Predicting Ratings on Google Local Reviews using Regression Techniques

Sreeja Madanambeti
014619930

Yamini Aalla
01262333

Xuan Shi
013856401

I. Introduction

Motivation

The reach of the web in this world has changed the world significantly. One of the developments of the internet is online reviews. People are no longer purchasing a product blindly or visiting a new place without reading a review. Reviews became more accessible through mobile applications. Most people nowadays opt to make decisions based on reviews. There is a massive amount of positive and negative reviews for every product or place. It is impossible to go through every reviewer experience, and it is a tiresome activity. Improving user experience is vital for customer satisfaction. This inconvenience can be reduced if a rating is predicted for a place or business based on user search. Predicting rating will guide the consumer to make a quick decision based on a search.

Objective

We will explore the google local dataset and develop a robust model for predicting ratings. We will develop the best possible rating prediction system using text analysis and regression techniques.

Structure of report

The system design and implementations chapter confer about high level representation of preprocessing steps, algorithms used, technologies and tools, and data flow in the system. Proof of concepts and evaluations discusses data preprocessing and methodology followed for the project. The discussion conclusion chapter addresses the decisions made, difficulties faced during our work, things that worked and didn't work for our predictions, and the conclusion and future scope for this project.

Related Work

Sasha et al. [1] built a sentiment summarizer; positive and negative sentiment is calculated on sentences and phrases in reviews. After sentiment analysis and aspect extraction, they have summarized the module that helps to obtain the overall summarization of review for a restaurant or any local service. Studying this paper helped in understanding the sentiment calculation on reviews.

Lizhen et al. [2] opted for a bag of words method for review rating prediction and structured it with root words and negators. They developed a ridge regression model to learn opinions. Statistics for opinion scores are derived and used as a feature for predicting the review. Studying this paper helped in understanding to extract new features to use in our predictions.

II. System Design and Implementation

A. Algorithms Explanation

1) Ridge Regression: Ridge is a regularization technique that reduces overfitting. Ridge regression uses L2 regularization. In ridge regression, the penalty added is equal to the square of the magnitude of the coefficient. We are using this model as it reduces multi-colinearity for better predictions. [3].

2) Multi-layer Perceptron: Multi-layer perceptron is a deep learning technique. This deep neural network can learn non-linear functions. It has hidden layers; layers of input nodes produce different set of outputs. We are using this algorithm to learn the non-linearity factor in data.

3) LASSO Regression: LASSO regression is a regularization technique that reduces overfitting. LASSO uses L1 regularization. In LASSO, Penalty added to the cost function equals the sum of absolute coefficients. We are using this model to reduce overfitting [3].

4) Random-Forest Regressor: Random forest is one of the ensemble methods, where estimators are large number of small independent decision trees. The predicted will be the average of several sub-samples of decision trees. We are using this regressor to predict the rating based on the average of multiple decision trees.

5) Multi Linear Regression: Multi-linear regression is a linear regression algorithm. It takes multiple variables instead of one variable to predict the outcome. We are using this regression algorithm as it estimates the relationship between more independent variables [4].

6) Gradient Boost Regressor: Gradient boosting is also an ensemble method. It produces a strong learner through a combination of weak learners in an iterative fashion. We are using this model because the accuracy of the predictive results is higher as it minimizes the predictive error altogether [5].

- Sentiment analysis
  In the reviews, there are many subjective words, such as hard, difficult, helpful, and opinions modified by very, quite, not. For rating prediction, "very helpful" is much stronger sentiment than "helpful". It is reasonable to assign a positive or negative weight to the opinion modifiers [2].
- Bag of words
  Bag of words is used in Natural Language Processing, to just keep multiplies of a text, disregarding grammar or

even word order. The occurrence of each word is treated as a feature in the model.

B. Technologies and Tools

Programming Language: : Python.

Frameworks: : Pandas, Numpy, Collections, Sklearn, matplotlib, Seaborn,NLTK.

Tools: : Jupyter Notebook.

