Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 179 (2021) 533–540

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Computer Science and Computational Intelligence 2020

Predicting Sneaker Resale Prices using Machine Learning

Dita Raditya[a,1], Nicholas Erlin P[a], Ferarida Amanda S[a], Novita Hanafiah[a,b]

[a]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[b]*Computer Science Department, BINUS Online Learning, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

The sneaker resale industry has reached the value of $2 billion and it is expected to increase by 200% in the next 5 years. This has changed the purpose of sneakers from a wardrobe collection into a promising business opportunity. The aim of this study is to compare several algorithms and decide which one has a better performance in predicting sneaker resale prices. Different techniques like linear regression and random forest have been used to make the predictions using sneaker sales history data gathered from StockX. It turns out that both models produce an outstanding result with similar values. However, with further evaluation, it can be concluded that random forest has a better performance compared to linear regression in predicting sneaker resale price.

Keywords: Sneaker resale; machine learning; linear regression; random forest; prediction;

1. Introduction

Over the course of history, sneaker has vastly changed over time, evolving from just a rubber plimsoll designed to protect our feet into a multi-billion-dollar industry and a significant fashion staple. The rise of sneaker culture has brought a value of $55 billion a year into the sneaker market. Sneakers have been a status symbol for decades now, but nowadays, it has been tremendously redefined by millennials with their desire for comfort and the constantly rising trend of athleisure[1]. In the world of sneakers, a very common yet haunting term that is used in many places is "reselling". According to Cowen & Co, the estimated valuation of the sneaker resale market is at $2 billion now and by 2025, it is predicted to reach more than $6 billion. The enormous development of the sneaker marketplace with

[1]* Corresponding author. Tel.: +62-878-0506-2001.
E-mail address: dita.raditya@binus.ac.id

1877-0509
10.1016/j.procs.2021.01.037

applications like StockX and GOAT provides a platform to buy and sell brand new or even used sneakers has been the secondary reason why sneaker prices are skyrocketing aside from the "Hype".

The "Hype" can only be explained as a marketing strategy and also how "trendy" the item is based on the number of customers that buy the shoes. It's a simple business law of supply and demand. With high demands and low supplies, people tend to inflate the prices as the shelves are selling out. While the market is growing and more people are starting to see this phenomenon as a business opportunity, as some of the sneaker prices rose immensely. For example, in an extreme case, the Nike Air Yeezy is a collaborated shoe between Nike and the popular rapper, singer, and songwriter Kanye West which retailed for $250 USD. It's publicly known that Kanye is now signed with Adidas and has stopped his contract with Nike, so the shoes are never going to be made again. Nowadays, you can get a new pair for around $7400 USD. That is a ground-breaking 3000% increase in price.

However, sneaker resell prices have been meticulously known to be highly unpredictable and being able to predict the prices would be quite the advancement. In this study, we are conducting our experiment to predict the resell prices using machine learning techniques. The ones we are focusing on are Random Forest and Linear Regression and at the end of this experiment, we would be able to find out which technique is more suitable for this data.

2. State of the Art

A variety of methods have been used to predict future prices using machine learning. Some researchers have successfully used four different Artificial Neural Network (ANN) algorithms: Support Vector Machine (SVM), Linear Regression, K-Nearest Neighbors (KNN), and Multi-layer Perceptron[2] to predict the price of second-hand cars[3]. Other researchers also use machine learning in their study by proposing Linear, Lasso, and Gradient Boosting Regression to create a housing cost prediction model[4]. It is believed that information technology, especially machine learning, has become an essential part in every aspect of businesses industry[5].

Linear regression becomes one of the most used algorithms to perform price and sales prediction models. There is also a statement saying that a linear regression model is a statistical method to predict the relationship between two variables: dependent and independent variables[6]. Linear regression attempts to model the relationship by fitting the linear equation to experimental data. Many researchers have used linear regression to create various prediction models and applied linear regression to predict ground vibration[7] and to predict commodity trading price[8].

