Capstone Project Final Report

Predicting Yelp Restaurant Ratings

1 Background

This project explores the possibilities to predict restaurant ratings using Yelp reviews. The research is based on Yelp's publicly available data set for education purposes.

The data set includes information on various types of businesses from a number of cities and regions in the U.S. and other countries. In order to perform meaningful research that can be readily interpreted by non-experts, the project examines data specifically about restaurants in Phoenix, Arizona. This filtering was made simply due to the availability of data on Phoenix (3,515 restaurants and 302K reviews about them).

By interpreting the influence of features a restaurant has on Yelp (items such as price level, location, availability of alcohol, parking and ambiance, among others), provides an objective set of criteria to predict what rating the restaurant is likely to receive. This part of the analysis uses regression models to determine 'learn' the relationship between features and ratings on a training set, and then predict ratings for a test data set.

However, ratings given by customers reflect their subjective experience of the restaurant they are reviewing – experiences such as whether they liked the food, how well they were served and how they felt about the price paid for what they got. This part of the analysis uses natural language processing (NLP) methods to process the reviews text and determine the relationship between words/phrases contained in a review and the reviewer's score.

2 Problem and Client

Factors that make a restaurant successful restaurant is often an enigma, even for experienced restaurateurs. What is certain is that customer satisfaction and loyalty are critical components perception and loyalty. When selecting a place to visit, users are increasingly relying on online services such as Yelp as a source of information. It is therefore ever more important for restaurateurs to build their understanding of what constitutes a high rating on Yelp, and what actions they can take to improve their ratings on Yelp.

As the Yelp data set is strictly intended for educational purposes, there client in this project is a hypothetical one. Indeed, the main goals of the project are to:

- Build an understanding of the relationship between restaurant features and ratings

- Develop tools for processing written reviews and inferring information from them, which includes both a general sentiment (positive/negative) and an actual Yelp rating (from 1 to 5 stars). Once developed, this tool can easily be scaled and reused on any written information about any restaurant.
- Draw conclusions about what aspects contribute to having excellent ratings (as well as very poor ratings), by selecting the top-rated and bottom-rated restaurants and identifying the elements that are common to each group.

The answers obtained from the exercise are aimed to help restaurateurs improve their operations in line with what customers expect and value.

3 Data Source and Data Processing

The Yelp Dataset is published by the business review service Yelp for academic research and educational purposes.

After filtering for restaurants, there are approximately 52,000 restaurants with approximately 2.9 million user reviews related to them. Filtering for Phoenix, Arizona only provides a set of approximately 3,515 restaurants and 302K reviews.

The raw data is available in six of files in *.json* format, of which two are relevant for the project:

- **business.json -** the records for individual businesses
- **review.json -** the records for reviews users wrote about businesses

The files are text files (UTF-8) with one json object per line, each one corresponding to an individual data record.

Extracting relevant data and processing for analysis was performed through the following steps:

Features data for regression analysis. The *business.json* file includes features about each individual restaurant, which can be directly converted to a pandas data frame using the *json_normalize* function from the pandas library. This yields 90+ columns that describe the restaurant's price range, meals served, availability of alcohol, ambiance, opening hours, location, parking, number of reviews and average rating rounded to the nearest half-star. The attributes were processed so that all categories are described as numerical values, forming the *X-matrix* of features. There is a large proportion of NaN values, which are replaced by column-wise averages. The averaged ratings are separated into the *y-vector* of target values. The output for this exercise are the *X-matrix* and *y-vector*, split into training and test sets prepared for regression analysis using a variety of methods.

NLP processing of reviews text. The *reviews.json* file holds the information about each review, including the business_id it refers to, text of the review written by the user and star rating given by the

user. Each review from the set of businesses pertaining to restaurants located in Phoenix is loaded into a data frame for analysis.

