Predicting Restaurant Ratings

IVAN TODOROVIC

OCTOBER 2017

EXECUTIVE SUMMARY

- This project employs a number of machine learning methods to predict restaurant ratings, builds tools for analysis of reviews and draws conclusions from reviews about what users most like and dislike
- The data used in the analysis is from Yelp's education and research data set
- Two distinct methods are used to predict reviews:
  - A restaurants' features (such as meal availability, parking, location, opening hours) are used to predict the average rating given by users
  - The text of user reviews is processed using natural language processing (NLP) techniques and converted to a vector format, which in turn is used to first train and then predict reviews
- The project will build tools that will help answer questions of importance to restaurateurs:
  - The NLP part will enable unrated reviews from any source to be processed to provide a 1-5 star rating prediction
  - The features part will help in understanding what features contribute to restaurants achieving high or low ratings
  - The top and bottom-rated restaurants can be further analyzed for commonly appearing words in order to identify what reviewers like, and conversely, what they most dislike

DESCRIPTION OF THE DATA SET

- The Yelp Dataset is published by the business review service Yelp for academic research and educational purposes – therefore it cannot be used for commercial purposes
- After filtering for restaurants reviews only, the data set includes 52,000 restaurants, with 2.9 million reviews written about them
- The data was further filtered for restaurants in Phoenix, Arizona only, in order to obtain a compact data set with mutually comparable restaurants. Phoenix data contain 3,515 restaurants and 302K reviews
- Two files are of interest in the project:
  - business.json – containing records for individual businesses (i.e. the features)
  - review.json – containing records for users' reviews about businesses (i.e. text of the reviews)

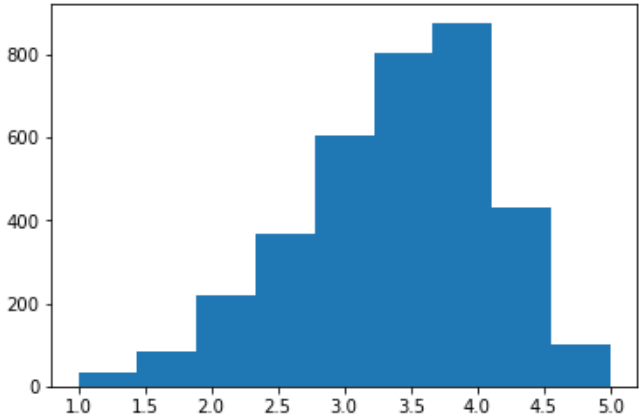
NLP PROCESSING OF REVIEWS

- The first step in the analysis is to perform NLP processing of reviews in order to train a model that can convert the text of a review into a quantitative score prediction, as shown below:

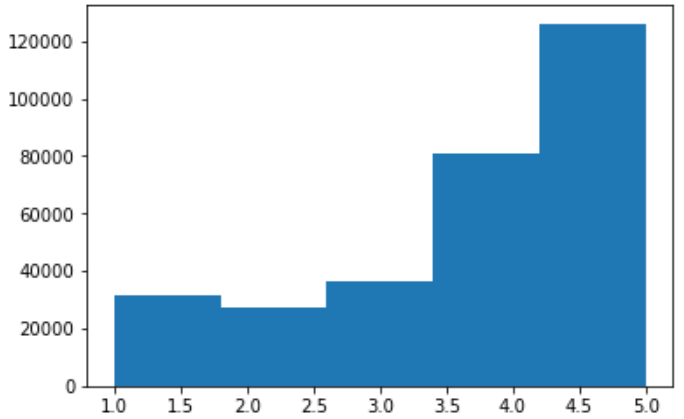
Step	Library / Function used	Description of tasks
1.	spaCy	▪ Sentence detection, tokenization and normalization of the corpus ▪ Converting tokens to lemmatized form
2.	Gensim / Phrases	▪ Detection of bi-grams on first pass and tri-grams on second pass ▪ Building a dictionary of lemmatized tokens
3.	Gensim LDA	▪ Building LDA modeling to identify to clusters representing top 50 ‘topics’ ▪ This enables a review to be summarized by topics
4.	Gensim word2vec	▪ Converting the corpus into a 100-dimensional vector representation of each word in the 80,000 word dictionary based on word embeddings ▪ Thus any text can be represented in the 100-D vector space using the trained word2vec model ▪ This representation is closely associated with the text’s meaning

