The Effect of Neighborhood on Yelp Rating of a Restaurant

Sahil Aggarwal
Amir Ali
Richard McGovern
Shreya Sabharwal

IMT 575 Data Science III: Scaling, Applications, Responsibility
Autumn 2018

EXECUTIVE SUMMARY

We discovered there was a statistically significant increase in performance metrics (accuracy, precision, recall) of a model after adding characteristics of a business's place and proximity to the feature set, lending support to the hypothesis that the neighborhood has a positive effect on the Yelp score of a restaurant.

INTRODUCTION

Introduction, Literature Review and Motivation

Yelp gathers a wide set of variables about businesses in metropolitan areas including restaurants and other services. They also gather 5-star ratings from reviewers who check-in to these establishments and rate the quality of their experience. These ratings can have a significant impact on future business performance when prospective patrons first use Yelp's service and see the highly rated ones first. So it is worth investigating which business characteristics correlate with a high Yelp score, but what about data unaccounted for by Yelp such as proximity and geographic characteristics? To what extent is a business's Yelp score pre-determined by its location relative to other businesses?

There is significant literature that states that Yelp ratings have a profound effect on the success of a business (Anderson and Magruder, 2011). To grow and improve their business, a lot of restaurants have a well-laid out plan to improve the visibility of their profile. These efforts include building a great profile, responding to customer reviews, using yelp metrics to track the traffic on their profile and targeted advertising to reach more customers. While all these efforts help increase positive reviews and (in some cases) lead to higher rating, we want to investigate if there are other factors related to a business's proximity that might result in higher reviews for a business on Yelp. There has been some research along the lines of analyzing the text reviews of customers to predict the Yelp rating of restaurants (Mingming, K. Maryam, 2014**).** One paper certainly talked about the impact of geographical neighborhoods on Yelp rating. They used latent factor models, and intrinsic and extrinsic characteristics of the business's neighbors to predict the yelp score(Longke et al). However, we didn't find any of the papers talking about the proximity to other businesses and diversity of businesses within a neighborhood.

Therefore, we are specifically interested in investigating - ***To what extent does the knowledge of a business's proximity and other geographic variables improve the accuracy of a model predicting its Yelp score?***

Dataset

The Yelp dataset consists of businesses, reviews and users data from 10 metropolitan areas. These files are available in JSON format and we are specifically interested in the business, reviews, and check-in files for Las Vegas area.

Below is the description of these files (Yelp Dataset):

- business.json - contains data about participating businesses including their location, neighborhood and business categories such as "restaurant."
- review.json - consists of reviews from users for each of the businesses in the business.json file.
- check-in.json - consists of check-in times for each day for each of the businesses.

METHODOLOGY

Data Cleaning and Preparation

We started with setting up an AWS RDS instance for postgreSQL and imported raw JSON files to the database using Python (psycopg library). There were a total of 22 business categories along with ~125 sub-categories. We combined ‘Food’ and ‘Restaurants’ into one category and disregarded the sub-categories, filtered the dataset for Las Vegas city and restaurants category which left us with 8453 records, i.e. 8453 unique restaurants in the city of Las Vegas. Since most of the JSON key-value pairs were nested, we flattened them in Python and kept only reasonable labels which were specific to restaurants as features such as restaurants delivery, restaurants take-out option etc. We converted the categorical features into binary, and imputed the missing values with the mean of the column values and ignored the features that had mostly null values.

The neighborhood feature was the basis for assessing the neighborhood characteristics of the restaurants. For the neighborhood labels that were missing we computed the shortest distance of the businesses to the centroids of each of the neighborhoods using their latitude and longitude features. We then assigned the neighborhood label for the missing values that had the shortest distance computed.

K-means clustering

We used the K-means algorithm to validate the Yelp neighborhood labels of restaurants based on the location (latitude/longitude). We computed the optimal value of k using the elbow method.

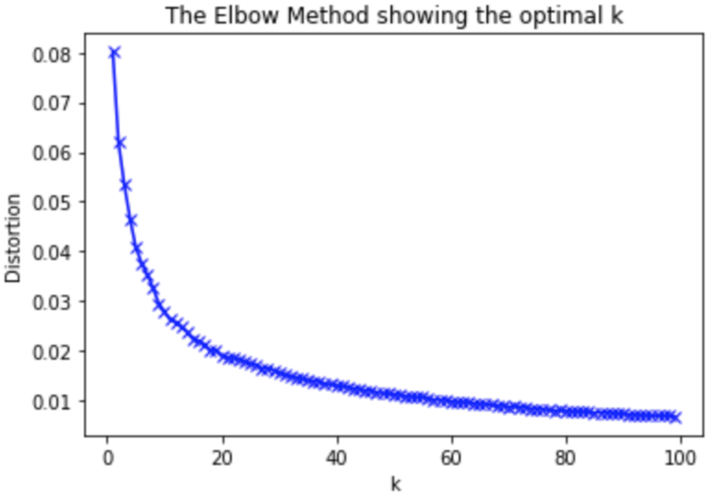


Fig: Elbow Method (K=16)

The optimal K of 16 validates the spread and relative spatial hierarchical levels manifested in Las Vegas neighborhoods. This informed that the neighborhood labels were pretty apt and we did not need another model based on k means labels to see how that differs in performance as opposed to the neighborhood labels model.

Feature Engineering

Features

We were able to extract features like the business_id, city, state, neighborhood, category, accepts_credit_cards, good for kids, outdoor_seating, delivery option, price range (1-4), reservations facility, takeout facility, parking facility, count of businesses in the neighborhood, diversity of business in the neighborhood', number of check ins, number of reviews posted, shortest distance to the Last Vegas Strip for every restaurant, shortest distance to movie halls, shortest distance to hotels, average rating of businesses within 100 metres and the Yelp Rating (stars) of a restaurant.