The analysis itself is performed using a previously trained Word2Vec model, which was obtained through the following steps[1]:

- The spaCy library was used to perform sentence detection, text tokenization and normalization, converting tokens to lemmas, part-of-speech recognition and identification of punctuation, stop-words and white space.
- Next, the gensim library's Phrases function was used to detect bi-gram and tri-gram phrases. A dictionary of over 80,000 words was built from the reviews text. The final corpus including bi-gram and tri-gram phrases was parsed to be stored in lemmatized form, excluding punctuation, white space and stop words but including the words and phrases that bring meaning to the text.
- The gensim Latent Dirichlet Allocation (LDA) model was used to group the lemmatized text data into 50 word clusters, modeling 50 topics most commonly described in the reviews. This gives the possibility to extract topics discussed in any review and thus represent the review text in a much simpler format.
- Finally the gensim Word2Vec function was used to convert the entire corpus into a vector representation of the entire dictionary of 80K+ words. Each word is represented as a vector in a 100-dimensional feature space. This will allow us to model each review into this 100-feature vector space.

All the training was performed using the full set of 2.9 million restaurant reviews available (regardless of location i.e. including the Phoenix reviews) in order to obtain the best possible model. The key outputs for this exercise are

- LDA topic model that allows us to extract the key topics of a review
- Word2Vec model for 80K+ words, that allows us to represent a word/sentence/review in the 100-D vector space

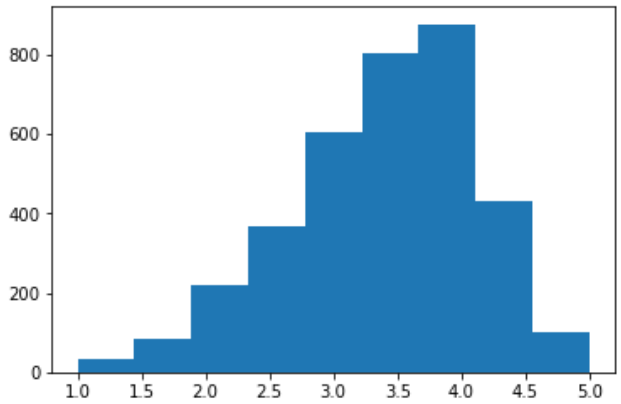
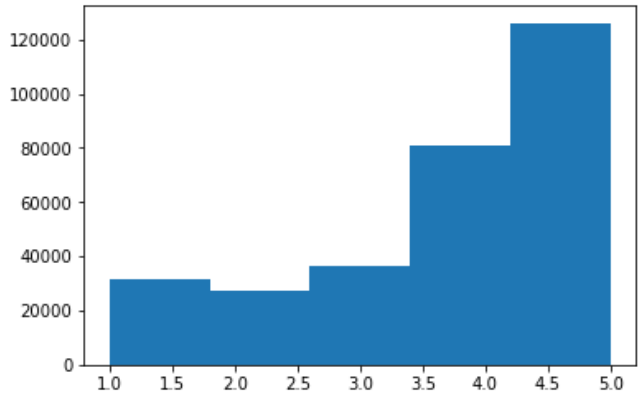
4 Predicting Restaurant Ratings

4.1 Predicting Average Ratings from Features

The features matrix after data cleaning includes 69 features for 3,515 restaurants, and the target ratings for each restaurant are averaged over the reviews written about them and rounded to the nearest half-star.

The distribution of source data in terms of star ratings is shown below:

1 The NLP processing to obtain the Word2Vec model and LDA model were performed following the PyData 2016 tutorial by Patrick Harrison "Modern NLP in Python".

Average star ratings for 3,515 restaurants	Individual ratings of 302K reviews
	

A number of prediction algorithms were tested, and their parameters tuned for optimization:

- Linear Regression
- Ridge Regression
- Random Forest Regression
- Gradient Boosting Regression
- MLP Regression

Each model was tested for the R^2 on the training and test sets, and the mean squared error (MSE) on the test set predictions. The summary of results is presented below:

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

The Linear and Ridge Regression models do not give a convincing result. The neural network MLP model does not perform any better, while the random forest model overfits the training data for a minor improvement in the test data score.