EXPLORATORY DATA ANALYSIS

Average star ratings for 3,515 restaurants



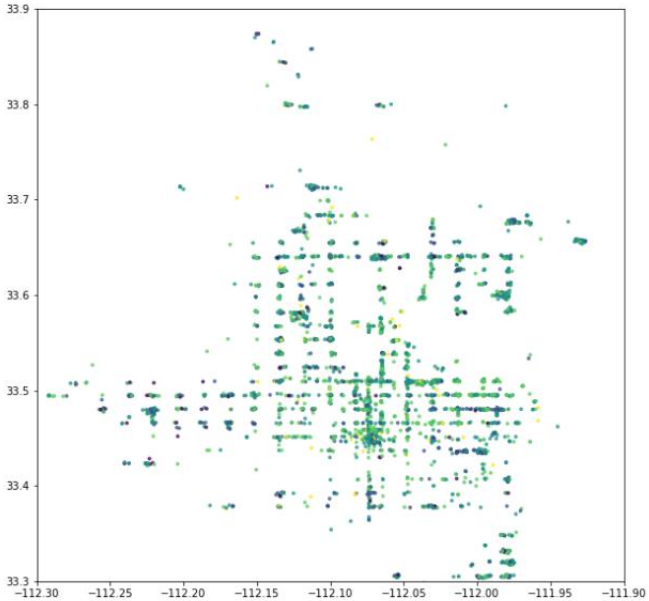
Individual ratings of 302K reviews



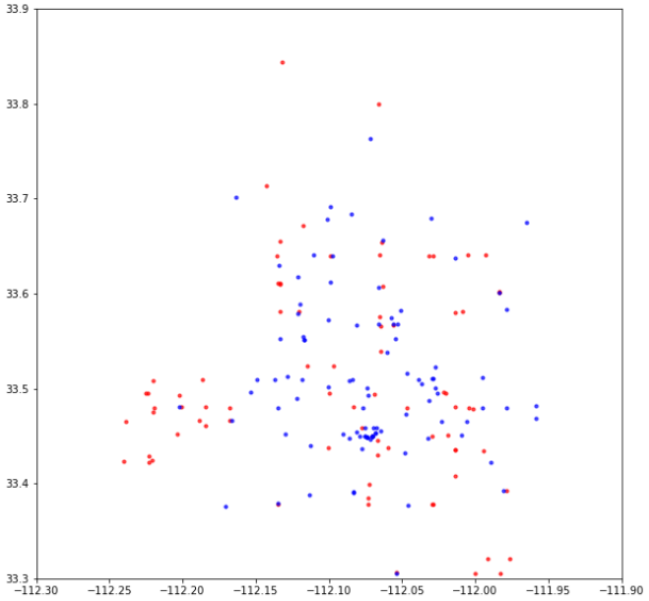
- Interestingly the individual reviews are highly skewed, with 5-star reviews by far the most common.
- Averaging reviews for each restaurant gives the left-hand-side histogram, which is rounded to the nearest half-star. The mean rating across all restaurants is 3.4 stars, so the 5-star reviews get somewhat balanced.

INFLUENCE OF LOCATION ON RATINGS

Location plot for all ratings



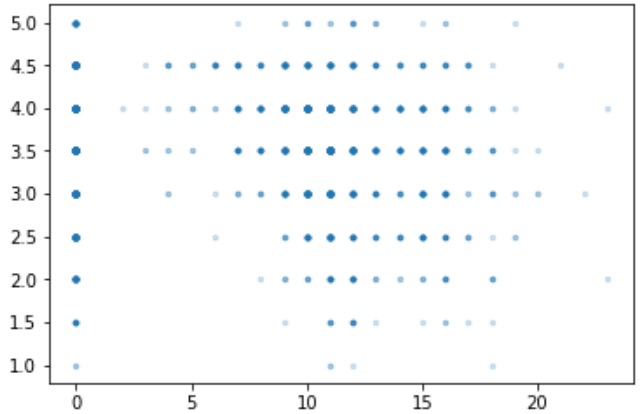
Location plot for 1.5-star (red) and 5-star (blue) ratings



- The best and worst-rated places can be right next to each other. Some patterns can be picked out (e.g. a cluster of 5-stars at -112.06/33.43 and a cluster of 1.5 stars in the lower left section) that may help predictions

INFLUENCE OF OTHER FACTORS

Total hours open on Sunday* vs. ratings



- Best-rated places need not stay open for long
- Poorly rated places are open longer in order to maximize footfall
- The missing data (0 hours) highlights the issue with feature predictions – lack of consistently available information