The world of forecasting has been dominated by linear methods for many decades since they are easy to develop and implement and they are also relatively simple to understand and interpret[9]. Other study also supports this statement by declaring that Linear regression is a widely used method with a simple form and strong ability to interpret and can be easily found in most statistical and economic textbooks[10]. It is also believed that over longer periods of time, linear regression sales prediction models may outperform the professional one[11].

Despite its simplicity, Linear Regression has several flaws that may affect the prediction model. Through their research, another author[12] stated that some difficulties occurred due to its inability to capture nonlinear relationships in the data. The other limitation of this model is its sensitivity towards overfitting, outliers, and multicollinearity. Careful and throughout data pre-processing is really needed before training to prevent any outliers and noise data in the model that will interrupt the training and testing process.

The other famous algorithm for prediction models is Random Forest, as being explained on another study[13] that Random Forest is a very popular algorithm for classification and regression due to ability to perform relatively high accuracy of prediction, built-in descriptor selection, and a method for assessing the importance of each descriptor to the model. Random Forest[14] works using randomness in the tree building process and the aggregation process and relies on a random sample of covariates and cases. This is the reason behind Random Forest's outstanding abilities including highly accurate predictions, robustness to noise and outliers, internally unbiased estimate of the generalization error, efficient computation, and the ability to handle large dimensions and many predictors.

Some researchers[15] used various machine learning algorithms on predicting uncertainty and found that the Random Forest model achieved an outstanding result of successfully giving the right prediction on 99% from the total input. In the political world, Random Forest has proven to be effective for Predicting Class-Imbalanced Civil War Onset Data by offering superior predictive power compared to several forms of logistic regression David[14].

However, Random Forest also has some drawbacks as the algorithm for prediction models. During the training process, a random forest will create a huge number of trees and then combine their results. This requires more complex computational power and resources and takes longer time on training period.

3. Methodology

A. Data Analysis and Exploration

In this study, a dataset from StockX is used from the StockX data contest from 2019 consisting of 99,956 sneaker sales (with 52 different sneakers), 150 extra data was also added due to lack of data variation as all of the data increased in price and none decreased in price. The 150 data is from the same source which is StockX as the recent sales were looked through. Within the dataset, 8 variables were used to determine the end results and those same variables are used to train the model. The variables are Order Date(the date the purchase was made), Brand(the sneaker brand), Sneaker Name(name of the sneaker which also consisted of its colorway), Sale Price(the price it was sold for in resale value), Retail Price(its original price), Release Date(the date the shoe was released), Shoe Size, Price Delta(the difference in the price of resale and retail), and Style Code(each shoe and colorway have a different style code, this is used to differentiate the various colorways) .

We explored the data by plotting the average sale price for each month, the average sale price for each shoe size, and the count of each sale price. We put these graphs to explore the data to see the spread of the shoe prices based on the features of shoe size, sale price by month, and number of shoes sold.

