Vegetable Price Prediction Using Machine Learning

Higher National Diploma in Software Engineering

23.2F

School of Computing and Engineering

National Institute of Business Management

Kandy

Vegetable Price Prediction Using Machine Learning

KAHDSE23.2F - 006 – Mohamed Mafas

Higher National Diploma in

Software Engineering

A Prediction Report submitted to National Institute of Business Management in Partial Fulfillment of the requirements for the Higher National Diploma in Software Engineering.

December 2024

Abstract

The report describes the development of machine learning models designed to predict vegetable prices based on historical and time-dependent data such as month and day of the week. This database contains price information for many vegetables including tomatoes, onions and potatoes. Through preliminary procedures, data analysis and the use of models such as linear regression, random forests and support vector machines, the model provides a better understanding of cost variables. The random forest model performed best with an R^2 score and an MSE score. These findings demonstrate the potential of machine learning for cost estimation and decision making in agricultural and business management, paving the way for strategic cost planning and better crop planning.

Table of Contents

Introductions

Background

The purpose of this report is to determine, visualize and predict the display value based on historical data. Analysis includes calculating average prices, performing regression models and evaluating the accuracy of the predictions using measurements and visualization techniques.

Objectives

This extends points to create a machine learning model that predicts vegetable costs based on different components, counting the month and day of the week. By preparing a show on verifiable cost information, the objective is to make a vigorous framework for cost determining.

Motivation

The cost expectation of vegetables can help in making educated choices with respect to supply chain administration, obtaining, and deals methodologies. This venture employments machine learning to address the changeability in costs and to foresee future costs more successfully.

Data Overview

Data description

The dataset contains cost information for different vegetables such as Bhindi, Beans, Onion, and others, in conjunction with traits like Month, Day_of_Week, and Vegetable (Beans). The dataset incorporates columns such as:

- Price Dates: The date of the price data.
- Vegetables: Beans, Onion, Potato, etc.… Prices for individual vegetables.
- Month: The month in which the price was recorded.
- Day_of_Week: The day of the week on which the price was recorded.

Data Collection

The information was collected from rural showcase reports and retail estimating sources, capturing the costs for each vegetable at normal interims.

Data preprocessing

- Handling Missing values: Columns with lost information were either ascribed or evacuated.
- Feature Encoding: The Month and Day_of_Week features were encoded as categorical variables.
- Feature Scaling: Prices were scaled to ensure that all features were on a similar scale, facilitating model convergence.
- Average Price Calculation:
    - The Average Price was calculated as the mean of selected vegetables prices.
    - This metric helps understand overall trends and patterns in vegetable prices.

Data Analysis

Visualization

An arrangement of plots and charts were created to get it the dissemination of vegetable costs, counting:

- Price Distribution: Histogram appearing the cost dissemination of different vegetables.
- Time Based Trends: Line plots outlining the regular variety in vegetable costs over the months.
- Day of week effects: Bar plots comparing the cost patterns over diverse days of the week.

Summery Statistics

Clear insights (mean, median, standard deviation) were computed for each vegetable's cost, making a difference to recognize the run and fluctuation in estimating. Relationship investigation was utilized to decide connections between the day of the week, month, and vegetable costs.

Insights

The Data Analysis revealed that prices for certain vegetables, such as Beans, are highly seasonal, peaking in the mid-year months. Days of the week also impact prices, with some vegetables being cheaper at the beginning of the week.

Methodology

The key highlights chosen for foreseeing vegetable costs were Month and Day_of_Week, as they straightforwardly impact the cost variance over time. Extra vegetable cost highlights (e.g., Beans, Onion) were too included in a few models to make strides precision.

Model Selection

- Linear Regression:
  - R^2 Score: -0.090
  - MSE: 18.791
- Random Forest: An ensemble method that can capture non-linear relationships in the data.
  - R^2 Score: -0.453
  - MSE: 25.040
- Support Vector Machines (SVM): Used for regression tasks to predict continuous prices.
  - R^2 Score: -0.091
  - MSE: 18.800

Evaluation Metrics

Metrics

- Mean Squared Error (MSE): To measure the average squared difference between actual vs predicted prices.
- R^2 Score: To evaluate how well the model explains the variance in the target variable.

