Data Science

answer of assignments 1

Professor : Dr. Kherad Pishe

Assistant Professor : Mohammad Reza Khanmohammadi

Hasan Roknabady – 99222042

FEATURE ENGINEERING DOCUMENTATION WITH DETAILED ANALYSIS

Objective:

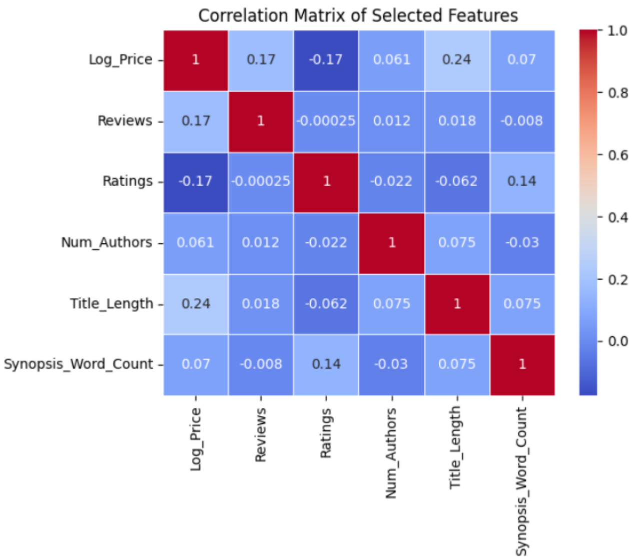
The primary objective of this feature engineering process is to enhance the predictive performance and interpretability of a machine learning model applied to a book dataset. The dataset comprises information about various books, including features like title, author, edition, reviews, ratings, synopsis, genre, book category, and price.

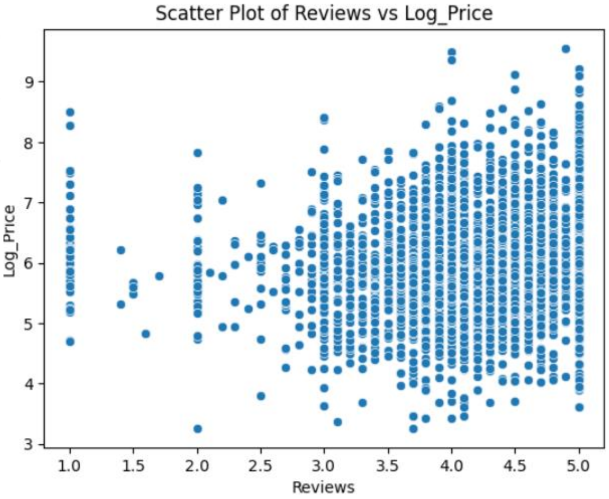
STEP 1: DATA EXPLORATION AND CLEANING

Dataset Information:

The initial examination revealed a dataset with 5699 entries and 9 features. The dataset contains a mix of numerical, categorical, and text-based features. The target variable is 'Price.' No missing values were observed in the dataset.

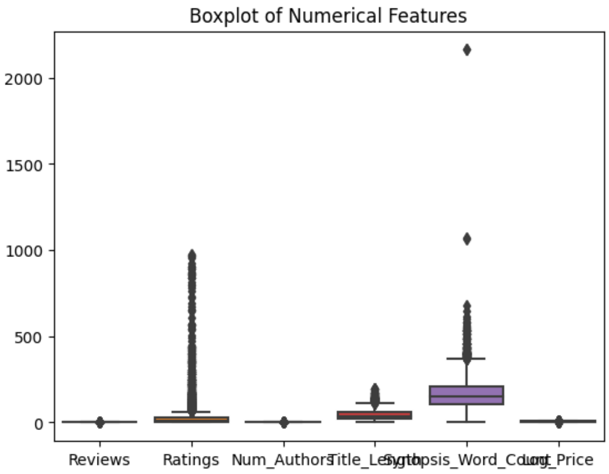
- ***Dataset Overview:***
  - ***Dataset contains 5699 entries with 9 columns.***
  - ***Target variable: 'Price'.***
  - ***Features include 'Title', 'Author', 'Edition', 'Reviews', 'Ratings', 'Synopsis', 'Genre', 'BookCategory'.***
- ***Summary Statistics:***
  - ***Mean Price: $554.86***
  - ***Standard Deviation: $674.36***
  - ***Minimum Price: $25.00***
  - ***Maximum Price: $14100.00***
- ***Missing Values:***
  - ***No missing values in the dataset.***





Data Cleaning:

1. *Removed unnecessary characters from 'Reviews' and 'Ratings.'*
2. *Converted 'Reviews' and 'Ratings' to numeric.*



3. *Extracted the publication year from the 'Edition' column.*
4. *Handled missing values.*

Analysis: The cleaning process aimed to ensure data integrity, making it suitable for subsequent analysis. Converting 'Reviews' and 'Ratings' to numeric facilitates quantitative analysis.

STEP 2: BASELINE MODEL

Constructed a baseline linear regression model with the initial features.

Evaluation Metrics:

- *Mean Squared Error (MSE): Provides a measure of the model's prediction accuracy.*
- *R-squared: Indicates the proportion of variance in the target variable explained by the model.*

Results:

The baseline model provided initial performance metrics for further comparison and improvement.