Alcohol availability and noise level

Category	Mean rating
None	3.38
Beer and wine	3.65
Full bar	3.53

Category	Mean rating
Quiet	3.53
Average	3.51
Loud	3.09
Very loud	2.86

- There is a preference for quieter places and those that serve beer and wine only
- Again the differences in ratings are subtle

* Although data for Sunday is shown here, the chart for any day of the week is largely similar

PREDICTION MODELING BASED ON FEATURES

Regression modeling approach:

- Features as the input parameters, a matrix of 69 features x 3,515 samples (relatively sparse as many values are not quoted), and ratings as the output (i.e. labels)
- Data split into 70% training and 30% testing sets
- R^2 and mean squared error (MSE) used for performance metrics

Prediction models used:

- Ordinary linear regression and ridge regression as linear models
- Random Forest method as a means of dealing with sparse data and non-linearity
- Gradient Boosting as an improvement on the Random Forest
- Multi-layer Perceptron (MLP) as a neural network based model

Results:

Model	R^2 Training Data	R^2 Test Data	MSE Test Data
Linear	0.21	0.20	0.54
Ridge	0.21	0.20	0.54
Random Forest	0.86	0.26	0.50
Gradient Boosting	0.68	0.34	0.45
MLP	0.22	0.22	0.53

- Gradient boosting gives the best performance on test data and does not overfit the training data
- However there is still significant bias in the training data, indicating that features alone are not sufficient for distinguishing good vs. poor ratings
- Reviews will provide the more subjective data that is missing in the features

PREDICTION MODELING BASED ON REVIEWS

Two modeling approaches used:

- Classification as an interpretation of reviews falling into one of five categories
- Regression as a numerical value quantifying the ‘goodness’ of a restaurant from 1 to 5
- The input parameter is a matrix of 100-D vector representation of each review, sampled over 302K reviews
- Data split into 70% training and 30% testing sets

Obtaining the vector representation of a review:

- Loading a review, converting to lemmatized tokens and removing white space, punctuation and stopwords
- Using the trained word2vec model to obtain the 100-D vector for each remaining token and averaging the vector over all tokens in the review

Performance metrics for classification:

- MSE on the test set
- Percentage of accurate predictions
- Percentage of ‘bad’ predictions (those that are off by two or more stars)

Performance metrics for regression:

- R2 on training data
- R2 on test data
- MSE on test data
- MSE on integer-rounded predictions to provide a direct comparison to the classification results

REVIEW-BASED PREDICTION RESULTS

Classification results:

Classifier Model	Accurate predictions	'Bad' predictions	MSE
Random Forest	54.0%	9.9%	1.22
Ada Boost	53.4%	6.7%	1.14
MLP Classifier	57.3%	5.5%	0.88

- The MLP Classifier achieves the best results:
  - Highest % of correct predictions
  - Lowest % of 'bad' predictions
  - Lowest MSE
- In each case the parameters were tuned to achieve the best possible results

Regression results:

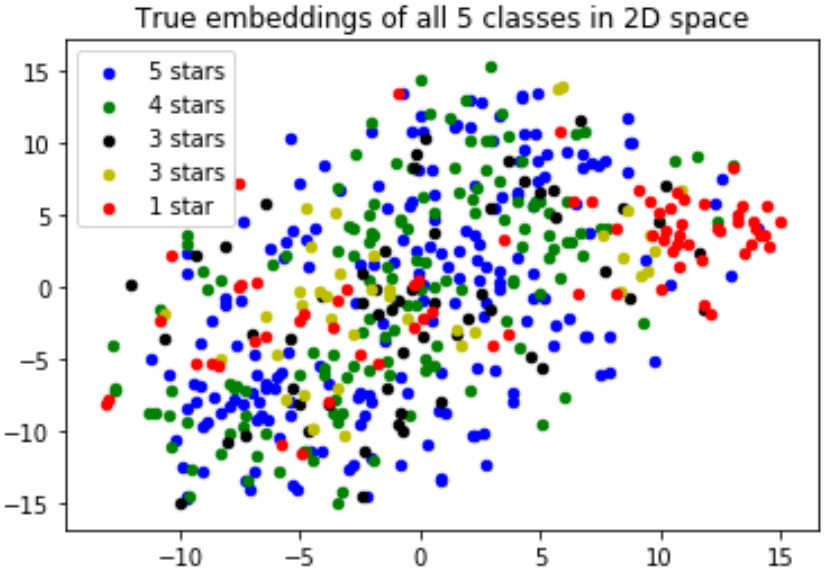
Regression Model	R^2 Training	R^2 Test	MSE	MSE (rounded)
Gradient Boosting	71.3%	61.7%	0.74	0.83
MLP Regressor	69.9%	67.7%	0.62	0.71