C. System Design

1) Data Flow: We loaded the places and reviews data of google local dataset. None values were removed and a number of reviews were selected to analyze. By merging and exploring data, important features were selected and added to the merged data. Then regression models and sentiment analysis were applied to the split data set. Regression models included Linear, Ridge, Lasso Regression, Gradient boost, Random forest and MLP. Finally, the models and methods were evaluated to conclude the one with best performance. Figure 1 showed the flowchart of the project.

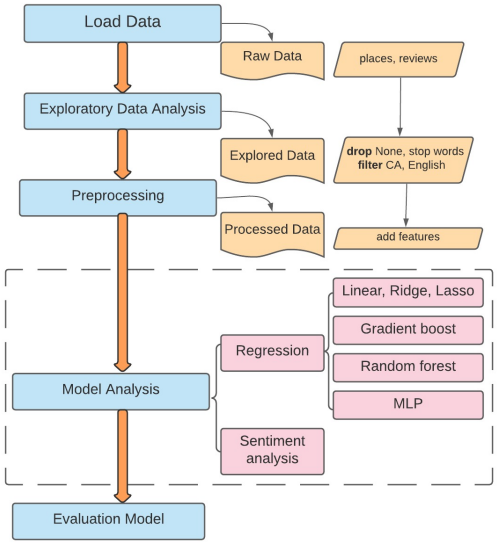


Fig. 1. Flow chart of project

2) Preprocessing:

- Extracted California places from places dataset and merged the features GPS and price to reviews dataset.
- For text analysis removed punctuation, stop words from 75,254 reviews and picked top 1500 words.

3) Project Implementation: Developed the four models considering different features in each scenario for predicting the ratings

- Scenario-1 Bag of words (BOW)
  In Scenario-1, we used Bag of Words to analyze the reviews. For review texts, capitalization, punctuation, and stop words were removed. After sorting the distinct words, the most popular words were selected. And then, regression models were applied to predict ratings considering popular words.

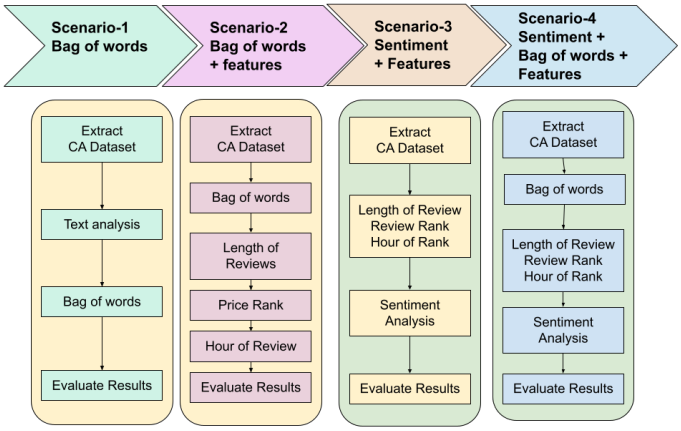


Fig. 2. Flowchart of Scenarios

- Scenario-2 (BOW + Features)
  As per our research, we had considered the features length of review text, time when the review was posted, Price Value additional to Bag of top 1500 Words, for review length calculated the length of review text, from given Unix review time extracted review hour, ordinal encoded the prices to values 1, 2 and 3 and trained by using various regression algorithms to achieve the error value less than the one generated from scenario one.
- Scenario-3 (Sentiment Analysis)
  One of the important aspects in text analysis of reviews, to predict the ratings is analyzing the review text. To analyze the review text, we are using sentiment analysis. We have implemented using bag of words in the previous scenario. We have split the review text into sentences, and split is applied for complete sentences. We have calculated the polarity and subjectivity using Text Blob. Text sentiment is calculated as positive and negative sentiments and excluded the neutral sentiments in the text. We are inputting these sentiment results to the regression algorithms.
  Sentiment-based rating prediction is implemented using several machine learning regression models considering features like positive sentiments, negative sentiments, hour, price level, and review length. Error metrics are measured for all the regression algorithms, and performance is reviewed on regression models.
- Scenario-4 (Advanced Model) In all three scenarios, we are not getting satisfactory results, so we developed an advanced model considering features Bag of 1500 words, length of review, an hour of review, ordinal encoded price

