**CMPE 257 Machine Learning**

**Project title- “Business reviews and Data analysis using Machine learning on yelp.”**

**Authors- Harshika Shrivastava**

**Tirupati Venkata Sri Sai Rama Raju Penmatsa**

**Sanika Vijaykumar Karwa**

**Swapna Kotha**

***Abstract***

**Feedback and reviews are assets for any successful business, it helps in increasing sales and also provide scope for improvement. In this project, we have tried to utilize machine learning models to get an understanding of business reviews with the Yelp dataset. We have tried to implement various models for the prediction of ratings, a user will give, based on reviews given by the user. We then tried to implement different models for the same task to understand how various algorithms such as linear models vs other models work. On top of that, we tried to increase the efficiency of the models by experimenting with datasets and by applying various techniques in the project.**

***Keywords: Machine learning models, Yelp dataset, review analysis***

# **Introduction**

In this project, we have implemented various techniques to predict the ratings a user will give according to the reviews given.We have also implemented other features like sentiment analysis, False review detection, and category prediction based on reviews. The main purpose of this project is to understand how learning works. We have implemented data preprocessing techniques to analyze the Yelp dataset[1] and to understand the pattern in the data. Also, in order to check whether learning is feasible or not we have tried to analyze various patterns in the data. The main purpose of this project is to implement a two-step learning approach that we have learned in class and to apply theoretical knowledge to experiments. We have implemented various algorithms such as decision trees, random forests, and gradient-boosting machines to obtain the result. Through this project, we learned many practical concepts of machine learning like preprocessing data, pattern analysis, implementing different models, and evaluating them. This project served as a stepping stone for us to enhance our analytical skills because it required an intense understanding of the problem and thinking creatively to solve it. Moreover, in order to implement various machine-learning algorithms we had learned many Python libraries and machine-learning concepts.

The project has given us an in-depth understanding of machine learning and its applications. It is also interesting to learn about how a simple task such as prediction can be implemented using various algorithms and techniques. This knowledge will certainly help us gain more insights into business analytics, where the most important aspect is to understand customer behavior.

# **About Dataset**

The data that we used comes from the Yelp Dataset Challenge [1]. It is a subset of Yelp’s businesses, reviews and user data that has been made publicly available for different uses. This dataset contains 5 JSON files namely, business, checkins, reviews, tip and user.

business.json - Contains business data along with location data, attributes and categories.

Review.json - Contains full review text data with the user\_id that wrote the review and the business\_id for whom the review is written.

User.json - Contains user data including the user’s friend mapping and all the other metadata that is associated with the user.

Checkin.json - Contains check-ins on a business

Tip.json - Contains tips written by a user on a business. They are written so as to convery fast suggestions and are generally shorter than reviews.

For our implementation, we have converted the JSON files to csv files. For few additional tasks, we have also taken a subset of data to do additional tasks i.e we have used Californian Restaurants only for the implementation of our models.

# **Method**

In this section, we are going to discuss various concepts that we need to understand in the implementation of the project.

1. **Data cleaning**

First, the files are in json format which are then converted into csv files for further use. Secondly, the null values present in the dataset are fairly insignificant in comparison to the actual data, so they have been removed giving a clean dataset. Thirdly, the top categories of the business and their reviews have been merged. This is done using the sqldf library which easily lets us import the dataframe into the database and allows us to do operations based on simple SQL statements. It lets us query any pandas Dataframe using SQL statements. Lastly, the CSV files are combined using the Business ID as an unique identifier key tpo combine the business dataset with their respective reviews.

For easier processing of models, we have used a smaller subset of data i.e taking the business classified as “Restaurants” in state “California”.

1. **Pre processing**

TF-IDF Transformer - This transformer stands for term frequency - inverse document frequency. TF-IDF is a measure which evaluates how important a word is to a document in collection of documents. This is done by looking at how many times a word appears in our document while also being attentive to how many times the same word appears in other documents which is the inverse document frequency. Hence, TF-IDF is a score which is applied to every word in every document.