The number of reviews, check ins, distances and the count aggregates are continuous variables, neighborhood and price range are multiclass and all our other variables are categorical binary variables. We also computed the exact average rating of a restaurant from review.json to analyze and interpolate results better.

New Features

New neighborhood features - diversity, count aggregates, distance from the Vegas strip, distance from each business category, average rating of businesses within 100 meters were engineered using the latitude longitude values and other count aggregates.

Business Diversity

We computed the "business diversity" of each neighborhood using the gini index coefficient over our discrete business categories (hotels, restaurants, shopping, education, etc). Intuitively, perhaps a restaurant that is in a neighborhood with more commercial diversity would have a higher Yelp score. The gini coefficient is computed over 'n' categories associated with all businesses (not necessarily all of which are represented in a given neighborhood), with $x_i$and $x_j$total businesses in categories $i$and $j$respectively, based on the following formula:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n}x_i}$$

This computes the sum of the differences of the number of businesses in each pair of categories -- basically the relative representation of each pair of business categories, normalized by total businesses. These relative mean absolute differences range from 0 to 1 and we normalized it so that 1 represents a perfectly diverse neighborhood (technically 1 - gini) where there are an equal number of businesses in each of the 21 different business categories. A neighborhood with a uniform distribution across all n=21 categories would have a maximal diversity score of 1, while a neighborhood with all businesses in one category would have the lowest. See the "**Variation of Yelp rating with Diversity**" figure.

Proximities and Distance Features

Using the location data for each business, along with the neighborhood labels, we generated more spatial features. We first computed distance of the nearest business of each category to each restaurant to see if this had some relationship with average Yelp rating. We also aggregated the counts of businesses from each category in each neighborhood a restaurant belonged to. However none of these were significant when predicting the Yelp score.

Next, we computed the *distance to nearest 5-star* hotel, restaurant, and "arts" category business for each restaurant. We also computed the average Yelp rating of all businesses within 100 meters of each restaurant. These features were the most predictive of Yelp score.

Exploratory Data Analysis

Variation of Yelp Score(continuous) with Price Range

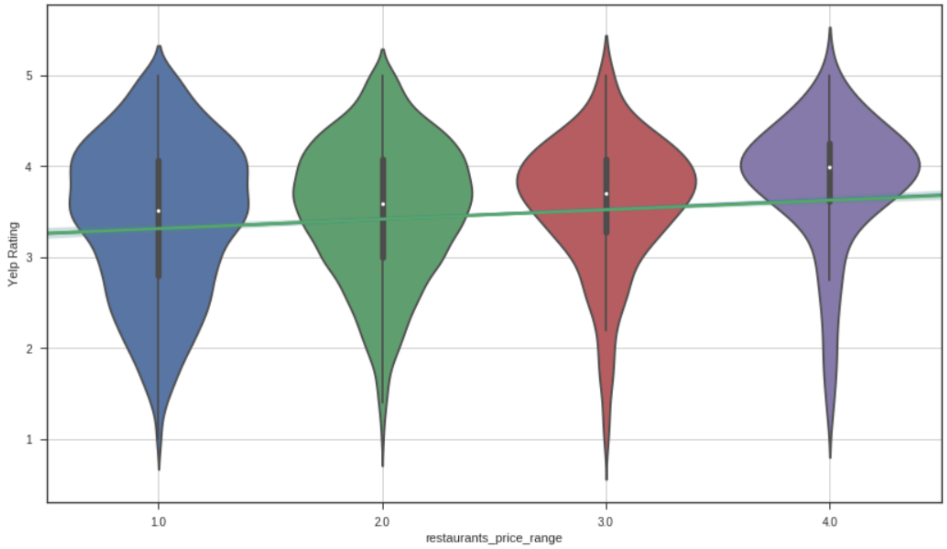


Fig: Yelp Rating vs Restaurants Price Range

Linear Regression Coefficients:

Slope: 0.12
Intercept: 3.27
R-squared: 0.0096

Restaurants that have a higher price range tend to have a higher average star rating, while lower price ranges tend to yield lower average star ratings.

Variation of Yelp score with Review count

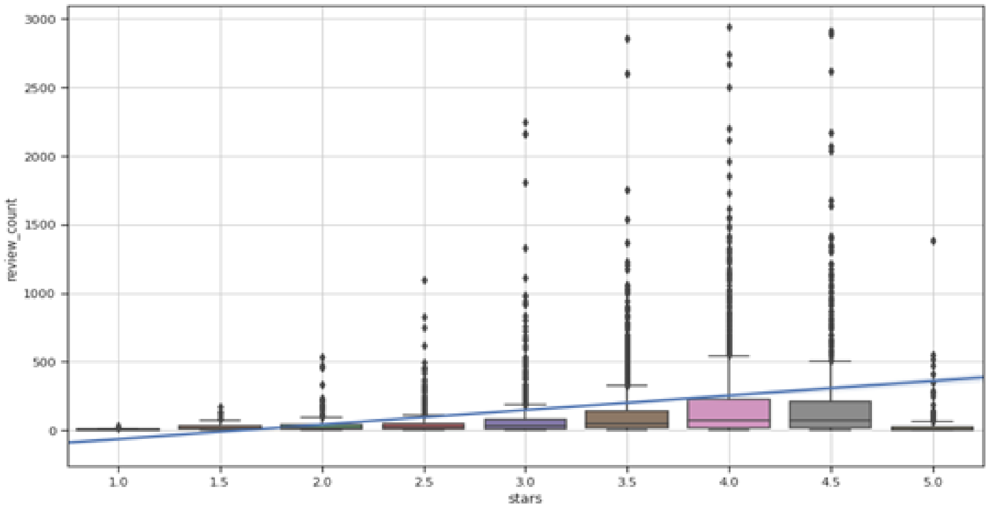


Fig: Yelp score vs Review count

slope: 52.87
intercept: -62.75
R-squared: 0.0355

To avoid outliers, we limited ourselves to restaurants that have less than or equal to 3000 reviews. However, most restaurants have reviews less than ~250. Interestingly, we observe that better restaurants (high yelp rating) tend to have higher number of reviews.

