Enhancing Product Rating Predictions on

NAME: HARI BAHADUR GHARTI MAGAR

STUDENT ID: 22075765

GOOGLE COLLAB LINK: https://colab.research.google.com/drive/1yB2Vb0olwTvfTVOGehipLv-khAr55Ue1?usp=sharing

ABSTRACT

- This case study explores how machine learning (ML) techniques can be applied to predict product ratings from customers on an e-commerce platform like amazon
- The dataset includes a number of variables, including rating count, discount percentage, discounted price, and actual price.
- Preprocessing involves renaming columns, filling in missing values, and standardizing numerical properties.
- Two different machine learning techniques are applied:
  Long Short-Term Memory (LSTM) networks for regression tasks to directly anticipate product ratings, and logistic regression for binary classification to predict whether a product earns a high rating (4.5 or above).
- Metrics like mean squared error (MSE) and confusion matrices are used to evaluate the performance of the models.

INTRODUCTION

- In today's e-commerce world, customer ratings play a vital role in shaping buying choices by signaling product satisfaction and quality. Businesses striving to refine their strategies understand the significance of accurately forecasting these ratings.
- Our study employs machine learning techniques to build models predicting product ratings using diverse attributes from pricing data to user reviews .
- We investigate two main methods: logistic regression for straightforward rating categorization and Long Short-Term Memory (LSTM) networks for detailed rating predictions due to their ability to capture complex temporal relationships.
- Our aim is to provide businesses with practical insights from predictive analytics, enhancing customer engagement and decision-making. By applying advanced machine learning algorithms, we aim to contribute to e-commerce analytics, supporting sustainable growth in the digital marketplace.

DATASET OVERVIEW

Our dataset, "amazon.csv," is extracted from
from Kaggle, is vital for our e-commerce predictive analytics study.
Link : dataset link

It contains product details like ID, name, category, prices, and reviews from Amazon.
We use this data to uncover patterns and insights for decision-making and strategy.

Before analysis, we carefully preprocess the dataset by handling missing values, renaming columns, and standardizing numerical features. This ensures data quality for training machine learning models.

I aim to provide actionable insights for businesses to improve customer satisfaction, optimize products, and succeed in the dynamic e-commerce landscape.

PROBLEM STATEMENT

- How can machine learning techniques be leveraged to predict customer ratings for products on an e-commerce platform, specifically Amazon?
- My main goal is to create models that predict customer ratings accurately using logistic regression and Long Short-Term Memory (LSTM) networks for which both contributed a lot to this case study.
- Logistic regression aids in categorizing products as high or low rated, providing clear insights into influential attributes through interpretable coefficients, and its performance is evaluated using metrics such as accuracy and precision.
- LSTM networks excel in capturing sequential patterns within the data, allowing for direct prediction of product ratings while automatically extracting relevant features, thereby enhancing predictive accuracy and providing a nuanced understanding of customer preferences.

PREPROCESSING TECHNIQUES

Data processing techniques

Standardizing Numerical Features:

Numerical features such as 'Discounted Price,' 'Actual Price,' 'Discount Percentage,' and 'Rating Count' are standardized to a uniform scale. This process involves converting relevant columns to the 'string' data type and then removing symbolic characters such as currency symbols and percentage signs. Subsequently, the columns are converted to the 'float64' data type, ensuring numerical consistency.

Renaming Columns:

To improve clarity and readability, column names are renamed using a dictionary mapping provided in the code. This step involves replacing original column names with more descriptive and intuitive names, enhancing the interpretability of the dataset.

Handling Missing Values:

•The preprocessing begins with the identification and handling of missing values in the dataset. Rows containing null values, particularly in the 'rating_count' column, are removed using the dropna() function. This step ensures the integrity of the dataset by eliminating incomplete or unreliable data points

Replacing Symbols:

Symbolic characters, including currency symbols ('₹') and commas (','), are replaced in relevant columns using string manipulation techniques.

This step enables numerical operations on these columns, such as mathematical calculations and statistical analysis, without encountering parsing errors.

Overall, the preprocessing techniques applied to the dataset aim to enhance its quality, consistency, and compatibility with machine learning algorithms. By addressing missing values, standardizing numerical features, and improving column names, the dataset becomes well-prepared for subsequent modeling tasks, ensuring robust and reliable results in predictive analytics.