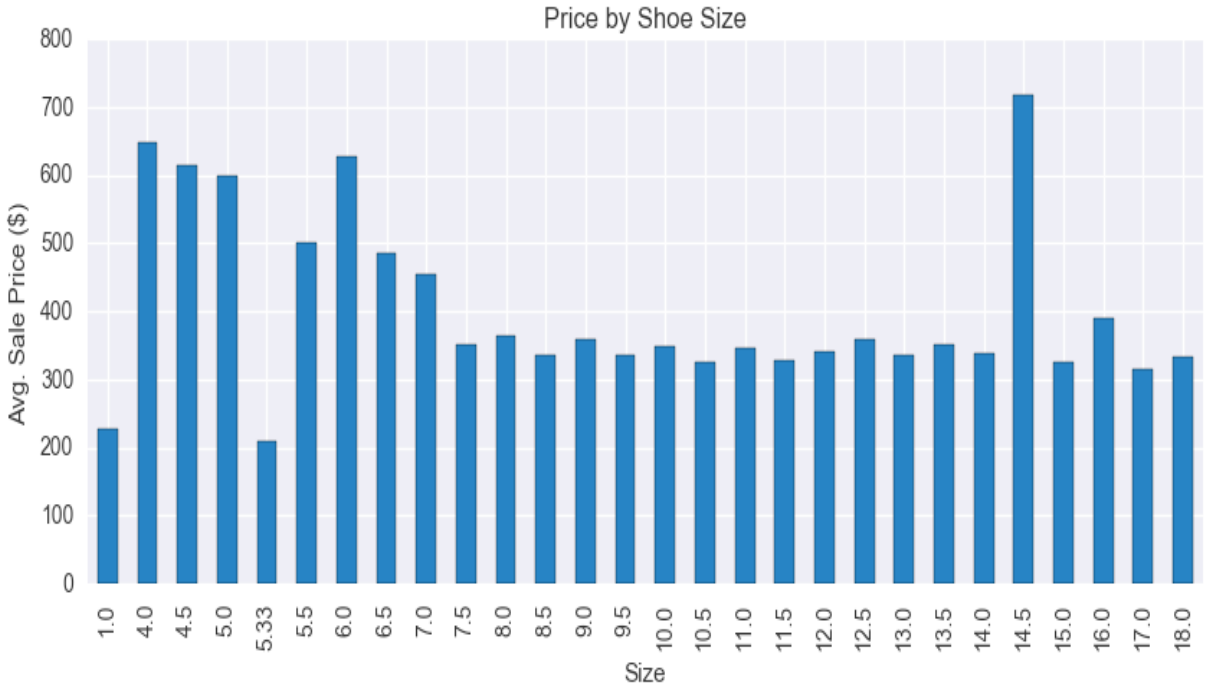

Fig. 1. Price by Shoe Size Graph

In Fig. 1, the shoe prices for each size are pretty even. There are a couple of data which stands out, for example, size 14.5 is considerably higher than the others whilst 1 and 5.33 is considerably lower.

Fig. 2. Average Sale Price by Month

In Fig. 2, the data is also even throughout with the exception of the month of September having a bit more sales than the other months.

Fig. 3. Count of Sale Price

In Fig. 3, most of the data shows that most of the shoes that are sold are below $500. It's not surprising as many people wouldn't want to spend too much on shoes but there are a select few out there that still buy as there are even a couple sales that go to $2000.

B. Feature Engineering

Cleaning up the dataset was done by creating additional variables, adjusting the variables' data type, and extracting the date part. By cleaning up the data, we discovered that around 50 data would not be compatible as there are some

shoe names with typos so they cannot be properly processed, some features of the data are NULL, so we decided to erase them. Afterwards, feature engineering is done on the dataset. To start, Erase the shoes that had sales below 50. If the sales are below 50, it is not enough data to properly process the data, not to mention it will not be accurate so all shoes that were below 50 were unusable.

New datasets was also created with the information for each shoes' name based on the purchase history dataset as additional variables were added in this dataset such as sale price data from 4 last purchases, rolling average of the last 4 observations, 'Intercept' which is for the OLS models as the model baseline and we fill this column with a value of 1 to incorporate the "intercept" into the model easier, 'Total days' which is the difference between the release date and the order date.

C. *Modelling*

1. Linear Regression

The predicted variables for this model are the sales price. For this model, we conduct a correlation matrix analysis between each variable and use that information to know the strength of the relationship between two variables. Based on the data shown in Fig. 4, the features with the highest correlation value towards Sneaker Sale Price were chosen for the model. Those features are Sneaker name, retail price, shoe brand, shoe line, 4 last sale prices (sale lagged), and rolling average of the last 4 observations.

The next step is to divide our clean dataset into a training set that has 80% of the total data and the test set consists of 20% from overall data. The last step is to train and test the model. A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept. This formula became the foundation of the Linear Regression training algorithm.