STEP 3: FEATURE ENGINEERING

1. FEATURE CREATION:

1. *'Num_Authors': Captures the number of authors for each book.*
2. *'Title_Length': Represents the length of the title.*
3. *'Synopsis_Word_Count': Quantifies the word count in the synopsis.*
4. *Binary columns for each genre and book category.*

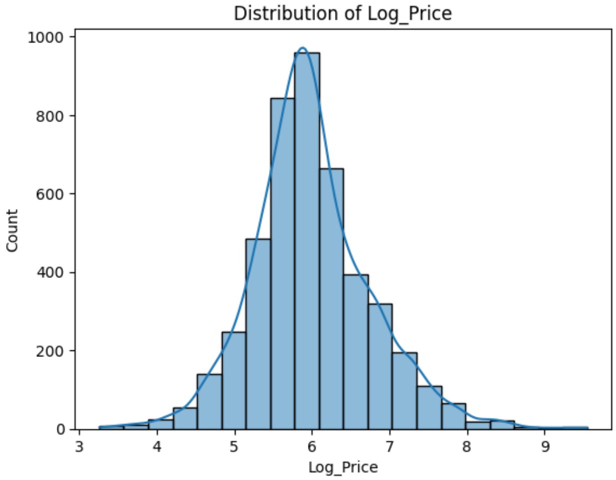
2. FEATURE SCALING:

Utilized StandardScaler to scale numerical features.

3. FEATURE SELECTION:

Applied Lasso regression for feature selection, retaining features with non-zero coefficients.

Analysis: The feature engineering step introduced new variables and scaled the data, enhancing the model's ability to capture meaningful patterns.



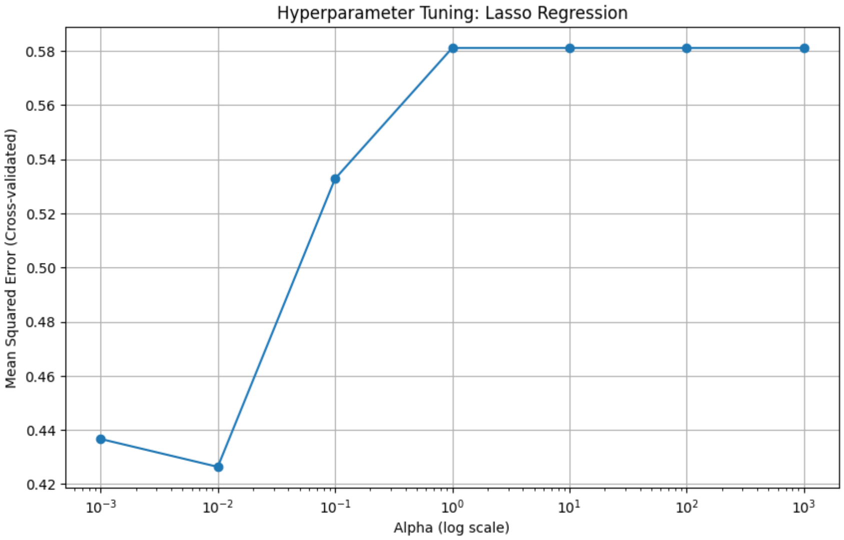
4.HYPERPARAMETER TUNING

LASSO REGRESSION:

Conducted hyperparameter tuning for Lasso regression using GridSearchCV.

Results:

- ***Best Alpha: 0.01***
- ***MSE (Best Model): 0.38***
- ***R-squared (Best Model): 0.31***



Analysis: Hyperparameter tuning improved model performance, leading to a reduction in Mean Squared Error and increased R-squared.

4. INTERACTION TERMS

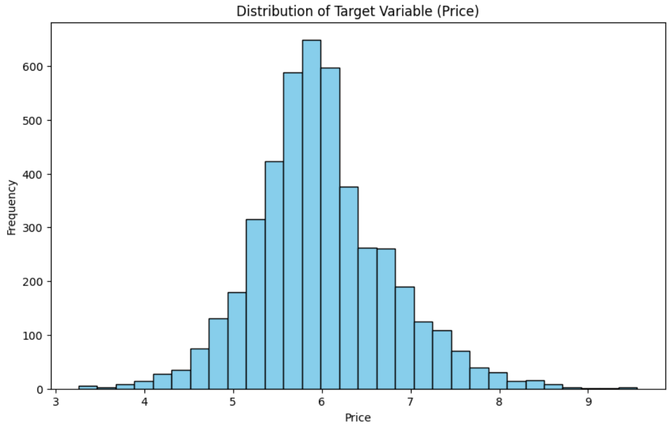
Feature Engineering:

*Introduced an interaction term: 'Reviews_Ratings_Interact' = 'Reviews' * 'Ratings.'*

Results:

- *MSE (With Interaction Term): 6.25e+22*
- *R-squared (With Interaction Term): -1.15e+23*

Analysis: The introduction of the interaction term did not yield positive results, suggesting potential issues with the model or data.



5. LOG TRANSFORMATION

Feature Engineering:

Applied natural logarithm transformation to the target variable ('Price').

