AIGS 1006 DEEP LEARNING

"Sneaker Data Analysis"

GROUP 3

Final Report

Members:

Vennila

Neha Parveen Mohammed

Sonali Jammichedu

Sonjeet Kaur

Sana Naseem

Loyalist College, Belleville, Canada

Table of Contents

Introduction

In recent years, the sneaker industry has witnessed remarkable expansion and evolution driven by a dynamic blend of style, technology, and cultural influence [5]. 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.

Within the expansive array of sneaker options available, understanding the features 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 are meaningful and have significant implications for various analysts, and end-users. 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.

Objective

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 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.

2. *Feature selection:*

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

    b. Applied label encoding technique to convert categorical data into numerical data.

    c. Used Random Forest for feature selection.

3. *Standard Scaling:* To ensure data is on the same scale and to improve accuracy we used standardization and normalisation technique.

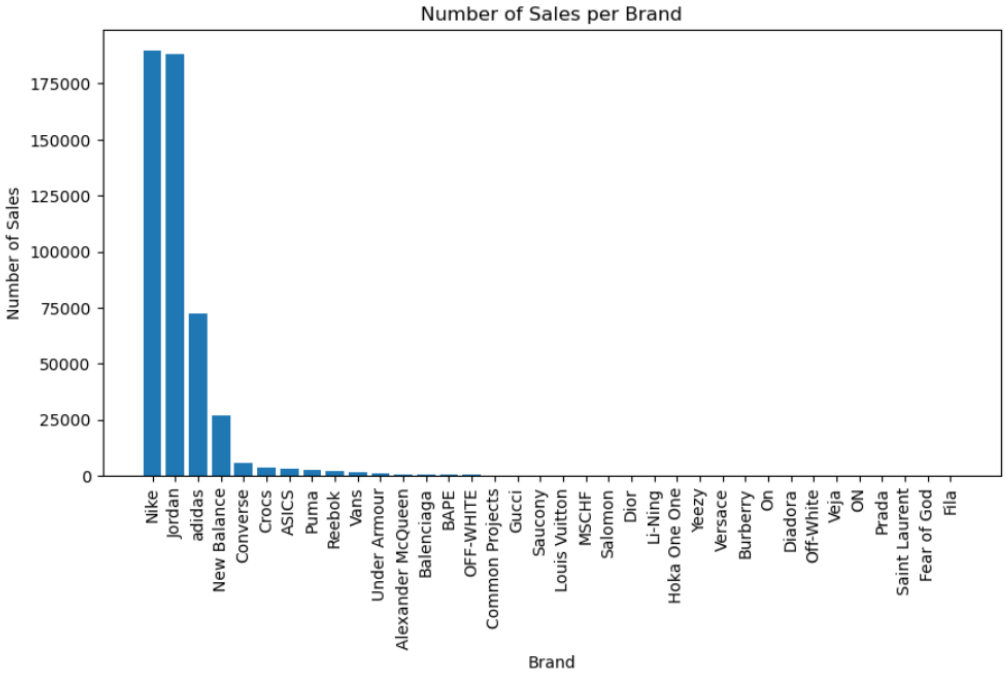
4. *Visualization*: By using Matplot and Seaborn libraries in python, the sneaker dataset can be transformed into statistical and graphical representation. It will give a clear understanding of data patterns, relationships, and trends which in turn will help to analyze the future predicted prices in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