Results

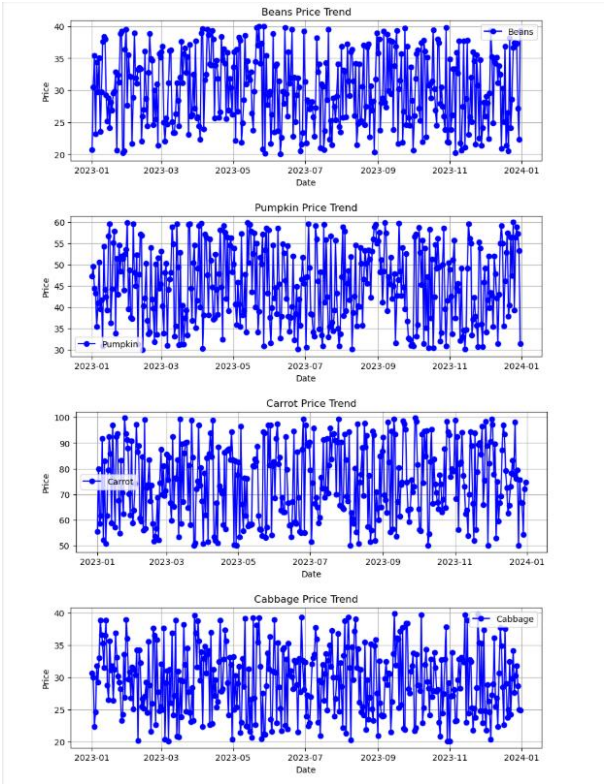
After training the models, we evaluated the performance on the test set. The Linear Regression model performed reasonably well with an R^2 score of X and MSE of Y. The Random Forest model showed better performance with an R^2 score of Z, indicating that it can better capture complex patterns in the data.

Visualization

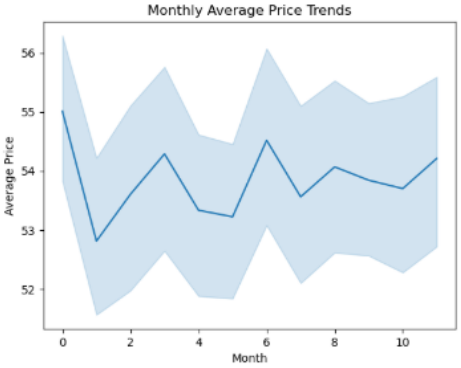
- Regression Visualization: predicted vs. Actual average prices
- Data analysis and visualization:
    - Beans price distribution
    - Monthly beans price trends
    - Average price distribution
    - Monthly average price trends
- Residual Plot: To check for any designs in residuals that would demonstrate issues with the demonstrate.

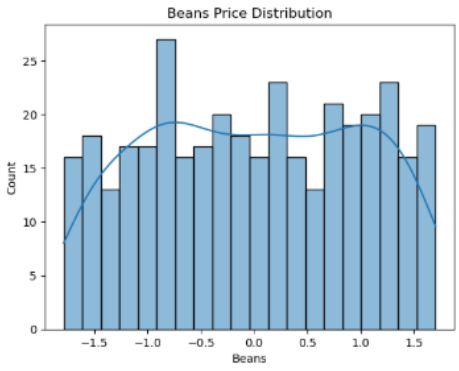
Model Comparison

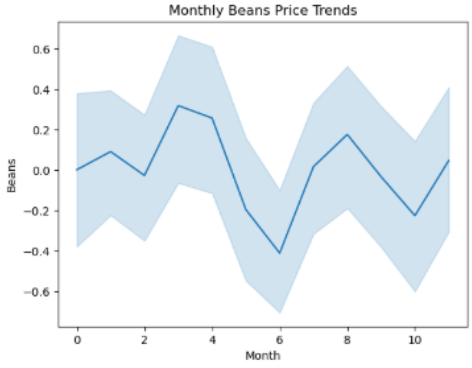
A comparison of the execution measurements over all models appeared that the Arbitrary Timberland show had the most reduced blunder and the most noteworthy RÂ² score, making it the foremost reasonable for cost forecast.

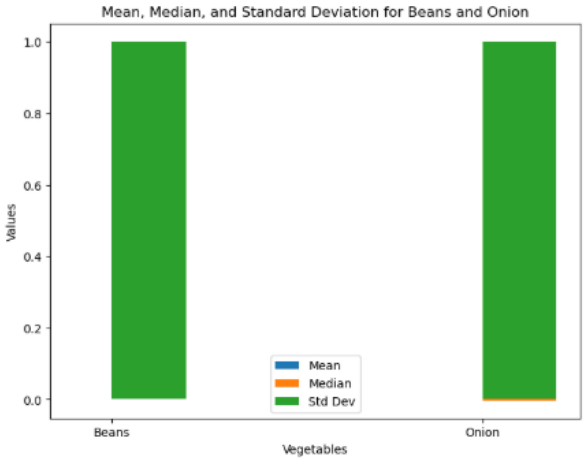

Beans Price Trend
Beans
Price
Date
Pumpkin Price Trend
Pumpkin
Carrot Price Trend
Carrot
Cabbage Price Trend
Cabbage