Variation of Yelp rating with Diversity

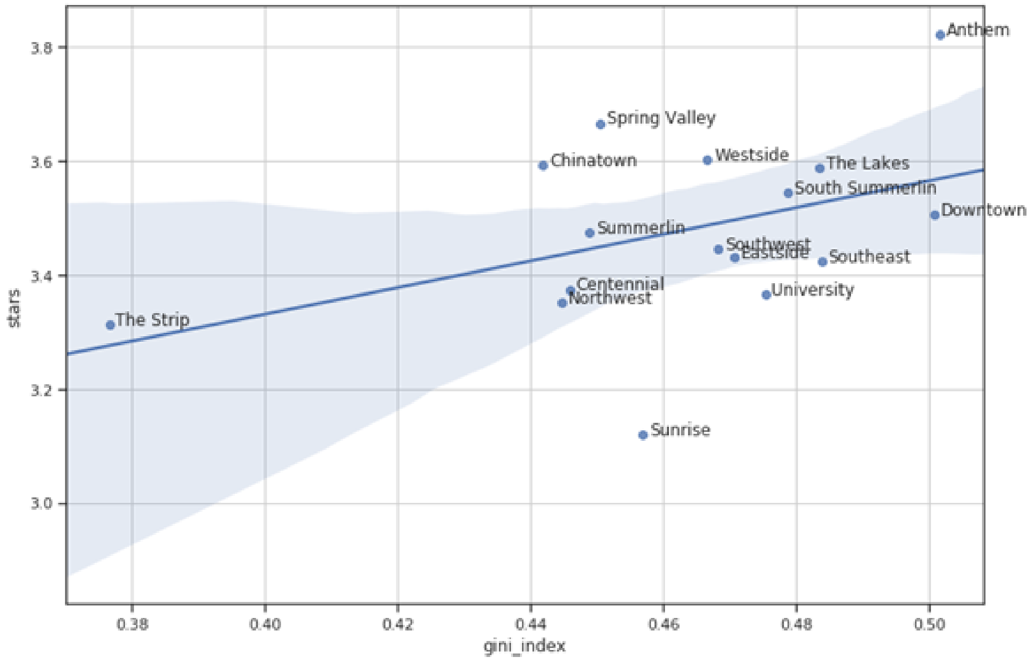


Fig: Yelp score vs Diversity

slope: 2.34

intercept: 2.39
R-squared: 0.1807

The slope for the regression is 2.34, i.e. on average for one unit increase in business diversity(going from a neighborhood with just restaurants to a perfectly diverse neighborhood with equal distribution of all 22 business types), the average rating tends to on average 2.34 points higher. We observe that neighborhoods that have higher gini index (more diverse business types) tend to have higher star ratings. This relationship probably is not causal. Previously, we observed that 'The Strip' which is the busiest neighborhood has the highest number of restaurants, reviews and check ins, but it is not very diverse in terms of other businesses types in the neighborhood. As a result, its collective yelp rating is almost the second lowest in Las Vegas.

Yelp score vs Avg rating within 100m

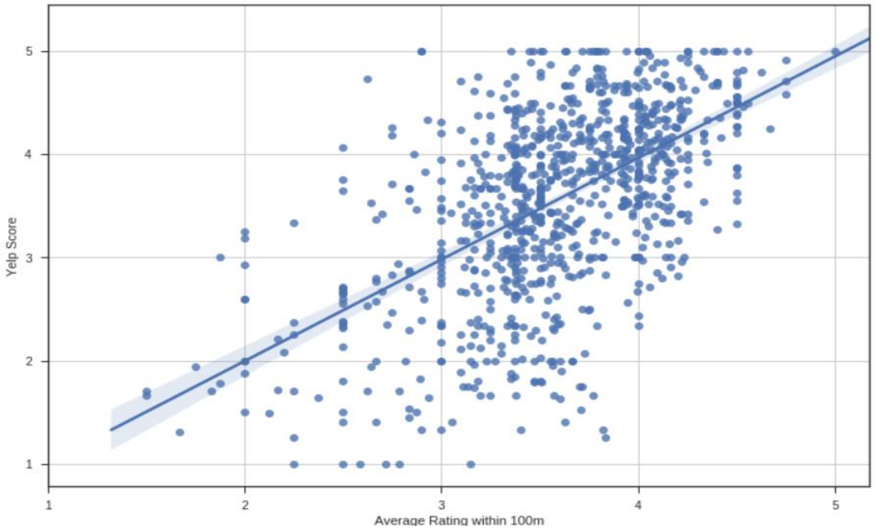


Fig: Yelp score vs Average Rating within 100m

slope: 0.86
intercept: 0.37
R-squared: 0.2404
The above visualization is for a particular neighborhood, Downtown. The plot is slightly noisy around yelp rating of 3 and 4, but we see a general positive trend. We see that the Yelp score of restaurants increases with the increase in average ratings of businesses within 100m radius. The slope is 0.86 which means that one unit increase in average rating within 100m will lead to an increase of Yelp score by 0.86.

Analysis Plan

We tested our hypothesis by performing classification to predict the Yelp score on two different models; one with only business characteristics and the other with both business and neighborhood characteristics. A large increase in accuracy from the first model to the second would inform that there is some signal in the neighborhood features. We tested the statistical significance of this accuracy difference by performing a t-test on the 5 fold cross validation results from both the models.

Modeling

To test our hypothesis that the restaurants neighborhood features have effect on the Yelp score, we decided to start out by building a multinomial logistic regression model to assess if we could predict the Yelp score based on only business characteristics first.