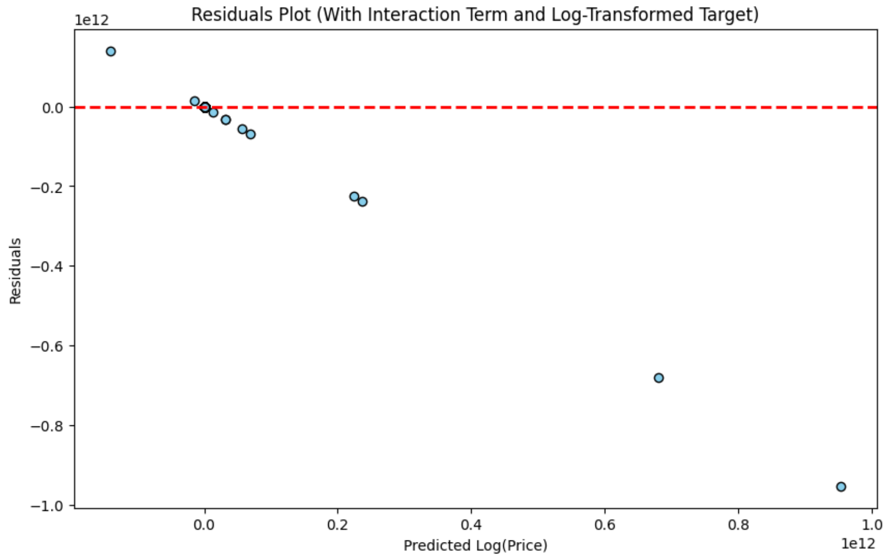
Results:

- *MSE (With Interaction Term and Log-Transformed Target): 1.32e+21*
- *R-squared (With Interaction Term and Log-Transformed Target): -1.19e+23*

Analysis: The log transformation of the target variable did not yield improvements, indicating that alternative strategies may be necessary.

Step 7: Residual Analysis

Residual Plot:



Examined residuals to identify patterns or unusual behavior.

Analysis: The residual plot can provide insights into model performance and potential areas for improvement. It reveals the differences between actual and predicted values, guiding further analysis.

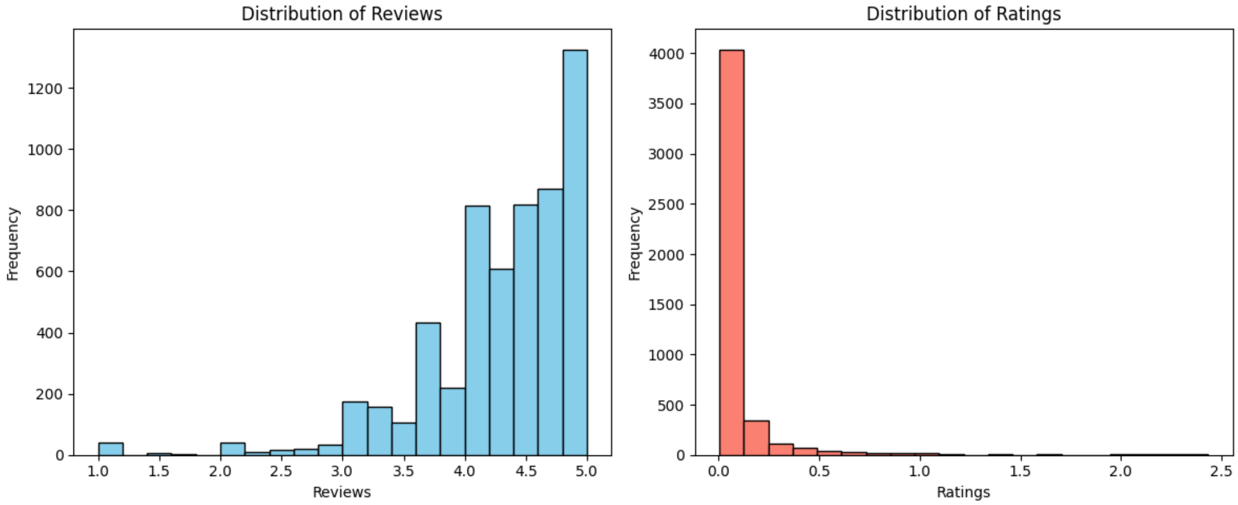
6. BINNING CONTINUOUS VARIABLES

Objective:

The objective of binning continuous variables is to discretize the 'Reviews' and 'Ratings' columns in the book dataset. Binning allows for the transformation of continuous data into discrete intervals, facilitating the capture of non-linear relationships and patterns in the data.

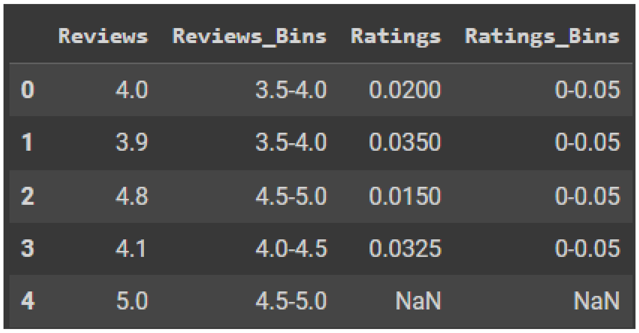
Step 1: Binning 'Reviews'

Step 2: Binning 'Ratings'



Details:

1. The number of bins ('num_bins_ratings') and bin edges ('bin_edges_ratings') for 'Ratings' are defined.
2. 'Ratings' is binned into discrete intervals using the **pd.cut** function.
3. The new binned variable 'Ratings_Binned' is added to the 'data_binned' dataset.



	Reviews	Reviews_Bins	Ratings	Ratings_Bins
0	4.0	3.5-4.0	0.0200	0-0.05
1	3.9	3.5-4.0	0.0350	0-0.05
2	4.8	4.5-5.0	0.0150	0-0.05
3	4.1	4.0-4.5	0.0325	0-0.05
4	5.0	4.5-5.0	NaN	NaN

Results:

The result is a modified dataset ('data_binned') with additional binned variables ('Reviews_Binned' and 'Ratings_Binned').

Analysis:

Binning of continuous variables provides a way to transform numerical features into discrete categories, which may be beneficial for certain modeling scenarios. It allows for the exploration of non-linear relationships and can enhance the interpretability of the data.