Technical Analysis

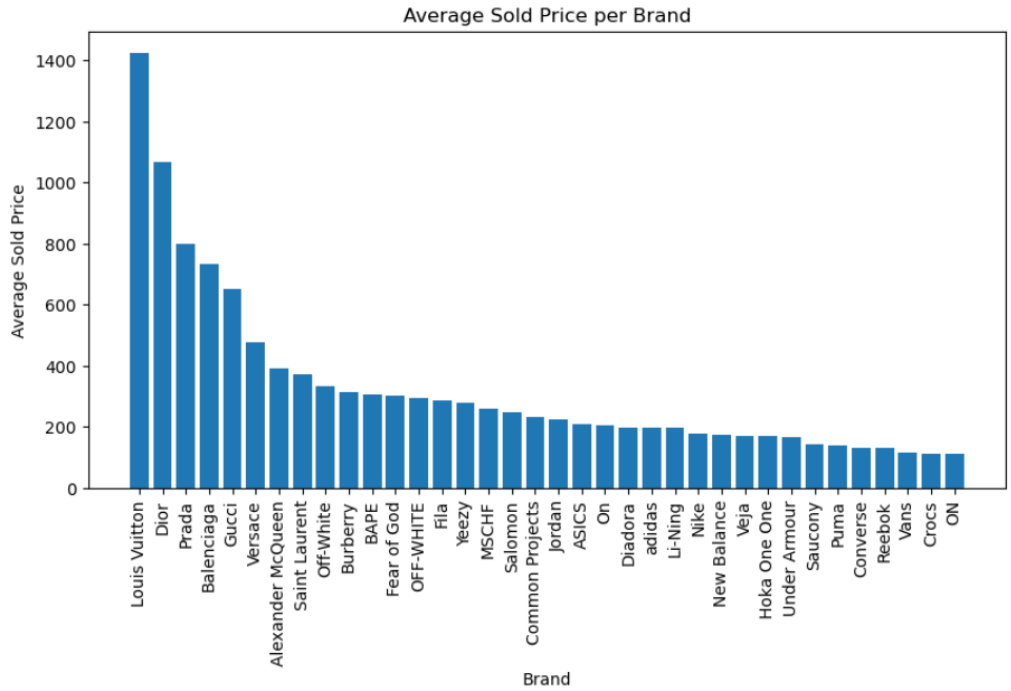
1. The sneaker dataset, which is given to us contains 14 million records, of which we took a subset of 500k to perform our analysis.
2. Initially we started experimenting with Random Forest algorithm, we observed a high mean square error, so later we used it for feature extraction.
3. Standardisation and Normalisation technique is used to improve the accuracy.
4. Feed Forward Network Algorithm is used to predict the future prices of sneaker.

Data Visualization Analysis

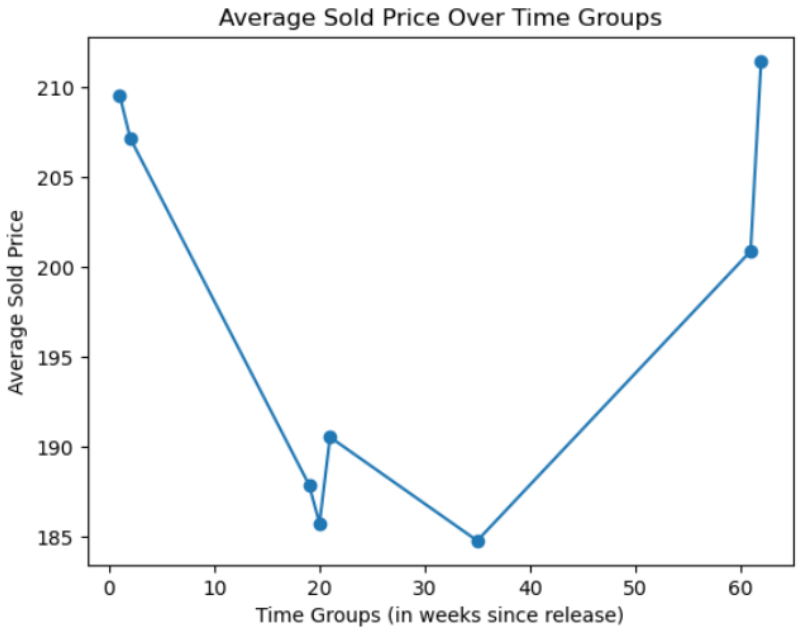
1. Below plot shows which brand has the highest number of sales.