Business Model

We had 8 unique classes in the target variable which was the Yelp rating varying between 1 to 5 stars. This model, albeit having some signal, performed bad (accuracy metric ~25%) probably because we had a class imbalance problem where most of records were concentrated on ratings of 3.5 and above. We tried different strategies to balance the target variable by using SMOTE and tuning the logistic regression model to account for class weights (class_weights = "balanced"). This however, did not lead to significant performance improvements in the model (~50% accuracy metric). We then applied stacking by oversample the minority class and building 6 different logistic regression models; then the predictions rom these were fed into a Random Forest Classifier (another model 'stacked' upon the previous ones). We however, did not see a lot of improvements. This lead us to switch to a different strategy of making the target variable continuous (by adding jitter in ratings using a fraction from the number of reviews per business) and performing OLS regression. This still did not give us better results where we achieved a mean squared error of 0.8 and an R^2 of 0.07.

After running into this bottleneck, we changed our strategy to bin the target variable (Yelp score) into two classes and perform a binary logistic regression. After experimenting with different mapping strategies (where to split the target), we performed binary classification using Logit on Yelp rating where score <=3.5 is 0 and score > 3.5 is 1. We based this decision on interviewing ourselves and our peers where most of them mentioned that they consider a restaurant to be good it has a rating higher than 3.5 on Yelp. This gave us an accuracy of 61%. We then proceeded to tune the hyperparameters for Logistic Regression, by adding class weights (accuracy = 61.5%) and performing cross validation. We performed regularization using 'l2' norm as none of coefficients were very small, so it did not make sense to turn them to zero. We chose the optimal value of C (1/ λ) to penalize the coefficients (accuracy = 63%)

We then used Random Forest Classifier to predict the Yelp score. Although this gave us a higher accuracy (~63.5) we observed during cross validation that it was overfitting. We then used Gradient Boosting Classifier which gave us a similar accuracy (63.5) but did not overfit. We decided to set this as our accuracy for the business characteristics only model.

Business and Neighborhood Model

We followed a similar approach for the neighborhood model, where we binned the target variable, performed Logistic Regression (accuracy of ~67.8%) and performed regularization to choose the best value of λ for the 'l2' norm.

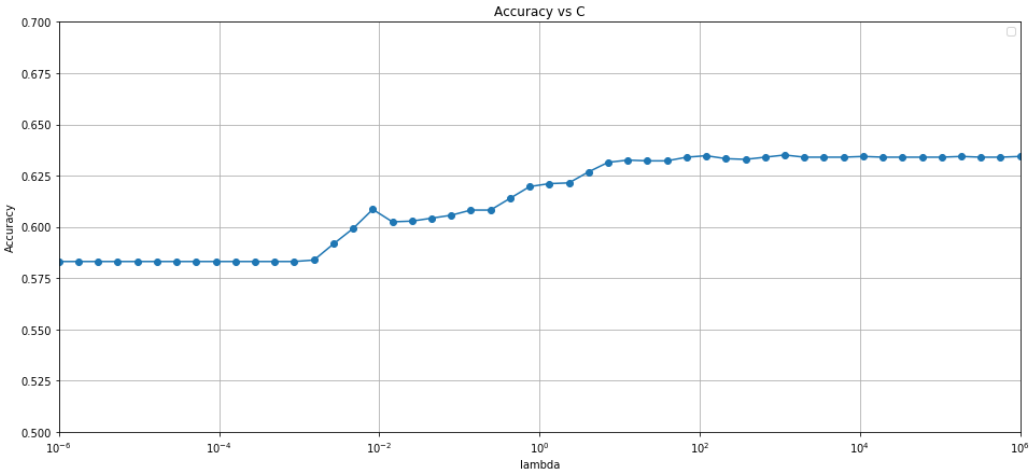


Fig: Choosing the best value of C for Logistic Regression - Neighborhood Model

This gave us an accuracy of 68.5%, a ~1% increase in accuracy. We then performed forward selection, starting with a null model and adding features that improve the performance (mean accuracy after cv) of the model. We found the model best performs with 17 features, giving us an accuracy of 70.2%.

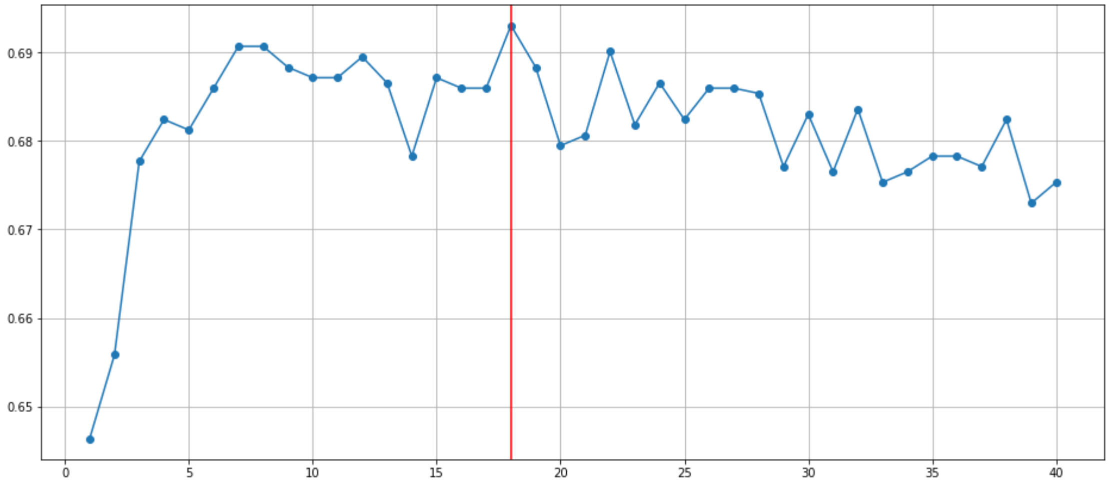


Fig: Forward Selection Results- Accuracy vs No of features - Red line on highest accuracy

We then performed Random Forest (71%, but overfits). We tried to tune the hyperparameters using Grid Search CV to get rid of overfitting, but that did not help us decrease the in-sample accuracy. We used Gradient Boosting Classifier (71.2%) which gave us similar results without overfitting validated by 5 fold cross validation.