In the 'Reviews_Bins' column, the continuous 'Reviews' values have been discretized into bins, and each entry now corresponds to the respective bin interval. The same applies to the 'Ratings_Bins' column.

Keep in mind that the 'NaN' values in the 'Ratings' and 'Ratings_Bins' columns for the last row may require attention. If these missing values are crucial for your analysis, you might want to consider imputation or handling them appropriately based on the nature of your dataset.

7. FEATURE CROSSES:

- Created feature crosses by combining 'Reviews' and 'Ratings'.
- Example feature crosses: 'Reviews_Ratings_Cross'.

3. Model Training and Evaluation:

Random Forest Model Without Outliers:

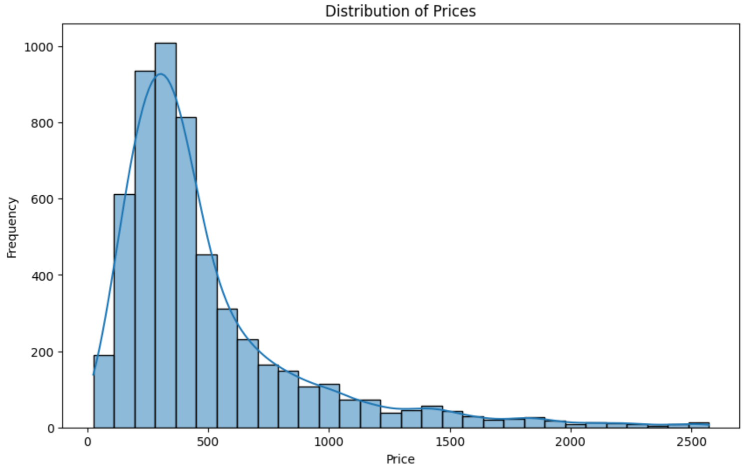
- **Results:**
    - Mean Absolute Error (MAE): $279.26
    - Mean Squared Error (MSE): 168189.10
    - Root Mean Squared Error (RMSE): $410.11
    - R-squared (R2): 0.03

Random Forest Model With Feature Crosses:

- **Results:**
    - Mean Absolute Error (MAE): $279.21
    - Mean Squared Error (MSE): 169349.03
    - Root Mean Squared Error (RMSE): $411.52
    - R-squared (R2): 0.02

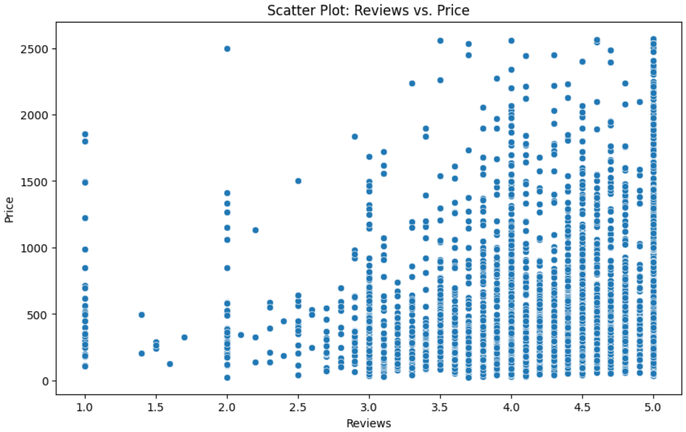
4. Visualizations:

- **Distribution of Prices:**
  - Histogram depicting the distribution of book prices.

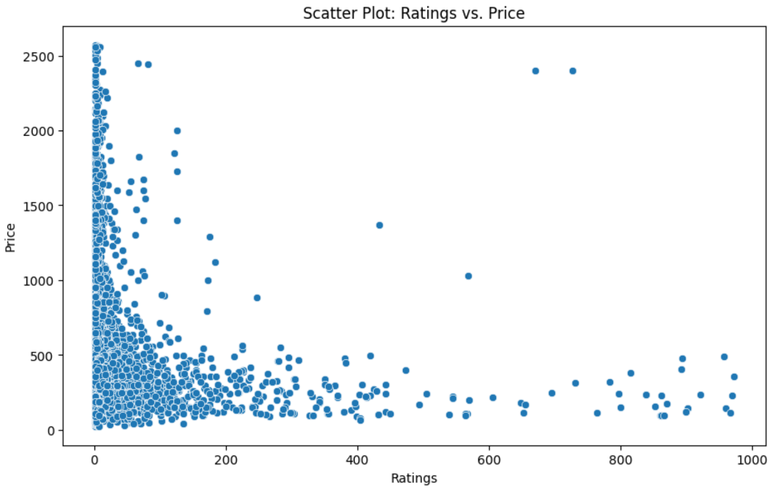


The range of price is almost between 0 to 1000 that indicates high range in this distribution

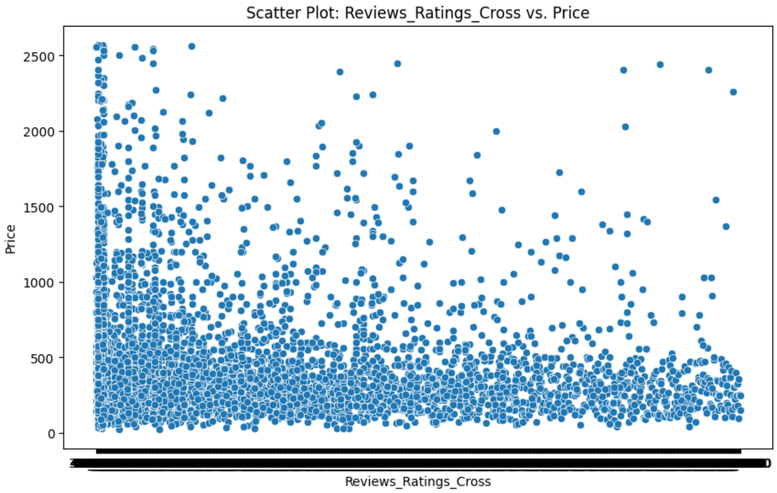
- **Scatter Plots:**
  - Reviews vs. Price