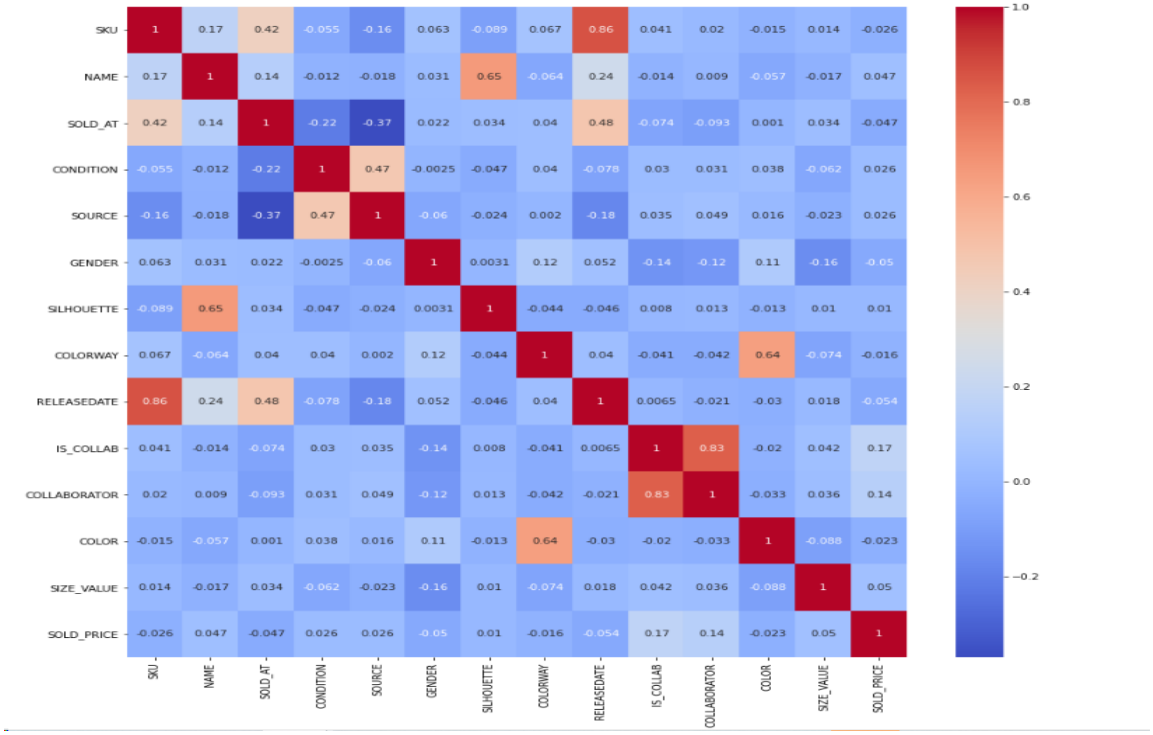
2. Below plot shows Average sold price per brand



3. Below plot shows Average sold price per week.



4. Heatmap are used to show relationship between two variables one plotted on each axis. Each square shows the correlation between the variables. If the value is closer to 0 there is no linear trend between two variables. The close to 1, it is more positively correlated.



Python Libraries Used

The key reasons why Python is good for AI and ML is its expansive ecosystem of libraries such as TensorFlow, PyTorch, Keras, and scikit-learn. These libraries provide pre-built functions and modules, drastically reducing development time and effort [4]. Additionally, they offer a wealth of resources and community support, empowering developers to confidently tackle complex AI, ML, and deep learning tasks [4]. This ease of use and rapid development play a vital role in AI, ML, and DL projects, where experimentation and iteration are fundamental processes.

Library	Usage
Pandas	Work with relational/ labeled data. Provides operations for manipulation numerical data and time series.
Scikit- learn	For predictive data analysis o sklearn.ensemble -> RandomForestRegressor o sklearn.preprocessing -> LabelEncoder o sklearn.model_selection-> train_test_split o sklearn.linear_model -> LinearRegression o sklearn.metrics -> mean_squared_error, r2_score o sklearn.preprocessing -> StandardScaler
NumPy	To work with n-dimensional arrays, provides high level of mathematical functions
MatPlotLib	For creating static, animated, and interactive visualizations in Python
Seaborn	For plotting statistical graphs in python
Tensorflow	o tensorflow.keras.callbacks -> ModelCheckpoint o tensorflow.keras.losses -> MeanSquaredError o tensorflow.keras.metrics -> RootMeanSquaredError o tensorflow.keras.optimizers -> Adam
StatsModels	Toolbox for checking linear regression model to check if the regression model is good or reliable.