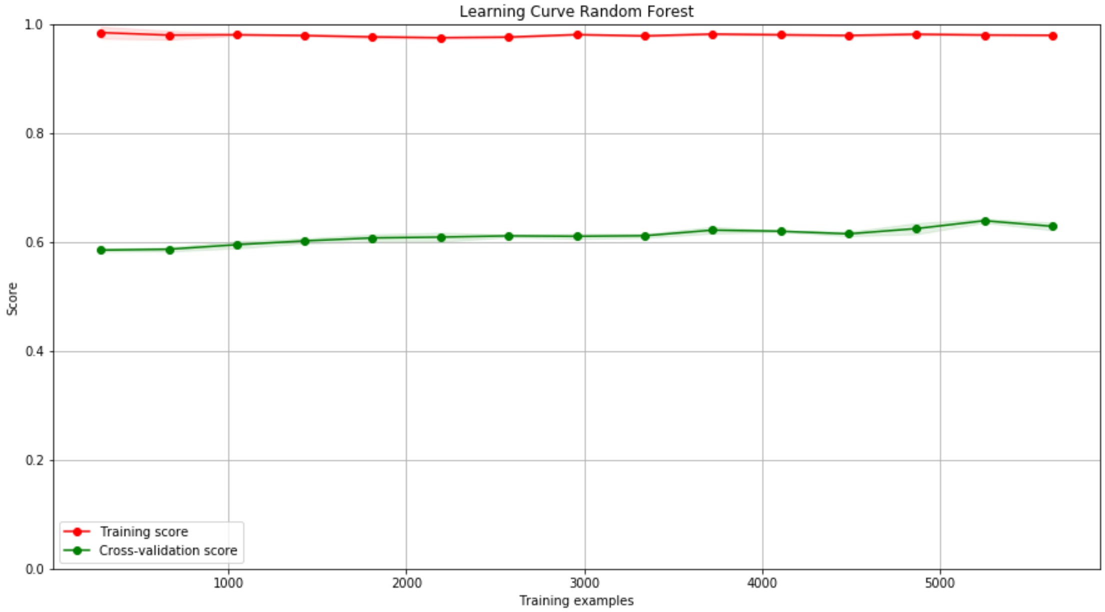


Fig: Random Forest Learning Curve - Overfitting

We then performed a t-test on the 5 fold test scores from both {business} and {business, neighborhood} models to test if the difference of the means is statistically significant. We received a p-value of 5.9e-05 which confirmed that the differences in the means in statistically significant.

RESULTS

Coefficients

The logistic regression model with business characteristics shows that most of the coefficients are statistically significant except for the “good for kids” and “restaurants take out” features. The “review count” feature had the highest positive coefficient (log odds = 7.83), which means that restaurants with larger number of reviews are more likely to have a Yelp rating of 4 or greater. Restaurants that offer delivery on average are 1.7 times more likely to have a Yelp score greater or equal to 4 as compared to restaurants to don’t offer delivery . Also restaurants that offer reservations are 1.7 times more likely to have a Yelp score greater or equal to 4 as compared to those that don’t offer reservations.

For the neighborhood model, the average rating within 100 meters had the highest coefficient (8.12) and was statistically significant. Other neighborhood features which were significant were distance to pet stores, distance to a 5 star hotel, distance to fitness centers, number of restaurants and hotels in the neighborhood and the business diversity.

Performance Metrics of both models

The neighborhood model had a significantly higher accuracy of 71% compared to the business model which had an accuracy of 63% (p-value of 5.9e-05). The neighborhood model also had higher precision (88 vs 62) which means that the neighborhood models positive predictions are more likely to be actual high rated Yelp restaurants as compared to the model which didn't include the neighborhood models . The neighborhood model also had more than double the recall of the business model (62 vs 27). This jump in recall means that adding neighborhoods features drastically improved the model's ability to correctly classify high rated Yelp restaurants.

