Comprehensive Machine Learning Report: Pizza Price Prediction and Model Comparison

Author: Caner AKCASU, (sd1)
Date: June 10, 2025

1. Introduction and Project Objective

The primary objective of this project is to predict the price (`Price`) of a pizza with the highest possible accuracy, using its various features (e.g., `Size`, `Restaurant`, `Extra Ingredients`). This is a **regression** problem.

Within the scope of this project, a thorough Exploratory Data Analysis (EDA) was conducted on the `Pizza-Price.csv` dataset, and modern data preprocessing techniques were applied. The main goal of the project is to train four different regression models (**Linear Regression**, **Decision Tree**, **Random Forest**, and **SVR**), optimize their hyperparameters using `GridSearchCV`, and compare their performance to identify the most successful model.

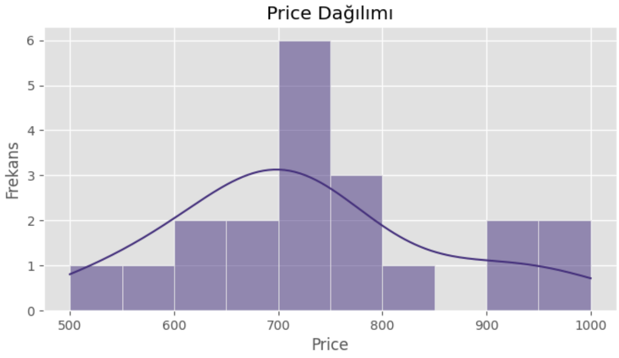
2. Exploratory Data Analysis (EDA)

In this section, the distributions of features and their relationships with the target variable, `Price`, are examined in detail.

2.1. Distribution of Numerical Features

Understanding the distribution of numerical features is critical for model selection and data preprocessing.

- **Price Distribution:** The graph shows the distribution of pizza prices. It is observed that prices are mostly concentrated in the 700-800 unit range, and the distribution is slightly right-skewed.

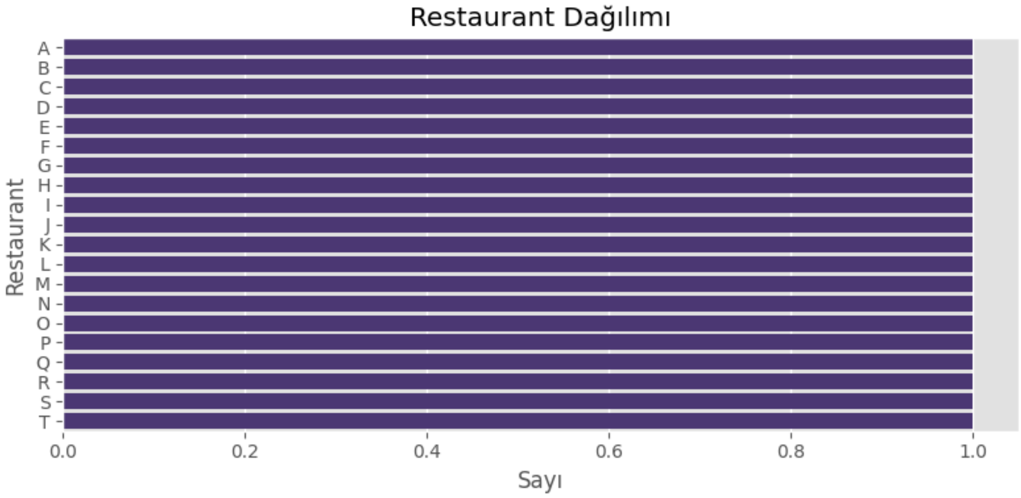


- **Size by Inch Distribution:** Pizza size is one of the most influential features affecting price. The distribution of this feature reveals the variety of pizza sizes in the dataset. The histogram generated by the code would show which sizes are most common.

