Amazon Review Rating Prediction Engine presentation

Author：Jiacheng Gu

Project Background and Objectives

E-commerce platforms need to predict the ratings of new products in advance and optimize pricing and recommendations.

Objective: To predict product ratings ranging from 1 to 5 by using machine learning methods, combining structured and text features.

Business value: Help merchants identify potential bestsellers, optimize inventory and marketing resources, and enhance user satisfaction.

Load data

```
Dataset loaded successfully!
Dataset basic information:
Data shape: (1465, 16)
Columns: ['product_id', 'product_name', 'category', 'discounted_price', 'actual_price', 'discount_percentage'
er_name', 'review_id', 'review_title', 'review_content', 'img_link', 'product_link']

Data types:
product_id            object
product_name          object
category              object
discounted_price      object
actual_price          object
discount_percentage   object
rating                object
rating_count          object
about_product         object
user_id               object
user_name             object
review_id             object
review_title          object
review_content        object
img_link              object
product_link          object
dtype: object

Missing values:
product_id            0
product_name          0
category              0
discounted_price      0
actual_price          0
discount_percentage   0
rating                0
rating_count          2
about_product         0
user_id               0
user_name             0
review_id             0
review_title          0
review_content        0
img_link              0
product_link          0
dtype: int64
```

Data cleaning

The fields such as rating, actual_prices, and discount_price are discount_percentage to numerical types

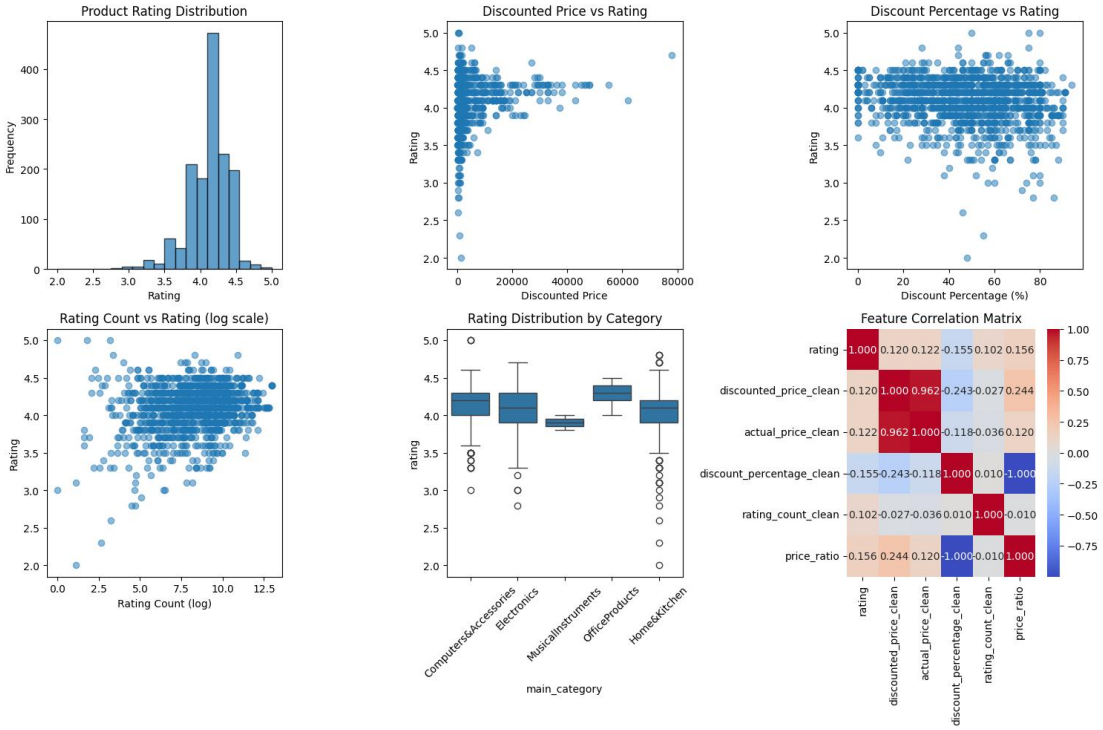
Newfeatures:['discounted_price_clean','actual_price_clean','discount_percentage_clean','rating_count_clean','price_raduct_length',