- The MLP Regressor achieves the best results
- As there is no overfit, the model also scales well and can be used for any restaurant review
- Regression outperforms classification, possibly due to the ordinal interpretation of labels (i.e. 5 > 1 vs. 'class 5' != 'class 1'
- Overall the MLP Regressor does the best job at interpreting the vector representation into ratings

REVIEW EMBEDDING

Visualizing ratings in 2-D space:

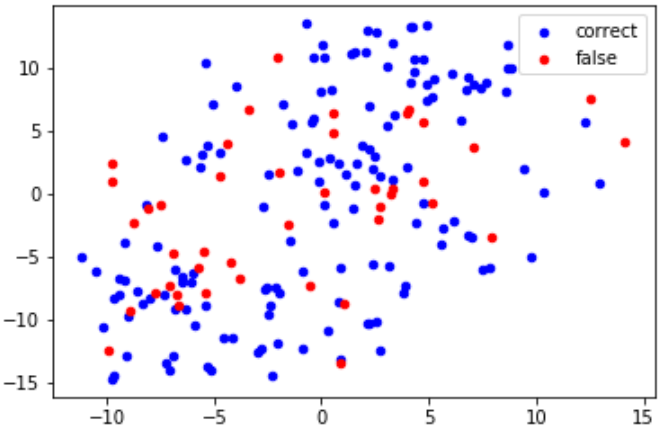
- Dimensionality reduction from 100-D to 2-D space may help visualize the embedding of reviews
- Here a sample of 500 test reviews was reduced to 2-D space using the t-SNE method
- Due to the extremity of the reduction it is difficult to spot patterns on a 2-D graph
- Picking one class at a time may help in understanding how well a methd works



REVIEW EMBEDDING

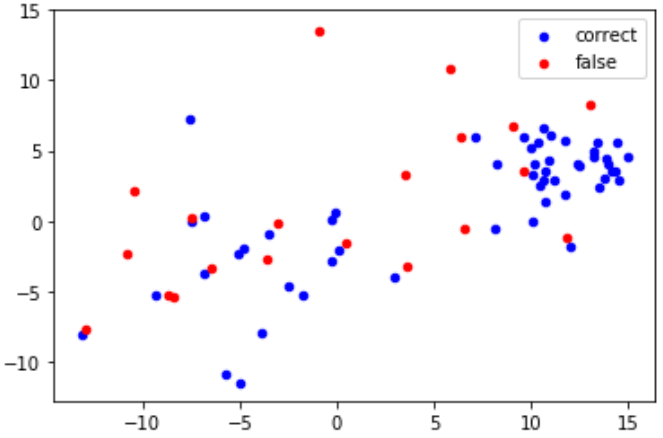
Correct and false classification of 5-star ratings for the MLP Classifier:

- 5-star ratings are not grouped in any particular way in 2-D space. Nevertheless the MLP Classifier does a fairly good job of identifying them
- The MLP Classifier seems to perform equally well at both ends of the spectrum



Correct and false classification of 1-star ratings fpr the MLP Classifier:

- 1-star ratings tend to be clustered in the upper right end of the 2-D chart and most of these are correctly predicted. The classifier does less well with 1-star ratings that are mixed with others



