The Prediction of Big Mart Sales

Final Report

Practical Machine Learning - DSCI - 6601

Yeganeh Safari

Mahtab Mohammadi

Abstract

In today's retail environment, retail giants are tapping into data analytics to optimize inventory and foresee consumer demand. The BigMart dataset offers a rich trove of data, encompassing both independent and dependent variables, critical for precise forecasting and insightful discoveries. Leveraging machine learning techniques such as random forests and linear regression, this project strives to elevate inventory management and drive data-driven decision-making. As a data science student, I want to build a model that predicts item sales and analyzes what factors impact those sales. I will use the Big Mart dataset, which has many features and a reliable method for prediction. I aim to achieve highly accurate results. The insights gained can guide decisions to boost sales.

Data DescriptionData Description

The dataset comprises various food items retailed across diverse markets. A comprehensive data dictionary is provided, delineating eight distinct features, encompassing four categorical and four numerical attributes, with a total of 8,523 observations.

Data Dictionary:

Item_Identifier: Unique Number assigned to each Item

Item_Weight : the amount of food(weight in g)

Item_Fat_Content: Item Fat Content (low-fat or regular)

Item_Visibility: Placement value of each Item: 0 - Far & Behind 1 - Near & Front

Item_type: Item utility (fruits, dairy, meat, etc.)

Item_MRP: Price of the Item

Outlet_Identifier: Unique Outler Name

Outlet_Establishment_Year: Year of Outlet Establishment

Outlet_Size: Size of the Outler

Outlet_Location_Type: Tier of Outlet Location

Outlet_Type: Type of item outlet utility in markets (supermarkets or groceries)

Item_Outlet_Sales: Total Sales of the Outlet

Correlation Map

The presented visualization illustrates the correlation plot of all variables within the dataset. It is discernible from the plot that sales exhibit a notably higher correlation with "item_MRP," denoting the price of each item. This observation implies that the sales generated by each outlet for a specific item predominantly depend on the price assigned to that item. Consequently, a discernible relationship exists between sales and item price.

Conversely, the remaining variables demonstrate negligible correlation, as evidenced by their proximity to zero on the correlation scale.

Feature Distribution

In the analysis of feature distributions, box plots were employed to visually represent the data. Notably, certain distributions exhibited outliers, signifying potential anomalies, while others adhered to a normal distribution pattern.

These findings contribute valuable insights into the dataset's characteristics and merit further consideration in our overall analysis. The distribution of the features, specially the ones with outliers (item visibility and outlet sales) led us to standardize the dataset.

Training Model

- **Linear regression**
- **Random forest**
- **Decision tree**

The investigation involved the application of three distinct predictive modeling algorithms: linear regression, random forest, and decision tree. These models were selected based on their suitability for handling a numerical dataset, allowing for a comprehensive exploration of the data's underlying patterns.

To facilitate a rigorous evaluation, the dataset was strategically partitioned, allocating 80% for model training and the remaining 20% for testing and validation. This division ensures that the models are exposed to diverse patterns within the data during the training phase and subsequently tested on unseen data, providing a robust assessment of their performance and generalization capabilities.

The incorporation of diverse algorithms enhances the study's analytical depth, as each model brings unique perspectives and strengths to the predictive task. This multifaceted approach contributes to a more thorough understanding of the dataset's characteristics and facilitates the identification of patterns that may be captured differently by each algorithm.

Hyper Parameter Tuning

In the pursuit of optimizing model performance, hyperparameter tuning was conducted specifically for the random forest and decision tree algorithms. Hyperparameter tuning involves systematically exploring a predefined set of hyperparameter values to identify the configuration that maximizes model efficacy.

For the random forest and decision tree models, key hyperparameters such as tree depth, number of trees (for random forest), and splitting criteria were systematically varied and assessed. This meticulous process aims to identify the hyperparameter combination that yields the most favorable model performance.

The benefits of hyperparameter tuning are manifold. Firstly, it enhances model generalization by identifying optimal configurations that mitigate overfitting, ensuring that the model performs well on unseen data. Additionally, hyperparameter tuning can lead to improved model interpretability and efficiency, as it refines the model's internal settings to align more closely with the underlying patterns within the data. Ultimately, this iterative optimization process contributes to the development of more robust and accurate predictive models.