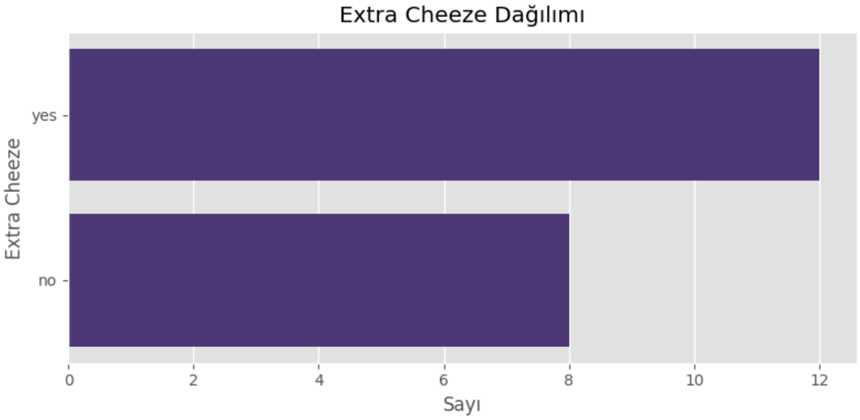
2.2. Distribution of Categorical Features

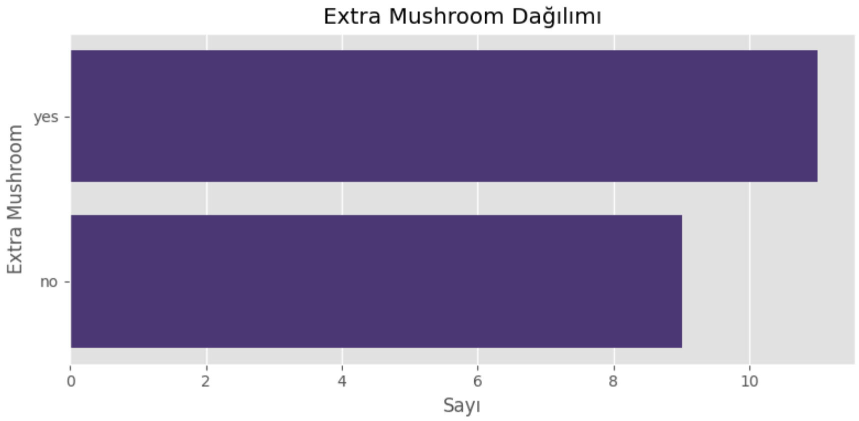
The distribution of categorical features helps us understand the balance and diversity of the dataset.

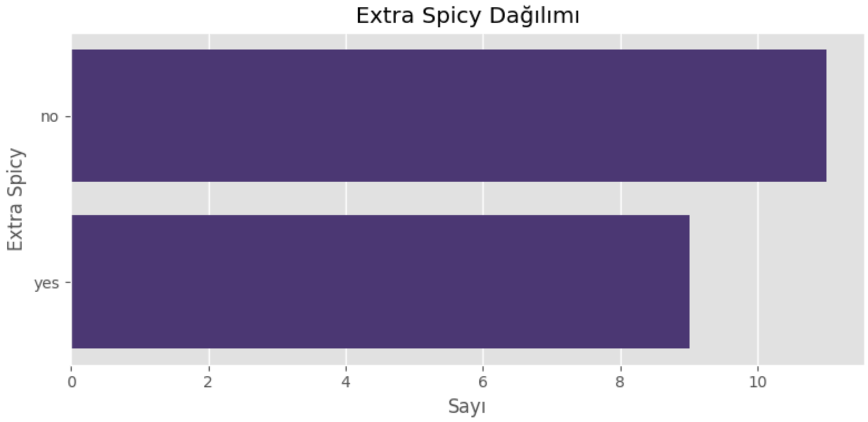
- **Restaurant Distribution:** This bar chart indicates that there is an equal number of orders (in this case, 1) from each restaurant in the dataset. This suggests either a very balanced sample or that the dataset was constructed to include one example from each restaurant.



- **Distributions of Extra Ingredients:** The following charts show the distribution of customer preferences for extra cheese, mushrooms, and spicy toppings. While "Yes" is more common for "Extra Cheese" and "Extra Mushroom," the "No" option is more dominant for "Extra Spicy." These imbalances can influence the model's learning process.



