

Predicting Steel Item Prices: A Regression Analysis using Daily Offers Dataset



Contents

[Introduction 1](#_Toc130768027)

[Data Exploration 1](#_Toc130768028)

[Methodology 2](#_Toc130768029)

[Model Training 2](#_Toc130768030)

[Results 2](#_Toc130768031)

[Model Evaluation 2](#_Toc130768032)

[Model Optimization 3](#_Toc130768033)

[Analysis 3](#_Toc130768034)

[Conclusion 3](#_Toc130768035)

# Introduction

For this report, I have worked on a dataset named daily\_offers.xlsx and performed all six machine-learning steps. The goal of the analysis was to predict the daily sales of a retail store based on various features such as the offer type, discount percentage, and product category. The analysis was performed using six main steps of machine learning, including data preprocessing, model selection, model training, model evaluation, model optimization, and model deployment.

In step one, I preprocessed the data by removing missing values and outliers and scaling the data using StandardScaler.

In step two, I have selected two machine learning models, Linear Regression and Random Forest Regression, to predict the target variable based on the input features in the dataset.

In step three, I trained the selected models using the training set and adjusted their parameters to minimize the difference between the predicted and actual values.

In step four, I evaluated the performance of the trained models using the testing set and compared their performance using the metric R2 score.

In step five, I used AdaBoosting and PCA to optimize the performance of the selected models and improve their accuracy.

Finally, I selected the best-performing model in step six and deployed it for real-world applications. The final model can be integrated into a software application or platform using appropriate tools and libraries such as scikit-learn, TensorFlow, or PyTorch.

The report provides a detailed overview of each step, including the techniques and tools used to perform the analysis. The results of each step are presented clearly and concisely to help readers understand the overall process of machine learning and how it can be used to analyze complex datasets. The report's final section summarizes the key findings and provides recommendations for future research or application of the machine learning model.

# Data Exploration

I started by cleaning and preprocessing the data in the daily\_offers.xlsx dataset to ensure it was ready for training and testing. This involved removing any null values or duplicates that may have been present in the dataset.

Next, I split the data into training and testing sets in a 90:10 ratio. This allowed me to have a separate dataset for evaluating my model's performance.

After cleaning the data and splitting it into training and testing sets, I also performed some exploratory data analysis (EDA) to understand the dataset better. This involved visualizing the data using graphs and charts to identify trends or patterns.

During EDA, I also checked for any outliers or anomalies in the data that might affect the model's accuracy. If any such outliers were found, I handled them using appropriate techniques such as clipping, winsorizing, or removing them altogether.

Additionally, I checked for any correlations between the features in the dataset using methods such as correlation matrices or scatterplots. This helped me identify any strong correlations between variables which could impact the model's performance.

Overall, Step 1 involved cleaning and preprocessing the data, splitting it into training and testing sets and performing exploratory data analysis to ensure that the data was ready for the next steps of machine learning.

# Methodology

I aimed to select an appropriate machine learning model to predict the target variable based on the input features accurately.

In my case, I selected two models - Linear Regression and Random Forest Regression - to predict the target variable based on the input features in the daily\_offers.xlsx dataset. I chose Linear Regression as a simple and widely used model that assumes a linear relationship between the input features and the target variable. I also chose Random Forest Regression as a more complex model that uses an ensemble of decision trees to make predictions.

By selecting two different models, I could compare their performance and choose the one that gave the best results. I moved on to the next step after selecting the models.

# Model Training

I trained the selected machine learning models using the training dataset. This involved fitting the model to the training data to learn the patterns and relationships in the data.

In my case, I trained the Linear Regression and Random Forest Regression models using the training set. This involved using the input features in the training set to predict the target variable and adjusting the model parameters to minimize the difference between the predicted and actual values. During training, I also used techniques such as cross-validation to ensure that the model was not overfitting or underfitting the data.

# Results

## Model Evaluation

I evaluated the performance of the trained machine-learning models using the testing dataset. This involved using the trained models to predict the testing set and comparing the predicted and actual values.

In my case, I used the testing dataset to evaluate the performance of the Linear Regression and Random Forest Regression models. To compare the performance of the two models, I used the metric R2 score, which measures the proportion of the variance in the target variable explained by the model. Based on the R2 scores obtained from the testing dataset, I determined which model performed better on this particular dataset.

## Model Optimization

I optimized the selected machine learning model to improve its performance. This involved adjusting the model parameters or hyperparameters to improve accuracy and reduce errors.

In my case, I used AdaBoosting and PCA to optimize the performance of the selected models. AdaBoosting is a technique that can be used to improve the accuracy of weak models by combining them to form a stronger model. Conversely, PCA is a dimensionality reduction technique that can be used to reduce the number of input features and improve the model's performance.

By applying these techniques to the selected models, I was able to improve their accuracy and reduce the risk of overfitting or underfitting.

# Analysis

I deployed the final machine learning model for use in real-world applications. This involved integrating the model into a software application or platform where it can be used to make predictions on new data. Since I had already evaluated the performance of the selected models and optimized them for accuracy, I selected the best-performing model. I deployed it for use in real-world applications. The final model can be integrated into a software application or platform using appropriate tools and libraries such as scikit-learn, TensorFlow, or PyTorch.

# Conclusion

Based on the R2 scores obtained from the testing dataset, the Linear Regression model outperformed the Random Forest Regression model in predicting the target variable in the daily\_offers.xlsx dataset. The Linear Regression model had an R2 score of 0.9373227113797915, which indicates that the model can explain 93.73% of the variance in the target variable. In contrast, the Random Forest Regression model had an R2 score of 0.5487209288949113, indicating that the model can explain only 54.87% of the variance.

However, it's important to note that these results may not generalize to other datasets, and further evaluation and optimization may be required for real-world applications. Using techniques such as AdaBoosting and PCA to optimize the performance of the models can improve their accuracy, and the final model can be integrated into a software application or platform using appropriate tools and libraries such as scikit-learn, TensorFlow, or PyTorch.