Case Study: Predicting the Success of Bestselling Books

Objective

The objective of this case study is to develop a regression analysis model to predict the success of bestselling books based on identified features. Success is defined by user ratings, and the features considered include the number of reviews, price, year of publication, and genre.

Methodology

1. **Data Collection**
    - The dataset consists of bestselling books with the following features:
        - **Name**: Title of the book
        - **Author**: Author of the book
        - **User Rating**: Average user rating of the book
        - **Reviews**: Number of reviews the book has received
        - **Price**: Price of the book
        - **Year**: Year the book was published
        - **Genre**: Genre of the book (Fiction or Non Fiction)
2. **Data Preprocessing**
    - Dropped irrelevant columns ('Name' and 'Author') as they are not numerical and challenging to encode meaningfully for regression.
    - Encoded the categorical variable 'Genre' into numerical values (0 for Fiction, 1 for Non-Fiction).
    - Standardized numerical features to ensure they are on a similar scale(not useful in this case).
3. **Feature Selection**
    - Selected the following features for the regression model:
        - **Reviews**: Number of reviews
        - **Price**: Price of the book
        - **Year**: Year of publication
        - **Genre**: Encoded genre (1 for Fiction, 0 for Non-Fiction)
4. **Model Development**
    - Split the data into training and testing sets (80% training, 20% testing).
    - Applied a linear regression model to predict user ratings based on the selected features.
    - Evaluated the model using Mean Squared Error (MSE) and R-squared (R2R^2R2) metrics.
5. **Model Evaluation**
    - **Mean Squared Error (MSE)**: 0.053
    - **R-squared (R2R^2R2)**: 0.0815
    - The low R^2 value indicates that the model explains only 8.15% of the variance in user ratings, suggesting that other factors not included in the model may significantly influence user ratings.

6. **Coefficient Analysis**
    - **Reviews**: Coefficient of -0.0229 indicates that more reviews are associated with slightly lower user ratings, possibly due to more reviews capturing a wider range of opinions.
    - **Price**: Coefficient of -0.0173 indicates that higher prices are slightly associated with lower user ratings.
    - **Year**: Coefficient of 0.0596 suggests that more recent books tend to receive higher user ratings.
    - **Genre (Non Fiction)**: Coefficient of -0.0333 indicates that non-fiction books tend to receive slightly lower user ratings compared to fiction books.

Findings

1. **Impact of Reviews**
    - While the number of reviews is an important factor, the slight negative coefficient suggests that a higher number of reviews can be associated with a more diverse range of opinions, leading to a marginally lower average rating.
2. **Impact of Price**
    - Higher prices are associated with slightly lower user ratings. This could be due to higher expectations from readers when paying more, which might lead to harsher reviews if expectations are not met.
3. **Impact of Publication Year**
    - More recent books tend to receive higher ratings. This might be due to contemporary relevance, improved writing quality, or marketing strategies that resonate well with current audiences.
4. **Impact of Genre**
    - Fiction books generally receive higher user ratings compared to non-fiction books. This could be due to the escapism and entertainment value provided by fiction.

Conclusion

The regression analysis model provides some insights into the factors contributing to the success of bestselling books. However, the low R^2 value suggests that there are other significant factors influencing user ratings that were not captured in this model. Future work should consider incorporating additional features such as book length, author popularity, book cover design, marketing efforts, and social media presence to improve the model's predictive power.