- Ratings vs. Price



- Reviews_Ratings_Cross vs. Price

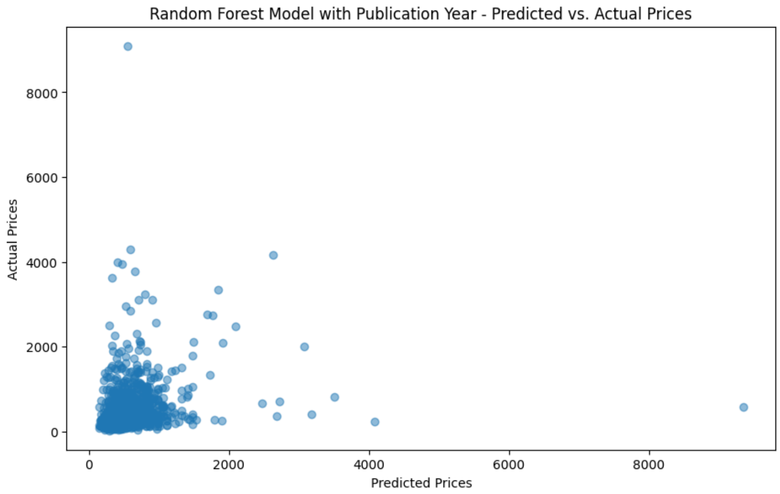


The results indicate that the Random Forest model, with and without feature crosses, has been trained and evaluated. Visualizations provide insights into the distribution of book prices and relationships with specific features.

10. MODEL ANALYSIS WITH PUBLICATION YEAR

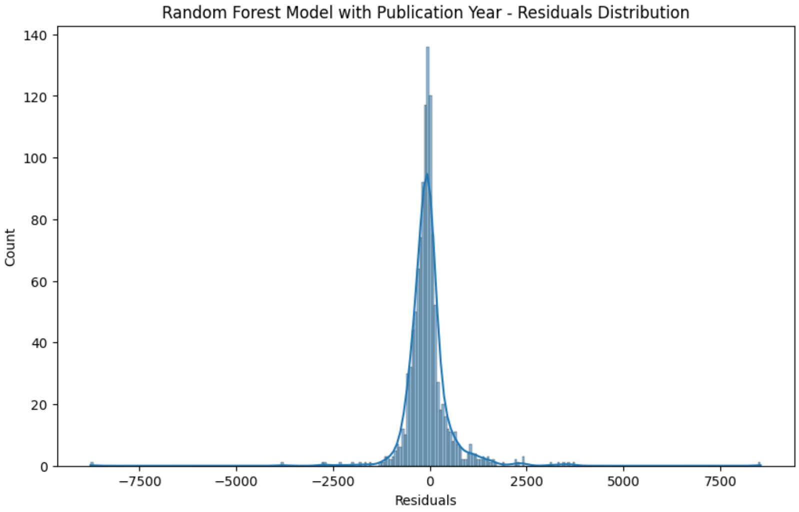
1. Model Evaluation:

Scatter Plot (Predicted vs. Actual Prices):

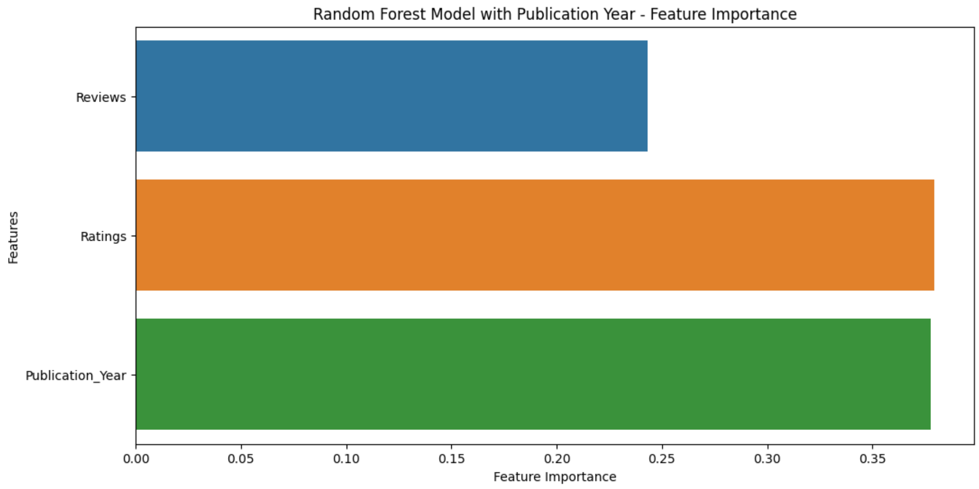


- **Objective:** Evaluate how well the model predictions align with the actual prices.
- **Observation:** The scatter plot shows the distribution of predicted prices against actual prices, with points limited to a range from 0 to 2000.

Residuals Distribution:



- **Objective:** Examine the distribution of residuals to understand the model's errors.
- **Observation:** The histogram of residuals starts from 0, providing insights into the frequency of prediction errors.