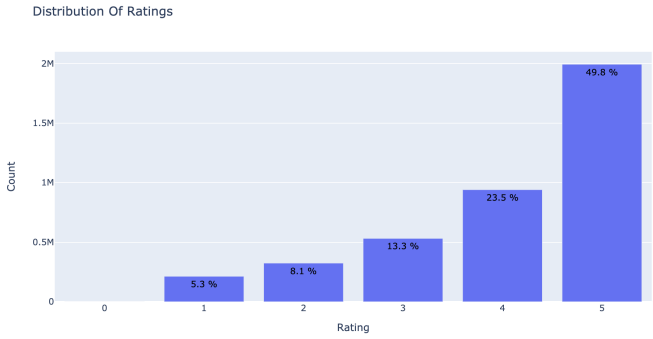
values, and sentiment analysis to all the review text. As a successful implementation, we have obtained a lower error value compared to all the other scenarios.

D. Regression Algorithms Implementation

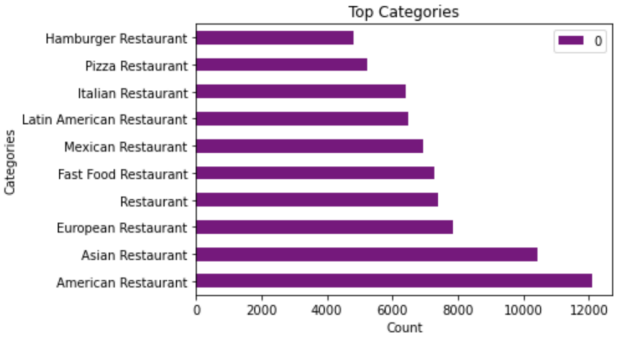
- Linear Regression
  We have used linear regression, it takes multiple variables as input. The input features for the last scenario Bag of 1500 words, length of review, an hour of review, ordinal encoded price values, along with sentiment analysis to all the review text. We have imported Linear regression from Sklearn, and the fit function is used to train the model; weights will be adjusted to data values till better accuracy is achieved. After training, we are using our Linear model for predictions. We have validated the predictions using a validation set. We have evaluated the model using error metrics. We opted to compare the performance using the Root mean square error (RMSE). In our advanced model (scenario 4), Linear regression has an RMSE of 0.9004.
- Ridge Regression
  Ridge regression is a regularization technique, and it does L2 normalization. To reduce overfitting, it adds a penalty to tune the error function. We have alpha values as 1.0 for initialization, and the larger alpha represents strong regularization and fit intercept as false. We have chosen these values after tuning several parameters; these parameters yielded the best results through tuning. We have evaluated the model using Mean absolute error (MAE), Mean squared error (MSE), Root mean square error (RMSE). As per the results of all the error metrics stated above, Ridge regression is one of the top-performing models with an RMSE of 0.9001.
- Random Forest Regression
  Random forest ensembles a large number of small and independent decision trees. Parameters used are n-estimators, max-depth, and random-state. The n-estimators is the number of trees in the forest; we have tuned it to 500. Maximum depth of a tree as two and randomness as zero. All the parameters are fine-tuned for their best prediction results. The evaluation metrics used are MAE, MSE, RMSE. The RMSE of the Random forest regressor is 1.0441.
- Multi Layer Perceptron
  Multi-Layer perceptron (MLP) learns the non-linearity in the data. Parameters of MLP used in predicting the rating are alpha, learning-rate-init, activation, random-state, max-iter. We have used rectified linear unit function(relu)as activation function as the other function tanh didn't yield good results. The regularization parameter alpha is 0.03. Initial learning rate to control step size as 0.04. Random number generation as 0.1 to initialize weights and bias. The maximum number of iterations is 600. The evaluation metrics used are MAE, MSE, RMSE to evaluate MLP. The RMSE of the MLP algorithm is 1.0264.
- LASSO Regression
  LASSO regression is a regularization technique, and it does L1 normalization. To reduce over fitting, it adds a penalty to tune the error function. We have alpha values of 0.25 for initialization; the alpha represents regularization and fit-intercept as false. Serval parameters are tested during tuning and chose the above due to less error. The evaluation metrics used are MAE, MSE, RMSE to evaluate LASSO. The RMSE of the LASSO algorithm in our advanced modeling is 1.2648.
- Gradient Boost Regressor
  Gradient boosting implements a gradient boosting decision tree algorithm. We have used a gradient boosting tree to tune our model. Several hyperparameters are used, including n-estimators, min-samples-leaf, max-depth, min-samples-split, max-leaf-nodes, max-features, learning- rate, random-state. We have tuned parameters by heuristics. The number of estimators is 100, the minimum number of samples at leaf node is 1, maximum depth is 3, a minimum number of samples is two to split a node. The number of leaf nodes is unlimited; the learning rate is one, and the random state is zero. This library boasts computational capabilities, and it is one of the best performing libraries for our regression problem. The RMSE of boosting model is 0.8871.

