Application of the KDD Methodology for Predicting Avocado Prices: A Case Study Using the Avocado Prices Dataset

October 21, 2024

Abstract

This research paper explores the application of the Knowledge Discovery in Databases (KDD) methodology to predict avocado prices using the publicly available Avocado Prices dataset from Kaggle. The KDD process, consisting of five phases—Data Selection, Data Preprocessing, Data Transformation, Data Mining, and Interpretation/Evaluation—was meticulously followed. The aim was to understand price patterns and predict future prices using regression models. This study outlines each phase in detail, inclu...

1 Introduction

The Knowledge Discovery in Databases (KDD) methodology provides a structured approach for discovering meaningful patterns from large datasets. In this case study, we apply the KDD process to the Avocado Prices dataset from Kaggle to understand and predict avocado prices. The project involves data exploration, feature engineering, model building, and evaluation of results, aimed at identifying key factors influencing price variations.

2 Data Selection

In the data selection phase, the relevant columns were identified from the dataset, focusing on features that contribute to price prediction, such as `Total Volume`, `Type`, `Region`, and `Date`. Initial exploration indicated the importance of these features based on their potential influence on avocado prices.

2.1 Dataset Description

The dataset consists of various columns representing different attributes of avocado sales across multiple regions, including date, type (conventional or organic), and total volume sold. The target variable is `AveragePrice`, representing the average selling price of avocados. The selected features were further analyzed for relevance and potential transformation in subsequent phases.

3 Data Preprocessing

This phase involved handling missing values, dropping irrelevant or redundant columns, and checking for outliers.
Handling Missing Values: No missing values were found in the dataset, making it ready for transformation.
Outlier Detection: Outliers were identified but retained to preserve data integrity.

4 Data Transformation

In the data transformation phase, feature engineering techniques were applied, including encoding categorical variables, extracting temporal features, and scaling numerical features.
Encoding: The `Type` column was label encoded, and the `Region` column was one-hot encoded to convert categorical data into numerical form.
Correlation Analysis: A correlation matrix was generated to identify relationships between numerical features (see Figure 1).

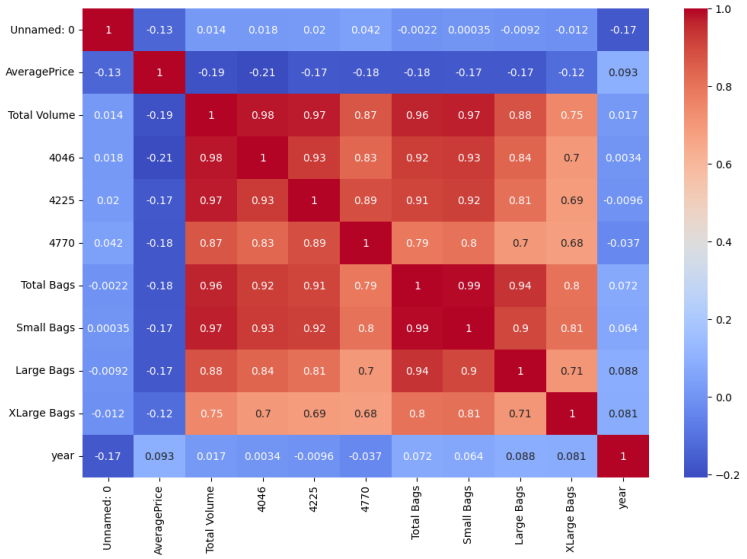

Figure 1: Correlation Matrix of Numerical Features

5 Data Mining

In this phase, various regression models were trained, including Linear Regression, Decision Trees, Random Forests, and Gradient Boosting.
Feature Importance: Feature importance analysis was performed using the Random Forest model to identify the most influential features affecting avocado prices (see Figure 2).

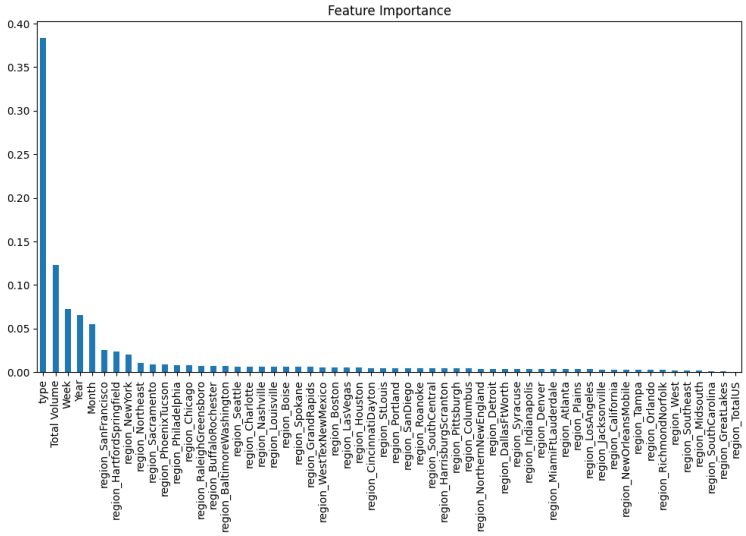

Figure 2: Feature Importance from Random Forest Model

6 Interpretation and Evaluation

The evaluation focused on comparing the actual and predicted prices, analyzing residuals, and determining model effectiveness.

Model Performance: The actual vs. predicted prices plot helps to visually assess model accuracy (see Figure 3).

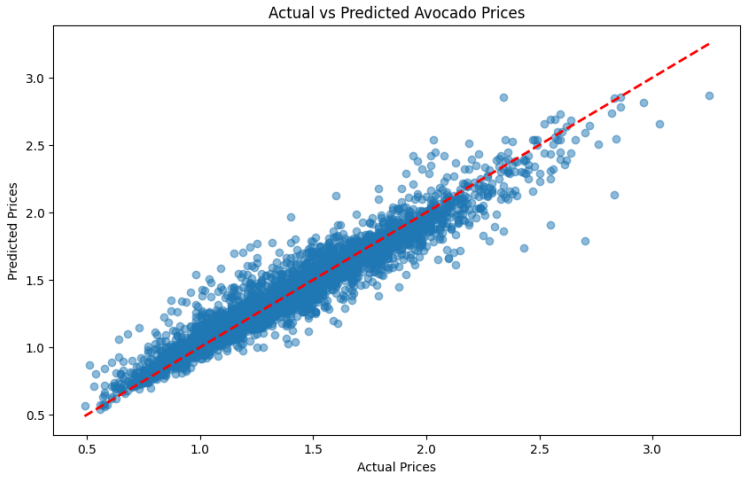

Figure 3: Actual vs. Predicted Avocado Prices

7 Conclusion

The KDD methodology proved effective in predicting avocado prices, providing valuable insights into the factors affecting price variations. Future work could involve integrating

additional external data sources, such as weather patterns or economic indicators, to improve model accuracy further.

Acknowledgments

The author would like to thank the creators of the Avocado Prices dataset on Kaggle for providing the data for this study.

References

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37–54.

[2] Neuromusic. (2018). Avocado Prices dataset. Retrieved from `https://www.kaggle.com/datasets/neuromusic/avocado-prices`.