# Predicting Shoe Prices Using Machine Learning Algorithms

Akshar Sinha

## Abstract

This paper explores the application of machine learning algorithms in predicting shoe prices and the correlations between a shoes' material, color, brand, type, gender, and size and its price. Using a dataset of 5000 shoe entries from Kaggle, the data was preprocessed and turned into numeric values for modeling. The models used include Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Regression. The results show that the Support Vector Regression model had the lowest Mean Squared Error, which means that it can accurately predict shoe prices and find the correlation between a shoes' make and price.

## 1. Introduction

### Motivation

The footwear industry is a large part of the global economy, with sneaker real prices increasing 12.3% since January of 2020 according to Footwear Distributors (Gracia, 2023). Shoe prices have been fluctuating due to factors such as brand popularity, material, design, and market demand. These fluctuations are a significant challenge for both consumers and retailers who need to make informed purchasing and pricing decisions. Being able to accurately predict the price of sneakers is important as it can influence consumer behavior and corporate strategies. This could lead to increased sales in certain brands of shoes, helping corporations in inventory management. In an industry valued at billions of dollars globally, even slight improvements in price prediction accuracy can have substantial financial implications.

## Background

Predicting consumer product prices has always been a complex task due to the multitude of factors influencing prices. Old price prediction methods relied heavily on historical data analysis, market trends, and expert judgment. While these methods have been somewhat effective, they often fall short in capturing the complex and non-linear relationships between various factors that influence prices such as consumer preference. These old methods usually are data driven neglecting the impact that consumer has on the market. Recent advancements in artificial intelligence (AI) and machine learning (ML) offer new opportunities to accurately predict price. ML can process vast amounts of data and identify patterns that are not apparent through traditional analysis, making it a powerful tool for predicting consumer product prices.

## Problem

Accurately predicting shoe prices is a multifaceted problem that involves understanding and modeling the various factors that influence prices. These factors include brand reputation, material quality, design trends, and market demand. Traditional methods are often inadequate for capturing the complexity of these relationships. This paper aims to address this problem by leveraging AI to improve price prediction accuracy. By using machine learning algorithms, we can analyze large datasets and identify patterns that traditional methods may overlook, leading to more accurate and reliable price predictions.The question the paper revolves around is: How accurately can AI predict shoe prices using various machine learning algorithms?

## Solution

(Author 1, Author 2, year of study 1, Author 1, Author 2, year of study 2)

AI, particularly machine learning, provides a deeper analysis and prediction on consumer product prices. Several studies have demonstrated the effectiveness of AI in various prediction tasks, including stock market forecasting, real estate pricing, and demand forecasting. In the context of shoe price prediction, machine learning algorithms can analyze various features such as brand, model, material composition, and market trends to predict prices accurately. This paper presents an AI-powered shoe price prediction system developed using Python and machine learning techniques. By leveraging the power of data and advanced algorithms, our system aims to create reliable predictions of shoe prices, helping consumers and retailers make informed decisions in the dynamic footwear market.

## Contribution

This paper presents an AI-powered shoe price prediction system using the following models:

- Linear Regression: A fundamental statistical method that models the relationship between a dependent variable and one or more independent variables.

- Decision Tree: A model that uses a tree-like graph of decisions and their possible consequences to predict outcomes.

-Random Forest: An ensemble learning method that constructs multiple decision trees and merges them to improve accuracy and control overfitting.

- Gradient Boosting: A machine learning technique that builds models sequentially, with each new model attempting to correct errors made by the previous models.

- Support Vector Regression (SVR): A type of support vector machine that supports linear and non-linear regression.