The Gradient Boosting model yields the best results, with parameters set to:

- (n_estimators=200, loss= 'ls', max_depth=5, learning_rate=0.1, max_features='log2')

Notably, limiting the max_features parameter has a positive effect on performance. Also, the model is sensitive to the max_depth parameter: a higher tree depth of e.g. 10 tends to overfit the model to the training data and is inaccurate on the test data, while a lower depth of e.g. 3 does not provide sufficient complexity to distinguish between relatively similar data points.

The standard deviation of the raw training data is 0.82 and the prediction errors have a standard deviation of 0.67. Hence, considering this result, the scatter plot of actual vs. predicted test data and histogram of prediction errors, we can conclude that the gradient boosting model does have some predictive power and picks up a general trend of lower vs. higher scores, but it does not provide an accurate and reliable way of predicting average ratings based only on features.

The next section will look at the more subjective aspects of the review in order to improve the predictions.

4.2 Sentiment Analysis and Prediction

This section considers the text data from reviews in order to learn and predict the sentiment of a review, as a positive (marked with a 1) or negative (marked with a 0). The training and test data for this was taken from the 302K Phoenix restaurant reviews.

The data was split in the following way:

- Reviews rated 4 and 5 stars were labeled as ‘positive’, with a value of 1
- Reviews rated 1 and 2 stars were labeled as ‘negative’, with a value of 0
- Reviews rated 3 stars were excluded from the data as they are indecisive

The data now included ca. 265K reviews and these were split into training and test sets. The following classification algorithms were used to train and predict reviews for positive/negative sentiment:

- Random Forest classifier
- MLP classifier

Each review was represented as a 100-feature vector using the trained Word2Vec model. In order to obtain a single vector for each review (given that reviews vary in length), an average vector was derived for each review. This was performed through the following steps:

1. Using spaCy to parse the review by converting words to lemmatized tokens and remove punctuation, white space and stop words
2. For each remaining token in the review, the Word2Vec model was used to represent the token in vector form
3. For all tokens in the review, the vectors were summed and divided by the number of tokens in order to obtain a mean vector that represents the ‘meaning’ of the review in 100-D vector space

4. The vector representation of the reviews is therefore converted to an X-matrix of 100-D vectors for each review, and the y-vector is a 1/0 representation of the review sentiment.

The training data was used to learn the relationship between the vector representation and review sentiment, and used on the test data to predict sentiment.

Results of the analysis are presented below:

Model	Proportion of accurate predictions	Precision / Recall on positive reviews	Precision / Recall on negative reviews
Random Forest Classifier (500 estimators)	91.6%	97.7% / 92.1%	70.3% / 89.5%
MLP Classifier (solver='adam', alpha=1e-3)	94.2%	96.3% / 96.2%	86.8% / 89.5%

The neural network based MLP classifier gave the best result, after some tuning of the *alpha* parameter, solver used and the learning rate. The default parameters give an accuracy score of 94.0%, and this was increased to 94.2% by increasing the value of *alpha* and using the 'adam' solver. Notably, the MLP Classifier performs better with respect to the precision score on negative reviews. As negative reviews represent 12.2% of the test data, accurately capturing the negative reviews is the critical task for the classification model.

The purpose of this exercise was to build a tool that can distinguish between positive and negative reviews, as a (potentially more robust) alternative to predicting actual ratings. In order to enforce a decision between positive and negative sentiment, it is necessary to remove the indecisive 3-star ratings from the data. However the model can be used to test any review for sentiment, and in case of reviews that are not conclusively positive or negative, the model will make a decision between these two classes.

It is also worth noting hat the relatively high scores are somewhat helped by the fact that the Word2Vec model was trained on data that also includes the test reviews (it was trained on the whole set of 2.9 million restaurant reviews in the Yelp data set). Hence a set of completely new, unseen reviews may well include words or phrases that are not included in our existing Word2Vec dictionary that would not be convertible by the Word2Vec model, resulting in a loss of meaning and a potentially lower prediction accuracy.