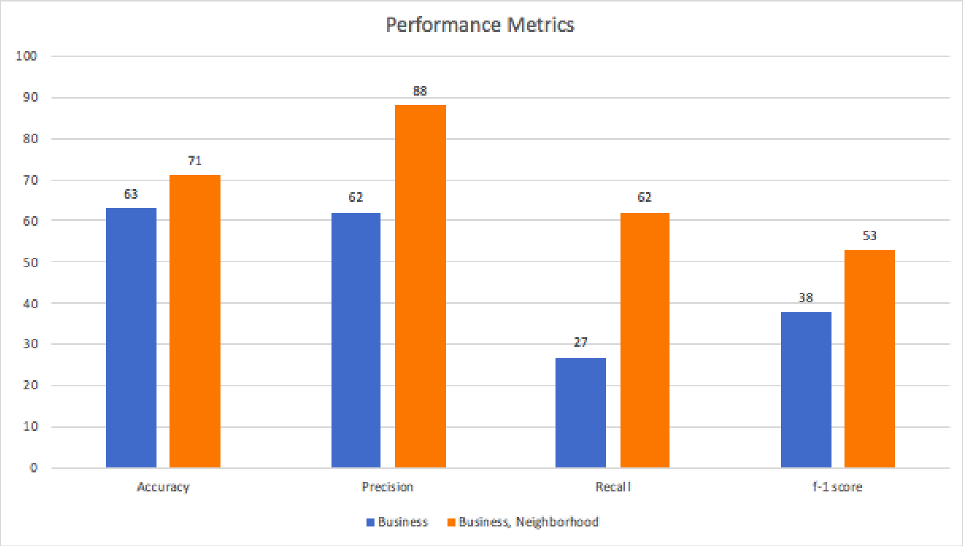


Fig: Comparison of performance metrics for the two models

ETHICAL CONSIDERATIONS

Here we outline the potential beneficiaries as well as those adversely impacted by results from this model and further consequences. Stakeholders include local businesses to the area in which the model is applied, in this case, Las Vegas, but especially those rated on Yelp. While most studies show little correlation between business performance and Yelp score, business could nonetheless be impacted by a sudden change in rating suggested by the model, especially as services like Yelp gain credibility. In particular, we may anticipate negative impacts in either case of our model predicting Yelp scores that are higher or lower than reality. Misinformation always has the potential to be destructive, especially for a business that must adapt to a changing competitive environment. Predicting scores too high could result in a business not investing more resources to improve their quality of service or areas of provision, resulting in persistence of actual issues and the business suffering. Conversely, predicting scores too low could result in a business investing resources in making changes that are not necessary, and could ultimately lead to a self-fulfilling prophecy as they become worse.

Another important set of stakeholders to consider are customers and others who frequent these businesses. Customers have a lot to potentially gain from a more competitive environment engendered by this model. As these measures are made more visible, businesses will try harder to improve and hopefully listen more to customers both individually and in aggregate. However, customers can also be harmed by decisions made after implementation of these models. They may depend on services provided by businesses with an estimated lower Yelp score. If other customers see these lower scores and consequently tend to avoid them, the businesses will suffer and may eventually fold, depriving their customers of that service. This matters especially for customers who for reasons including (dis-)ability, racism, and classism only have access to these businesses projected as under-performing.

Finally, we must also consider the impact on residents of the neighborhoods and local areas of these businesses. If Yelp scores of these businesses is projected lower, and this model output is believed, actual Yelp scores may decrease, and business performance in local aggregate might suffer. In such a case, the commercial value of these neighborhoods may depreciate, further reproducing existing structural socio-economic harms. Residents with limited accessibility to services they need might be most impacted by this overall devaluation as business is taken elsewhere to other parts of the city that are not within wheelchair-walking distance, or accessible by transit.

LIMITATIONS

Although we found some support for our hypothesis, our process to arrive at this conclusion has important limitations to recognize. This analysis was performed on Las Vegas, which is a major tourist destination and has characteristics very different from other cities in the United States. Consequently our models would require more feature engineering and re-thinking to be applied in other cities.

Another important limitation is that the relationship between business performance and Yelp score is very spurious. We are not suggesting any of these high-signal features could support better business performance, but it is easy to come to this conclusion, especially given how networked ratings like Yelp may reproduce this error at scale.

We were also missing some important features which may have improved model output significantly, but we did not account for. These include time series forecasting and accounting for multiple confounding factors to do causal inference of restaurant Yelp scores after periods of data collection. We did not conduct any text analysis of the reviews, which may have had relationships with these spatial features. Many of the business characteristics variables were missing a lot of data that had to be imputed or dropped. And all distances were computed through Euclidean metrics (“as the crow flies”) and did not use more realistic measures, such as along a network of streets, sidewalks, and transit lines.

FUTURE WORK

For future considerations, we recommend filling the gaps in spatially aggregated feature engineering and reconstructing new models for different cities. New features could be accounted for such as population demographics for a neighborhood from census data (including income, ethnicities, and other groups), business density within proximities, diversity for sub-categories of restaurants (accounting for different ethnic foods), and estimations of distance along different networks. We also recommend recomputing all of the aggregate measurements with other spatial levels such as the k-means clusters and 100-meter proximities.

Finally, we recommend attempts to reproduce and expand this model to different cities with more heterogeneity in their neighborhoods and businesses. Different features might be more predictive of Yelp score in cities that are not such high tourist destinations, and instead have customers that live there year-round. In order to distinguish some of the more subtle patterns within these highly smooth spatial aggregates and simple regression, we recommend neural networks to learn more complex features and decision boundaries between businesses with higher and lower Yelp scores. Text analysis of reviews using tf-idf, sentiment analysis, and latent allocation of customer complaints by neighborhood could also provide higher precision in aggregate Yelp assessments of businesses.