Best Parameters From Model Tuning

The graph depicting the relationship between the maximum depth of the decision tree and model performance reveals a nuanced pattern. Notably, within the range of maximum depths from 5 to 6, the model attains its optimal performance, as evidenced by the convergence of the training score and cross-validation score. This convergence implies that the model generalizes well to unseen data, striking a balance between complexity and simplicity.

However, beyond a maximum depth of 6, a discernible divergence emerges. While the training score continues to increase, indicative of the model's ability to fit the training data more closely, the cross-validation score starts to decline. This divergence is indicative of overfitting, where the model becomes excessively tailored to the idiosyncrasies of the training data and struggles to generalize to new, unseen data.

The diminishing cross-validation score suggests that the model's performance on new data diminishes as the complexity of the model increases beyond a certain point. This underscores the importance of selecting an optimal maximum depth to achieve the best trade-off between model complexity and generalization. In this context, a maximum depth between 5 and 6 emerges as the most favorable, where the model achieves high performance on both the training and cross-validation datasets, indicative of robust predictive capabilities

Predicting Regression

The performance metrics for our linear regression model are as follows:

1. R-squared Score (R²): The R-squared score is a measure of how well the model explains the variability in the target variable. A score of 0.57

indicates that approximately 57% of the variance in the dependent variable is explained by the model. A higher R-squared score suggests a better fit of the model

2. Root Mean Squared Error (RMSE): The RMSE represents the average magnitude of the errors between the predicted and actual values, considering both the direction and magnitude of the errors. With a value of 1069, it indicates the average deviation of your model's predictions from the true values. Lower RMSE values indicate better predictive accuracy, and 1069.22 provides a measure of the average size of the errors in our model's predictions.
3. Mean Absolute Error (MAE): The MAE is another metric that quantifies the average magnitude of errors between predicted and actual values. A value of 791.83 suggests that, on average, our model's predictions deviate by approximately 791.83 units from the actual values. Like RMSE, lower MAE values indicate better predictive performance.

In summary, the linear regression model demonstrates a moderate level of explanatory power ($R^2 = 0.57$), and the RMSE and MAE values provide insights into the average magnitude of prediction errors. Interpretation of these metrics should be done in consideration of the specific characteristics and requirements of your dataset and the context of the predictive task.

Predicting Random Forest

The R-squared score of 0.55 indicates that the random forest model explains approximately 55% of the variance in the dependent variable.

The RMSE of 1097.69 provides an average measure of the deviation between the predicted values and the actual values. This value suggests that, on average, the predictions from the random forest model deviate by approximately 1097 units from the true values.

The MAE of 764.48 indicates the average magnitude of errors between the predicted and actual values. This value suggests that, on average, the random forest model's predictions deviate by approximately 764 units from the actual values.

Predicting Decision Tree

And as we can see, the results for the decision tree are 0.61 for R squared, 1027.73 for Root Mean Squared Error and 721.64 for Mean Absolute error.

Comparing The Models

- R-squared Score (R^2): The decision tree model has the highest R^2, indicating the best explanatory power among the three models.

- Root Mean Squared Error (RMSE): The decision tree has the lowest RMSE, suggesting smaller average prediction errors compared to the other models.
- Mean Absolute Error (MAE): The decision tree also has the lowest MAE, implying smaller average absolute prediction errors.

Overall Impression: The decision tree model appears to outperform both linear regression and random forest in terms of R-squared, RMSE, and MAE. It demonstrates higher explanatory power and yields more accurate predictions on average.

Conclusion And Future Work

In the context of future work, it is recommended to explore advanced modeling techniques, specifically XGBoost and neural networks, to potentially enhance predictive performance beyond the capabilities of the current models. The incorporation of these sophisticated algorithms can unveil complex patterns within the data, contributing to improved accuracy and robustness in predicting sales.

Additionally, expanding the dataset beyond its current one-year timeframe is advised. A more extensive dataset can provide a broader representation of diverse patterns and trends, enhancing the models' ability to generalize and adapt to various scenarios. This extension is particularly crucial for capturing seasonality and accounting for special occasions, such as Black Friday or Christmas, where sales exhibit distinctive patterns. Acknowledging and incorporating these events into the modeling process can lead to more accurate predictions, considering the heightened sales activity during such periods.

In summary, future endeavors should involve the exploration of advanced models, the expansion of the dataset to encompass a more extended timeframe, and a nuanced consideration of special occasions to refine the predictive capabilities, ensuring a more comprehensive and accurate understanding of sales dynamics.