## 2. Dataset

### Description

The dataset used in this study consists of 5000 shoe entries, each with 7 features. These features include information about the brand, model, material composition, and price of the shoes. The dataset provides a comprehensive overview of various shoe attributes, making it suitable for training and evaluating machine learning models. Each entry in the dataset represents a unique shoe, with detailed information on its characteristics and pricing.

| | Brand | Model | Type | Gender | Size | Color | Material | Price (USD) |
|---|---|---|---|---|---|---|---|---|
| 0 | Nike | Air Jordan 1 | Basketball | Men | US 10 | Red/Black | Leather | $170.00 |
| 1 | Adidas | Ultra Boost 21 | Running | Men | US 9.5 | Black | Primeknit | $180.00 |
| 2 | Reebok | Classic Leather | Casual | Men | US 11 | White | Leather | $75.00 |
| 3 | Converse | Chuck Taylor | Casual | Women | US 8 | Navy | Canvas | $55.00 |
| 4 | Puma | Future Rider | Lifestyle | Women | US 7.5 | Pink | Mesh | $80.00 |

**Figure 1**: First 5 Rows of the Raw Dataset

### Source

The dataset was sourced from Kaggle, a popular platform for data science and machine learning competitions (Rattanaporn, 2023). Kaggle datasets usually have high quality and comprehensiveness, making them ideal for machine learning projects. The specific dataset used in this study was chosen because of the number of features included and the relevance to the task of shoe price prediction. By using a well-curated dataset from this reputable source, this ensures that our models are trained on accurate and reliable data.

## Inputs

Key input features in the dataset include brand, shoe model, material, gender, brand, and type (e.g., running, skate). These features were selected based on their potential impact on shoe prices. For example, brand reputation can significantly influence consumer perception and pricing, while material composition affects production costs and durability. By including a diverse set of features, we aim to capture the various factors that influence shoe prices and improve the accuracy of our predictions.

## Output

The output variable in this study is the shoe price, recorded in USD and the correlation table. This output is the main goal for our machine learning models. By predicting the price of each shoe, we aim to provide valuable insights for consumers and retailers. Accurate price predictions can help consumers make informed purchasing decisions and assist retailers in setting competitive prices and managing inventory. One limitation was that the dataset didn't provide a price year, so it is unknown how inflation would increase these prices relative to the current year.

## Data Types

The dataset contains both numerical and categorical data. Numerical data includes price, while categorical data includes features such as brand and material composition. To make the data suitable for machine learning algorithms, categorical variables were encoded into a numerical format using one-hot encoding. This preprocessing step ensures that the data is in a format that can be effectively utilized by machine learning models.

## 3. Methodology / Models

### Preprocessing

Preprocessing is an important step in preparing the dataset for machine learning. It involves cleaning the data, handling missing values, and encoding categorical variables. In this study, we used various preprocessing techniques to ensure that the dataset was consistent and suitable for modeling. First, we removed any duplicate entries to avoid redundant data. Next, we handled missing values by either removing incomplete entries or imputing values based on the median or mode of the respective feature, ensuring that the dataset was complete and free of inconsistencies.

### Dataset Cleaning

Dataset cleaning involved several steps to ensure the quality and consistency of the data. We began by examining the dataset for any missing or erroneous values. Missing values can significantly impact the performance of machine learning models, so they were either removed or imputed. If one category has more values than the others, this would invoke a bias in our models, creating an inaccurate prediction of shoe price. For example, if the dataset had more ‘Nike’ shoes, then the correlation would be higher with ‘Nike’ shoes and create an ‘Nike’ based prediction for our price.  For numerical features, we used the median value for imputation, while for categorical features, we used the most frequent value. This approach ensured that the imputed values were representative of the dataset. Additionally, we standardized numerical features to have a mean of zero and a standard deviation of one, which helps improve the performance of certain machine learning algorithms.

### Variable Encoding

To make the categorical variables suitable for machine learning algorithms, we used one-hot encoding to convert them into a binary format. One-hot encoding creates a new binary column for each category, indicating the presence or absence of that category. For example, the brand feature was converted into multiple binary columns, each representing a different brand. This approach allows machine learning algorithms to process categorical data effectively and capture the relationships between different categories.

## Data Visualizations

Data visualization played a crucial role in understanding the relationships between different features and the target variable. We used correlation heatmaps to visualize the strength and direction of the relationships between features. We also used different xy graphs to present the relationship between different aspects of the shoe and its price. The correlation heatmap (see Figure 2) shows that certain features, such as type and material, have a strong correlation with the price. Visualizations helped us identify important features and provided insights into the data that guided our modeling approach.

