Introduction

In this project, I aimed to predict star ratings of Amazon Movie Reviews using machine-learning techniques that do not involve deep learning. My goal was to experiment with different classification algorithms and feature engineering techniques to achieve high accuracy in prediction. The data provided included `train.csv` and `test.csv,` containing review scores and corresponding metadata.

During my research, I drew inspiration from various articles and papers that explored sentiment analysis using logistic regression. One such reference is the paper titled "A Sentiment Analysis of Food Review using Logistic Regression," which provided insights into using logistic regression for sentiment analysis, including feature extraction techniques, data preprocessing, and model evaluation. This paper guided my approach, particularly in using logistic regression as a simple yet effective method for binary classification and extending it to multiclass problems by treating each rating level as a separate class.

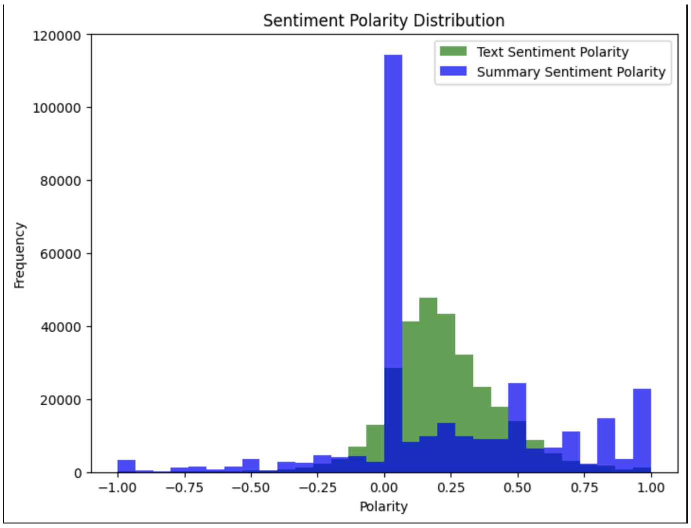
Data Preprocessing

Initially, I loaded the `train.csv` dataset, followed by dropping any rows with missing values. To ensure representative sampling and efficient training, I randomly shuffled the dataset and generated three stratified samples of 300,000 rows each for experimentation. During preprocessing, I handled missing values by filling empty text or summary fields with default values to maintain data consistency.

Feature Engineering

To improve model performance, I applied several techniques. One key aspect was stratified sampling, which ensured representative data across all classes despite the imbalanced distribution. I also engineered features like sentiment polarity, subjectivity, and review length to enrich the model's input data. Using Naive Bayes transformations as additional features boosted the model's ability to differentiate between classes. Additionally, I combined different initialization methods and applied hyperparameter tuning for logistic regression, which improved the model's accuracy and stability. The feature engineering phase included extracting features like review length, text sentiment polarity, summary sentiment polarity, text subjectivity, and summary subjectivity. I processed the text features using the TextBlob library, which provided sentiment and subjectivity scores that were later used in the models. Additionally, I created features representing Naive Bayes probabilities for each class (1-5 stars) by fitting the Naive Bayes transformation on the review text and summary. The pipeline included steps for both TF-IDF transformation and logistic regression as the classifier.

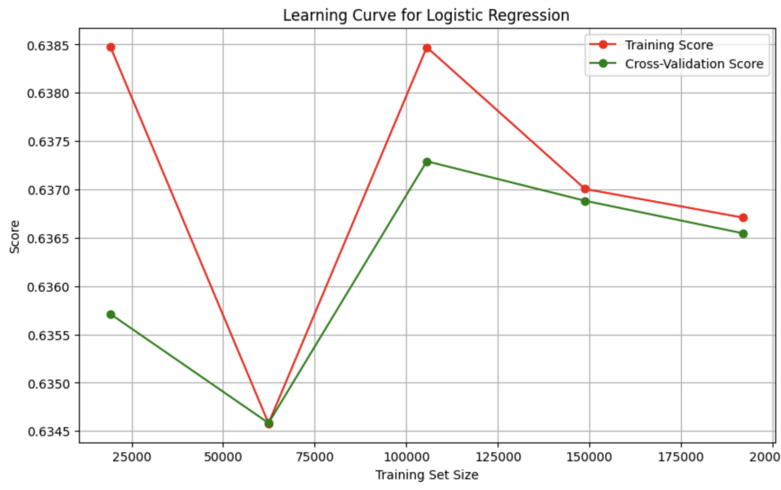
The paper on sentiment analysis of food reviews highlighted the importance of extracting useful features such as sentiment polarity, which was instrumental in my feature engineering process. Similar to the paper's approach, I used TF-IDF vectorization to represent the review text, and I implemented logistic regression as the main classification model, leveraging its interpretability and efficiency.



The second graph, Sentiment Polarity Distribution, illustrates the distribution of polarity scores for both text and summary fields. The text sentiment polarity is mostly centered around neutral values, with a significant number of reviews having a polarity score close to zero, indicating neutral sentiment. The summary fields show a similar pattern but with slightly higher peaks at both ends, indicating that summaries tend to express sentiment more strongly compared to the full text.

Model Training

I divided the training data into training and test sets using an 80-20 split to facilitate model validation. Logistic Regression was used as the primary classifier, which I trained with a maximum of 500 iterations. Additionally, I implemented cross-validation to evaluate the generalizability of the model. The model achieved an accuracy score of approximately 63.8%.