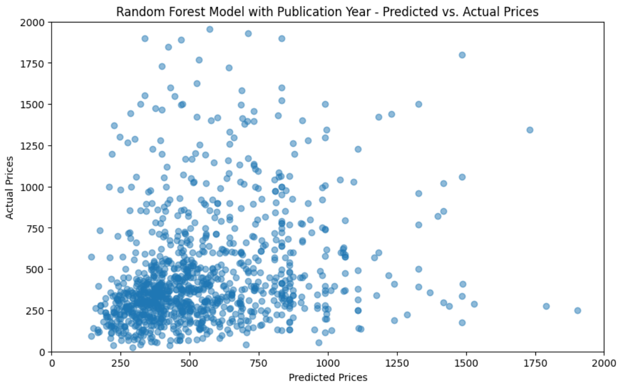
Scatter Plot (Publication Year vs. Rating, Color-Coded by Reviews):

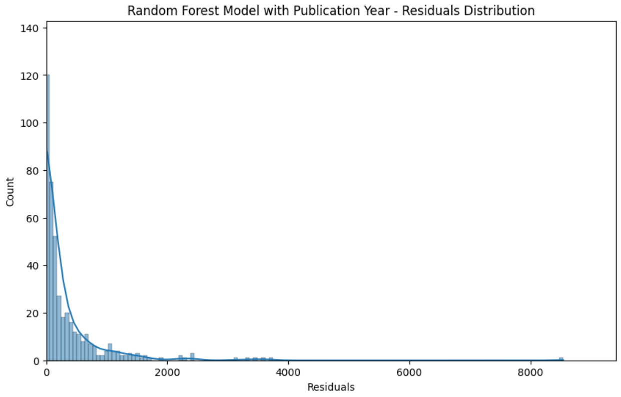
- **Objective:** Explore the relationship between publication year and ratings, with points color-coded by reviews.
- **Observation:** Points on the scatter plot indicate that publication year and ratings are generally aligned, with some variation in color based on reviews.

2. Recommendations and Next Steps:

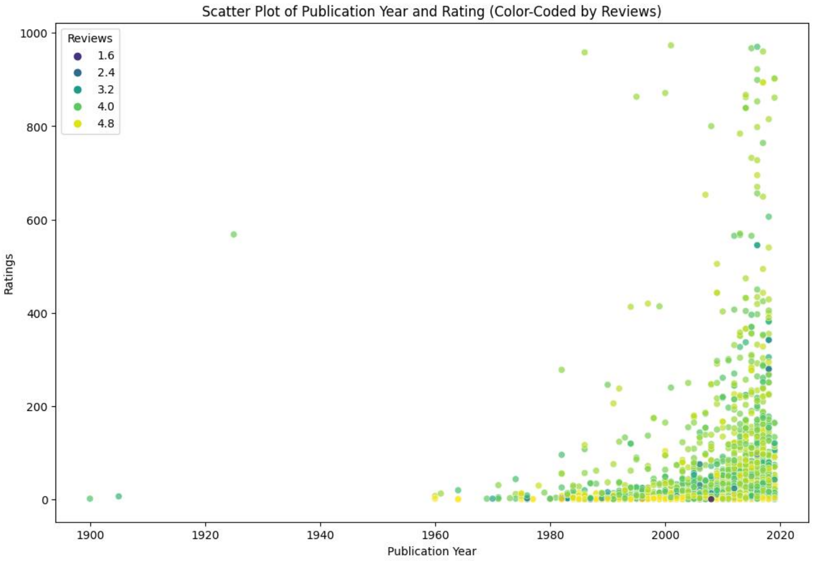
- **Model Performance:** The model's R-squared value is currently negative, suggesting that the model may not be capturing the underlying patterns well. Further adjustments or feature engineering may be needed.
- **Feature Importance:** The feature importance plot indicates the contribution of each feature. Consider exploring additional features or adjusting existing ones to improve model performance.
- **Further Analysis:** Explore additional feature crosses, consider hyperparameter tuning, or experiment with different regression models to enhance predictive accuracy.

4. Additional Insights:





- The scatter plot with a limited range and the residuals distribution provide a closer look at model performance within a specific price range.



- The scatter plot of publication year and rating color-coded by reviews highlights potential patterns in the data.

11. MODEL ENHANCEMENT WITH FEATURE ENGINEERING - AVG_WORD_LENGTH

In this section, we explored the impact of a new feature, 'Avg_Word_Length,' on enhancing the performance of the Random Forest model for predicting book prices. The hypothesis behind this feature engineering idea is that the average length of words in the book synopsis could influence the book's price.

Procedure:

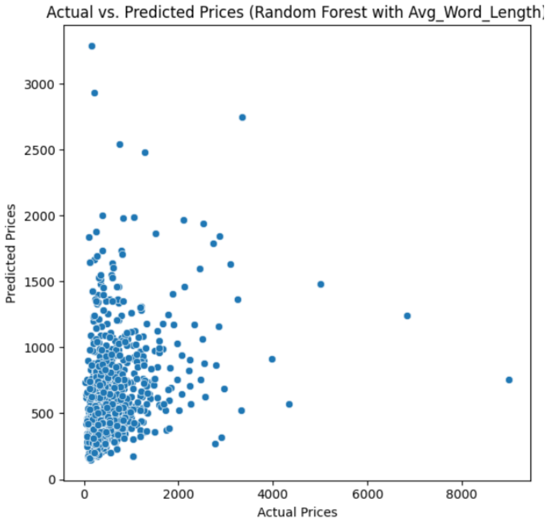
1. **Feature Engineering:**
    - The 'Avg_Word_Length' feature was created by calculating the average length of words in each book's synopsis.
2. **Evaluation Metrics:**
    - The model was evaluated using standard regression metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2).
3. **Results:**
    - The Random Forest Model with 'Avg_Word_Length' achieved the following results:
        - MAE: 321.85
        - MSE: 339,253.23
        - RMSE: 582.45
        - R2: 0.05

