



MACHINE LEARNING

.....

- Naive Bayes, Random Forests, Decisions Trees and Logistic Regression were all attempted to make predictions.
- Logistic Regression consistently delivered the best results, followed by Naive Bayes.
- Logistic Regression Cross-Validation had AUC means of over 90%.
- Confusion Matrix precision of unhelpful ratings were over 80%, and precision of helpful ratings were over 90%.
- Hyperparameter C tuned as 0.007742636826811269.

RESULTS

- Validation sets returned 91% accuracy.
- Test sets returned 88% accuracy.
- Changes in min_df and max_df led to minimal gains.
- Star ratings were predicted with greater accuracy. This can be used to flag users who gave the wrong amount of stars.
- Deep learning did not initially outperform Logistic Regression, but more tests and reviews deliver better results.
- The helpful rating metric can be applied to any product that counts votes (traditionally thumbs up / thumbs down).
- The next step is to use a similar pipeline to determine helpful reviewers.

