Goodreads Average Rating Prediction Using a Model

Enoch Foli-Degbesse [Data Analytics A21]

Contents

1. Introduction

Nowadays, a massive number of reviews is available online. Besides offering a valuable source of information, these informational contents generated by users, also called User Generated Contents (UGC) strongly impact the purchase decision of customers. As a matter of fact, a recent survey (Hinckley, 2015) revealed that 67.7% of consumers are effectively influenced by online reviews when making their purchase decisions. More precisely, 54.7% recognized that these reviews were either fairly, very, or important in their purchase decision making. Relying on online reviews has thus become a second nature for consumers. Therefore, models able to predict the user rating are critically important companies to stay in business, to deliver quality service to your clients.

In this project, two (2) models were used (Linear Regression and Random Forest Regressor) which will be trained on the GoodReads datasets. The performance of the model will be tested and eventually the best model will be selected thanks to accuracy metrics,

2. Objective

The objective of using **Goodreads book** dataset in predicting book's rating therefore is to achieve the following:

- Data analysis (data processing, data cleaning, exploratory analysis, plots of relevant attributes)
- Feature selection (feature engineering, feature pruning, choice justification)
- Model training (motivation for selected model, comparison of different models)
- Model evaluation (evaluation metric, results interpretation)
- Project hosting on GitHub: **https://github.com/enochdefoli/Book-Rating-Prediction-Model**

3. Scope

The scope of the project is to develop a Jupyter Notebook detailing all the steps taken and as well as the python code written to achieve the listed objectives.

4. Work Approach

The project implementation adopted the following work approach to ensure that the scope of the project is successfully implemented:

4.2. Data Cleaning

This step, also known as pre-processing, is aimed at cleaning the data. Indeed, text data such as the "title" and "author" columns contain unnecessary and redundant characters and must be normalized. More precisely, the objective of the data cleaning consists in:

I. Removing special characters
II. Formatting the text
III. Removing extra whitespace
IV. Removing non-ascii words
V. Check for invalid values in the dataset such as null values: - records for language_code and text_review_count dropped.
VI. Merged languages with similar codes (e.g en-US, en-GB, en-CA, etc.)

4.3. Feature Selection

At this stage we are going to identify which columns are needed in the dataset, to take the unwanted ones, and if possible, transform the possible ones:

I. 'Publication_date' transformed into year and month, and dropped afterwards
II. Picking the first names whenever there are multiple 'authors' delimited by "/ (" , and the same approach applied to 'title'
III. Drop the column 'isbn' and 'isbn13' since it has no relevant impact on the average rating.
IV. Encode the 4 top 'languages_code' (i.e. eng, fre, ger, spa)
V. Numerical representation of the' title' and 'author' attribute to identify distinct values.

4.4. Plot of Relevant Attributes

I. Graph to have a clear picture of the distribution of the attribute 'average rating'
II. To know the distribution of the 4 topmost languages in percentage.
III. Identify which Books have the Highest rated ratings.
IV. Identify authors who had the highest rating, and on which books.
V. Relationship between Rating and number of pages with respect to the language code.

4.5. Training Model

In the training phase we classifier the cleaned dataset into training and test. The training data is then given to the model to learn from, to be later, able to accurately predict test data. To implement the training step, the Python programming language and its Scikit-learn libraries were used.

Note that the different model (Linear Regression Model and Random Forest Regressor) has been trained with both on the same transformed dataset.

The choice of these models is because we want to determine the relationships between the Target and the Features, and secondly the Target is a continuous variable.

4.6. Evaluation

This step enables to measure the performance and test the effectiveness of the trained model. In other words, we can see whether the models learned some general principles and is able to predict an accurate outcome on new unseen instances.

To perform the evaluation of the different models, the following metrics were used:

a) R-squared (R^2)
b) Mean Squared Error (MSE)
c) Root Mean Squared Error (RMSE)

4.7. Result

The result of both models on the same dataset are presented below.

Linear Regression Model

R-squared (R^2) is 0.61
Mean Squared Error (MSE) is 0.056
Root Mean Squared Error (RMSE) is 0.23

Random Forest Regressor

R-squared (R^2) is 0.99
Mean Squared Error (MSE) is 0.0007
Root Mean Squared Error (RMSE) is 0.026

5. Challenges

However, one challenge that cannot be overlooked is the issue of class imbalance. The datasets are relatively skewed in terms of class distribution. This issue is particularly acute for the case of target variable. For instance, there are significantly more reviews with 3, 4 and 5-star ratings than there are reviews with 1 and 2 stars.

6. Conclusion

The lower value of root mean squared error (RMSE) closer to 0 preferred the better model performance, conversely, the higher value of R-square(R2) closer to 1 show that the regression line fits the data well and the model performance is better. Meaning a strong relationship between the dependent and independent variable.

Also, mean square error (MSE) is the average of the square of the errors. The larger the number the larger the error which means there is bigger variation between the test result and the predicted result. Therefore, the Random Forest Regressor model is far better than Liner Regression model.