'category_depth']

Text feature: TF-IDF extracts product description keywords (100 dimensions)

```
==================================================
3. Data Preprocessing
==================================================
Processing target variable rating...
Rating data type after conversion: float64
Number of missing values in rating: 1
Removed missing values: 1465 -> 1464 rows

Rating statistics:
count    1464.000000
mean        4.096585
std         0.291674
min         2.000000
25%         4.000000
50%         4.100000
75%         4.300000
max         5.000000
Name: rating, dtype: float64
Processing price data...
Creating derived features...
Data preprocessing completed!
New features:
['discounted_price_clean', 'actual_price_clean', 'discount_percentage_clean', 'rating_count_clean', 'price_ra
duct_length', 'category_depth']
```

Data Exploration and Visualization (EDA)



We can find:

Rating distribution: Most products are rated between 4 and 4.5 points

Discounts and Ratings: The ratings of high-discount products fluctuate greatly

The more comments there are, the more concentrated the ratings will be

We can find five main_category

The correlation among the features is weak, and there is also a negative correlation

Feature engineering

Multi-source feature fusion enhances the expressive ability of the model, especially text features have made significant contributions to score prediction

```
==================================================
5. Feature Engineering
==================================================
Feature data type check:
  discounted_price_clean: float64, missing values: 0
  actual_price_clean: float64, missing values: 0
  discount_percentage_clean: float64, missing values: 0
  rating_count_clean: int64, missing values: 0
  price_ratio: float64, missing values: 0
  absolute_savings: float64, missing values: 0
  product_name_length: int64, missing values: 0
  about_product_length: int64, missing values: 0
  category_depth: int64, missing values: 0
Processing product description text features...
TF-IDF feature extraction successful, number of features: 100
Final feature count: 110
Sample count: 1464
Invalid values in feature matrix: 0
Infinite values in feature matrix: 0
```

Model construction and training

Training set/test set division: 8:2, hierarchical sampling

Standardization: Use StandardScaler for linear regression

Training models: Linear Regression, Random Forest, Gradient Boosting, XGBoost

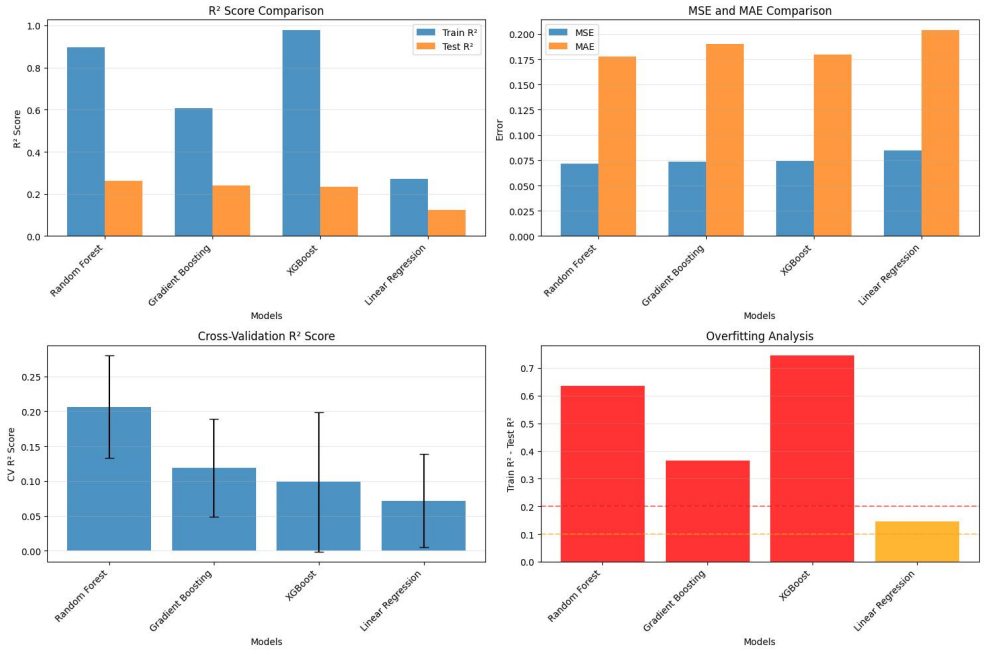
The number of samples in the training set/test set: 1171/293

I compared multiple mainstream regression models, considering both the interpretability of linear models and introducing an integration method to improve accuracy.