E. Visualization

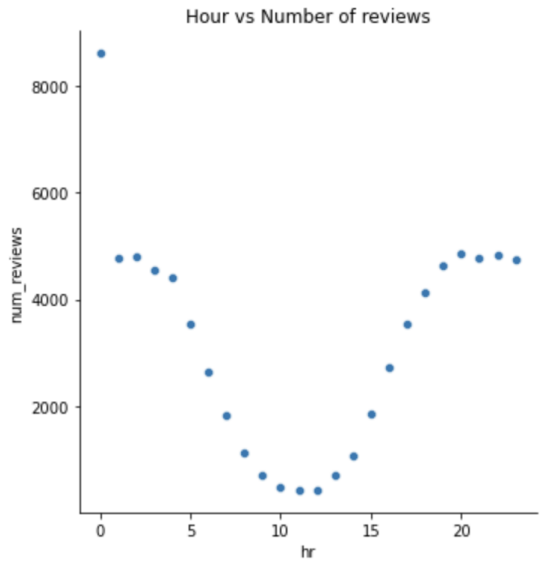
- Rating Distribution:
  From the figure distribution of ratings for business places we can see 49.8 % of all ratings are 5 and very few for 0,1,2.



- Top 10 Categories:
  From 847 total categories in the review dataset, the top 10 categories are plotted, and observed American Restaurants are more popular than other categories.

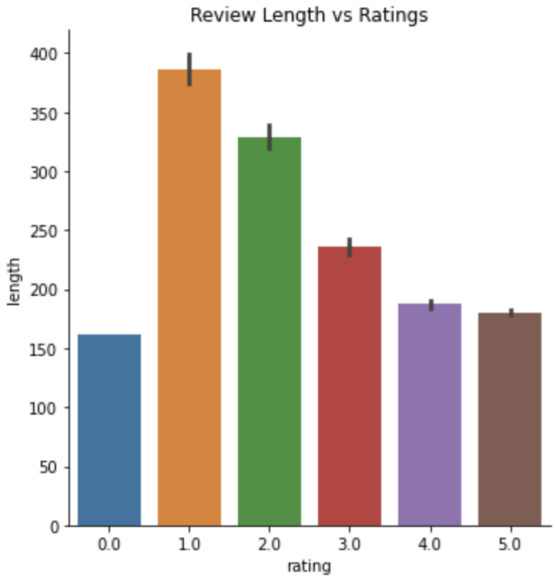


- Review Hour vs Ratings:
  We have observed many reviews were posted at midnight and the lowest no of reviews posted during the middle of the day around 12-3 PM.