Count Vectorizer - As machines do not understand characters and words, we need to convert them into numbers to be understood. CountVectorizer is a method of converting textual data to numerical data. It converts collection of text documents to a vector of term/token counts. It also enables preprocessing of text data prior to generating vector representation.

Stemming - This is a technique which is used to extract the base form of words by removing affixes from them. It is used to process unstructured text and simplifies the task of analyzing the processed task. It used the stem of the word. We have tried using stemming in our project bu did not continue using it as it was affecting the performance of the model.

Lemmatization - This technique uses the meaning behind the words to extract the lemma. To do this, it has a detailed dictionary which the algorithm can look through to link the form back to its lemma. It is necessary to look at the morphological analysis of each word.

1. **Improving models by data balancing**

Imbalanced data is the type of data which has uneven number of distributions in the target dataset. It causes a severe skew in the class distribution and this bias can influence many algorithms excluding the minority class entirely. In our dataset, the imbalance is caused by number of reviews wherein the reviews “5” is the majority class.

To address this problem, we can randomly resample the training dataset. The two main approaches for the same are oversampling and undersampling for imbalance classification. Random oversampling involve duplicating examples from minority class and adding them to training dataset and random undersampling involve deleting examples from the majority class and deleting from training dataset.

# **Implementation and Example analysis**

**The functionality of the project** - Our project has following functions.

*1. Star rating prediction*- In this function, we are training our models on yelp dataset.

2.*Tagging prediction-* In this function, if any review is given as user input, we predict the category of business it belongs to.

3*. Sentiment analysis*- In this function, we are classifying the sentiments of the user based on the review given about any business.

Sentiment analysis:

We initially started with sentiment analysis to get an overview of review classification into two main categories which are positive and negative sentiments. The reviews which are loaded from the JSON file are loaded into a data frame. For sentiment analysis, we have only used the two attributes of the review\_text and review\_start. The new attribute label is added to the data frame according to the star rating. The reviews with a star rating above 3 are labeled as 1 and the reviews below three are labeled as 0. Then the text reviews are converted to a vector on the frequency of the occurrence of each word using count vectorization while tokenizing the words both unigrams and bigrams in order to perverse the sentiment of the review. The model is trained by using logistic regression with liblinear analyzer The model is able to predict the sentiment with an accuracy of 93.7%. We also extended the sentiment analysis model with three sentiments positive sentiment, neutral sentiment, and negative sentiment. The reviews with a star rating above 3 are labeled as 2 and the reviews below three are labeled as 1 and reviews with a rating equal to 3 are labeled as 0. This model is able to predict the sentiment with an accuracy of 86.6%.

# **Result comparisions**

**A.Comparing various models for Rating Prediction:**

We have trained various models for the task of rating prediction based on the review given by the user. The following results show the comparison between various classification models for the same task.