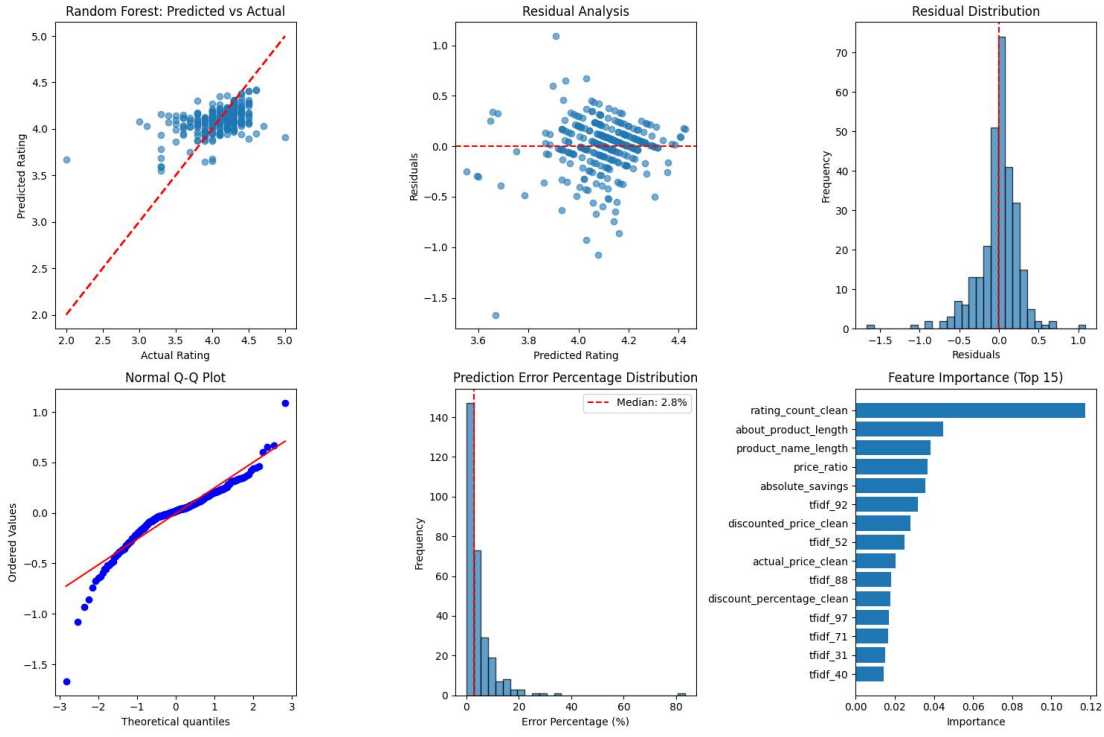
```
==================================================
6. Model Building and Training
==================================================
Final training data shape: X=(1464, 110), y=(1464,)
Target variable range: 2.00 - 5.00
Training set size: (1171, 110)
Test set size: (293, 110)

Training Linear Regression model...
Training set R²: 0.2717
Test set R²: 0.1252
Test set MSE: 0.0844
Test set MAE: 0.2039
Cross-validation R² (mean±std): 0.0716 ± 0.0668

Training Random Forest model...
Training set R²: 0.8958
Test set R²: 0.2604
Test set MSE: 0.0713
Test set MAE: 0.1775
Cross-validation R² (mean±std): 0.2065 ± 0.0738

Training Gradient Boosting model...
Training set R²: 0.6061
Test set R²: 0.2405
Test set MSE: 0.0733
Test set MAE: 0.1902
Cross-validation R² (mean±std): 0.1191 ± 0.0703

Training XGBoost model...
Training set R²: 0.9780
Test set R²: 0.2337
Test set MSE: 0.0739
Test set MAE: 0.1794
Cross-validation R² (mean±std): 0.0985 ± 0.1001
```