REFERENCES

F. Mingming, K. Maryam. “Predicting a Business Star in Yelp from Its Reviews Text Alone” . 5 January, 2014

H. Longke, S. Aixin, L. Yong. “Your Neighbors Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction“

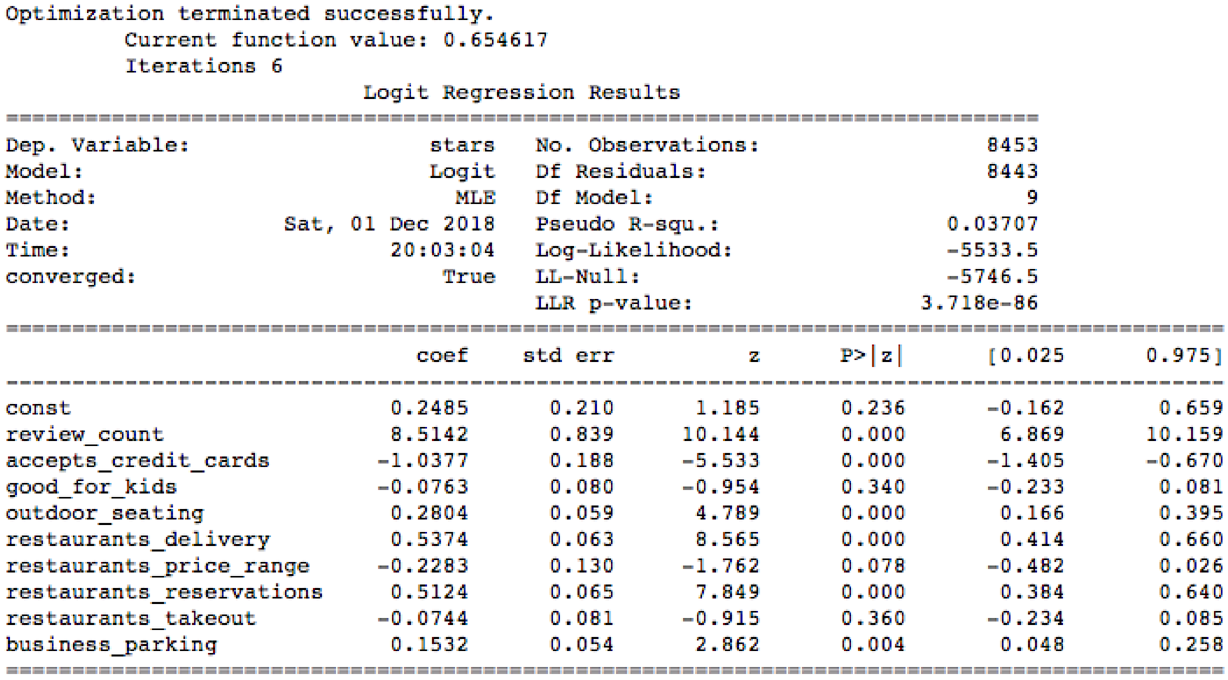
M. Anderson and J. Magruder. “Learning from the Crowd.” The Economic Journal. 5 October, 2011.

Yelp Dataset Documentation. https://www.yelp.com/dataset/documentation/main

T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," in *Neural Computation*, vol. 10, no. 7, pp. 1895-1923, 1 Oct. 1998.doi: 10.1162/089976698300017197, URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6790639&isnumber=6790370