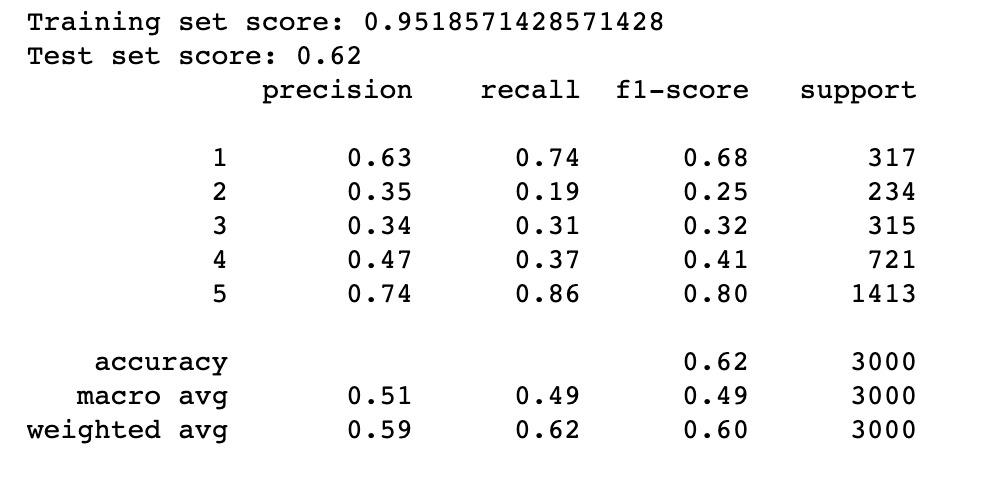
**Using Full dataset:**

1. Linear SVC

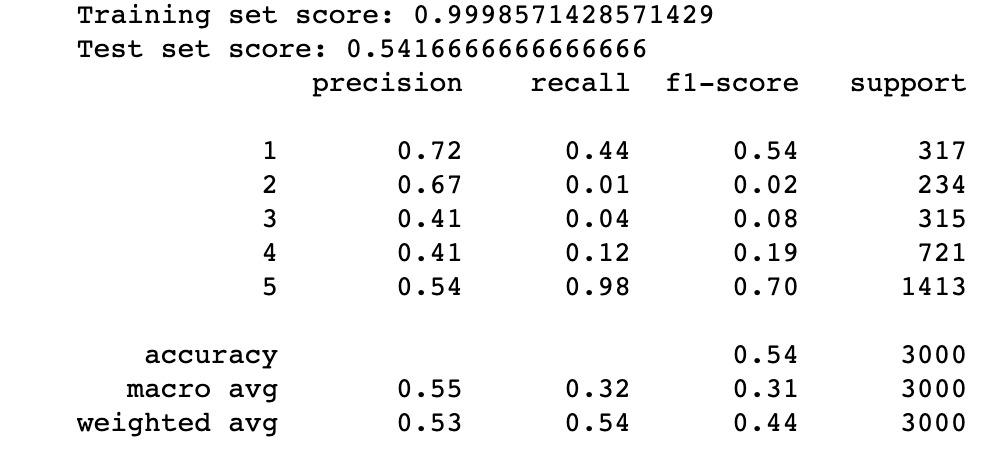


**Using Half dataset:**

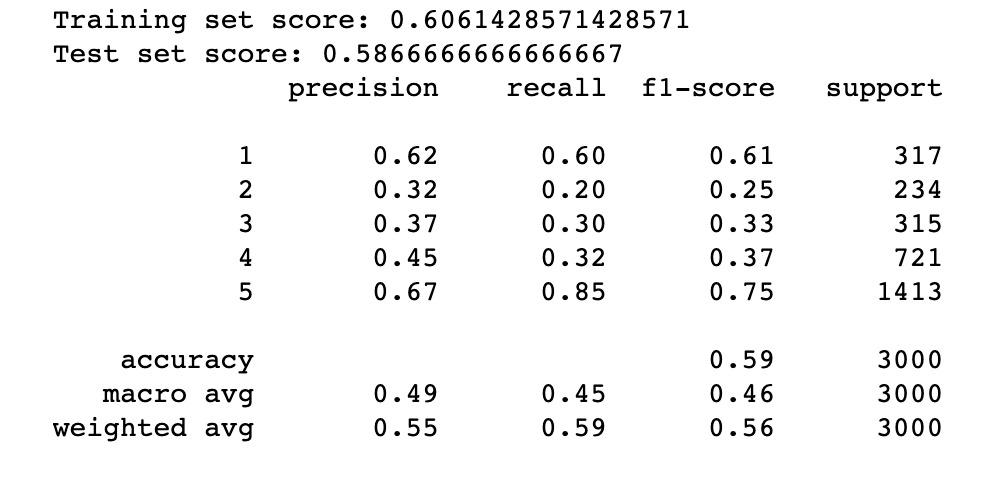
1. Linear SVC



2. Random Forest

****

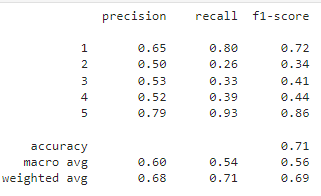
3. AdaBoost



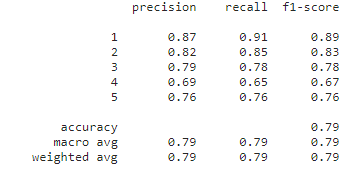
**B. Improving Rating prediction models:** One of the major concerns our models have faced was dealing with imbalanced data. As our models were getting trained on imbalanced data, they were not performing efficiently.

Below results show comparison between F1 score of various classes, before balancing and after balancing.

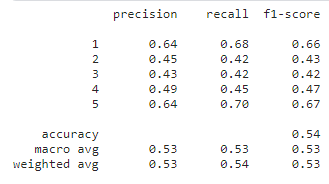
1. F1 score and precision before Balancing data.