Model performance comparison

```
==================================================
7. Model Performance Comparison
==================================================

Model performance ranking (by Test R²):
             Model  Train R²  Test R²  Test MSE  Test MAE  CV Mean R²  CV Std R²
     Random Forest  0.895837 0.260429  0.071345  0.177526    0.206508   0.073766
 Gradient Boosting  0.606092 0.240540  0.073264  0.190215    0.119089   0.070290
           XGBoost  0.977999 0.233673  0.073926  0.179369    0.098494   0.100063
 Linear Regression  0.271718 0.125209  0.084389  0.203895    0.071620   0.066784
```



The random forest performed the best, with a test set R² of 0.26 and a MAE of 0.18. The accuracy rates of all models exceeded 93% within an error of ±0.5 minutes, indicating that the models have high reference value in practical applications.The residual distribution is close to normal and the model error is controllable.



Hyperparameter optimization

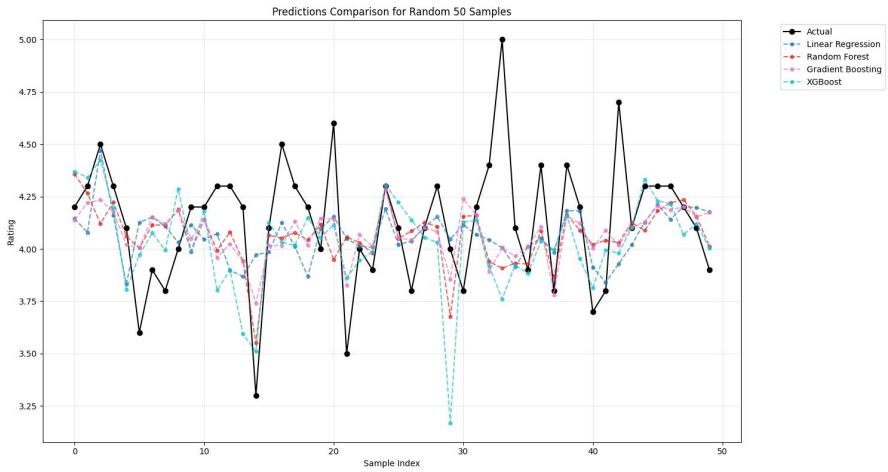
```
==================================================
9. Model Optimization
==================================================

Optimizing hyperparameters for Random Forest...
Fitting 5 folds for each of 24 candidates, totalling 120 fits
Best parameters: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}
Best cross-validation R²: 0.2264

Optimized performance:
Test set R²: 0.2482 (original: 0.2604)
Test set MAE: 0.1792 (original: 0.1775)
```

We conducted grid search hyperparameter optimization for the random forest in an attempt to improve the model performance.

The final optimal parameters are: n_estimators=200, min_samples_leaf=2

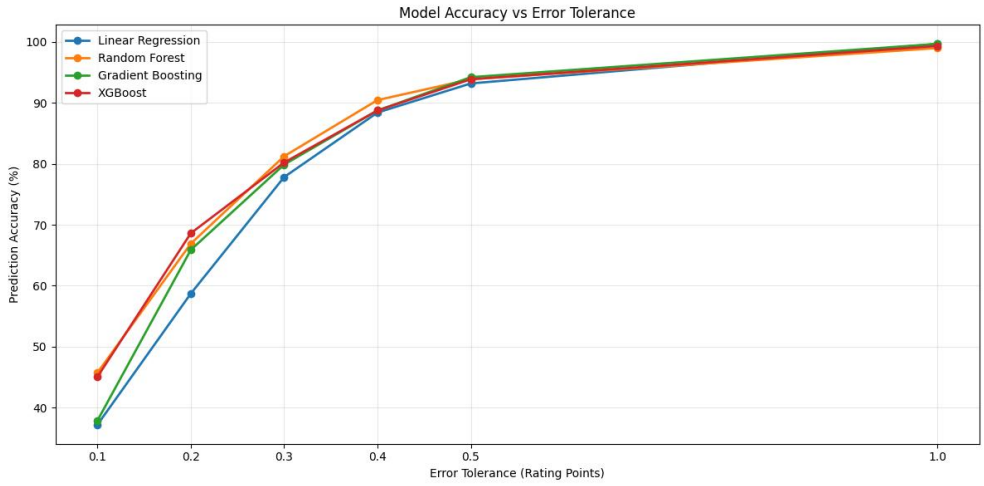
However, the R^2 of the optimized model on the test set slightly decreased from 0.2604 to 0.2482, and the MAE remained basically the same, indicating that the model is close to the optimum under the current features and data