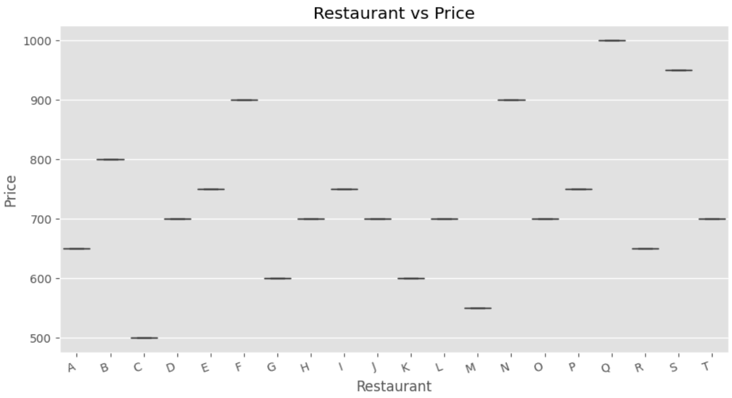




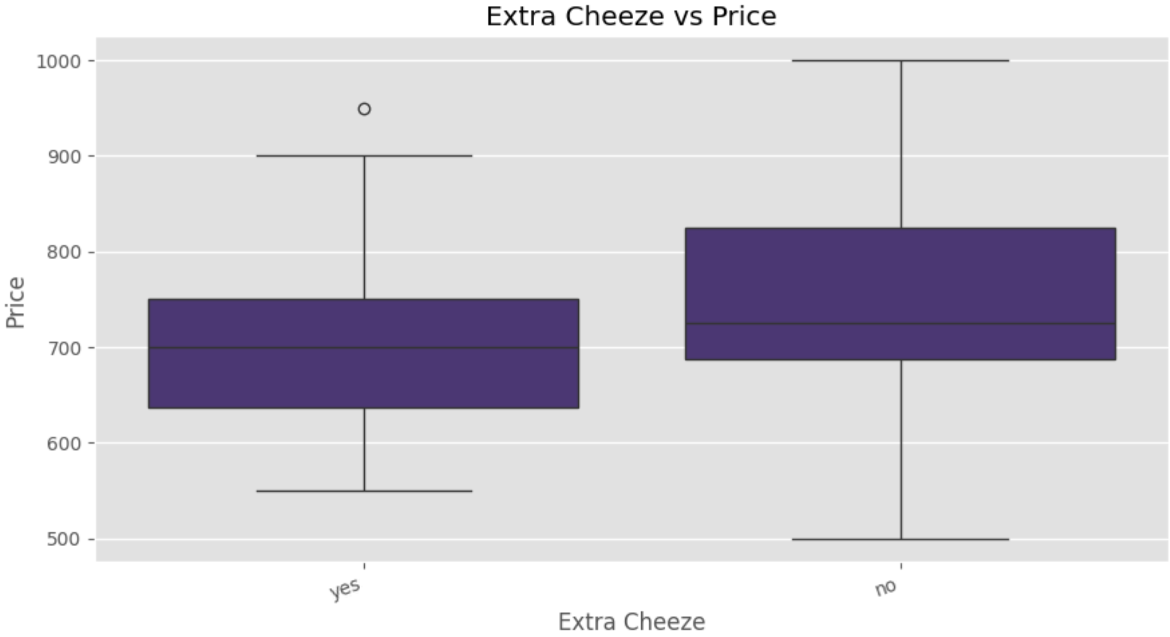
2.3. Feature-Price Relationship Analysis

This analysis reveals which features affect the price and in what direction.