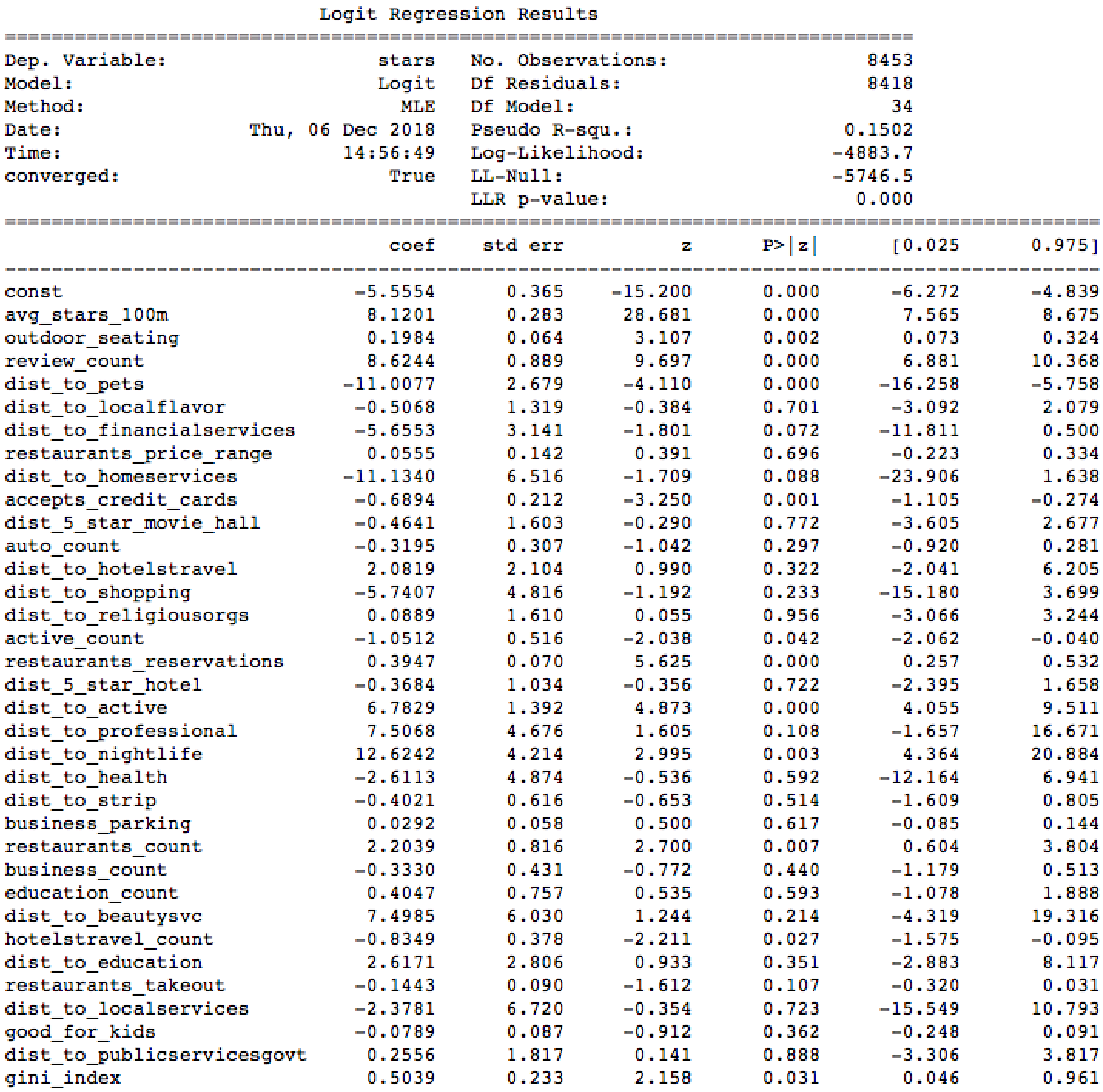
APPENDIX

Results from the Logit model on Business Characteristics



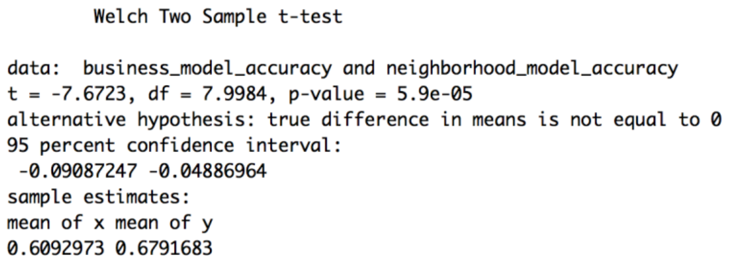
```
Optimization terminated successfully.
         Current function value: 0.654617
         Iterations 6
                           Logit Regression Results                           
==============================================================================
Dep. Variable:                  stars   No. Observations:                 8453
Model:                          Logit   Df Residuals:                     8443
Method:                           MLE   Df Model:                            9
Date:                Sat, 01 Dec 2018   Pseudo R-squ.:                 0.03707
Time:                        20:03:04   Log-Likelihood:                -5533.5
converged:                       True   LL-Null:                       -5746.5
                                        LLR p-value:                 3.718e-86
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                        0.2485      0.210      1.185      0.236      -0.162       0.659
review_count                 8.5142      0.839     10.144      0.000       6.869      10.159
accepts_credit_cards        -1.0377      0.188     -5.533      0.000      -1.405      -0.670
good_for_kids               -0.0763      0.080     -0.954      0.340      -0.233       0.081
outdoor_seating              0.2804      0.059      4.789      0.000       0.166       0.395
restaurants_delivery         0.5374      0.063      8.565      0.000       0.414       0.660
restaurants_price_range     -0.2283      0.130     -1.762      0.078      -0.482       0.026
restaurants_reservations     0.5124      0.065      7.849      0.000       0.384       0.640
restaurants_takeout         -0.0744      0.081     -0.915      0.360      -0.234       0.085
business_parking             0.1532      0.054      2.862      0.004       0.048       0.258
============================================================================================
```

Results from the Logit model on Business Characteristics and Neighborhood Characteristics



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  stars   No. Observations:                 8453
Model:                          Logit   Df Residuals:                     8418
Method:                           MLE   Df Model:                           34
Date:                Thu, 06 Dec 2018   Pseudo R-squ.:                  0.1502
Time:                        14:56:49   Log-Likelihood:                -4883.7
converged:                       True   LL-Null:                       -5746.5
                                        LLR p-value:                     0.000
=============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------
const                        -5.5554      0.365    -15.200      0.000      -6.272      -4.839
avg_stars_100m                8.1201      0.283     28.681      0.000       7.565       8.675
outdoor_seating               0.1984      0.064      3.107      0.002       0.073       0.324
review_count                  8.6244      0.889      9.697      0.000       6.881      10.368
dist_to_pets                -11.0077      2.679     -4.110      0.000     -16.258      -5.758
dist_to_localflavor          -0.5068      1.319     -0.384      0.701      -3.092       2.079
dist_to_financialservices    -5.6553      3.141     -1.801      0.072     -11.811       0.500
restaurants_price_range       0.0555      0.142      0.391      0.696      -0.223       0.334
dist_to_homeservices        -11.1340      6.516     -1.709      0.088     -23.906       1.638
accepts_credit_cards         -0.6894      0.212     -3.250      0.001      -1.105      -0.274
dist_5_star_movie_hall       -0.4641      1.603     -0.290      0.772      -3.605       2.677
auto_count                   -0.3195      0.307     -1.042      0.297      -0.920       0.281
dist_to_hotelstravel          2.0819      2.104      0.990      0.322      -2.041       6.205
dist_to_shopping             -5.7407      4.816     -1.192      0.233     -15.180       3.699
dist_to_religiousorgs         0.0889      1.610      0.055      0.956      -3.066       3.244
active_count                 -1.0512      0.516     -2.038      0.042      -2.062      -0.040
restaurants_reservations      0.3947      0.070      5.625      0.000       0.257       0.532
dist_5_star_hotel            -0.3684      1.034     -0.356      0.722      -2.395       1.658
dist_to_active                6.7829      1.392      4.873      0.000       4.055       9.511
dist_to_professional          7.5068      4.676      1.605      0.108      -1.657      16.671
dist_to_nightlife            12.6242      4.214      2.995      0.003       4.364      20.884
dist_to_health               -2.6113      4.874     -0.536      0.592     -12.164       6.941
dist_to_strip                -0.4021      0.616     -0.653      0.514      -1.609       0.805
business_parking              0.0292      0.058      0.500      0.617      -0.085       0.144
restaurants_count             2.2039      0.816      2.700      0.007       0.604       3.804
business_count               -0.3330      0.431     -0.772      0.440      -1.179       0.513
education_count               0.4047      0.757      0.535      0.593      -1.078       1.888
dist_to_beautysvc             7.4985      6.030      1.244      0.214      -4.319      19.316
hotelstravel_count           -0.8349      0.378     -2.211      0.027      -1.575      -0.095
dist_to_education             2.6171      2.806      0.933      0.351      -2.883       8.117
restaurants_takeout          -0.1443      0.090     -1.612      0.107      -0.320       0.031
dist_to_localservices        -2.3781      6.720     -0.354      0.723     -15.549      10.793
good_for_kids                -0.0789      0.087     -0.912      0.362      -0.248       0.091
dist_to_publicservicesgovt    0.2556      1.817      0.141      0.888      -3.306       3.817
gini_index                    0.5039      0.233      2.158      0.031       0.046       0.961
```

T-test results



```
	Welch Two Sample t-test

data:  business_model_accuracy and neighborhood_model_accuracy
t = -7.6723, df = 7.9984, p-value = 5.9e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09087247 -0.04886964
sample estimates:
mean of x mean of y 
0.6092973 0.6791683 
```