Predicting Avocado Prices: A Comparative Analysis of Machine Learning Models

Sangram Jagtap

September 30, 2023

Abstract

This paper presents an analysis of a dataset containing avocado prices and corresponding features, aiming to build predictive models for forecasting these prices. Different machine learning models, including Linear Regression and Random Forest, are evaluated based on their performance metrics. The study offers insights into the importance of various features and discusses potential deployment scenarios.

1 Introduction

Avocado prices can vary significantly based on various factors like type, region, and seasonality. Accurately predicting these prices can offer immense value to various stakeholders in the avocado industry. This paper provides an in-depth analysis using the CRISP-DM methodology and evaluates the performance of different predictive models on the task.

2 Methodology

The CRISP-DM methodology, encompassing Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, was employed. The dataset was sourced from Kaggle and contained features like date, type, region, and different avocado PLU codes. Two predictive models, Linear Regression and Random Forest, were trained and evaluated.

3 Results

The performance metrics for the models are as follows:

Table 1: Performance Metrics Comparison

Metric	Linear Regression	Random Forest
MSE (Training)	0.0664	0.0022
MSE (Test)	0.0664	0.0154
R^2 (Training)	0.5912	0.9864
R^2 (Test)	0.5866	0.9042

4 Discussion

The Random Forest model outperformed Linear Regression in terms of all performance metrics. However, there was a noticeable difference between training and test metrics for the Random Forest, suggesting potential overfitting. Feature importance analysis revealed that the type of avocado (organic or conventional) played a significant role in determining its price.

5 Conclusion

The study successfully applied machine learning models to predict avocado prices, with the Random Forest model showing promising results. The insights from the analysis can aid various stakeholders in the avocado industry in making informed decisions.

6 Future Works

- Exploration of advanced models and ensemble techniques.
- Incorporation of more recent data for improved model relevance.
- Fine-tuning of model parameters to further enhance accuracy and mitigate overfitting.

References

1. Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press.

2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

3. Kaggle. (2018). Avocado Prices Dataset. Retrieved from Kaggle website.