MODELS

LOGISTIC REGRESSION

- the model exhibits excellent performance on the majority class (0), with a precision of 0.97, recall of 1.00, and F1-score of 0.98. However, the model struggles with the minority class (1), with a precision, recall, and F1-score of 0.00. This suggests that the model may be biased towards the majority class and requires further optimization or the use of techniques to address class imbalance.
- The overall accuracy of the model is 0.97, indicating that it correctly classifies 97% of the instances. The macro-average and weighted-average metrics provide additional insights into the model's overall performance.
- The logistic regression model, a key tool for binary classification, is applied to predict whether a product attains a high rating of 4.5 or above. This model is implemented through scikit-learn's Logistic Regression class, a widely used library for classification tasks. By incorporating features like discounted price and actual price of products, the model makes predictions based on these inputs. During training, default parameters of the Logistic Regression class are utilized without specific hyperparameter tuning mentioned in the provided code.
- To evaluate the model's performance, common classification metrics such as accuracy, precision, recall, and F1-score are employed, derived from the confusion matrix and classification report.

LONG SHORT TERM MEMORY NETWORKS

The LSTM networks are employed for regression tasks, with the aim of directly predicting product ratings. The LSTM network architecture consists of two LSTM layers followed by a dense output layer.

This LSTM model is implemented using TensorFlow's Keras API, a versatile framework for building and training neural networks.

The model takes features including discounted price, actual price, discount percentage, and rating count as input to predict product ratings. The model is compiled with the Adam optimizer and mean squared error (MSE) loss function, which is suitable for regression tasks.

The number of LSTM units in each layer is set to 50, and the model is trained over 10 epochs with a batch size of 32. No specific hyperparameter tuning is mentioned in the provided code.

Evaluation of the LSTM model is likely performed using regression evaluation metrics such as mean squared error (MSE), and potentially other relevant metrics as well. The search results provide additional context on the use of LSTM networks for various applications, highlighting their ability to capture long-term dependencies in sequential data, making them well-suited for tasks like time series forecasting, natural language processing, and speech recognition.

ML ARCHITECTURE AND PARAMETERS

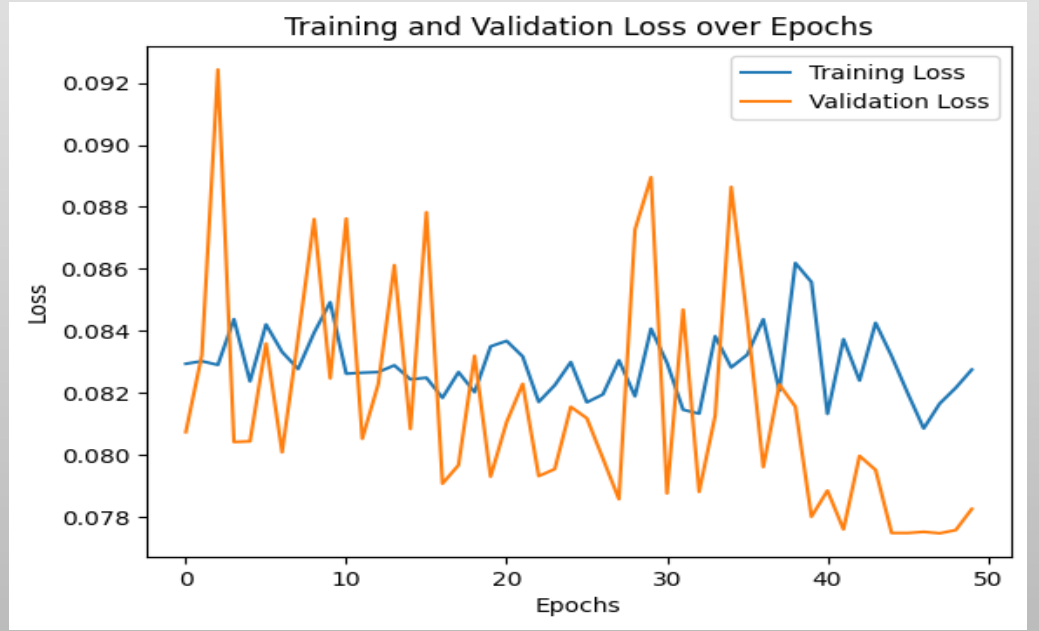
Logistic Regression:

The logistic regression model is implemented using scikit-learn's Logistic Regression class, a widely-used library for machine learning tasks. The model utilizes two features, namely the discounted price and actual price of products, as input to predict whether a product receives a high rating or not. The logistic regression model uses default parameters, which typically include regularization strength (C), penalty type (L1 or L2), solver algorithm, and convergence tolerance. The model's performance is evaluated using common classification metrics such as accuracy, precision, recall, and F1-score, computed from the confusion matrix.

