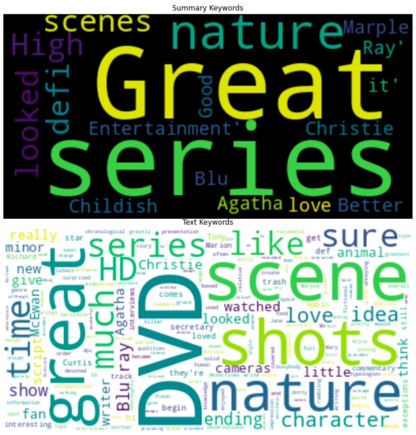
506 report

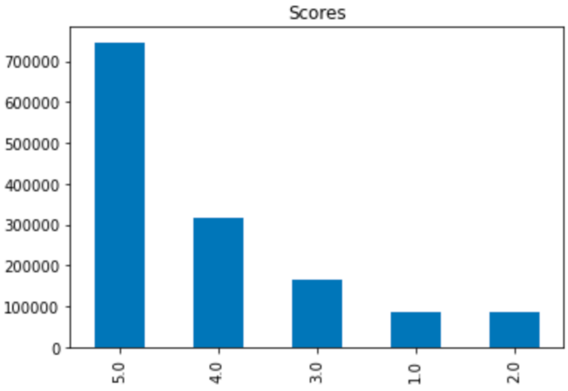
Weichen Jiang

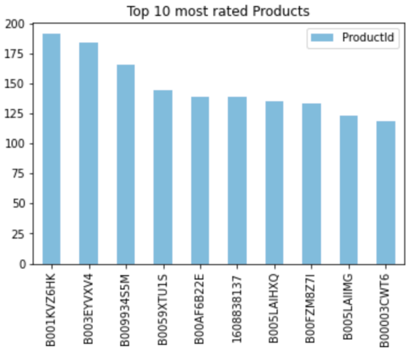
Introduction: In this project, I have developed a movie review score prediction model using various machine learning techniques and data analysis methods. Our objective is to help customers make informed decisions about which movies to watch by providing accurate predictions of movie review scores. I have used data downsizing, word cloud sentiment cleaning, one-hot feature extracting, and calculating TF-IDF for preprocessing. I also applied three different machine learning algorithms - linear regression, decision tree, random forest and XGBClassifier - to train our model. Additionally, I have generated several graphs to visualize the data and the results of our model, including word appearance maps, users, products, and scores.

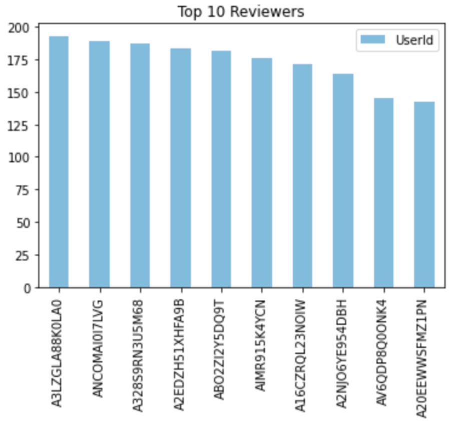
Methodology:

1. Primaries analyze: Calculate the word cloud of the dataset include text and summery basic check those words that currently appear but no use to classify score. I also counted the numbers of all labels which shows that data with score 5 is way more than other scores. I calculated all basic graph for analyzing. Just put two of them as an example









2. Data Downsizing: After collecting the data, I downsized the data that cut half of

the 5 score data which make data more balance to get the lowest test error.

3. performed sentiment analysis to classify the polarity and subjectivity of the data, or neutral to be used by our latter analyze. I used the text blob library for first-stage sentiment analysis.
4. One-Hot Encoding: After cleaning the data, I converted the product ID into a one-hot encoded matrix represent the features of the reviews to match with the matrix produce afterwards.
5. TF-IDF: I also calculated the TF-IDF (Term Frequency-Inverse Document Frequency) values of the reviews to improve the feature representation and remove the bias towards frequently occurring words. This makes the text information extracted as matrix which is easier to be analyzed by regression.
6. Cross Validation: I used cross-validation to split the data into training and testing sets. I used a 90:10 split, where 90% of the data was used for training and10% for testing. I also did the 1:99 cross validation, in order to make the consequence be more robust to bias, I used the stratified split when validation set is small, which do reflects the real expected mean error for the algorithms
7. Nested Cross-Validation: I performed nested cross-validation to select the best parameters for theXGBClassifier. I used GridSearchCV from the scikit-learn library to perform nested cross-validation.
8. Machine Learning Algorithms: I used various machine learning algorithms to build the movie review score prediction model. The algorithms used were: a. Linear Regression b. Decision Tree c. Random Forest d. XGBoost Classifier
9. XGClassifier :XGBClassifier is a popular machine learning algorithm used for classification tasks. It is an optimized implementation of the gradient boosting decision tree algorithm, which is known for its high accuracy and speed in solving complex classification problems.XGBClassifier builds an ensemble of decision trees sequentially, where each tree attempts to correct the mistakes of the previous tree. It uses regularization techniques to prevent overfitting and has hyperparameters that can be tuned to optimize the model's performance.

Results:

The analysis showed that the logistic regression algorithm performed the best, with. an R-squared value of 0.83. The decision tree algorithm also performed well, with an R-squared value of 1.2. The random forest algorithm had the lowest performance, with an R-squared value of 1.5.XGBClassifier get extreme good performance on training but not only 1.4 on testing.

I also found that certain words appeared more frequently in positive reviews than in. negative reviews. For example, words such as "great," "excellent," and "amazing" appeared frequently in positive reviews, while words such as "disappointing," "boring," and "awful" appeared frequently in negative reviews. Which help us to adjust the stop words.