CS 229 - Project Milestone
Machine Learning for Yelp rating prediction

Frederik Johan Mellbye - frederme
Pranav Bhardwaj - pranavb
Nicolas Bievre - nbievre

November 15, 2019

1 Introduction

Founded in 2004, Yelp is one of the most widely used restaurant and merchant information platforms in the United States. Users are encouraged to contribute to the Yelp community by leaving reviews of their experience with businesses they visit. These reviews are shared with friends and open to the public on the platform. In order to gain insights on this wealth of data Yelp released a data set via the 13^{th} round of its data challenge. The Yelp data set consists of over 6 million reviews of nearly 200,000 businesses from 1.7 million users. This includes business information, user information, and reviews containing star ratings.

2 Goal

We aim to use a variety of different machine learning algorithms to predict the rating a user will give a business. In a first approach, we frame this as a binary classification problem, where a positive label 1 corresponds to a "good" rating (5 stars), and a negative label 0 corresponds to a "bad" rating (< 5 rating).

3 Dataset

3.1 Rating Distribution

We made the decision to frame this problem as a binary classification task based on the distribution of ratings in the dataset:

Rating (Stars)	Proportion
1	15%
2	8%
3	11%
4	22%
5	44%

By aggregating the ratings 1-4 as negative labels and isolating the ratings of 5 as positive labels, we observe a near class balance of 44% positive labels and 56% negative labels. Prediction of positive labels can be thought of as recommendations to users on businesses they will like. This will be the first step to our recommendation model.

3.2 Features

The business entities in the Yelp dataset are described in detail by 54 features going from the category of the business to the category (Restaurant,....) to rather or not a restaurant is appropriate for kids. The Yelp users are described by 22 features. As a first step we only include 21 features to best describe the pair (user, business). This first selection was based on business meaning of the feature but we plan to rigorous feature selection in the following weeks.

3.3 Train / Validation / Test Split

The current train-validation-test split is such that the set of users and businesses in each split is disjoint. This decision was made to avoid selection bias. The size of our data sets following the split is listed below:

	Unique Users	Unique Businesses	Number of Reviews	Five Star Reviews (%)
Train	284,793	39,484	2,476,843	38.3%
Validation	153,642	34,858	296,241	50.1%
Test	269,028	75,904	328,428	54.1%

In the future, we plan to switch to a temporal train-validation-test split, where we observe the first n years of data and train our model. Then we "implement" our recommendation system, and see how our binary classification tasks generalizes to future years data in the test set.

4 Data Processing

We begin by using individual user and business information to predict if the user will give the business a 5 star rating. To do this we aggregate the user, business, and review information provided by Yelp in separate files. The only review file information used is the star rating. The current train-validation-test split is such that of the sets include disjoint users and businesses.

5 Initial Results

5.1 Models

Several models have been used for binary classification task at hand. In this initial phase of modeling, we are more interested in baseline performance, so no hyperparameter tuning or cross-validation was done. We simply trained using our training set and tested with a positive predictive threshold of 0.5 on the test set.

5.2 Results

Below several metrics on the predictions created by the trained models on the validation dataset:

	Logistic Regression	Decision Tree	Random Forest	AdaBoost
Accuracy	49.89%	75.61%	74.46%	76.62%
F1	0.01%	73.37%	71.08%	75.03%
Precision	46.67%	81.00%	82.17%	80.73%
Recall	0.00%	67.05%	62.63%	70.08%

We believe that the logistic regression model performs poorly due to the lack of linear relationship between our predictors and rating. Different users have different preferences, so it is not clear if a predictor like "User review count" should have a positive or negative coefficient.

6 Future Directions

Reasonable accuracy has been achieved using tree based models which can represent nonlinear functions of our parameters. We believe that that performance can be increased by leveraging more information and the community structure within in the data set.

6.1 Business Features

Yelp dataset provides categories and attributes for each business. Examples of categories are "Restaurants" and "Shopping". Examples of attributes are "Parking Outside" and "WiFi". In the current models, we do not use this information as predictors, as there are a very large number of them stored in a format not immediately usable for models.

In the future we will select subsets of all possible categories and attributes in the dataset and one-hot encode them as predictors. The sheer amount of training examples present in the data set supports the addition of hundreds of features. We believe bias will be reduced without much adverse affect to variance.

6.2 Community Structure

The main advantage of the Yelp dataset the community structure in which the reviews of some may influence others. In particular Yelp friendships can play a critical role in the notoriety since some is more likely to like a place that his/her friends liked too. Hence our idea to explore graph Machine Learning.

6.3 Recommendation Model

So far we focused on predicting if a user will like a business or not, but we plan to explore modeling a ranking of businesses for a user in order to have a recommendation system.

7 Contributions

- Data Processing: Frederik Mellbye
- Modeling: Nicolas Bievre, Pranav Bhardwaj
- Exploratory Data Analysis: Nicolas Bievre, Pranav Bhardwaj, Frederik Mellbye