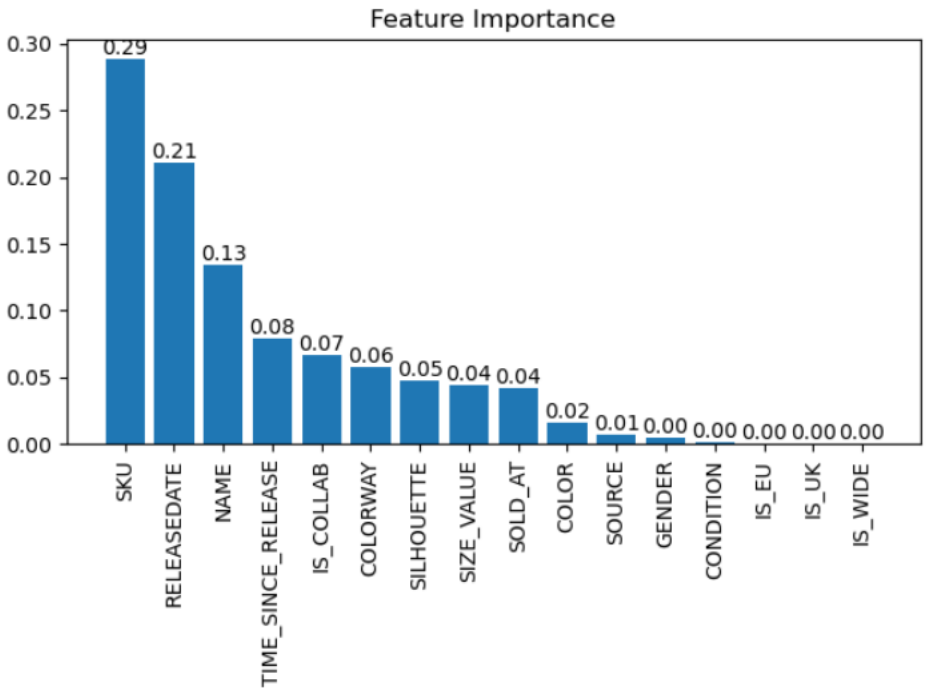
Final Deliverables

Approach to the problem

1. Approach 1: *Random Forest*

We used Random Forest Algorithm to determine the importance of the given features and their impact on predicting the Price.

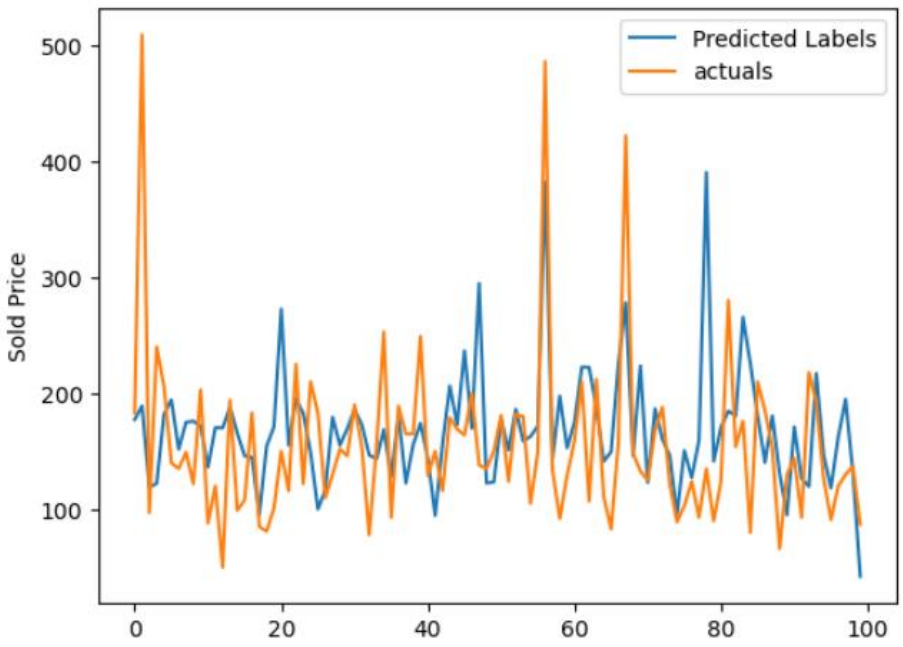
Below plot shows the feature importance of the properties of sneaker data. From the below graph it is seen that SKU has the highest importance out of the others.