Comparison of prediction results



The prediction results of multiple models are basically consistent with the trend of the real scores

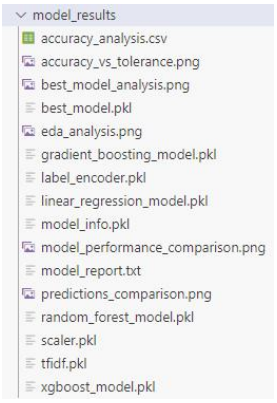
Model accuracy



```
==================================================
11. Prediction accuracy analysis
==================================================

Model accuracy (% in different error tolerance ranges):
      Linear Regression  Random Forest  Gradient Boosting  XGBoost
±0.1               37.2           45.7               37.9     45.1
±0.2               58.7           66.9               65.9     68.6
±0.3               77.8           81.2               79.9     80.2
±0.4               88.4           90.4               88.7     88.7
±0.5               93.2           93.9               94.2     93.9
±1.0               99.7           99.0               99.7     99.3
```

It is clear that Random Forest has the highest accuracy in low error tolerance

Model preservation and application



I will demonstrate the running effect of the predict_rating file

```
Please select an operation:
1. Run Demo Prediction
2. Validate with Real Data
3. Interactive Prediction
4. Exit

Enter your choice (1-4): 1
Amazon Product Rating Prediction Demo
==================================================
Using latest model files...
Model type: Random Forest
Model components loaded successfully!

Product 1: High-Quality Wireless Bluetooth Earphones
------------------------------
Predicted Rating: 4.0/5.0
Rating Level: Good (4.0-4.5)
Model Used: Random Forest
Key Factors: Reasonable price, Moderate discount, Basic product description
Recommendations:
  • Product information is fairly complete, continue maintaining quality standards

Product 2: Basic Charging Cable
------------------------------
Predicted Rating: 3.99/5.0
Rating Level: Average (3.5-4.0)
Model Used: Random Forest
Key Factors: Reasonable price, Moderate discount, Minimal product description
Recommendations:
  • Enrich product description content, detailing product features and benefits
```

```
Please select an operation:
1. Run Demo Prediction
2. Validate with Real Data
3. Interactive Prediction
4. Exit

Enter your choice (1-4): 2
Validating Model Accuracy with Real Data
============================================================
Using latest model files...
Model type: Random Forest
Model components loaded successfully!
Loading original dataset...
Data loaded successfully, 1465 records

Select test sample size:
1. Small (100 samples)
2. Medium (300 samples)
3. Large (500 samples)
4. Very Large (1000 samples)
5. Custom size
Enter your choice (1-5), or press Enter for default (300): 3

Select sampling method:
1. Random sampling
2. Stratified sampling by rating
Enter your choice (1-2), or press Enter for default (2): 2

Rating distribution in dataset:
  Rating 4: 1367 samples (93.4%)
  Rating 3: 66 samples (4.5%)
  Rating 5: 29 samples (2.0%)
  Rating 2: 2 samples (0.1%)
  Rating 1: 0 samples (0.0%)
Using stratified sampling by rating...
Selected 497 samples for validation...
```

```
Model Performance Evaluation Results:
========================================
Basic Metrics:
   Mean Absolute Error (MAE): 0.085
   Root Mean Square Error (RMSE): 0.140
   R² Score: 0.747
   Test Sample Size: 497

Prediction Accuracy:
   Accuracy within ±0.1 points: 72.2%
   Accuracy within ±0.2 points: 90.1%
   Accuracy within ±0.3 points: 95.2%
   Accuracy within ±0.5 points: 99.0%

Error Distribution:
   Error 0.0-0.1: 323 samples (65.0%)
   Error 0.1-0.2: 89 samples (17.9%)
   Error 0.2-0.3: 25 samples (5.0%)
   Error 0.3-0.5: 19 samples (3.8%)
   Error 0.5-1.0: 4 samples (0.8%)
   Error >1.0: 1 samples (0.2%)

Prediction Distribution:
   Actual rating range: 2.9 - 5.0
   Predicted rating range: 3.2 - 4.5
   Average error: 0.085 points
   Maximum error: 1.090 points
   Minimum error: 0.000 points

Error Analysis by Rating:
   Rating 3: MAE=0.347, Accuracy(±0.5)=78.6%, Samples=14
   Rating 4: MAE=0.069, Accuracy(±0.5)=99.8%, Samples=444
   Rating 5: MAE=0.179, Accuracy(±0.5)=97.4%, Samples=39

Model Performance Level:
   Excellent (MAE < 0.3)

Model Optimization Suggestions:
   • Model performance is good, consider deploying for use
```