4.3 Predicting Ratings from Reviews

In order to predict actual rating reviews, the same Word2Vec representation of review meaning was used, this time including the 3-star reviews as well. The y-matrix of target results includes the rating given by the user as an integer between 1 and 5. A classification model was used again as a way of

predicting the actual rating and classify reviews into 5 distinct groups. The classification models tested are: Random Forest Classifier, Ada Boost Classifier, and MLP Classifier.

Regression models are also be appropriate here and would provide more precise information about a review (e.g. a 3.6 and 4.4 would both be classified as a 4-star rating in the classification model). The regression models used are the Gradient Boosting Regression (GBR) and MLP Regression (MLPR).

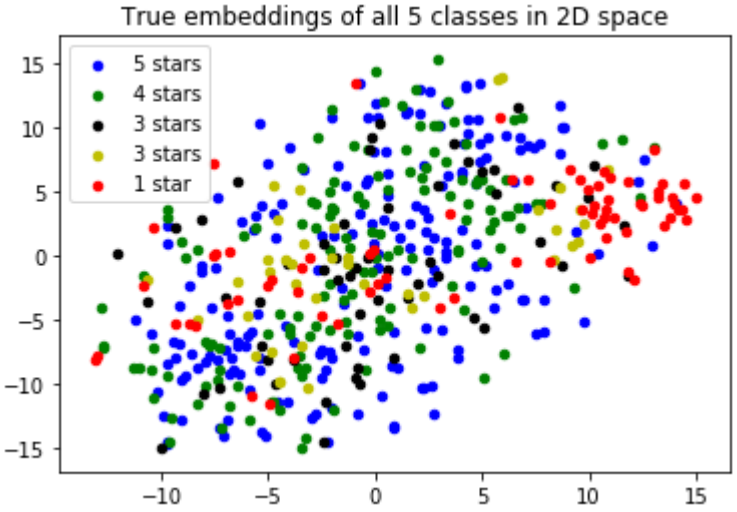
Classification

The results obtained on the classification tests are presented below:

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

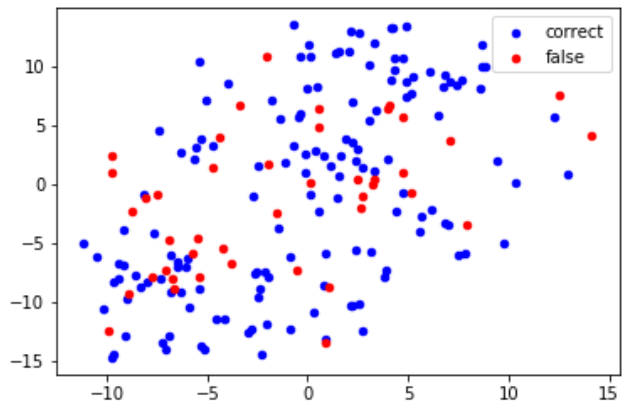
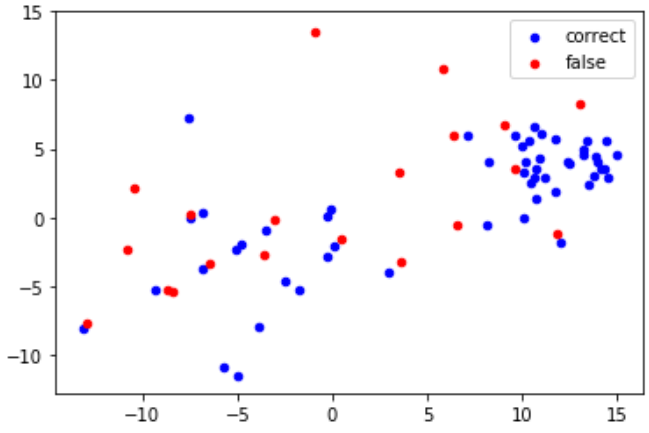
Accuracy on test data was measured using three metrics: the percentage of accurate predictions, percentage of 'bad' predictions (where the prediction is off by two or more classes) and the mean squared error (MSE).