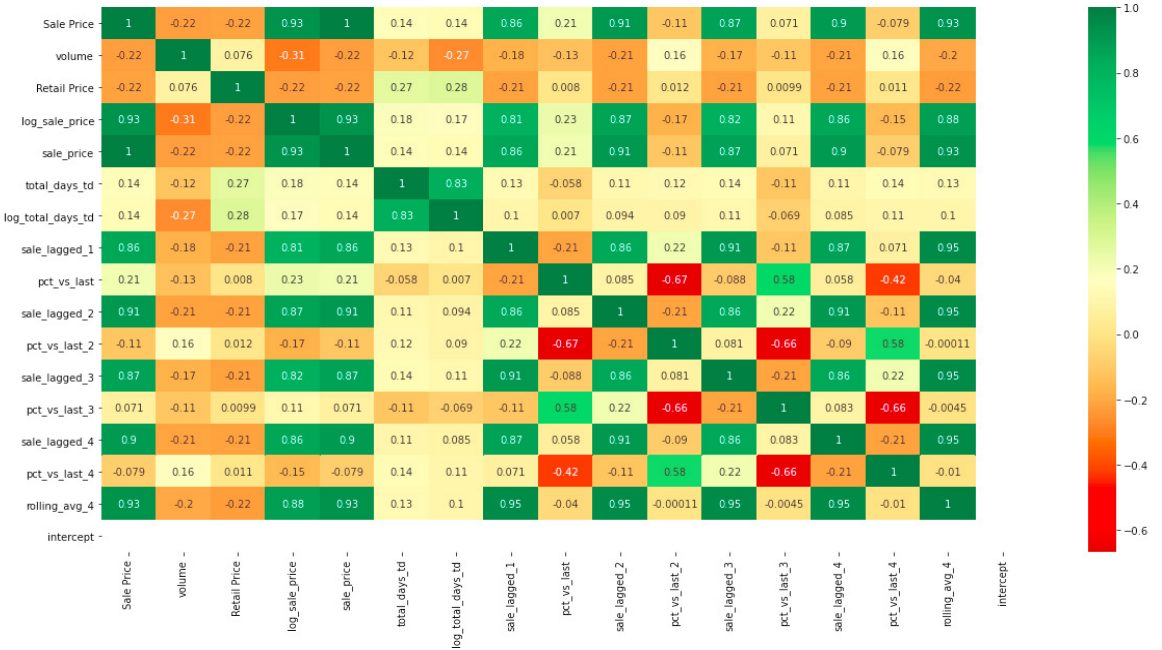

Fig. 4. Features Correlation Matrix

2. Random forest

For random forest, the same steps that we use in Linear Regression were followed. Using the same information that was received from the previous correlation matrix and applied it as well in this model. Since this model will also

predict the sale price of the sneaker, the same features to predict the sale price as in Linear regression were used. Those features are Sneaker name, retail price, shoe brand, shoe line, 4 last sale prices (sale lagged), and rolling average of the last 4 observations.

For this model, the sour clean dataset was divided into a training set that has 80% of the total data, and the test set consists of 20% from overall data. The value of N-estimators was set to 1000 and Max-depth into 5. The last step is to train and test the model. Here is the illustration of Random Forest pseudocode that is being applied on this model.

Precondition: Number of total tree $\boldsymbol{T}$, features $\boldsymbol{F}$, training set $\boldsymbol{S} = (x_1, y_1), \ldots, (x_n, y_n)$

```
function RandomForest ()
    R ← 0
    for i = 1 to T do
        S(i) ← Randomly sample the training data from S
        ri ← TreeLearn(S(i))
        R ← R ∪ {ri}
    end for
    return R
end function
function TreeLearn (S)
    At each node:
        Calculate best split point
        f ← selected feature F with highest information gain
        Split node in f
        Create f child nodes of S
    return tree
end function
```

4. Results and Discussion

In order to see the result of the tested model, some statistical measures are being used as a comparison of both models. Those measurements are R-squared (R^2) and Mean Squared Error (MSE). R^2 is a comparison of the residual sum of squares with the total sum of squares. Total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line. On the other gun, MSE used to measure the average of the squares of the errors obtained from the average squared difference between the estimated values and what is estimated.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{1}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{2}$$

Table 1. Test Model Result

	R2 Train	R2 Test	MSE
Linear Regression	0.9318	0.9286	12562
Random Forest	0.9320	0.9234	13474

As clearly seen from Table 1. that the R2 score for both models are very similar and surprisingly high. But we can see that the R2 score for the training set is slightly higher for both models than the test set, this might be the sign of small overfitting that occurs on both models. So, we try to apply Cross-Validation to prevent this issue. Here is the table with the R2 score of both models after using K-fold cross-validation with K = 10.