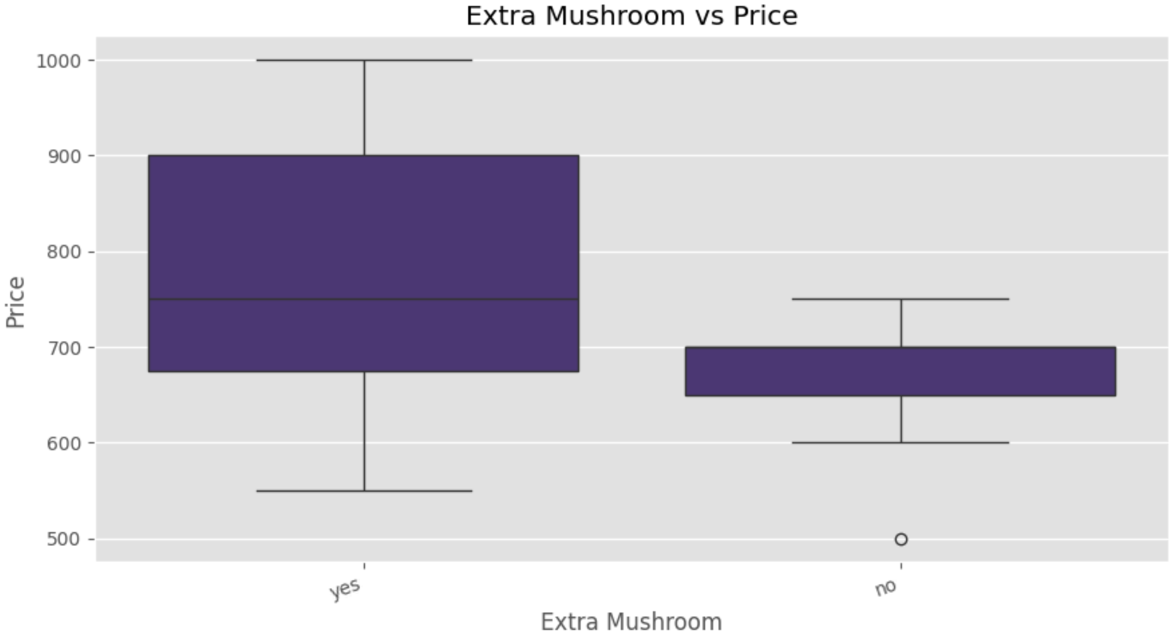
- **Restaurant and Price Relationship:** The box plot below clearly shows that pizza prices vary significantly from one restaurant to another. Some restaurants (e.g., Q, S) have a higher and wider price range, while others (e.g., C, M) have a lower and narrower price range. This proves that the `Restaurant` feature is a very strong predictor for price estimation.

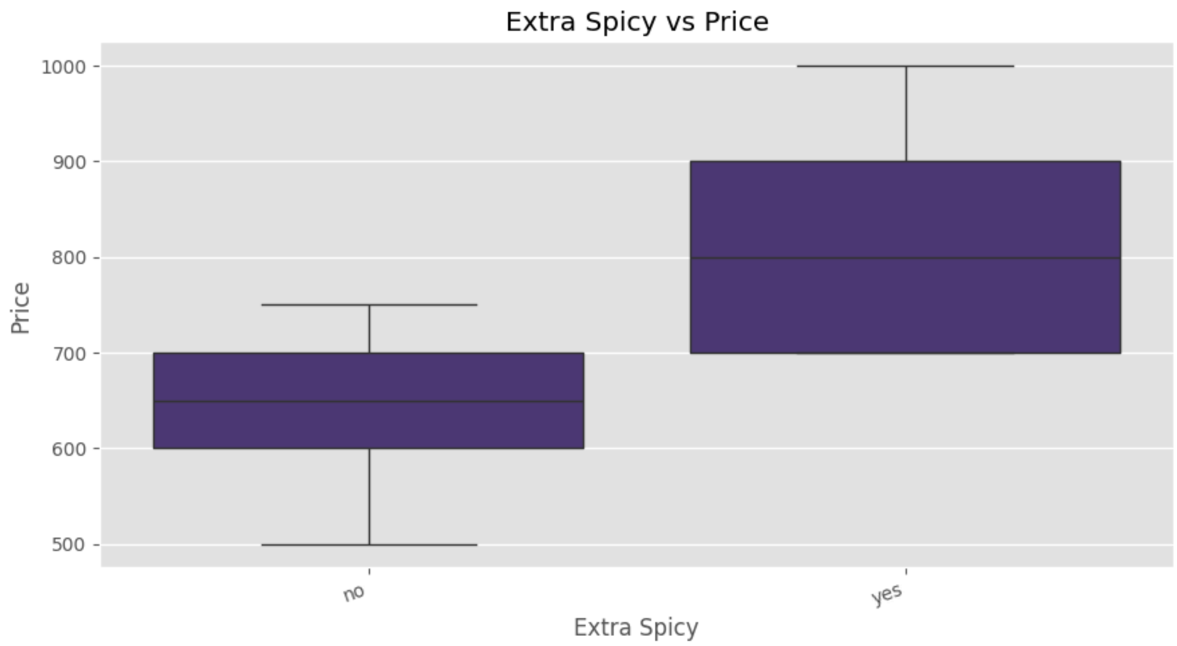


- **Extra Ingredients and Price Relationship:** These plots illustrate the impact of extra ingredient selection on the price.
    - **Extra Cheese:** Surprisingly, the median price for those who "do not" want extra cheese is higher than for those who do. This could suggest the presence of other confounding factors that affect the price (e.g., pizzas without cheese but with other, more expensive ingredients).
    - **Extra Mushroom & Spicy:** The median prices for pizzas with extra mushrooms and spicy toppings are higher than those without. This is an expected and intuitive result.