1. F1 score and precision after performing data balancing using Oversampling.

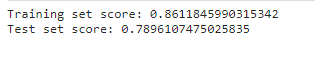


1. F1 score and precision after performing data balancing using undersampling



When we implemented balancing the imbalance data, we used various methods for performing sampling of data. Below is the result comparison. Test scores for oversampling and undersampling for the LinearSVC model is shown below.

1. For oversampling using Random oversampling

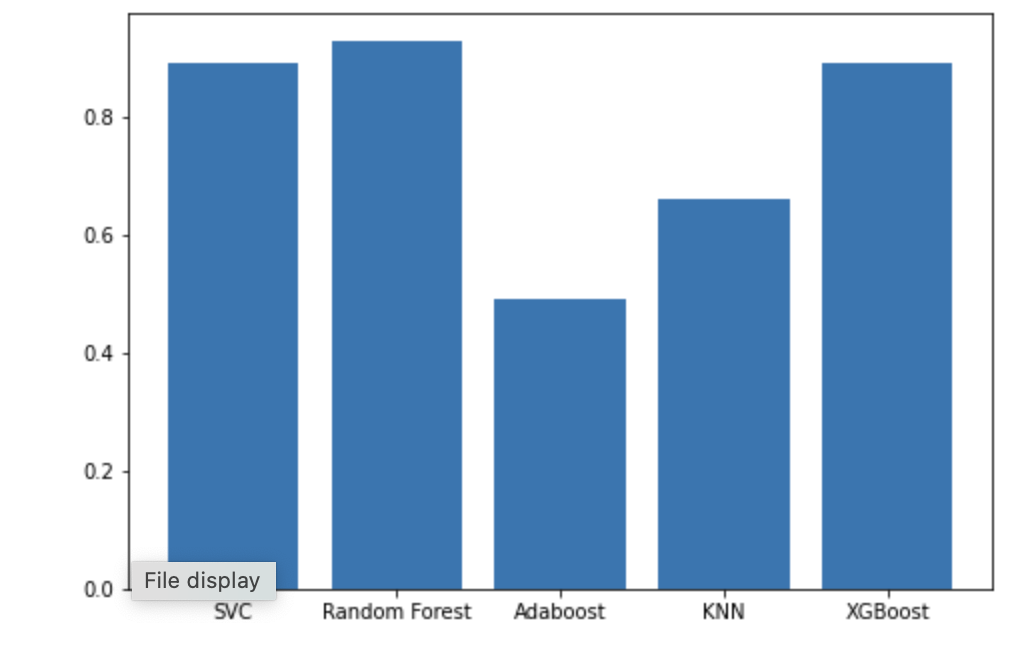


1. For undersampling using Random undersampling



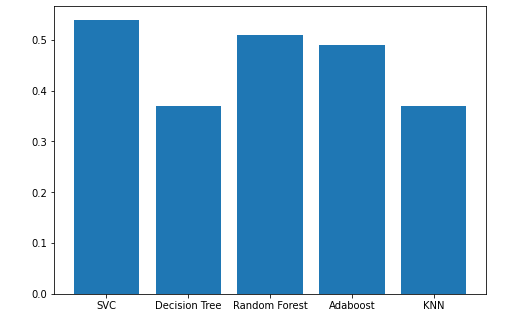
From the above scores, we can see that for linear models oversampling is good approach than undersampling.