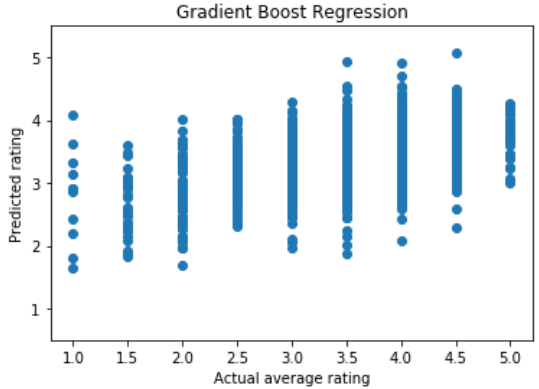
REVIEW VS. FEATURES-BASED PREDICTIONS

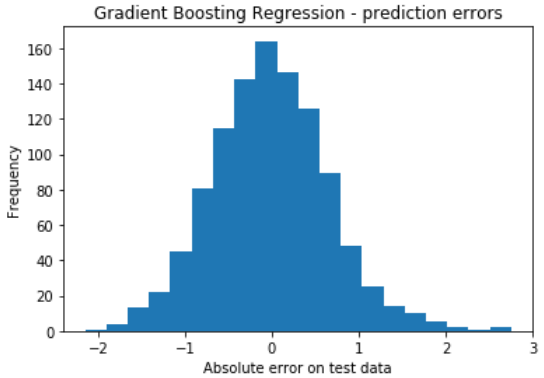
Feature-based predictions:

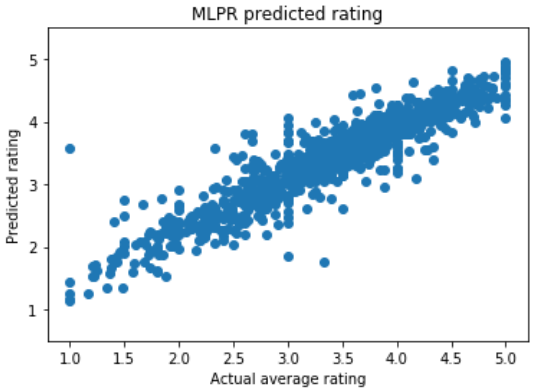
- Vague predictions with only minor differences across the spectrum
- Standard deviation of errors of 0.68 stars
- Lacks the ability to distinguish between good and poor ratings

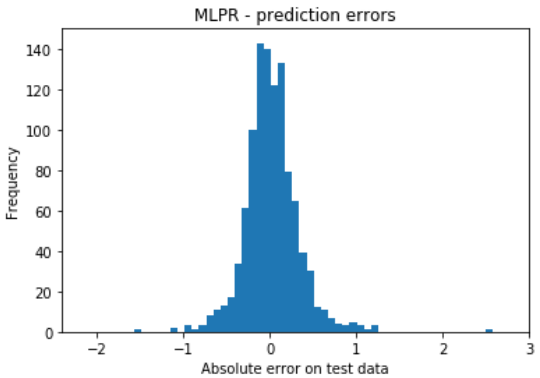
Review-based predictions:

- Ratings for a restaurant are obtained by averaging all reviews about it (on average 86 per place)
- This allows for a far more accuracy, with a standard deviation of errors at 0.31 stars









MOST COMMON WORDS

- Next, the top-rated and bottom-rated restaurants are analyzed for most commonly appearing words
- Results are presented as the words for which there is a large difference in frequency over the sampled reviews
- This enables the identification of items that customers most like and dislike

Topic group	Common words in positive reviews	Common words in negative reviews
Food quality related	Delicious, great, amazing, fresh, flavor, perfect, excellent, fantastic, awesome, tasty, ingredient, homemade, favorite, love	Bad, cold
Service / staff related	Friendly, owner	Order, time, go, ask, get, service, say, minute, come, location, take, table, server, bar, manager, eat, wait, sit, tell, hour, leave, pay, want, kid, experience, $, think
Actual food / drink	Pizza, breakfast, donut, coffee, toast, ice, sandwich, egg, gyro, wine, pancake, soup, crust, roll, cream	Drink, burger, steak, chicken, Mexican, cheese, salsa

CONCLUSIONS

- Two methods are developed for predicting restaurant ratings:
  - A feature-based model, which results in vague predictions with a standard error of 0.68 stars. Hence features do not hold sufficient information about what makes a restaurant receive positive vs. poor ratings
  - A review-based model that uses NLP methods and word embeddings to uncover the ‘meaning’ behind a review. Averaged over a number of reviews for a restaurant, this provides a far more accurate and useful prediction, with a standard error of 0.31 stars
- Review-based analysis can be scaled up to provide a Yelp-like rating on any restaurant based on text, reviews or blog posts written about them
- Restaurants that are consistently rated highly are usually praised for their quality of food (something ‘special’ that customers enjoy) and their friendly service
- Restaurants with consistently low ratings are criticized by reviewers for poor customer service, long waiting times and inadequately prepared food
- Comfort food and drinks such as pizza, donuts, coffee and wine are more often in positive reviews, while bar service and meats are commonly in negative reviews, i.e. more difficult to get right