Table.2 Test Model Results with Cross Validation

	R2 with CV	MSE with CV
Linear Regression	0.9060	12513
Random Forest	0.9330	9730

In the initial testing, Linear Regression performs better than Random Forest but when Cross Validation is put into play, Random Forest has higher scores than Linear Regression. This also affects MSE for both models, especially random forest, as we can see that MSE for random forest decreased by 3000.

When using Random Forest, the R2 Train is slightly higher than the R2 Test which indicates a little overfitting. After performing cross-validation the R2 Score is higher and the MSE is lower, so we concluded that we could fix the overfitting using cross-validation. With Linear Regression, however, the R2 Decreased.

5. Conclusion

The sneaker resale industry has become a multi-billion-dollar worth industry and by the next 5 years, it is expected to reach more than $6 billion. This phenomenon has encouraged many people to take this as a business opportunity. However, sneaker resell prices have always been known to be unpredictable. It will be quite the advancement to be able to predict the future prices of sneakers. In this study, we are conducting our experiment to predict the resell prices using two machine learning techniques which are Linear Regression and Random Forest, and comparing both methods to decide on which one has better performance when predicting sneaker resale prices.

Linear Regression has been widely used in price and sales prediction models as it is quite easy to understand and interpret. However, this method can be too sensitive when handling overfitting, outliers, and multicollinearity, not to mention its inability to capture non-linear data. Another well-known algorithm in machine learning is Random Forest, which is popular for classification and regression with its ability to perform a relatively high accuracy when predicting. However, this method is more advanced when compared to Linear Regression as it requires more computational resources to process.

From the training and testing process that we conducted for both models, we can obtain the R^2 and MSE scores as the statistical measurements in order to know the performances from both models. Based on the results, a conclusion could be made that the R^2 and MSE score of both methods are extremely similar, although Linear Regression has a slight edge compared to Random forest, when we apply Cross-Validation, it fixes Random Forest's trait of being prone to overfitting and even improves the R^2 score of Random Forest, even exceeding Linear Regression's R^2 score. The MSE score for Random Forest also decreases by around 30% due to the appliance of Cross-Validation. Because of Random Forest's ability to handle outliers which cannot be found in Linear Regression, it makes it more suitable to work with complex data just like the one used in this study, and the fact that there is not much pre-processing needed such as data scaling and outlier removal to provide a great results, is another "plus" for Random Forest. Those statements brought us to a conclusion that the Random Forest model with 10 folds of Cross-Validation performs better compared to the Linear Regression model in predicting sneaker resale prices. With further research and improvements, we believe that in the future, this model could be properly used by sneaker resellers to assist them on getting more insight about the resell prices of upcoming sneakers.

6. Future Works

Sneakers have been known to be notoriously difficult to predict. For example, the tragic death of Kobe Bryant skyrocketed the prices of his sneakers. When Kanye West left Nike to partner with Adidas, it made his Nike Air Yeezy shoes increase even more. Events like this are the kind of predictions that the system will not be able to make by sheer numbers alone. For most if not all shoes, using machine learning techniques are enough, yet the same cannot be said for other cases. For future researches, adding a social media API will allow the model to readjust the prediction prices using keywords that are linked with some shoes, like for example 'Air Jordan' with Michael Jordan, 'Yeezy' with Kanye West, and many others. By using this, we can more accurately predict prices when certain events happen in the world.

Other opportunities for future research emerge when a larger dataset is applied. The dataset used for this experiment was relatively small in terms of variation as we only put fifty different shoes. By using more data from other shoes, the model can predict a wide variety of sneakers which would vastly improve the model itself. With more dataset though, keep in mind that a more powerful system would be recommended to process all the data.

