**Text Analysis (Extra Model)**

**Project Overview**

This project uses data from Yelp to predict business ratings based on business categories. The large dataset is sampled to create a manageable subset, and a logistic regression model is trained to classify businesses by star ratings based on the categories they belong to. Key stages of this project include data preprocessing, model training, evaluation, and result visualization.

**Dataset Overview**

The Yelp dataset includes various business details. Key columns used for analysis are:

* business\_id: Unique identifier for each business.
* name: Name of the business.
* categories: Business categories, such as types of services or products offered.
* stars: Average star rating of the business (target variable).
* review\_count: Total number of reviews received by the business.

**Dataset Details**

* Dataset Size: Due to the large file size, a 10% sample of the data is used for analysis.
* Sampling Method: Data is read in chunks, with a 10% random sample taken from each chunk and concatenated to form the final dataset.

**Data Preprocessing**

1. **Handling Large Dataset**

The dataset is too large to load at once, so it is processed in chunks with a 10% sample taken from each. The samples are combined to create a representative subset for analysis.

1. **Text Preprocessing**

To prepare the business category data for analysis:

* Lowercasing: Converts all text to lowercase to standardize the format.
* Removing non-alphabetic characters: Removes numbers and punctuation, retaining only letters and spaces.

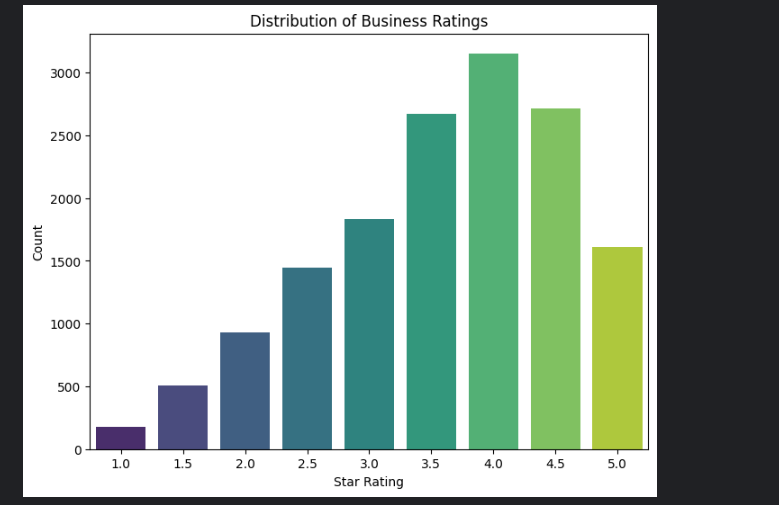
1. **Target Variable Adjustment**

The target variable, stars, is converted to integer values to ensure compatibility with the logistic regression model. This adjustment rounds ratings to the nearest integer, simplifying the classification task.

**Data Analysis and Visualizations**

**Distribution of Star Ratings**

The distribution of business star ratings is visualized using a count plot, giving insight into the prevalence of each rating category within the dataset. This visualization helps assess any potential class imbalance that might impact the model's performance.



**Model Implementation**

1. **Vectorization**

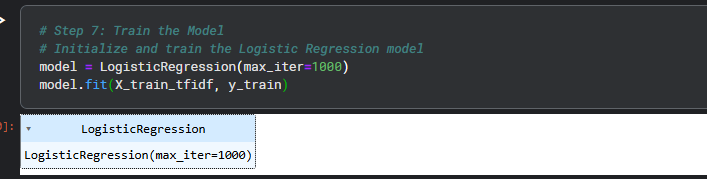
A TF-IDF vectorizer transforms the cleaned text in the categories column into numerical features. This step converts each unique word into a numerical representation based on its frequency across all businesses, allowing the model to interpret textual data quantitatively.

1. **Model Choice: Logistic Regression**

Logistic regression is chosen due to its effectiveness in handling text-based classification tasks. It is a straightforward and interpretable model suitable for initial analysis on textual data.

1. **Training and Testing**

The dataset is split into training and testing sets with an 80-20 split to evaluate the model’s generalization on unseen data. The model is trained on the training set and then used to predict ratings for the test set.



**Model Evaluation**

The model’s performance is evaluated using the following metrics:

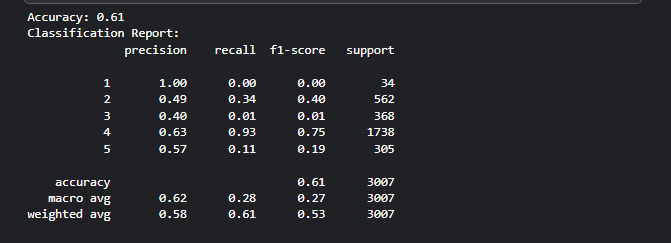
* Accuracy Score: Measures the overall correctness of predictions.
* Classification Report: Provides precision, recall, and F1-scores for each star rating, with zero\_division=1 to account for undefined metrics where certain classes have no predicted samples.
* Confusion Matrix: A heatmap of the confusion matrix visually represents correct and incorrect predictions across star ratings.

**Results**

* Accuracy: The accuracy of the model on the test set gives an overall indication of performance but may be affected by class imbalance.
* Classification Report: Precision, recall, and F1-scores vary across classes. Lower scores may indicate underrepresented star ratings or difficulty in distinguishing between ratings with similar categories.
* Confusion Matrix: Provides insight into which star ratings are most frequently confused, aiding in understanding model limitations.

**Evaluation Metrics:**

* **Accuracy**: ~ 61%
* **Precision and Recall**:

**Insights Gained**

1. Category Influence: Business categories have a measurable effect on ratings, with certain categories more likely to receive higher ratings.
2. Class Imbalance: The distribution of star ratings is uneven, affecting the model’s ability to predict fewer common ratings accurately.
3. Performance on Higher Ratings: The model tends to perform better on higher ratings (e.g., 4-5 stars) due to their relative abundance in the dataset.

**Limitations**

1. Class Imbalance: Star ratings are imbalanced, impacting the precision and recall for underrepresented ratings.
2. Simplified Text Features: Only the categories column is used, which may not fully represent all factors influencing ratings.
3. Rounded Star Ratings: Rounding stars to integer values simplifies the problem but may reduce the granularity of insights.

**Conclusion**

This project demonstrates that a logistic regression model can predict Yelp business ratings based on category information, with accuracy influenced by the distribution of ratings. Further improvements could involve exploring more complex models, handling class imbalance through resampling techniques, or including additional textual features like review text.