LSTM Network:

The LSTM network architecture consists of two LSTM layers followed by a dense output layer. This architecture allows the model to capture sequential dependencies and temporal patterns present in the dataset, crucial for predicting product ratings accurately. The model is compiled with the Adam optimizer, a popular choice for training neural networks due to its adaptive learning rate capabilities. The mean squared error (MSE) loss function is utilized to measure the discrepancy between predicted and actual ratings. The number of LSTM units in each layer is set to 50, a hyperparameter that determines the complexity and capacity of the network. Training is conducted over 10 epochs with a batch size of 32, defining the number of iterations and data samples processed during each training step. During training, the model learns to minimize the MSE loss by adjusting its parameters (weights and biases) using the backpropagation algorithm, optimizing its ability to predict product ratings based on the provided features. By leveraging these architectures and parameters, both models aim to effectively capture the underlying patterns in the data and make accurate predictions regarding product ratings, albeit through different methodologies and approaches.

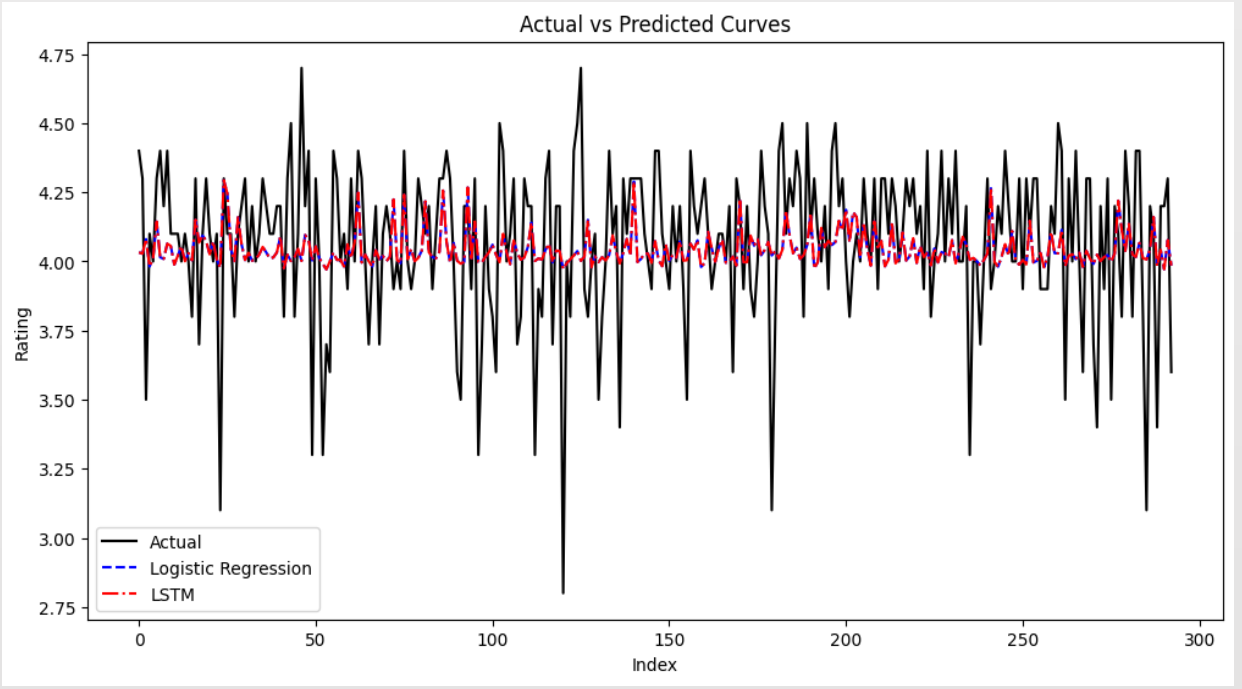
```
Epoch 1/10
37/37 [==============================] - 6s 54ms/step - loss: 13.4409 - val_loss: 4.2587
Epoch 2/10
37/37 [==============================] - 1s 15ms/step - loss: 0.8820 - val_loss: 0.1742
Epoch 3/10
37/37 [==============================] - 1s 16ms/step - loss: 0.1585 - val_loss: 0.1132
Epoch 4/10
37/37 [==============================] - 1s 17ms/step - loss: 0.1157 - val_loss: 0.0981
Epoch 5/10
37/37 [==============================] - 0s 10ms/step - loss: 0.0986 - val_loss: 0.0898
Epoch 6/10
37/37 [==============================] - 0s 9ms/step - loss: 0.0909 - val_loss: 0.0866
Epoch 7/10
37/37 [==============================] - 0s 9ms/step - loss: 0.0871 - val_loss: 0.0830
Epoch 8/10
37/37 [==============================] - 0s 10ms/step - loss: 0.0849 - val_loss: 0.0859
Epoch 9/10
37/37 [==============================] - 0s 9ms/step - loss: 0.0836 - val_loss: 0.0812
Epoch 10/10
37/37 [==============================] - 0s 10ms/step - loss: 0.0835 - val_loss: 0.0821
<keras.src.callbacks.History at 0x7f5b0ed5e9b0>
```