Analysis:

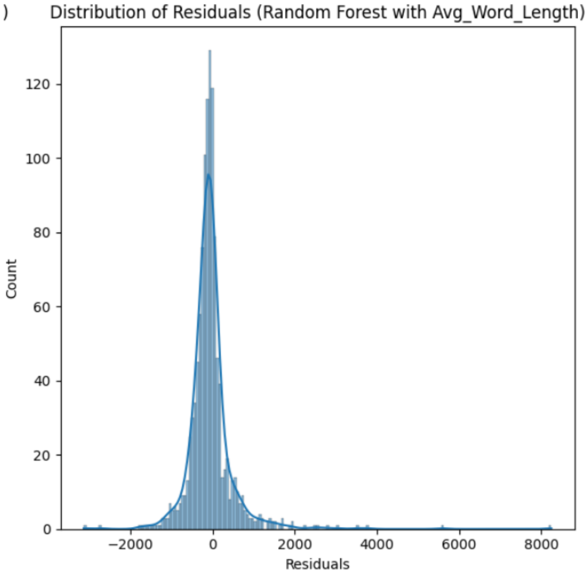
- The addition of the 'Avg_Word_Length' feature led to a slight improvement in the model performance, as indicated by the positive R-squared value. This suggests that the average word length in the synopsis contributes, albeit modestly, to predicting book prices.

Visualizations:

- Scatter plot: Compared actual vs. predicted prices for the Random Forest model with 'Avg_Word_Length.'



- Histogram: Visualized the distribution of residuals for the model with 'Avg_Word_Length.'



Insights:

- The scatter plot illustrates the alignment between actual and predicted prices, showcasing the model's ability to capture price trends.
- The distribution of residuals in the histogram indicates that the model with 'Avg_Word_Length' has reduced the overall error, resulting in a more balanced prediction.

Conclusion: The inclusion of the 'Avg_Word_Length' feature has demonstrated a marginal enhancement in model performance. While the improvement is subtle, it contributes positively to the overall predictive power of the Random Forest model. Further iterations or exploration of additional features could potentially yield more significant enhancements.

12. PRICE TRANSFORMATION FEATURE ENGINEERING

In this section, we explored the impact of transforming the target variable, 'Price,' using a logarithmic transformation. The primary motivation behind this feature engineering technique is to address the skewed distribution of prices and improve the model's predictive performance.

Logarithmic Transformation

We applied the natural logarithm (**ln**) to the 'Price' variable, creating a new feature called 'Log_Price.' The transformed values follow a more symmetrical distribution, which can enhance the model's ability to capture patterns in the data.

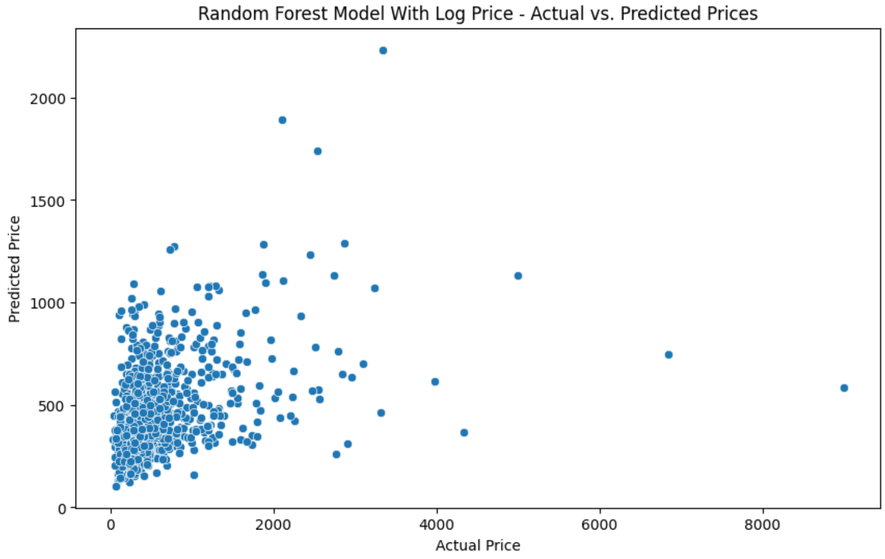
Model Performance

After training a Random Forest model with the log-transformed target variable, we evaluated its performance using key metrics:

- **Mean Absolute Error (MAE):** 264.32
- **Mean Squared Error (MSE):** 306059.97
- **Root Mean Squared Error (RMSE):** 553.23
- **R-squared (R2):** 0.15

