CSC 180 Section 1 Project 1

Project 1: Yelp Business Rating Prediction using TensorFlow

Alexei Godfray - 302898713

Suyesh Shrestha – 303861311

Due at 10:30 am, Monday September 29, 2025

1) Problem Statement
    a) Our problem statement was to find a Fully Connected Neural Network model that would be able to predict a business's average star rating on yelp, through its reviews.
    b) The dataset we used was the Yelp Dataset provided by our instructor, and our focus was on the business Json dataset's review count, and the review Json dataset's review text.
    c)
2) Methodology
    a) Once we had the datasets, business, and review, we filtered out businesses that had less than 20 review counts and kept the ones that reached that threshold. In the review dataset, we cut out the columns that were unnecessary to our model, such as user_id and state. We filtered and manipulated the datasets till the business dataset had only the business id and its star rating, while the review dataset had only the business id and it is review text. Our method of joining these 2 datasets into one was to join them based on their business id's that they shared and had all the ratings' text that belonged to a business be into one "all reviews" field, as raw text.
    b) Once we had the one data frame, we split the data frame into testing and training data, using the train_test_split function.
    c) We then experimented with the training data through many iterations of models, with the testing data being our ground truth.
3) Experimental Results and Analysis

a) Throughout our many models and trainings, we were able to get the RMSE score of our models on the testing data to around 0.47 to 0.42 as the rmse.

b) We experimented with different layer and neuron configurations and implementing different optimizers.

c) We also changed some of the parameters in the Tf-idf stage by changing the max_features parameter and min_df value.

4) Task Division and Project Reflection

a) who is responsible for which part

i) Alexei and Suyesh worked together through discord call to facilitate the notebook with Alexei coding and Suyesh researching the given lab notebooks for code inspiration.

b) challenges your group encountered and how you solved them

i) One of our biggest challenges was to group up the raw text data of the different reviews that belonged to the same business into a single field in a pandas data frame.

c) what you have learned from the project as a team.

i) We learned about how to implement a regression model from a tf-idf vectorized data.

ii) The idea of getting a single continuous value from the combined raw text from many different text samples, was daunting, but once we understood the method of using the fcnn sequential model, and the tf-idf vectorizer.

iii) We learned about the importance of testing data, after we imported more data from the yelp dataset. We were able to get that desired 80-20 split on our data, for training and testing, we were better able to evaluate and test our model to get a more accurate RMSE score for that model.