**C.** **Comparing various models for Rating Prediction using Oversampling:**

****

**D. Comparing various models for Rating Prediction using undersampling:**

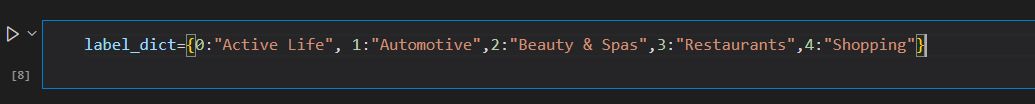
Here, we are comparing various methods with undersampling.



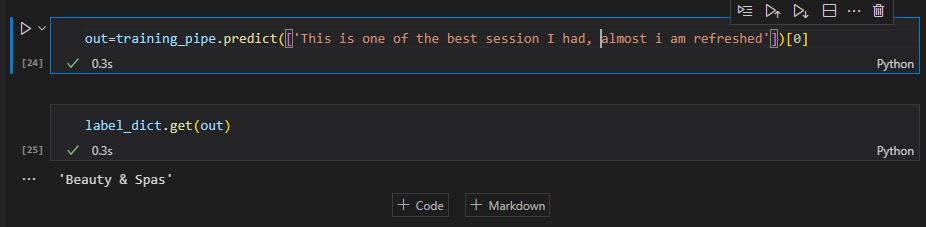
Fromthe above results, we can compare various models on the x axis we have various models and y axis shows F1scores.

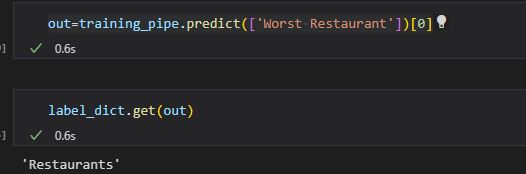
**E.Tagging model results:**

When checking results for tagging model, we have divided our categories into 5 categories which are very different from one another as shown below-



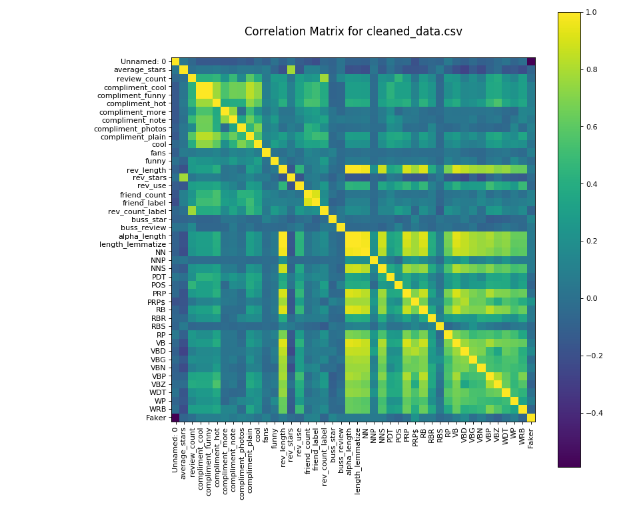
The following figures show how our model is able to predict the categories, based on the review given in the text input.

****

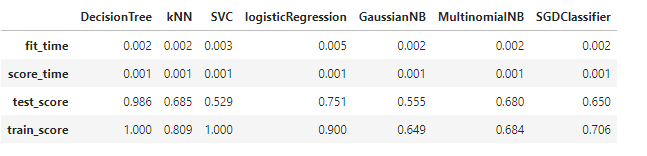
****

**F. False review Detection:** When we tried to understand the Yelp Dataset, we realized that many of the businesses have fake reviews as well. We have implemented various classification algorithms to classify whether the review is fake or real.