Here, the MLP classifier gave the best results, with the highest proportion of accurate prediction, lowest proportion of 'bad' predictions and the lowest MSE. In this case the neural network method outperformed decision-tree based ensemble methods



Further, results of the 100-D vector representation were reduced to 2-D using t-SNE dimensionality reduction to reveal the embedding of classes. The size of the sample was reduced to 500 data points selected from the test data set, to reduce the computation time of the t-SNE algorithm and to facilitate visualization. The embedding for different classes is difficult to distinguish in 2D space, so we can

consider the correct and false classifications of individual classes. Results shown below are for the MLP classifier::

Correct and false classification of 5-star ratings	Correct and false classification of 1-star ratings
	

Regression

Aside from classification, the star ratings can also be represented by a regression model, where a higher rating indicates a 'better' restaurant. The scaling is a human-interpretable score and does not have a mathematical logic behind it, however for models used (MLP Regression and Gradient Boosting Regression), this should not have a significant impact on results. The results of the prediction models are presented below:

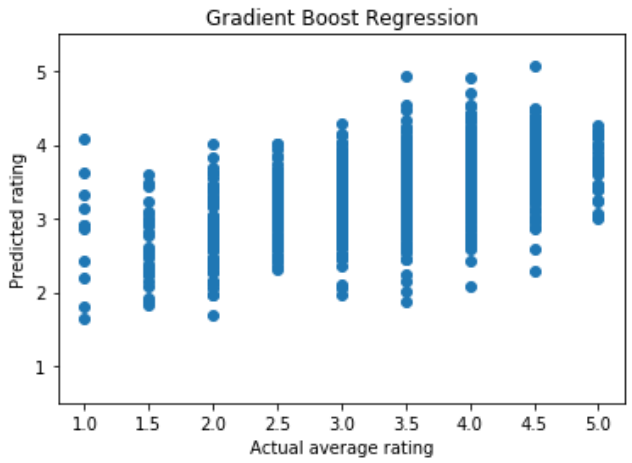
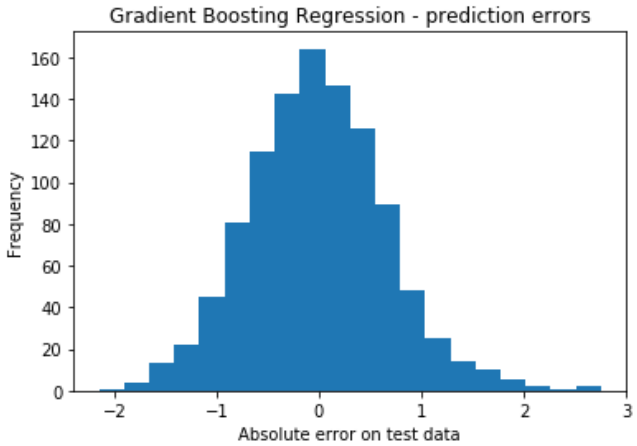
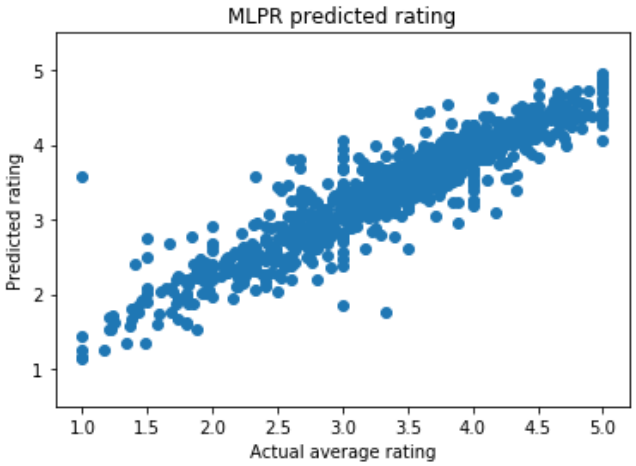
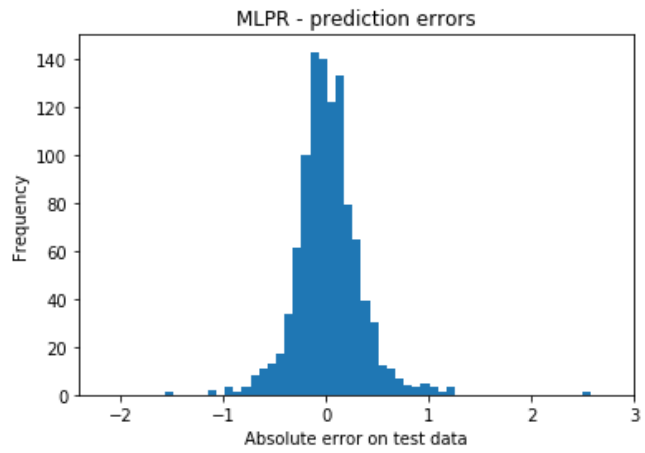
Regression model	R^2 on training data	R^2 on test data	MSE	MSE (integer-rounded)
Gradient Boosting	71.3%	61.7%	0.74	0.83
MLP Regression	69.9%	67.7%	0.62	0.71

