 45.87309. It is clear from the findings that on special cases, there are circumstances where the rarity of a specific shoe may shoot up its sale price to be significantly higher than what it retails for. For this specific case, our model may not be accurate enough to calculate certain market trends such as shoe rarity caused from limited time drops or hype created due to influencer endorsements.

For future reference, this model can be further tuned with more aggressive random searches for hyperparameter tuning. There could be potential for more features such as added market trend variables to further explain this. Variables such as shoe rarity could be added to assess the overall impact of shoe rarity for predicting the retail price of sneakers, which could help reduce MSE. Additionally, the outliers can undergo more sophisticated algorithms such as more ensemble models to further reduce the impact of outliers.

ACKNOWLEDGEMTNS/ETHICAL CONSIDERATIONS

Ethical considerations are critical to this analysis, ensuring that the model’s predictions and insights are transparent. The model must not reinforce existing biases in the data. For instance, if certain brands or styles are undervalued, the model could inadvertently learn and perpetuate these biases, leading to unfair outcomes. Transparency is equally important, requiring clear explanations of how the model makes predictions. This includes thoroughly documenting the features used and outlining the algorithm’s decision-making process. Protecting privacy is essential—any personal information within the dataset must be anonymized,

and all data usage should adhere to relevant privacy laws and regulations. Finally, the potential impact on stakeholders must be carefully considered. The model’s predictions can significantly affect various groups, including buyers, sellers, and the broader sneaker industry. For instance, buyers might be influenced by

the model’s recommendations, sellers might adjust their pricing strategies based on the predictions, and the industry might see shifts in demand. Ensuring that the model benefits all parties involved and supports a fair and equitable marketplace is vital.

REFERENCES

“How to Remove a Character in an R Data Frame Column?” TutorialsPoint, www.tutorialspoint.com/how-

to-remove-a-character-in-an-r-data-frame-column.

“XGBoost R Tutorial — Xgboost 2.1.3 Documentation.” XGBoost Documentation — Xgboost 2.1.1 Docu-

mentation, xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html.

“XGBoost in R.” Statology, www.statology.org/xgboost-in-r/.

“Xgboost in R: How Does Xgb.cv Pass the Optimal Parameters into Xgb.train.” Stack Overflow,

stackoverflow.com/questions/35050846/xgboost-in-r-how-does-xgb-cv-pass-the-optimal-parameters-into-

xgb-train.

“How to Tune Hyper Parameters Using Random Search in R? -.” www.projectpro.io/recipes/tune-hyper-parameters-random-search-r.

ProjectPro, 22 June 2021,

“StockX Sneaker Data Contest.” Kaggle: Your Machine Learning and Data Science Community,

www.kaggle.com/datasets/hudsonstuck/stockx-data-contest.Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00