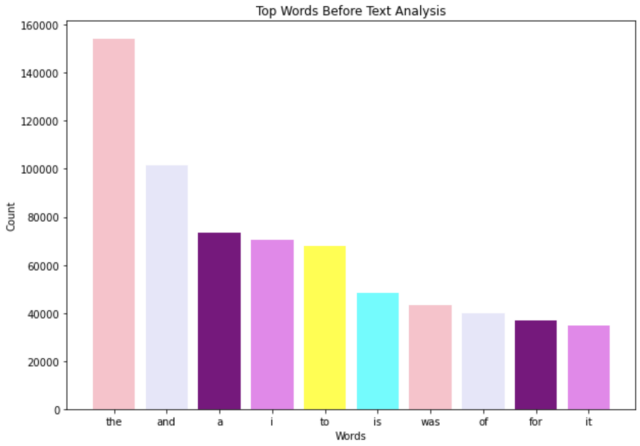


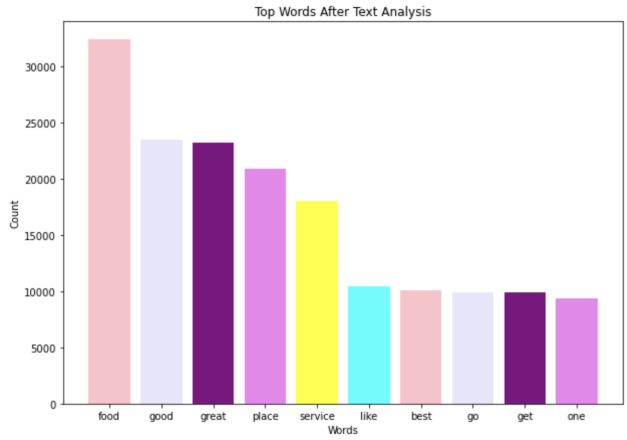
- Price Rank Analysis:
  We have observed that ratings were higher for businesses with low prices (1) and average ratings least for places with high prices.
  It is also clear that at hours 7-10 AM, ratings posted are deficient in number.

- Review Length vs Rating:
  People giving low ratings are writing length reviews around 300-350 words. Majority of the people are writing in the range of 50-150 words.



- Text Analysis:
  Before Performing text analysis the most popular three words are the , and, a.
- After Performing text analysis the most Popular three words are Food, Good and Great.





III. Experiments and Evaluation

A. Dataset Description

Google's local reviews dataset has reviews of businesses all over the world. It contains three distinct datasets Places Data with a size of 276MB includes 3,116,785 local business, User Data with the size of 178MB has 4,567,431 users and Review Data has 11,453,845 reviews with the size of 1.4GB. Dataset has reviews in different languages and places all over the world. For our project, we have chosen the region of interest as California and the Language as English. After filtering with these requirements, we have 4,000,000 reviews and 48865 local businesses.

- Link to Dataset

http://cseweb.ucsd.edu/~jmcauley/datasets.html#google_local

B. Preprocessing Decisions

We have both numerical and text representation in our dataset.

1) Data Cleaning:

- Filtering California places from Places Data file by giving GPS coordinates Latitude and Longitude values as 32.32, 42, -124.26,-114.8, respectively.
- From California Data dropping all the instances with empty price and GPS.
- From reviews data file filtering, all the values with gPlusPlaceId add the features price and GPS from California data set to review data set.
- Dropping all the non-empty and null instances from the review data file.
- After filtering the data, we have 48865 places and 76254 reviews.

2) Preprocessing for Text Data:

- For the text data in reviews removed capitalization, all the stop words in English and punctuation.
- Before filtering, we had a total count of 60554 unique words where the and a are the top three famous words. After the filtering, we had a total count of 60415 unique words. We have considered the Top 1500 bag of words. Where food, good and great are the top three famous words.
- Created a dictionary with word and its count as key value pair and transformed to a set.

3) Transformation of Data: After data extraction and cleaning we want to add features review length, price and hour.

- We have calculated the length of review and added review length to data set.
- Converted the Unix review time to data time format by using pandas date time and obtained hour of review.
- Encoded the price level to numerical values and added to the data set.

C. Evaluation methodology

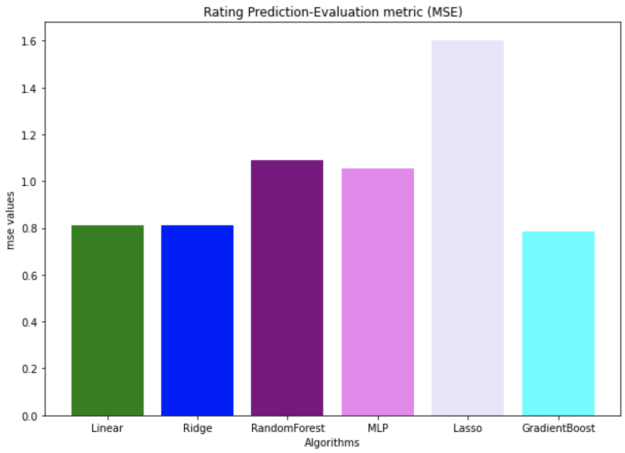
We have split the data set to training set, validation set and test set in the ratio of 70:15:15.

