Predicting Amazon Review Ratings Using K-Nearest Neighbors and User-Centered Features

---

Introduction

This project aimed to develop a predictive model for Amazon Movie Review star ratings using metadata and user-centered features. The provided starter code suggested the **K-Nearest Neighbors (KNN) classifier**. Although attempts were made to test other classifiers, the computational requirements were prohibitive for timely execution. KNN was therefore chosen as the final model for its ability to capture patterns efficiently in user behavior with minimal tuning requirements (Altman, 1992).

Feature Engineering.

The features engineered for this model include:

1. **User_Avg_Score**: This feature represents the average score a user has given across all their reviews. Previous studies on recommender systems indicate that a user's historical rating tendencies can predict future ratings effectively (Beel et al., 2016). Users with consistently high average scores, for instance, are likely to give higher ratings in future reviews.
2. **User_Negative_Count**: This feature tracks the number of 1-star ratings a user has provided. Research indicates that users who frequently rate products negatively exhibit distinct behavioral patterns, which can impact future review predictions (Aciar et al., 2010). By identifying users with a high count of negative ratings, the model can make more conservative predictions for these users.
3. **Multiple_Negative_Reviews (Binary)**: This binary feature flags users who have given more than one 1-star rating, acting as a simplified indicator of negative tendencies without emphasizing exact counts.

These features were engineered to capture user tendencies that influence rating behavior, enabling the model to leverage user-specific patterns rather than generic metadata alone.

Model Evaluation

The model achieved an accuracy score of approximately **0.8154** on the test set, indicating that KNN was effective in distinguishing between different review ratings based on user-centered features.

Limitations

Despite achieving a local testing accuracy of **81.5%**, the model performed significantly worse in the Kaggle competition, with an accuracy closer to **20%**. Several factors may have contributed to this discrepancy:

1. **Over Reliance on Average Score**: The User_Avg_Score feature, while useful in capturing user tendencies, may have led to unintended bias in the predictions. By basing predictions on a user's average past ratings, the model could struggle to generalize to new ratings, especially if the test set includes users with shifting behavior or outlier reviews. Studies indicate that while average rating behavior is a useful feature, it can also risk oversimplifying individual preferences (Yang et al., 2017).
2. **Data Drift and Feature Instability**: The Kaggle test set may have a different distribution of ratings or user behaviors compared to the local training set. If users in the test set exhibited behaviors inconsistent with those seen in the training data, this could have reduced the model's ability to make accurate predictions.
3. **Class Distribution Differences**: The Kaggle test set may have had a different class distribution from the local test set, potentially impacting model performance. KNN's reliance on local neighborhood patterns may lead to inaccuracies if the rating distribution shifts significantly between training and test sets.
4. **Lack of Model Flexibility**: KNN, being a non-parametric algorithm, does not learn explicit patterns but instead bases predictions on the closest neighbors. In scenarios with large datasets or varying distributions, KNN can become less reliable. This limitation, coupled with potential overfitting to user average scores, might have led to the model's reduced performance in the Kaggle competition.

Summary and Conclusions

In this project, a K-Nearest Neighbors model was implemented to predict Amazon review ratings based on user-centered features. The final model achieved an accuracy of **0.8154** on the test set. Despite attempts to explore other classifiers, KNN was ultimately chosen for its computational efficiency and simplicity, aligning with research that supports its effectiveness in numerical analysis (Altman, 1992). However, the significant discrepancy in Kaggle accuracy highlights limitations in using average scores as predictors and suggests the need for more flexible, adaptive models in future work.

(The highest kaggle submission includes the starter code, this is a report written on the issues and research behind the hypothesized model based on user behavior.)

References

Aciar, S., Zhang, D., Simoff, S., & Debenham, J. (2010). Informed recommender: Basing recommendations on consumer product reviews. *IEEE Intelligent Systems*, 22(3), 39-47.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175-185.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). *Recommender systems handbook*. Springer.