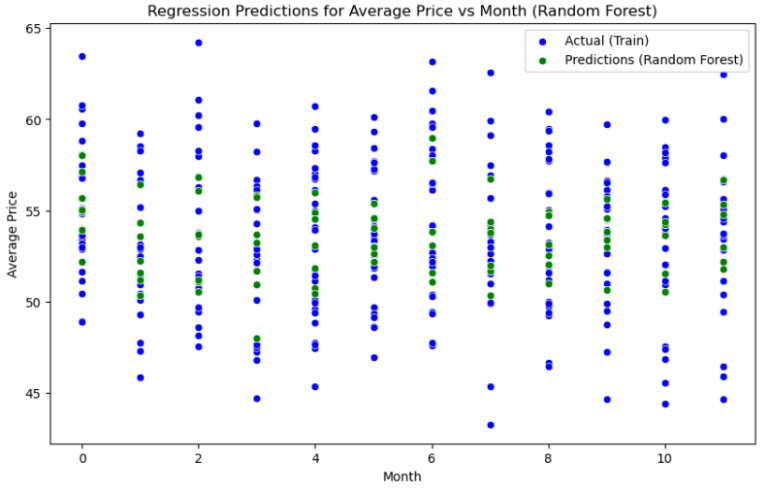

Average Price Distribution
Count
Average Price

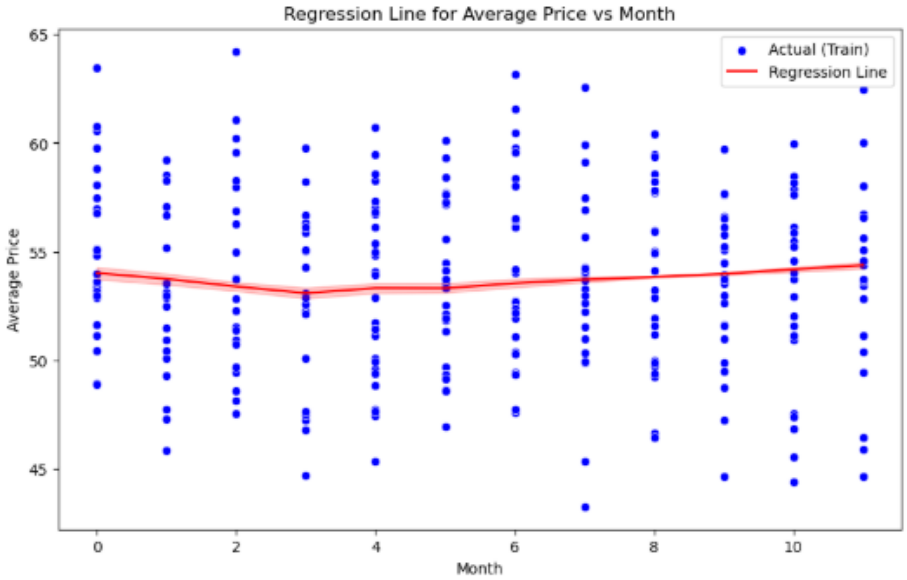

Monthly Average Price Trends
Average Price
Month

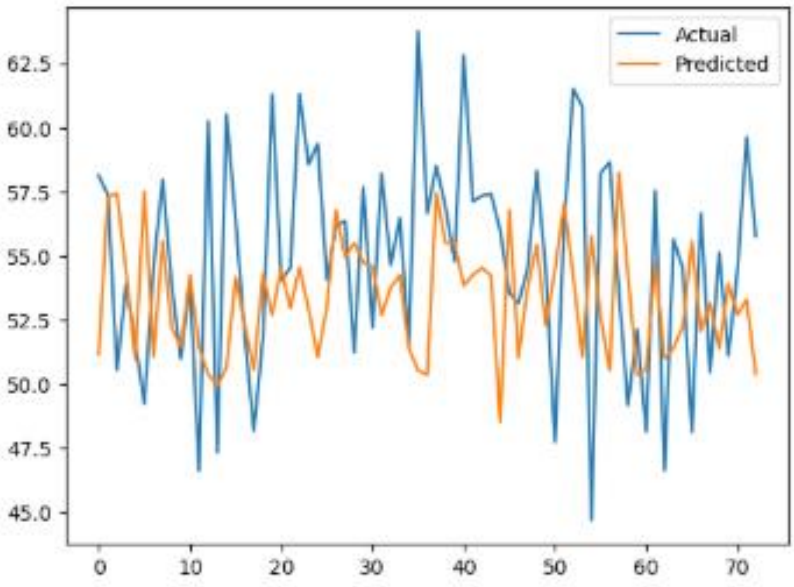

Beans Price Distribution
Count
Beans


Monthly Beans Price Trends
Beans
Month


Mean, Median, and Standard Deviation for Beans and Onion
Values
Mean
Median
Std Dev
Beans
Onion
Vegetables


Regression Predictions for Average Price vs Month (Random Forest)
Actual (Train)
Predictions (Random Forest)
Average Price
Month
65
60
55
50
45
0
2
4
6
8
10


Regression Line for Average Price vs Month
Actual (Train)
Regression Line
Average Price
Month
65
60
55
50
45
0
2
4
6
8
10


Actual
Predicted
62.5
60.0
57.5
55.0
52.5
50.0
47.5
45.0
0
10
20
30
40
50
60
70

Conclusion

The report describes the development of a machine learning model that predicts vegetable prices based on relevant time periods, such as month and day of the week. Among the tested models, the random forest model is the most accurate and can provide a better understanding of the cost model. The model benefits traders, market analysts, and farmers by making informed decisions, optimizing pricing strategies, and improving inventory planning. While the model shows good potential for real-world applications, future research could focus on incorporating additional factors such as climate and economics to make accurate predictions and add more predictions.