For evaluating the performance imported mean square error,root mean square error and mean absolute error from sklearn.

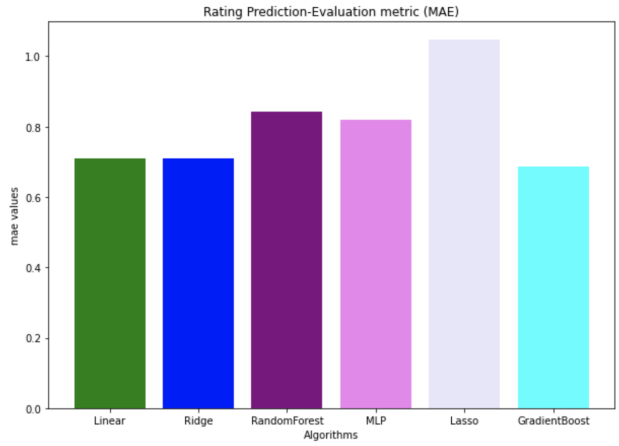
- Mean Square Error(MSE): It squares the difference between original and predicted value and calculates the average.
- Mean Absolute Error(MAE): MAE averages the difference between actual predicted value.
- Root Mean Square Error(RMSE): RMSE is the square root of mean square error.

D. Evaluations Comparison

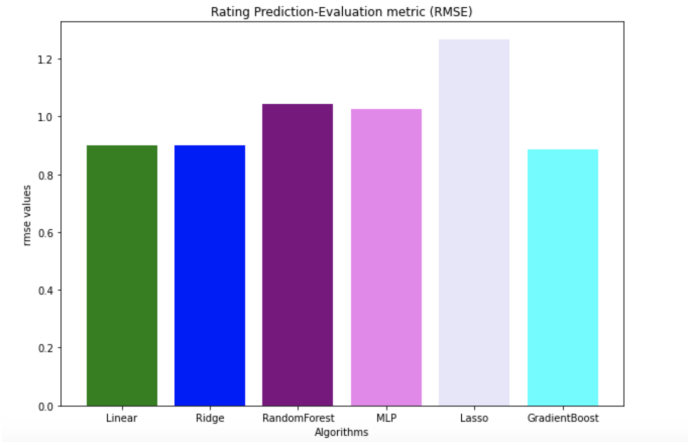
- Scenario 4 Performance evaluation for regression algorithms using Error Metric(MSE)



- Scenario 4 Performance evaluation for regression algorithms using Error Metric(MAE)



- Scenario 4 Performance evaluation for regression algorithms using Error Metric(RMSE)



E. Results Analysis

- In scenario1, Bag of 1500 words is passed as a feature. Machine learning algorithms like Linear regression, Ridge and LASSO regression, Random Forest, MLP, Gradient Boost regressor are applied. Models are evaluated on error metrics such as MAE, MSE, RMSE. We can observe the scenario one results from the table given below; MLP regressor is the best working model when considering only text analysis.

Scenario 1

ALGORITHMS	RMSE	MSE	MAE
LINEAR REGRESSION	0.92577	0.85706	0.72515
RIDGE REGRESSION	0.92555	0.85666	0.72504
RANDOM FOREST REGRESSION	1.13506	1.28837	0.91248
MLP REGRESSION	0.90927	0.82835	0.70898
LASSO REGRESSION	1.22711	1.50579	1.05256
GRADIENT BOOST REGRESSION	0.91091	0.82976	0.70882

- The features used for scenario 2 are the length of review text, the time when the review was posted, Price Value additional to Bag of the top 1500 Words. When we compare the results from the scenario 2 table, it is evident that Gradient boost regressor is the top-performing model as it's RMSE (0.9087), MAE(0.6969), MSE(0.8076). The second best is Multilayer perceptron as per the result in scenario 2.

