Analysis on Yelp Dataset

By Amar Shivaram Pallassana Gopalakrishnan

Datasets Used

The datasets used were the following ones:

- yelp_academic_dataset_business
- yelp_academic_dataset_review

Formulated Usecases

1. Check if the restaurant is open or not
2. Sentiment Analysis
   a) Check the sentiments of the review
   b) Intensity based sentiments of the review
3. Predict the rating of the given review

Usecase 1

- The usecase makes use of the yelp_business dataset.
- The target attribute is *is_open*.
- The columns which can contribute to the opening of a restaurant are filtered
- There is an imbalance in the target attribute and hence appropriate sampling is done with SMOTE before running the model.
- Logistic regression is run for various parameters.
- The maximum accuracy obtained is around 55%.

Other Usecases

Usecase 2

- This usecase is of 2 parts
  - Checking a review is a positive review,negative review or neutral review
  - Checking the intensity of the sentiments.
- The yelp_review dataset is taken and the *text* attribute is used as reviews.
- **TextBlob** package is used for checking the sentiments and **Vader** package is used to check the intensity of sentiments.
- Before passing the review into the package for checking sentiments, appropriate text cleaning is done to remove stopwords, punctuations, numbers etc.

Other Usecases

Usecase 3

- The usecase uses the yelp_review dataset with the attributes *text and stars* to predict the rating based on the review text.
- We treat this as a binary classification and hence we take only ratings, 1 and 5 and then continue the process.
- We construct models using SVM, Naïve Bayes, Logistic regression.
- Appropriate hyperparameter tuning is done through *gridsearch command in sklearn*.
- The appropriate parameters are chosen and the model is built and evaluated.
- We conclude that support vector machine provides us with a maximum accuracy

Other Usecases

Concepts implemented

- Exploratory data analysis
- Sampling with SMOTE
- Text cleaning
- List comprehension
- Hyperparameter tuning with *gridsearch*
- Model selection and evaluation.

Thank you