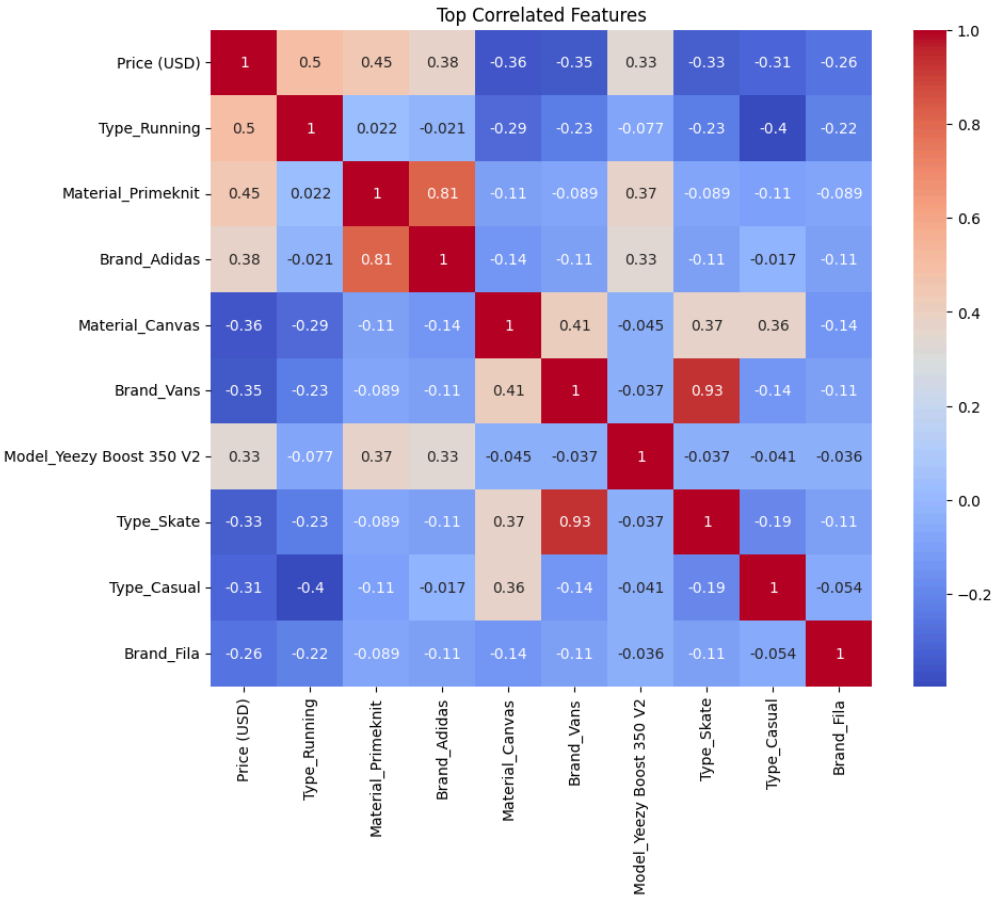


**Figure 2:** Correlation heatmap displaying top correlated features


**Figure 3:** Graph displaying the relationship between shoe size (x) and price (y)

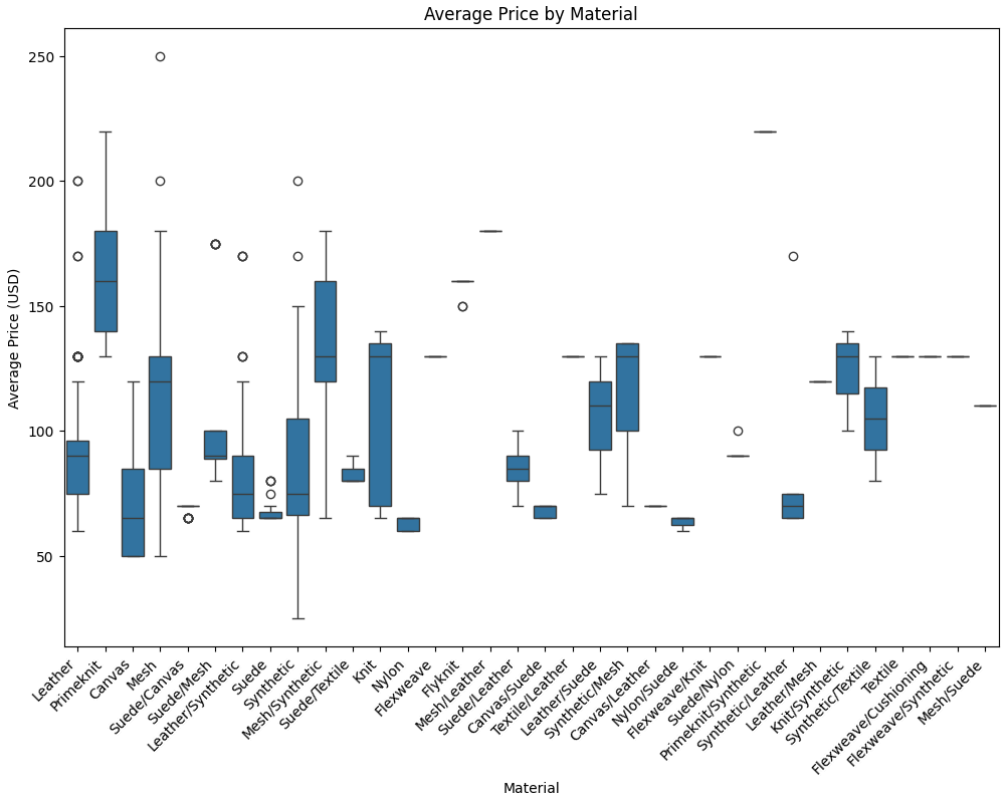