In this case the MLP Regression model works best with the following parameters:

- MLPRegressor(solver='lbfgs', alpha=0.0001, hidden_layer_sizes = (1000,), learning_rate='adaptive')

For an individual review, this model gives an MSE of 0.62, or 0.71 when rounded to the nearest integer value (as a direct comparison to the classification models). In either case this is a more accurate result than the classification models, where the lowest MSE was 0.88 using the MLP Classifier.

5 Features vs. Reviews Prediction Results

FEATURE-BASED PREDICTIONS	
Scatter Plot of Predictions vs. Actual Data	**Histogram of Prediction Errors**
	 Mean = -0.025, Std = 0.68
REVIEW-BASED PREDICTIONS AVERAGED OVER ALL REVIEWS ON A BUSINESS	
Scatter Plot of Predictions vs. Actual Data	**Histogram of Prediction Errors**
	 Mean = 0.023, Std = 0.31

With the feature-based and review-based models fully prepared, the next step is to compare their relative performance in predicting overall ratings for a business. As a reminder, the feature-based model predicts the ratings directly from features, and the review-based model predicts ratings for each individual review and then averages the predictions over all reviews for that business. The number of reviews on a business can vary between 1-2 reviews to >500 reviews; however on average there are approximately 86 reviews per business (302,403 reviews / 3,515 businesses). Hence the review-based

model is expected to be more accurate as the larger sample size will reduce the variance of the averaged prediction error.

Further, the two models were tested on the same test sample of 1,055 businesses. The training/test split from the features analysis was reused in the review-based analysis and the review model was trained and cross-verified on the training set, and tested on the unseen test set. Hence the results from the two models are directly comparable.

As is evident from the scatter plots and histograms, the prediction-based model provides far greater prediction accuracy for the average Yelp rating of a business, provided there are multiple reviews written about the business.

6 Most Common Words

In order to obtain a deeper understanding of the data set from a human perspective, we can identify what words most commonly appear in the top-rated and bottom-rated restaurants and therefore what aspects are most commonly praised/criticized by reviewers.

To achieve this, the data was regrouped to separate the following:

- 50 top-rated businesses, each with a minimum of 20 reviews: giving a set of 7,110 reviews
- 120 bottom-rated businesses, each with a minimum of 20 reviews: giving a set of 7,637 reviews

As there are more 5-star ratings than 1-star ratings, it is necessary to modify the selections in order to obtain two comparable data sets. Next, the reviews were parsed using spaCy to:

- remove punctuation, white space and stop words
- extract the remaining words in lemmatized form
- count the number of times they appear in each data set – the top-rated restaurants and the bottom-rated ones

The most meaningful interpretation of this data is to look at words where the difference in the number of appearance is the largest – i.e. words that more commonly appear in reviews for top-rated businesses than bottom-rated ones, will give some meaning to what reviewers praise in the restaurants they like best, and vice-versa for the most poorly scored businesses.

