Project Documentation

Objective

The objective of this project was to build a Trend Analysis Machine Learning (ML) model to predict the future popularity and sales trends of various product categories using given dataset features like product ratings, prices, discounts, and sales history.

Methodology

1. Data Preprocessing

- **Loading Data**: The dataset was loaded into a pandas DataFrame from a CSV file.
- **Cleaning Data**: The data was cleaned to handle missing values and correct anomalies in features like 'Rating in Stars' and 'Sold Count'. Non-numeric entries were removed or corrected.
- **Feature Conversion**: Features were converted to appropriate numeric formats for analysis, including converting price information from strings to floats.

2. Feature Engineering

New features were derived from existing data to enhance the model's predictive capability:

- **Discount Amount**: Computed as the difference between original and current prices.
- **Discount Percentage**: Calculated as the percentage reduction from the original price.
- **Price to Sold Ratio**: Designed to understand the relationship between the price and the number of units sold.
- **Rating Effectiveness**: A product of rating and the count of ratings, giving a weighted sense of customer satisfaction.

3. Model Selection and Training

- **Model Selection**: A Random Forest Regressor was chosen for its effectiveness in handling non-linear data and its robustness against overfitting.
- **Training**: The model was trained using the prepared features, excluding the direct 'Sold Count' in later iterations to prevent data leakage.

4. Model Evaluation

- **Metrics**: Mean Squared Error (MSE) and R-squared were calculated to assess the model's performance. A high R-squared value indicated that the model could explain a significant variance in the dataset.

5. Predictions and Future Forecasting

- **Hypothetical Data Prediction**: The model was used to predict sales for a hypothetical new product based on provided features.
- **Feature Importance**: Feature importance was analyzed to understand which factors most influence sales trends, revealing insights into the effectiveness of ratings and pricing strategies.

Findings

- **High Impact of Ratings and Sales History**: The model identified that the number of sales and customer ratings (count and effectiveness) significantly influenced the predictions.
- **Importance of Feature Engineering**: Engineered features like discount amount and price ratios provided valuable insights and improved the model's predictive accuracy.

Instructions to Execute the Code

1. **Environment Setup**: Ensure Python and necessary libraries (pandas, sklearn, numpy) are installed.

2. **Data Loading**: Load the data using pandas from the CSV fil

Conclusion

This project demonstrated how machine learning could be leveraged to predict product sales trends from historical data and various product features. Future work could explore more sophisticated time-series models or deep learning approaches for further enhancement.