**Figure 4:** Box and Whisker plot of the Average Price by Material

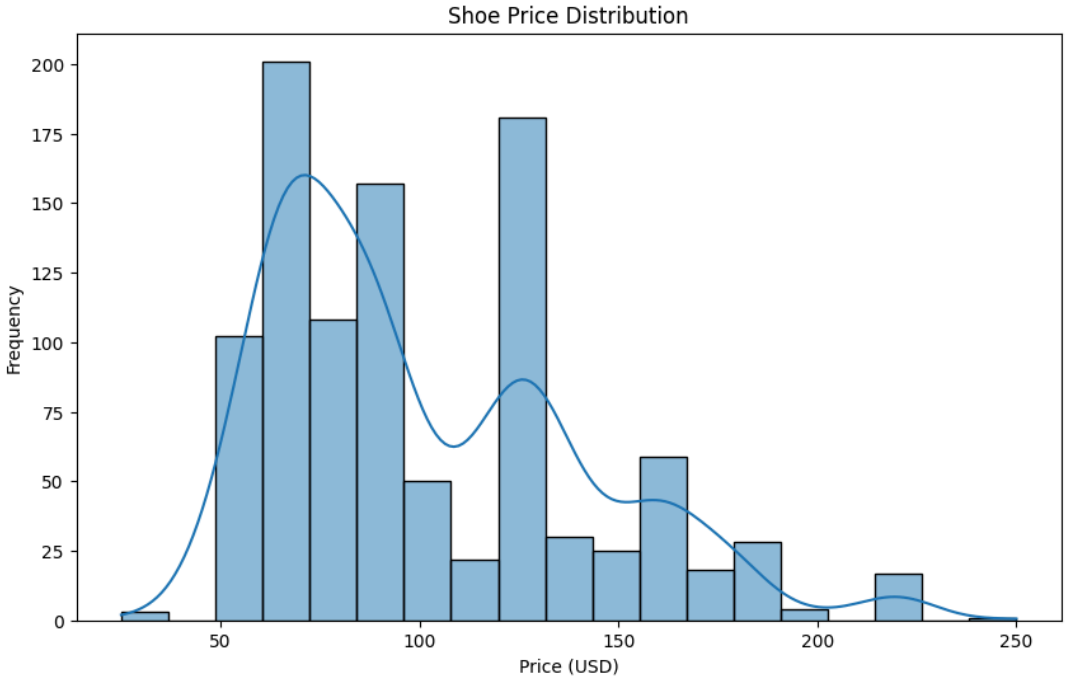


**Figure 5:** Price Distribution

## AI Model Building

The following models were built and evaluated:

- Linear Regression: This model was chosen for its simplicity and ability to provide a baseline for comparison with more complex models.

- Decision Tree: This model was chosen for its interpretability and ability to handle both numerical and categorical data.

- Random Forest: This method was chosen for its ability to improve accuracy and control overfitting by averaging multiple decision trees.

- Gradient Boosting: This model was chosen for its ability to improve accuracy by sequentially building models that correct errors made by previous models.

- Support Vector Regression (SVR): This model was chosen for its ability to handle non-linear relationships and provide accurate predictions.

## 4. Results and Discussion

## Hyperparameter Tuning

Hyperparameter tuning is essential for optimizing the performance of machine learning models. We used GridSearchCV to perform an exhaustive search over a specified parameter grid for each model. This technique allowed us to find the best combination of hyperparameters that maximize model performance. For Linear Regression, no hyperparameters were tuned as it has no adjustable parameters. For the Decision Tree, we tuned parameters such as `max_depth`, `min_samples_leaf`, and `min_samples_split`. For Random Forest, we tuned parameters including `max_depth`, `min_samples_leaf`, `min_samples_split`, and `n_estimators`. For Gradient Boosting, we tuned parameters such as

`learning_rate`, `max_depth`, `n_estimators`, and `subsample`. For SVR, we tuned parameters including `C`, `epsilon`, and `kernel`.

Table of Results

The performance of each model was evaluated using metrics such as $R^2$ and Mean Squared Error (MSE). The best parameters and corresponding scores for each algorithm are presented below:

| Algorithms | $R^2$ | MSE | Best Parameter |
|---|---|---|---|
| Linear Regression | N/A | 1.42e+27 | {} |
| Decision Tree | -213 | 230.59 | {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5} |
| Random Forest | -203 | 226.74 | {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 50} |
| Gradient Boosting | -163 | 162.41587184074308 | {'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 200, 'subsample': 0.8} |
| SVR | -218 | 148.42609363870739 | {'C': 10, 'epsilon': 1, 'kernel': 'linear'} |

## Discussion

The results indicate that the SVR model achieved the lowest MSE, indicating the highest accuracy in predicting shoe prices. This suggests that SVR is particularly effective at capturing the complex relationships between features and the target variable. In contrast, Linear Regression performed poorly, likely due to the complexity of the relationships in the data that it cannot capture. The Decision Tree and Random Forest models showed moderate performance, with Random Forest slightly outperforming Decision Tree due to its ensemble nature. Gradient Boosting also performed well but was slightly outperformed by SVR. These findings highlight the importance of selecting appropriate models for specific prediction tasks and the potential of advanced machine learning algorithms in improving prediction accuracy.

## 5. Conclusions

## Contribution

This study presents an AI-powered shoe price prediction system that utilizes various machine learning algorithms to improve prediction accuracy. By leveraging data from Kaggle and employing advanced preprocessing and modeling techniques, we developed a system that can accurately predict shoe prices. This system has the potential to revolutionize pricing strategies in the footwear industry by providing more accurate and reliable price predictions.

## Key Result

The key result of this study is that the SVR model provided the most accurate predictions with the lowest MSE. This indicates that SVR is well-suited for

predicting shoe prices and can capture the complex relationships between features and the target variable. The findings demonstrate the potential of machine learning in improving price prediction accuracy and highlight the importance of selecting appropriate models for specific tasks.