2. Approach 2: *Feed Forward Neural Network*

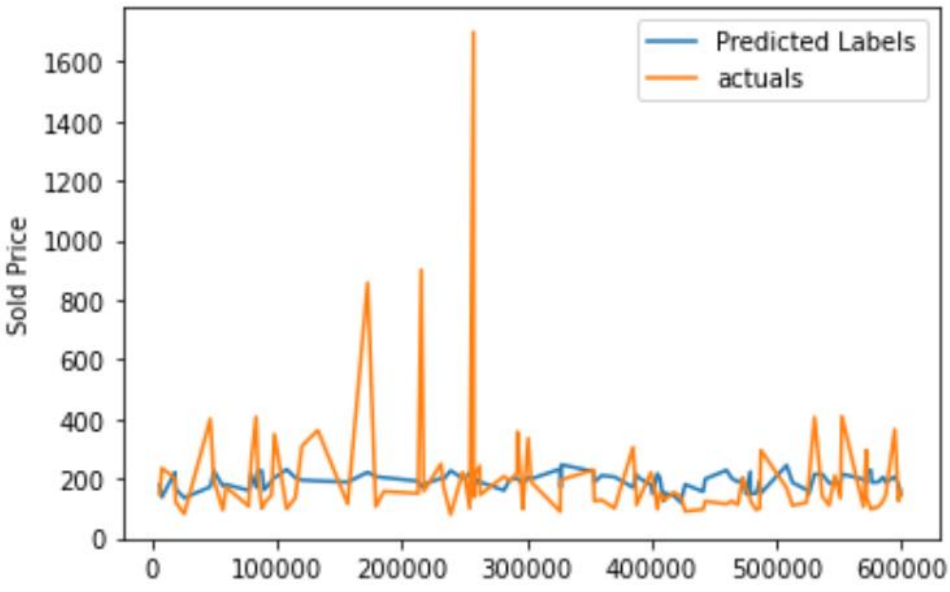
We used feed forward network because it is feasible to handle the nonlinear and complex data.

Below plot shows the actual vs predicted labels.

