CS 6375
Project Status Report:

Rating Prediction Using Reviews (Yelp Dataset)

Names of students in your group:

ADRITA DUTTA: axd172930
SANTHOSH MEDIDE: sxm174930
PRATIMA: pxl180030
CHIRAG SHAHI: cxs180005

Number of free late days used: **0**
Note: You are allowed a total of 4 free late days for the entire semester. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

PROBLEM STATEMENT:

Predict the ratings given to a business based on the reviews given and suggest this ratings to the user as an auto fill.

Dataset used : YELP dataset (https://www.yelp.com/dataset)

Original Dataset Description: The original dataset contains 6 json files, business.json, reviews.json, user.json, checkin.json, tip.json, photo.json

Dataset Components used: For the purpose of our project we have used the following json files from the yelp dataset
business.json, reviews.json, user.json

Dataset size: 3GB

Final Features:
Naming convention used: Filename_columnname

review_id	review_stars	review_funny_upvotes	review_useful_upvotes	review_cool_upvotes	total_tokens	compound_score_review	user_avg_stars	user_yelping_since	user_review_count

Features explanation:

1.Review_id
Found in : review.json
Explanation: Id of review

2. **Review_stars: (class: Star values: 1 - 5)**
Found in: review.json
Explanation: Ground Truth

3. review_funny_upvotes
Found in: review.json
Explanation: Upvotes that the review received

4. review_useful_upvotes
Found in: review.json
Explanation: Upvotes that the review received

5. review_cool_upvotes
Found in: review.json
Explanation: Upvotes that the review received

6. total_tokens
Calculated using reviews found in reviews.json
Explanation: Count of number of words in a review

7. compound_score_review
Calculated using reviews found in reviews.json
Explanation: sentimental score of the review

8. user_avg_stars
Found in: review.json
Explanation: Upvotes that the review received

9. User_yelping_since
Found in : user.json
Foreign Key: User_id
Calculated using the date given in users.json
Explanation: User has been member of yelp (number of days)

10. User_review_count
Found in : user.json
Foreign Key: User_id
Explanation: Number of reviews given by user

PRE-PROCESSING THE DATA

1. Extraction of relevant data from the 3 datasets (users, business and reviews)
2. Joining the relevant columns using Python and Spark SQL to form a single dataset.
3. Running python scripts to generate extra features like compound_score_review, total_tokens and yelping_since.
4. Choosing the training set of size 5000 data points and saving the data in MySQL database for better querying.
5. Graphed correlation matrix and scatterplot to draw insights from the data.
6. To refine the feature selection, used FeatureSelection library of sklearn.

    - F-classif : Compute the ANOVA F-value for the provided sample. ANOVA stands for Analysis of Variance.
    - Select_K_best : Computes the best k features
    - Chi2 : Compute chi-squared stats between each non-negative feature and class. This score can be used to select the n_features features with the highest values for the test chi-squared statistic from X.
    - RFE : Feature ranking with recursive feature elimination.
    - Mutual_info_classif : Estimate mutual information for a discrete target variable

7. A temporary model has been built using Decision Tree and Neural Networks using 5000 data points and achieved a 51% test error on them.

Steps to follow next:

1. DECIDING BETTER FEATURES to improve the testing and training results.
2. Building a temporary model on the chosen dataset (5000 data points) for easy training and testing purposes.
3. ALGORITHMS WE WILL BE USING TO BUILD MODELS (3-6 among the below mentioned algorithms): Decision Tree, Neural Networks, Support Vector Machines, Gaussian Naïve Bayes, Logistic Regression, K-Nearest Neighbours, Bagging, Random Forest, AdaBoost
4. FINALLY, TRAIN ON BIG DATASET