AIGS 1006 DEEP LEARNING

“Future Prediction of Sneaker Prices using Machine Learning.”

GROUP 3

Members:

Vennila

Neha Parveen Mohammed

Sonali Jammichedu

Sonjeet Kaur

Sana Naseem

Loyalist College, Belleville, Canada 2023

Table of Contents

1. Introduction……3
2. Literature Review……4
3. Objectives……5
4. Solution Approach……5-6
5. Progress Made……6
6. Technical Analysis……6-7
7. Challenges……7
8. Next Steps……7
9. References……7

INTRODUCTION

In recent years, the sneaker industry has witnessed remarkable expansion and evolution driven by a dynamic blend of style, technology, and cultural influence. We find ourselves in an era of data-driven, decision making and the ability to understand the intrinsic pattern within the sneaker market is crucial for business, marketers and enthusiasts alike. As the industry continues to expand and diversify, the need for informed decision making becomes increasingly critical.

With in the expansive array of sneaker options available, understanding the factors that drives consumer choices and market trends is essential. So, the sneaker analysis aims to address these needs by exploring patterns within a curated dataset and seeking to unravel the fundamental dynamics that define sneaker market. Through this exploration we hope to deliver actionable insights for business and marketers with the knowledge necessary for informed decision making in this ever - evolving industry.

The insights gained from this analysis carry significance for various stake holders. The sneaker dataset encapsulates a diverse range of information, including sales data, products attributes.

So, for the sneakers data we are using advanced analytical methods, aiming to address specific objectives such as explore and analyse consumer preferences within the sneaker market, identifying styles, colours, brands etc. Utilize historical data to identify emerging trends, providing potential insights to potential future market trends.

The scope of the analysis encompasses of Data availability, The analysis may focus on specific region or market segment, limiting the generalizability of findings to a broader global context.

As we conclude the analysis, the role of visualization emerges as crucial component in our sneaker market. The visual representation of trends, correlations, and insights not only simplify the complexities of the data but also empower stakeholders to grasp key takeaways**.**

LITERATURE REVIEW

The sneaker sale industry has been widely popular among people especially the young generation and is expected to increase rapidly. This has now changed from how we used to see sneakers as a wardrobe collection into a promising business opportunity.

We came across a very informative paper by Dita Raditya, Nicholas Erlin P, Ferarida Amanda S, Novita Hanafiah which focuses on how they have implemented Linear Regression and Random Forest Algorithms on sneaker sales history data gathered from StockX to make a future forecast of the sale price of the sneakers. Linear Regression is the most widely used algorithms to perform price and sales prediction models [1]. But linear regression has many drawbacks which could affect the future predicted prices, such as to capture non-linear relationships in the data [2]. It is also seen that linear regression is very sensitive towards overfitting, outliers and multicollinearity [1].

Another algorithm which is discussed is the Random Forest, which is very popular for classification and regression as it has the capability to perform relatively high accuracy of prediction, built-in descriptor selection and a method for assessing the importance of each descriptor to the model [1]. The working of the random forest uses randomness in the tree building process. It is also stated [1] that it has outstanding abilities including highly accurate predictions, robustness to noise and outliers and the ability to handle large dimensions of data and many predictors makes it suitable to be used for future price predictions with accuracy.

Some researchers [3] used many algorithms to predict uncertainty and found that Random Forest gives outstanding result by giving almost 99% right predictions from the total input. However, Random Forest also has some drawbacks [1] 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. During the modelling it was seen that by using linear regression and random forest the coefficient of determination i.e., R^2(R-Squared) was very high for both the models which might be the sign of overfitting which could be avoided by using K-fold cross-validation.

Hence, linear regression method can be too sensitive when handling overfitting, outliers and multicollinearity. Although random forest gives much accurate predictions as compared to linear regression, capable to handle both categorical and continuous variables, it can show feature importance, which can be useful for feature selection and interpretation and can work well with default parameters and requires less tuning, this is the reason why we have chosen to go ahead with Random Forest for modelling sneaker dataset.

OBJECTIVES

1. To perform data cleaning such as, deleting unwanted rows/columns, removing duplicate data, identifying and deleting the rows with data entry errors.
2. To perform Exploratory Data Analysis (EDA) on the given data to gain insights and understand the patterns, outliers, relationships, etc.
3. To create new features (if necessary, like, calculating the age of sneakers from the release date) and to identify main features and factors that influence the prices of the sneakers.
4. To develop a model (like random forest, gradient boosting, like regression etc.) to predict the sneaker prices in the future and clustering similar kind of sneakers based on their properties such as silhouette, condition, brand etc., to identify which categories perform similarly in the market.
5. To make use of sneaker data that help us to plot different visualizations like price distribution, trends model predictions etc., which can be very useful in identifying the most profitable sneaker categories or to predict future prices.

SOLUTION APPROACH

1. Data Cleaning:

a. To understand the data by inspecting its structure and identify null/undefined values, if any exists then removing the null values by row wise deletion.

b. To handle typographical errors if any along with that to avoid outliers to improve model performance and to ensure data is on the same scale by using standardization and normalisation technique.

2. Feature selection:

a. To analyze how well the sneaker data features are correlated to each other, to measure the strength and relationship between variables by using correlation coefficient.

b. To apply label encoding technique to convert categorical data into numerical data.

c. To use Random Forest which will help predicting the output of future predicted sneaker price with high accuracy even for a dataset which is very large in size.

d. Reasonable prediction is expected without tuning the hyper-parameter, thus solves the issue of overfitting in decision trees.

e. To detect the outliers by using boxplot along with Winsorization technique.

3. Visualization:

By using matplot and seaborn in python, the sneaker dataset can be transformed into statistical and graphical representation. It will give a clear understanding of data patterns, relationships and trends which in turn will help to analyze the future predicted prices in a much better way.

PROGRESS MADE

a. We started off by analyzing the data and its structure, cleaned the data.

b. Tried Implementing label encoder to all the features.

c. We split the data into train and test which comprises of 70% of train data and 30% of test data.

d. Currently we started exploring Random Forest and we are planning to experiment with other algorithms.

TECHNICAL AND DATA VISUALIZATION ANALYSIS

Technical Analysis:

- The sneaker dataset, which is given to us contains 14 million records, of which we took a subset of 200k to perform our analysis. We are focusing on the key features such as, brands, prices, release dates of the sneakers etc.
- We used dropna() function to clean the unwanted rows and we are working on removing the rows with typographical errors.
- We have started experimenting with Random Forest algorithm as it be used for both classification and regression tasks and as it can also handle the non-linear relationships between the features and the labels which is most common in the real-world data.

Data Visualization Analysis:

- We haven't plotted any graphs so far; however, we are planning to plot the visualizations of varying sneakers prices, popular brands etc.

CHALLENGES/ISSUES FACED

1. We are unable to read the data from S3 bucket to Jupyter (AWS Jupyter) directly.
2. We did research on performing data cleaning, but once we came to know that we have to delete all the rows which contained null values then we completed the data cleaning.

NEXT STEPS

1. We will be trying out different models like XG Boost, Arima, LSTM.
2. By using correlation, variance, Step down backward propagation from scikit learn, we will be using feature-selection module.
3. We will compare all outputs to get the different models.

REFERENCES

1. Raditya, D., & Hanafiah, N. (2021). Predicting sneaker resale prices using machine learning. Procedia Computer Science, 179, 533-540.

2. Nunno, Lucas. (2014). "Stock Market Price Prediction Using Linear and Polynomial Regression Models." 1–6

3. 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