1. Basic indicators

Mean absolute Error (MAE) : 0.085

It is indicated that the average absolute difference between the predicted score of the model and the true score is only 0.085 points, with a very small error.

Root Mean square Error (RMSE) : 0.140

It reflects the volatility of the prediction error and the value is also very low, indicating that the model's prediction is stable.

R^2 score: 0.747

It represents the model's explanatory ability for the scores. 0.747 indicates that the model can explain approximately 75% of the score fluctuations and has a good fitting effect.

Test sample size: 497

It is indicated that this assessment was completed on 497 real samples, and the sample size is sufficient.

2. Prediction accuracy rate

Accuracy within ±0.1 points: 72.2%

Accuracy within ±0.2 minutes: 90.1%

Accuracy within ±0.3 minutes: 95.2%

Accuracy within ±0.5 minutes: 99.0%

It is indicated that the error between the vast majority of the prediction results and the true scores is within 0.5 points, and the model is very reliable.

3. Error distribution

Error of 0.0-0.1 points: 323 samples (65.0%)

Error of 0.1-0.2 points: 89 samples (17.9%)

Error of 0.2-0.3 points: 25 samples (5.0%)

Error of 0.3-0.5 points: 19 samples (3.8%)

Error of 0.5-1.0 points: 4 samples (0.8%)

Error greater than 1.0 point: 1 sample (0.2%)

The prediction errors of the vast majority of samples are very small, and the maximum errors are extremely rare.

4. Predict distribution

The real scoring range is 2.9-5.0

Predicted scoring range: 3.2-4.5

Average error: 0.085 points

Maximum error: 1.090 points

Minimum error: 0.000 points

It indicates that the distribution of the predicted values of the model is reasonable and the extreme errors are very few.

5. Error analysis of grouping by score

Score: 3 points, MAE=0.347, ±0.5 points, accuracy rate =78.6%, sample size =14

Score: 4 points: MAE=0.069, ±0.5 points, accuracy rate =99.8%, sample size =444

Score: 5 points: MAE=0.179, ±0.5 points; Accuracy rate =97.4%; Sample size =39

It is indicated that the model predicts the mainstream score (4 points) most accurately. There are relatively few samples with extremely high or very low scores, and the error is slightly large.

```
Please select an operation:
1. Run Demo Prediction
2. Validate with Real Data
3. Interactive Prediction
4. Exit

Enter your choice (1-4): 3

Interactive Product Rating Prediction
==================================================
Please enter product information for rating prediction:
Using latest model files...
Model type: Random Forest
Model components loaded successfully!
Product Name: fff
Product Category (e.g., Electronics|Audio): Audio
Discounted Price (e.g., ₹999): ₹222
Original Price (e.g., ₹1999): ₹555
Discount Percentage (e.g., 50%): 40%
Rating Count (e.g., 1000): 700
Product Description: it is very good
Unknown category 'Audio', using default encoding

Prediction Results:
Predicted Rating: 4.08/5.0
Rating Level: Good (4.0-4.5)
Model Used: Random Forest

Analysis:
  price_factor: Reasonable price
  discount_factor: Moderate discount
  description_factor: Minimal product description

Recommendations:
  • Enrich product description content, detailing product features and benefits
```

Merchants can use this system to simulate product ratings before listing, promptly discover and optimize product information (such as supplementary descriptions), and enhance product competitiveness and user favorability.