References

1. Yeh, Mei-Chen, and Shao-Ting Yang. (2015) "A multimodality approach to predicting the popularity of sneakers." *2015 IEEE International Conference on Consumer Electronics – Taiwan.*
2. Peerun, Samiyah, Nushrah Henna Chummun and Sameerchand Pudaruth. (2015). "Predicting the Price of Second-Hand Cars Using Artificial Neural Networks." *Proceedings of the Second International Conference on Data Mining, Internet Computing, and Big Data*. 17–21.
3. Pudaruth, S Sameerchand. (2014). "Predicting the Price of Used Cars Using Machine Learning Techniques." *International Journal of Information & Computation Technology* **4 (7)**: 753–764.
4. Satish, G. Naga, Ch. V. Raghavendran, M.D.Sugnana Rao, and Ch.Srinivasulu. (2019) "House Price Prediction Using Machine Learning." *International Journal of Innovative Technology and Exploring Engineering Regular Issue* **8 (9)**: 717–722.
5. Syazali, Muhamad, Fredi Ganda Putraa, Achi Rinaldia, Lintang Fitra Utamia, Widayantib, Rofiqul Umamc and Kittisak Jermsittiparsertd. (2019) "Partial Correlation Analysis Using Multiple Linear Regression: Impact on Business Environment of Digital Marketing Interest in the Era of Industrial Revolution 4.0." *Management Science Letters* **9**: 1875–1886.
6. Zahari, Nazirul Mubin, Raja Ezzah Shamimi, Mohd Hafiz Zawawi, Ahmad Zia Ul-Saufie and Daud Mohamad. (2019) "Prediction of Future Ozone Concentration for Next Three Days Using Linear Regression and Nonlinear Regression Models." *IOP Conference Series: Materials Science and Engineering*.
7. Kara, Ram Chandar, Rama Sastry Vedala, and Hegde Chirant. (2016) "A Critical Comparison of Regression Models and Artificial Neural Networks to Predict Ground Vibrations." *Geotechnical and Geological Engineering* **35 (2)**: 573–583.
8. Liu, Jun, Yuan Tian, and Qing Yan. (2018) "Modelling and Forecasting of Commodity Trading Price." *Journal of Physics: Conference Series*.
9. Ahangar, Reza Gharoie, Mahmood Yahyazadehfar, and Hassan Pournaghshband. (2010) "The Comparison of Methods Artificial Neural Network with Linear Regression Using Specific Variables for Prediction Stock Price in Tehran Stock Exchange." *(IJCSIS) International Journal of Computer Science and Information Security* **7 (2)**: 38–46.
10. Gupta, Swati, and Pinki. (2018) "Sales Forecasting using Linear Regression and Support Vector Machine." *International Journal of Innovative Research in Computer and Communication Engineering* **6 (4)**: 3749–3755.
11. Al-Dmour, Ahmed H, and Rand H. Al-Dmour. (2018) "Applying Multiple Linear Regression and Neural Network to Predict Business Performance Using the Reliability of Accounting Information System." *International Journal of Corporate Finance and Accounting* **5 (2)**: 12–26.
12. Nunno, Lucas. (2014). "Stock Market Price Prediction Using Linear and Polynomial Regression Models." 1–6.
13. Palmer David S., Noel M. O'Boyle, Robert C. Glen, and John B. O. Mitchell. (2006) "Random Forest Models To Predict Aqueous Solubility." *Journal of Chemical Information and Modelling*. **47 (1)**: 150-158.
14. Muchlinski David, David Siroky, Jingrui He, and Matthew Kocher. (2015) "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data". *Political Analysis Advance Access* **24 (1)**: 87-103.
15. Coulston John W, Christine E. Blinn, Valerie A. Thomas, and Randolph H. Wynne. (2016) "Approximating Prediction Uncertainty for Random Forest Regression Models." *Photogrammetric Engineering & Remote Sensing* **82 (3)**: 189-197.