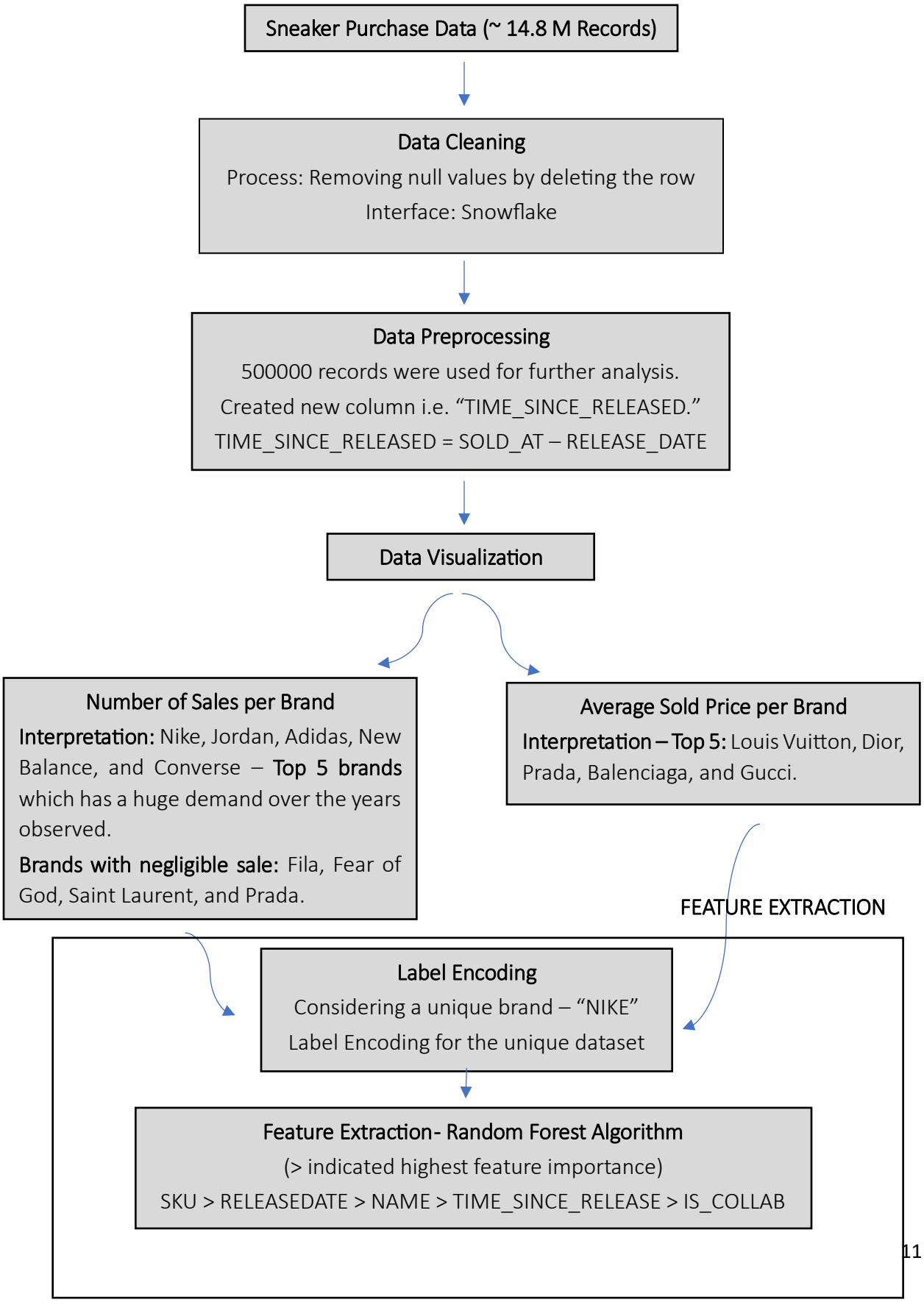


3. Approach 3: *Linear Regression*

Linear Regression is the statistical model which describes the relationship between dependent variable and one or more independent variables.



Logical Flow Analysis



Data Standardization

Transforms Data – to improve stability and accuracy.

Model Selection

Feedforward Neural Network

Splitting Data into train and test: 80%- train, 20%- test

Model Fitting

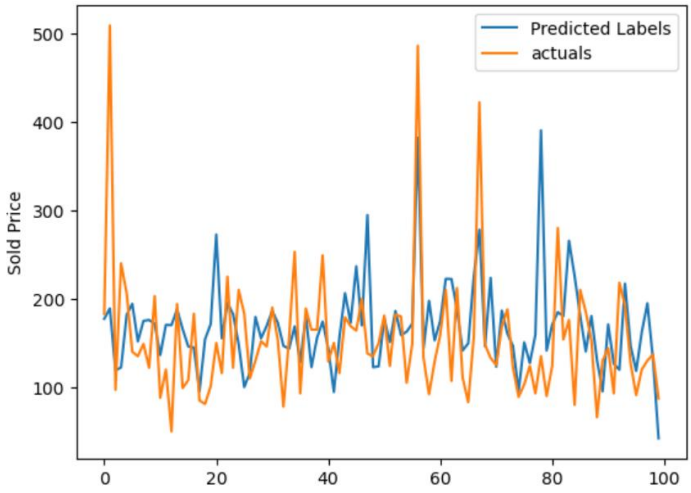
Checking how well the model adapts to the data on which it was trained.

Inverse Transform

Converting actual and predicted labels back to their original scale for calculating the mean squared error and for plotting.

Model Testing/ Validation with Test/ Set Validation Dataset

Mean Square Error – 168. 307



Challenges Faced

1. We were unable to read the data from S3 bucket to Sage Maker(AWS) directly.
2. Faced issues while selecting the right model for the prediction, as it was giving very high mean square error.
3. Faced issues while encoding the features.
4. Faced issues while using the standard scaler and while scaling back to the original values.

Conclusion

We tried to build the algorithm by using different models, out of which we selected random forest to determine feature importance and feed forward neural network to predict future prices. Due to the inconsistency in the data, we are getting a very high mean squared error when we model using linear regression that is why we went ahead with feed forward neural network.

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

4. 7 Reasons Why Python is Best for AI, ML, and Deep Learning (onix-systems.com)

5. How Nike Sneaker Collecting Has Shaped Culture and Lifestyle - Swappa Blog