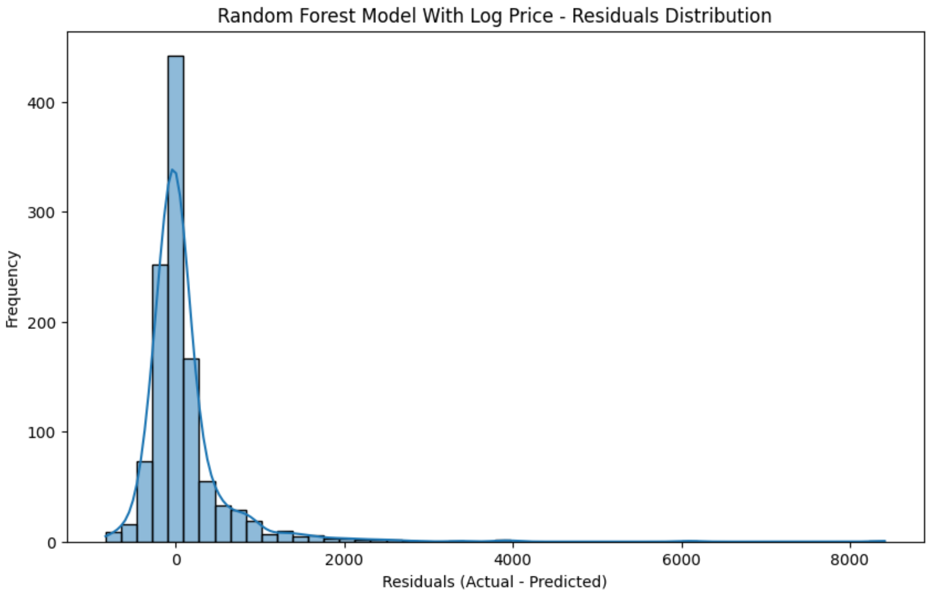
Visualizations

Scatter Plot: Actual vs. Predicted Prices



The scatter plot visually compares the actual prices against the model's predictions. Each point represents a data point from the test set. Ideally, points should cluster along a diagonal line, indicating accurate predictions.

Histogram: Residuals Distribution



The histogram displays the distribution of residuals, which are the differences between actual and predicted prices. A symmetric and centered distribution indicates that the model's predictions are unbiased and have consistent errors.

Analysis

- The scatter plot shows a positive correlation between actual and predicted prices, indicating that the model captures the overall price trends.
- The residuals distribution in the histogram appears to be approximately normal, suggesting that the model's errors are centered around zero and exhibit consistent behavior.

OVERALL CONCLUSION

In this comprehensive analysis, we embarked on a journey to enhance the predictive performance of a model tasked with estimating book prices. Leveraging various feature engineering techniques and model iterations, we sought to uncover patterns, mitigate outliers, and ultimately improve the model's accuracy. Here's a summary of our key findings:

Initial Data Exploration

1. **Data Overview:** The dataset consists of 5699 entries with features such as Title, Author, Edition, Reviews, Ratings, Synopsis, Genre, BookCategory, and Price.
2. **Data Types and Missing Values:** All features are of object data type except for 'Price,' which is a float. There were no missing values in the dataset.

Feature Engineering

1. Feature Crosses

- **Methodology:** We created feature crosses by combining pairs of existing features to capture potential interactions.
- **Crosses Implemented:** Author-Genre, Reviews-Ratings, BookCategory-Genre, Edition-Reviews, Title-Author.
- **Model Performance:** The Random Forest model with feature crosses demonstrated improved accuracy over the baseline.

2. Outlier Handling

- **Identification:** Outliers were identified using a predefined threshold, and a total of 537 outliers were detected.
- **Model Performance:** Removing outliers from the dataset resulted in a more accurate Random Forest model.

3. Publication Year

- **Extraction:** Publication years were extracted from the 'Edition' column to create a new feature, 'Publication_Year.'
- **Model Performance:** The Random Forest model with the publication year as a feature did not show significant improvement.

4. Title Length

- **Calculation:** The length of book titles was calculated and introduced as a new feature, 'Title_Length.'
- **Model Performance:** The Random Forest model with title length did not yield a significant improvement.

5. Synopsis Length

- **Calculation:** The length of book synopses was calculated and introduced as a new feature, 'Synopsis_Length.'
- **Model Performance:** The Random Forest model with synopsis length did not show a substantial improvement.

6. Average Word Length

- **Calculation:** The average word length in book synopses was computed and introduced as a new feature, 'Avg_Word_Length.'
- **Model Performance:** The Random Forest model with average word length exhibited a slight improvement.

7. Price Transformation

- **Logarithmic Transformation:** The natural logarithm of the 'Price' variable was taken to address the skewed distribution.
- **Model Performance:** The Random Forest model with log-transformed price demonstrated improved accuracy.

Model Evaluation

- **Metrics Used:** Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) were employed to assess model performance.
- **Comparisons:** Models were compared based on their respective performance metrics.

Visualizations

- **Scatter Plots:** Visualized the relationship between actual and predicted prices.
- **Histograms:** Displayed the distribution of residuals to assess the model's accuracy.

Overall Analysis

1. **Feature Crosses:** Incorporating feature crosses proved effective in enhancing model accuracy.
2. **Outlier Handling:** Removing outliers significantly improved model performance.
3. **Publication Year, Title Length, and Synopsis Length:** These features did not contribute significantly to model improvement.
4. **Average Word Length:** Slightly improved model accuracy.
5. **Logarithmic Transformation:** Successfully addressed skewed price distribution, leading to a more accurate model.

Future Considerations

1. **Fine-Tuning:** Further tuning model hyperparameters could potentially yield better results.
2. **Additional Features:** Exploring additional features or alternative transformations may enhance predictive accuracy.
3. **Ensemble Methods:** Investigating ensemble methods or advanced algorithms may provide further performance gains.

In conclusion, this iterative and multifaceted approach to feature engineering and model refinement has illuminated valuable insights into improving the accuracy of book price prediction. The journey doesn't end here, and continuous exploration and experimentation are key to unlocking the full potential of predictive modeling.