Below is the correlation matrix for the yelp dataset.

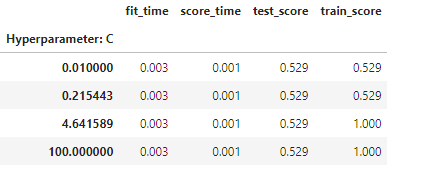


Below, the table shows the comparison of various algorithms for the task of classification on the same dataset. Table.. shows the Fit\_time comparison which is based on the time taken by the model for training on the training data. Train and test score shows, how our models have performed on training data and test data.



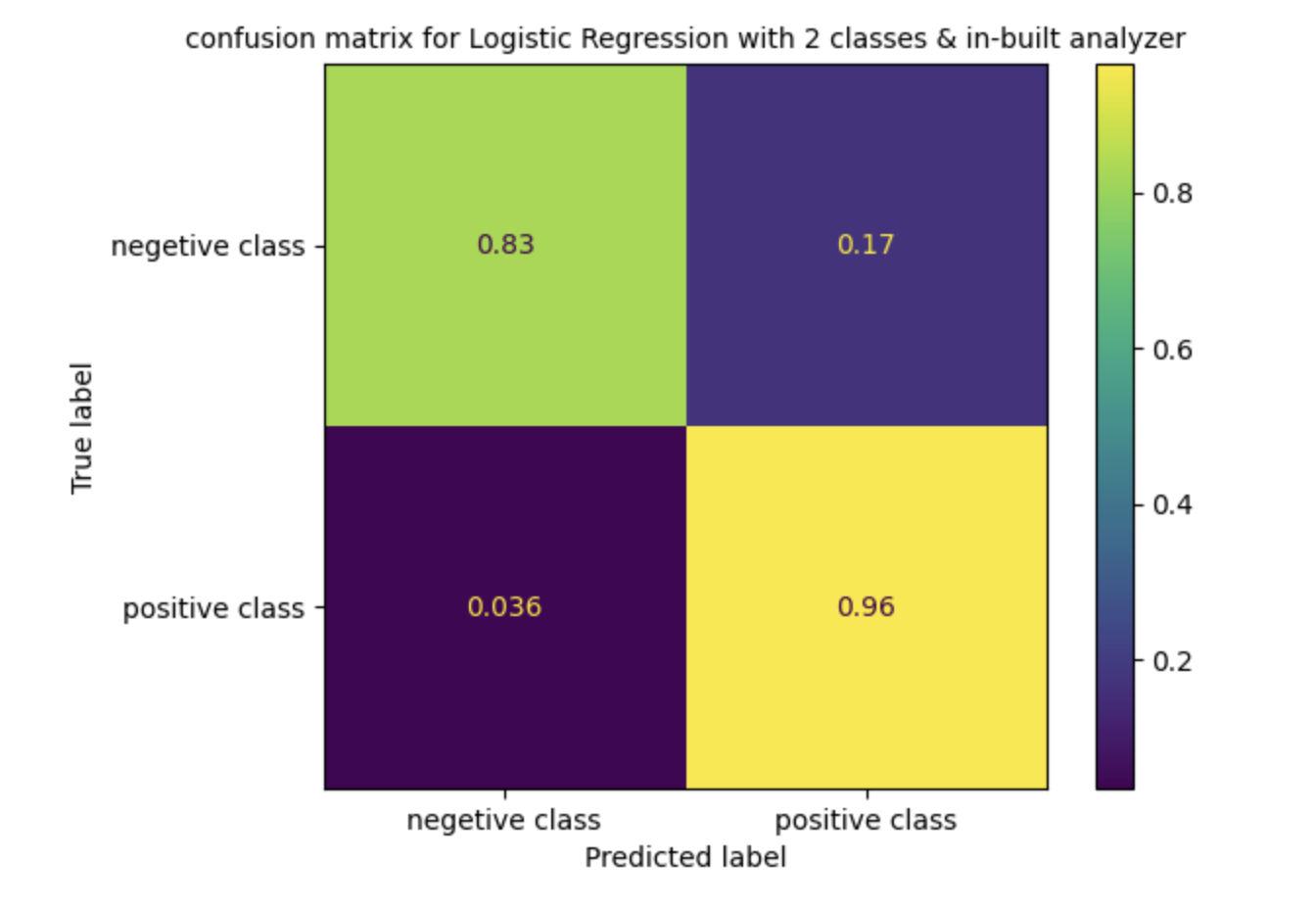
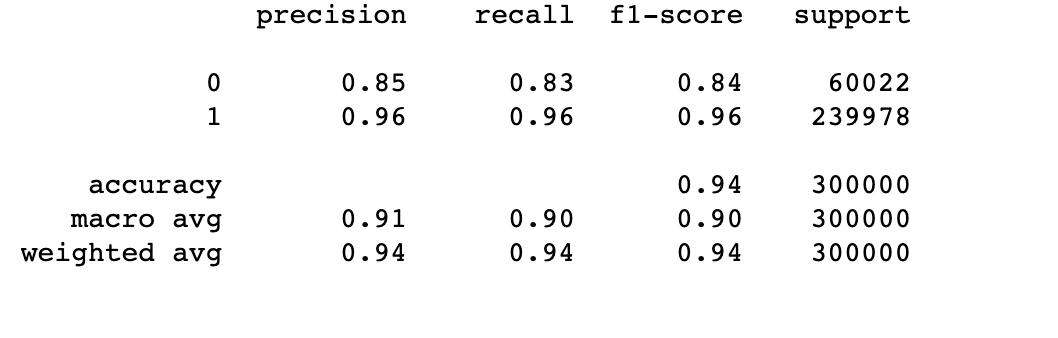
As from the table above it seems that the SVC model (support vector machine) is not giving very good accuracy. As the goal of this project is to understand linear models, we