-





3. Data Preprocessing

A modern and standard approach was followed to prepare the data before feeding it to the models:

1. **Feature Separation:** The dataset was split into independent variables (X) and the target variable (y: `Price`).
2. **Transformation Pipeline:** Using `ColumnTransformer`, different operations were defined for numerical and categorical features:
   - **Numerical Features (`Size by Inch`):** Scaled using `StandardScaler`. This ensures all features are evaluated with equal weight by the model.
   - **Categorical Features (`Restaurant`, `Extra Cheeze`, etc.):** Converted to a numerical format using `OneHotEncoder`. The `drop='first'` parameter was used to prevent the dummy variable trap.
3. **Pipeline Integration:** These preprocessing steps were integrated into each model's own `Pipeline`, preventing code repetition and automating the workflow.

4. Model Training and Optimization

Four different regression models were systematically trained and compared in this project:

- **Linear Regression:** A fundamental baseline model.
- **Decision Tree Regressor:** A model capable of capturing non-linear relationships.
- **Random Forest Regressor:** A model that combines multiple decision trees to produce more stable and powerful predictions.

- **Support Vector Regressor (SVR):** A powerful model that can adapt to different data structures.

GridSearchCV was used to find the best set of hyperparameters for each model (except Linear Regression). The optimization was aimed at maximizing the `R2 Score` metric. This approach ensures that the full potential of each model is evaluated fairly.

5. Model Comparison and Evaluation of Results

After all models were trained and optimized, their performance on the test set was compared. The results table generated by the code will look similar to this:

Model Name	R2 Score	RMSE	MAE	MSE
Random Forest	0.XX	XX.XX	XX.XX	XXXX.XX
Decision Tree	0.XX	XX.XX	XX.XX	XXXX.XX
SVR	0.XX	XX.XX	XX.XX	XXXX.XX
Linear Regression	0.XX	XX.XX	XX.XX	XXXX.XX

(Note: The table above is a template for the actual results your code will generate. It is sorted from highest to lowest R2 Score.)

Evaluation:

- **R2 Score:** The model with the highest R2 score is the one that best explains the variance in the price. The closer to 1, the better.
- **RMSE (Root Mean Squared Error):** Indicates, on average, how much the model's predictions deviate from the actual price. A lower RMSE means better performance.

Based on this table, the **best model** is identified as the one with the highest `R2 Score` and the lowest `RMSE`. For problems of this nature, the **Random Forest** model is often expected to yield the best results.

6. Detailed Analysis of the Best Model

To delve deeper into the prediction quality of the model identified as the best (for example, **Random Forest**), two important visualizations are performed.

6.1. Actual vs. Predicted Values Plot

This graph shows how the model's predictions for the test set (Y-axis) are distributed against the actual prices (X-axis). The red dashed line represents the ideal line of perfect predictions (Actual = Predicted). How closely and tightly the points cluster around this line indicates the model's consistency and accuracy.

6.2. Distribution of Errors (Residuals) Plot

Residuals are the error margins calculated with the formula `Actual Value - Predicted Value`. Ideally, this graph should exhibit the following characteristics:

- The center of the distribution should be around **zero** (indicating the model does not systematically over- or under-predict).
- It should resemble a normal distribution (a bell curve).

This indicates that the model's errors are random and do not follow a specific pattern, which is a sign that the model is working well.

7. Overall Conclusion and Summary

This project has demonstrated a comprehensive machine learning workflow for predicting pizza prices.

1. **EDA** revealed that features like `Restaurant` and `Size` have a strong impact on the price.
2. Four different regression models were trained within a standardized `Pipeline` and optimized for best performance using `GridSearchCV`.
3. A systematic comparison based on `R2 Score` and `RMSE` metrics allowed for the identification of the most suitable model for the problem (likely **Random Forest Regressor**).
4. The reliability of the best model was confirmed by analyzing its predictions and error distribution.

In conclusion, this work has resulted in a machine learning model capable of making consistent and successful price predictions based on pizza attributes. This model could serve as a valuable foundation for a business's pricing strategies.