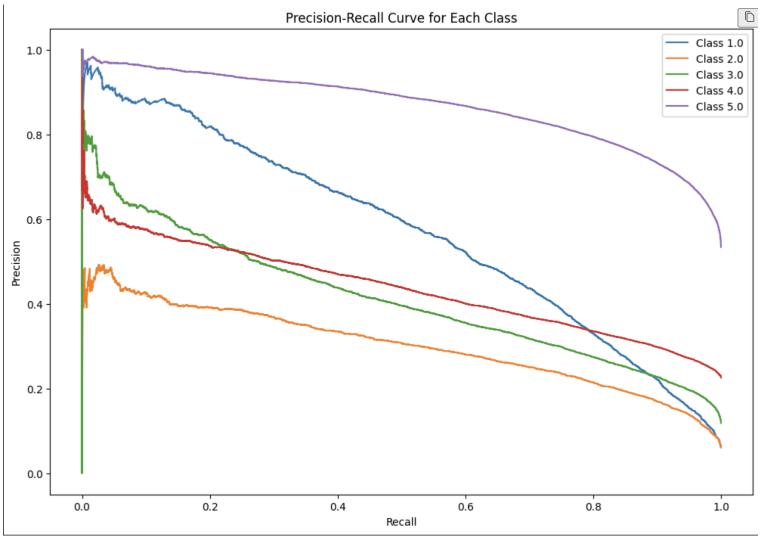
The Learning Curve for Logistic Regression graph provides an overview of the model's performance as a function of training set size, comparing training and cross-validation scores. The graph shows that the model initially overfits the training data with smaller sample sizes, but as the training set size increases, the training and cross-validation scores converge, indicating improved generalizability.

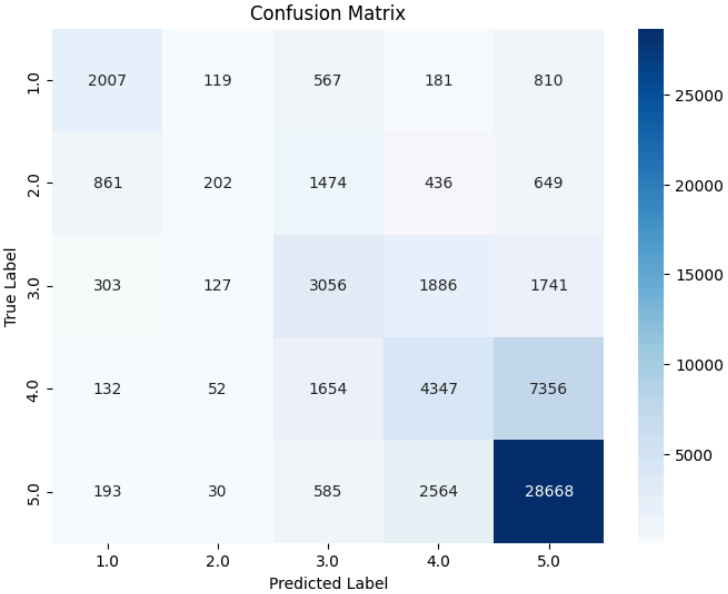
Evaluation and Results

After training, I used various metrics to evaluate the model, including accuracy, precision, recall, F1 score, and a confusion matrix. The classification report revealed that the model performed better on higher scores, especially the 5-star ratings, which correlates with the skewed class distribution. The model achieved a precision of 0.73 and a recall of 0.89 for 5-star ratings, while the recall for 2-star ratings was particularly low at only 0.06, indicating difficulty in predicting lower scores accurately.

The paper on sentiment analysis of food reviews also emphasized the use of logistic regression for interpreting model outcomes through metrics like precision, recall, and F1-score. Inspired by this, I carefully analyzed the model's confusion matrix and precision-recall curves to understand where the model struggles and which classes are often misclassified.

To further understand the precision and recall for each class, I plotted the Precision-Recall Curve for Each Class. The curve shows that the precision-recall trade-off is most favorable for 5-star ratings, while it deteriorates significantly for lower ratings, especially 2-star.





	1.0	2.0	3.0	4.0	5.0
1.0	2007	119	567	181	810
2.0	861	202	1474	436	649
3.0	303	127	3056	1886	1741
4.0	132	52	1654	4347	7356
5.0	193	30	585	2564	28668

The Confusion Matrix provides insights into the classes that are often confused by the model, such as 4 and 5-star ratings. The matrix shows that a substantial number of 4-star reviews were misclassified as 5-star, which is consistent with the class imbalance observed in the dataset.

Conclusion

This project explored several aspects of feature engineering, model selection, and evaluation techniques for the classification of Amazon Movie Reviews. The results showed that while logistic regression is effective for simpler scenarios, the class imbalance heavily affected the prediction accuracy for lower-rated reviews. Future work can include addressing the data imbalance using resampling techniques or experimenting with other ensemble-based classifiers.

References-

https://www.researchgate.net/publication/334654833_A_Sentiment_Analysis_of_Food_Review_using_Logistic_Regression
https://towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b
https://medium.com/@nirajan.acharya777/sentimental-analysis-using-linear-regression-86764bfde907
https://www.youtube.com/watch?v=My4JgIeFdWk