Scenario 2

ALGORITHMS	RMSE	MSE	MAE
LINEAR REGRESSION	0.92437	0.85446	0.72387
RIDGE REGRESSION	0.92414	0.85404	0.72376
RANDOM FOREST REGRESSION	1.13505	1.28834	0.91248
MLP REGRESSION	0.91041	0.82885	0.70126
LASSO REGRESSION	1.34261	1.80262	1.11610
GRADIENT BOOST REGRESSION	0.90870	0.80766	0.69692

- In scenario3, where features are positive sentiments, negative sentiments, hour, price level, and review length, the best performing model is again Gradient boost regressor with an RMSE(1.0478), MSE(1.0980), MAE(0.8416) scores.

Scenario 3			
ALGORITHMS	RMSE	MSE	MAE
LINEAR REGRESSION	1.07680	1.15951	0.87438
RIDGE REGRESSION	1.07685	1.15961	0.87428
RANDOM FOREST REGRESSION	1.05390	1.11071	0.85242
MLP REGRESSION	1.06969	1.14423	0.87002
LASSO REGRESSION	1.01359	1.29034	0.94791
GRADIENT BOOST REGRESSION	1.04785	1.098022	0.841645

- The fourth scenario is an advanced model where features are features Bag of 1500 words, length of review, an hour of review, ordinal encoded price values, along with sentiment analysis to all the review text. It is indispensable from Scenario 4 Table Gradient boosting outruns the other models and considered the winning model for predicting the ratings of google local reviews.

Scenario 4			
ALGORITHMS	RMSE	MSE	MAE
LINEAR REGRESSION	0.90040	0.81072	0.70972
RIDGE REGRESSION	0.90019	0.81034	0.70960
RANDOM FOREST REGRESSION	1.04410	1.09016	0.84350
MLP REGRESSION	1.02647	1.05364	0.81820
LASSO REGRESSION	1.264842	1.59978	1.04597
GRADIENT BOOST REGRESSION	0.887106	0.78695	0.68551

IV. Discussions & Conclusions

Best Decisions

- By combining sentiment analysis to Bag of words, and added features, review length, price rank, and hour of review gave best results.
- Considering different regression methods for Scenario 4, Gradient Boost Regression performed best.

Difficulties in Implementation

- Due to huge dataset it took nearly 7 hours for initial loading.
- We tried to implement polynomial regression but the Kernel was crashing.Besides, SVR regression didn't give acceptable RMSE scores.
- Splitting the sentences for analyzing sentiments considering different types of sentences was overwhelming.

Conclusion

- Considering variant features we observed interesting patterns in data like low price places are getting high reviews, and reviews are being posted are more at mid night.
- As per our analysis we learned review text is performing prominent role in rating prediction rather than considering individual features like length of review text,hour of review and price values.
- As a future work we can further improve this model by building recommendations for a given business place considering the best scenario implemented for rating prediction.

V. Task Contribution

NAME	TASKS
Sreeja Madanambetti	Data Analysis, Data Transformation, Performed Sentiment Analysis on review text and applied all regression models, Evaluation of Results, Project Report and Presentation.
Xuan Shi	Data Analysis, Data Cleaning , Performed Text Analysis and extracted additional features and applied all regression models, Evaluation of Results, Project Report and Presentation.
Yamini Aalla	Data Analysis, Data Cleaning and Visualization, Developed advanced model by applying all regression model, Evaluation of Results, Project Report and Presentation.

Git Hub Source Code

- https://github.com/sreeja3003/256-Project

References

[1] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar, "Building a sentiment summarizer for local service reviews," 2008.

[2] L. Qu, G. Ifrim, and G. Weikum, "The bag-of-opinions method for review rating prediction from sparse text patterns," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 913–921, 2010.

[3] L. Melkumova and S. Y. Shatskikh, "Comparing ridge and lasso estimators for data analysis," *Procedia engineering*, vol. 201, pp. 746–755, 2017.

[4] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 809–821, 2015.

[5] G. Ridgeway, "The state of boosting," *Computing Science and Statistics*, pp. 172–181, 1999.