also tried to tune the hyperparameters such as gamma C in the model and the below table represents various scores of the model performance.



**G. Sentiment Analysis**

Trained a model using regression sentiment prediction based on the review given by the user. The following shows the result metrics and the prediction for the sample test data



# **Conclusion**

After implementing various techniques we can conclude the following things about Yelp dataset and machine learning.

1. Do you have enough data to learn?

We have used yelp dataset for california restaurants and we trained our model on that and we were getting good prediction results with that. Further, we divided our dataset into half and trained our model on this half dataset.

In second case also we are getting good accuracy in prediction and hence we can say, if we are able to predict even with half dataset correctly, We have enough data to give good results.

2. Are you implementing Two step learning process?

Ans- Our project has models which are learning in a two step process.

From below scores for Linear SVC model, we can say that our Ein -> 0 as our training accuracy is 97%. And also, as our training and test scores are nearby we can conclude that our Eout -> Ein. Hence, we can say we have learned in a two step process successfully and learning is feasible for us because we can utilize that in getting good prediction.



3. How have you improved your models?

Ans- We have trained various models and compared them, we experimented with features and correlation matrix to improve efficiency. We then perform Data preprocessing again and experimented with different samples of data. We also used data balancing techniques to improve the model.

# **Appendix**

Project Demo link- <https://rajuptvs-cmpe-257-projectteam12-streamlit-deploy-test-mezmgu.streamlit.app/>

Clean Datasets- <https://drive.google.com/drive/folders/1ZZQRmJ9G19Y1m9oQK2_bPa39LfepY1hf?usp=sharing>

Sourcecode link- <https://github.com/harshika14/CMPE-257_ProjectTeam12>

# **References**

[1] Yelp Dataset- (link: <https://www.yelp.com/dataset>)