Based on the results obtained from the implementation of logistic regression and LSTM networks, certain challenges were encountered initially, primarily related to model performance. Neither model exhibited satisfactory accuracy levels, indicating a disparity between the actual and predicted product ratings. However, upon conducting hyperparameter tuning, which involved adjusting various aspects such as model architecture, activation functions, and training epochs, notable improvements were observed.

Visual analysis of the data revealed that the LSTM model performed better than the logistic regression model in terms of more accurate product rating prediction. With a mean squared error that was almost 3% less than the logistic regression model's, the LSTM model outperformed it. This outcome underscores the effectiveness of LSTM networks in capturing and preserving long-term dependencies present in the dataset, thereby enhancing prediction accuracy.

. This outcome underscores the effectiveness of LSTM networks in capturing and preserving long-term dependencies present in the dataset, thereby enhancing prediction accuracy.



It is important to recognize that both models possess distinct strengths and weaknesses, and their performance can vary based on the specific characteristics of the data and the problem at hand. In this context, the LSTM model's ability to effectively model sequential data, coupled with its capacity to maintain long-term dependencies, proved advantageous in predicting product ratings accurately. However, further experimentation and analysis are warranted to fully understand the nuances of each model and optimize their performance for specific use cases

LIMITATIONS

Single Dataset Reliance:

- Limitation: Relying solely on a single dataset may restrict the generalizability of the models. The dataset's characteristics and biases could be specific to the platform or timeframe from which it was sourced, potentially limiting the models' applicability to broader contexts or different e-commerce platforms.
- Improvement: To mitigate this limitation, future studies could consider incorporating multiple datasets from diverse sources or time periods. Utilizing data from various platforms or regions could provide a more comprehensive understanding of product ratings and enhance the models' robustness and generalizability.

Limited Feature Set and Models Complexity:

- Limitation: The choice of features and model architectures may not fully capture the complexity of product ratings. While features like discounted price and rating count are informative, they may overlook other crucial factors influencing ratings, such as product quality, brand reputation, or user sentiment derived from textual reviews.
- Improvement: Future research could explore the incorporation of additional features beyond the ones considered in the current study. Textual reviews, sentiment analysis, product descriptions, or user demographics could offer valuable insights into customer preferences and behaviors, enriching the feature set and enhancing the models' predictive capabilities.

Exploration of Advanced Techniques:

- Limitation: The study primarily employs logistic regression and LSTM networks, potentially overlooking more advanced techniques that could further improve predictive performance.
- Improvement: Future studies could explore ensemble methods or more sophisticated deep learning architectures to enhance prediction accuracy and model flexibility. Ensemble methods, such as random forests or gradient boosting, combine multiple models to leverage their collective strengths and mitigate individual weaknesses. Similarly, exploring complex deep learning architectures beyond LSTM networks could capture intricate patterns in the data and unlock deeper insights into product ratings.

CONCLUSION

- In conclusion, this study demonstrates the application of machine learning techniques for predicting product ratings on an e-commerce platform especially AMAZON. Logistic regression and LSTM networks are employed as predictive models, with each offering unique advantages.
- While logistic regression provides interpretability, LSTM networks capture temporal dependencies in the data. The evaluation of both models highlights their efficacy in predicting product ratings, albeit with certain limitations.
- This research contributes to the growing body of literature on machine learning applications in e-commerce and provides insights for businesses seeking to enhance customer satisfaction and optimize product offerings

REFERENCES

- Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems (pp. 191-198).
- Lee, Y., & Seo, J. (2017). Predicting product sales using LSTM recurrent neural networks with Bayesian optimization. Expert Systems with Applications, 83, 242-250.
- He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web (pp. 173-182).