A summary of differences is shown below:

Topic group	Positive reviews	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

Although the classification of these words and their interpretation are subjective in nature, we can attempt to draw some common-sense conclusions, using the domain knowledge of restaurants from the perspective of an average customer:

- The quality of food served is a major theme in mentioned in reviews of top-rated restaurants; the common theme appears to be an aspect (e.g. a meal) that customers love
- Service is not often praised as such and we can presume that good service is a given. The common themes seem to be that the owner’s dedication is what makes a place excel, and customers like friendly service
- Comfort foods/drinks such as pizza, donuts, coffee and wine have more mentions in the top-rated reviews, while people tend to be more critical of meats (chicken, steak, burger)
- The major theme for complaint seems to be poor service, due to customers having to wait, ask for things and/or their dissatisfaction with staff. The use of verbs indicates that bad reviews are often a narrative of what happened

7 Conclusion

7.1 Conclusions from the Existing Model

We have now developed two methods for predicting the average Yelp rating for a restaurant:

1. Predicting the average rating based on a restaurant's features, including its location, opening hours, attributes such as parking and ambiance and price level.
2. Predicting the rating of each review based on the review text and averaging across all reviews written about the restaurant. This was achieved using the gensim library’s word2vec model,

based on a 100-dimensional vector representation of each word, trained over the Yelp data set of 2.9 million restaurant reviews.

Both models were trained and tested on 3,515 restaurants in Phoenix, Arizona.

The feature-based model is only able to give a general trend and provide a vague prediction that is marginally better than simply assuming that a restaurant will have a mean rating of 3.4 stars. The best model here was Gradient Boosting Regression (GBR), which provided a MSE of 0.45 stars, and a standard deviation of the prediction error of 0.68.

The reviews-based model provides a better approximation. The MPL Regression algorithm gives a MSE of 0.57 on each individual review, and <0.1 when averaged over approx. 80 reviews per business. The standard deviation of prediction errors for a restaurant is 0.31 stars, which is a far more accurate and useful result.

From the business perspective, we can conclude that features do have an impact on ratings, and the most important features are location and opening hours (from feature importance of the GBR method). However, the subjective, human aspect of a restaurant, as captured by the reviews, is the more important factor in determining ratings. Hence the text-based predictive model can be used to analyze any text written about a restaurant and provide either a positive/negative sentiment judgment or predict a Yelp-style rating from 1 to 5 stars.

Further, we have identified words that most commonly appear in the top-rated and bottom-rated reviews. By looking at the words that have the most discrepancy between to and bottom reviews, some high-level conclusions can be made:

- A restaurant that wants top reviews needs to have a dedicated owner, stand-out dishes that customers love and friendly service
- A restaurant that has very poor ratings can be advised to review its level of customer service and pay attention to how their meat is prepared

The word2vec model can be visualized in 2D in order to identify the embedding of reviews by rating. However as the dimensionality reduction is very large (from 100 to 2 dimensions), the visualization does not provide a clear and easily interpretable classification.

7.2 Possible Uses and Further Suggestions

While the actual use of the data and models is restricted by the Yelp user agreement, the models are immediately usable and can be expanded to provide more insight:

- For a restaurant with poor or even average ratings, the data provides valuable insight into what customers love, as well as what they do not tolerate. This can help owner can make decisions about staffing, food choice and even restaurant concept

- The word2vec model can be used to parse any text review about any restaurant and provide either a sentiment prediction (positive/negative) or quantify the review in the form of a Yelp-style rating
- With a wider set of data, the model can be used to distinguish between customer preferences across different geographies and/or restaurant categories
- Using the existing data, further testing can be performed to improve the review-based model, e.g. a using deep learning package such as Keras, or refining the features model to increase accuracy and identify the critical features