## What Did You Learn

Through this study, we learned that AI, particularly SVR, can significantly enhance the accuracy of price predictions in the footwear industry. The results underscore the importance of preprocessing and hyperparameter tuning in optimizing model performance. Additionally, the study highlights the need for comprehensive datasets that include relevant features to improve prediction accuracy. By leveraging machine learning algorithms, we can gain valuable insights into the factors that influence shoe prices and develop more effective pricing strategies.

## Gaps

Despite the promising results, there are several gaps in the study that need to be addressed. One limitation is the size and diversity of the dataset. While the dataset used in this study was comprehensive, a larger and more diverse dataset could provide more accurate and generalizable predictions. Additionally, the study focused on a limited set of features, and incorporating additional features such as market trends and consumer reviews could further improve prediction accuracy. Additional work could analyze the feature importances to find the features that contribute the most and the least to the models' predictions. Future work should also explore the use of advanced algorithms such as neural networks to enhance prediction performance. Another limitation of this project is that the release date of the dataset and the years corresponding to the prices within the dataset are unknown. This is important because we are unable to apply an inflation factor thereby giving context across a time scale.

## Next Steps

Future work will focus on integrating real-time market data and exploring advanced algorithms like neural networks. Additionally, we plan to incorporate more diverse and comprehensive datasets to improve the generalizability of our models. By using real-time data and advanced algorithms, we aim to further enhance the accuracy and reliability of our shoe price prediction system. This will enable consumers and retailers to make more informed decisions in the dynamic footwear market.

## Conclusion

For the past few years, sneakers have become more prevalent in the economic market, making the ability to predict prices and find what impacts price invaluable. This work gives us an understanding of what impacts a shoes' price and how we can use this to make smart consumer decisions. In the correlation heat map, figure 2, we see that running shoes have a high correlation constant. This paper provides a baseline for the future of how corporations will price their sneakers and how consumers will purchase sneakers.

## Acknowledgments

We would like to thank Mr. John Basbagill for his help and mentorship through this construction and experimentation of this. We are also grateful for Inspirit AI giving us the opportunity to work on this project through the Independent Mentorship Program with John Basbagill.

## References

1) Garcia, Jolie. "Sneaker Prices Reach Highest Levels This Decade - FDRA." FDRA, April 13, 2023. https://fdra.org/latest-news/sneaker-prices-reach-highest-levels-this-decade/.

2) Kiattisak Rattanaporn. “Shoe Prices Dataset.” Kaggle.com, 2023. https://www.kaggle.com/datasets/rkiattisak/shoe-prices-dataset/data.
3) Raditya, Dita, Nicholas Erlin P, Ferarida Amanda S, and Novita Hanafiah. “Predicting Sneaker Resale Prices Using Machine Learning.” *Procedia Computer Science* 179 (2021): 533–40. https://doi.org/10.1016/j.procs.2021.01.037.
4) Satheesh, Vishnu. “Hyper Parameter Tuning (GridSearchCV vs RandomizedSearchCV).” Analytics Vidhya, April 18, 2021. https://medium.com/analytics-vidhya/hyper-parameter-tuning-gridsearchcv-vs-randomizedsearchcv-499862e3ca5.
5) Statista. “Sneakers - Worldwide | Statista Market Forecast.” Statista, 2024. https://www.statista.com/outlook/cmo/footwear/sneakers/worldwide.
6) Vihaan Miriyala. “Sneaker Price Prediction Using Machine Learning - Vihaan Miriyala - Medium.” Medium. Medium, April 20, 2024. https://medium.com/@vihaannagmiriyala/sneaker-price-prediction-using-machine-learning-b70a96d00388#:~:text=To%20help%20people%20accurately%20predict.
7) Zhang, Tony. “Predicting Sneaker Resell with Deep Learning.” Medium, September 22, 2020. https://medium.com/swlh/predicting-sneaker-resell-with-deep-learning-d3a78b144099.
8) Duyku, E., Guzel, M.S., Bostanci, E., and Askerzade, I. “A Machine Learning Based Approach For Price Estimation”. In *Proceedings of IAM, V.11, N.1, 2022, 50-61.*