Revamp Your Business

A Machine Learning Solution to Bridge the Gap for Businesses on Yelp!

Austin Adair

November 30, 2020

Abstract

Yelp! holds a lot of information on businesses, like reviews, user check-ins, how many times a user has reviewed, and much more. This information is crucial to businesses, but in some cases, there are establishments that have a gap between their number of check-ins and their star ratings. This means that the business has a high amount of customers, but with low review scores from those customers. Or that the business has a low amount of customers, but a really high review score. Either scenario poses a problem for the business and creates an ineffective business model. The goal of this analysis is to identify those that have either a high star rating but a low number of check-ins or a low star rating but a high number of check-ins. The normalized difference between the star rating and user check-ins of the business has been created as the response variable, and eight predictors from Yelp! the United States Census makes up the rest of the variables. Various regression models will be produced, cross-validated, and evaluated to find the best model. Then, the businesses can use this information to better understand their operations and brainstorm how to improve their star rating and or the number of user check-ins. However, there are limitations to this solution, which will be discussed throughout the report.

1 Introduction

1.1 The Business Problem

Running a business is one of the most difficult jobs in the market, with many factors that go into making one of these businesses a successful one. Between managing employees, branding, inventory management, and several other factors that go into the business, none are more important than getting customers into the business and what those customers thought of their experience. Without having a high customer base, and a high rate of service, the business will falter. But in some cases, there are businesses out there that have a gap between these two factors, meaning that the business has a high amount of customers, but with low review scores from those customers. Or that the business has a low amount of customers, but a really high review score. Either way this is a problem for the business, as either scenario is not sustainable for the long term survival of the business. If a business has a lot of customers, but bad reviews can lead to the short term popularity disappearing as customers won't come back due to poor service quality. And if a business has a low customer count, but high review scores can lead the business to go under due to a lack of sales.

1.2 Analysis Background

It is the main job of this analysis to help identify key factors that help to lead to this gap between the two factors. This is going to be done using data from Yelp!. Yelp is one of the top websites for business reviews for and by consumers. How this website works is that members of this particular site can leave overall "star" ratings of how they felt about their experience at a particular business. This value is between 0 and 5 and allows half number intervals between them. This website also allows members to write their own reviews of the business, with the details being completely up to the user of how in-depth those reviews are. Yelp also allows their users to "checkin" to a business, basically meaning that the user is openly saying that they have been to this particular business.

From the Yelp data, the review star rating, the amount of check ins, and individual reviews were collected for each business in the data set, as well as general information about each business. At this point, the target variable of the difference between star rating and check ins is not yet created. This was done by taking the difference of the normalized star ratings and the normalized check ins for each individual business. This variable is the target of the analysis done in this report, using regression methods to determine what factors within the Yelp data has some effect on this target variable.

1.3 Summary of Solution

In order to bridge the gap for businesses with high star ratings with low check-ins or vice-versa, low star ratings but high check-ins, several different regression analysis methods were performed on the data set. As mentioned above, the response variable was the normalized difference of the star ratings and check ins for each business. There were eight predictor variables included in the analysis, four of them developed from a sentiment analysis of the reviews for these companies. Seven different models were created, cross-validated and then compared on the R-squared and

mean squared error. The best model, a multiple linear regression model, can be used by these businesses on Yelp to better understand their business model and make adjustments to improve their number of check-ins, star-ratings, or both. However, there are some limitations to this solution which will be discussed throughout the report.

2 Data Collection, Preparation, and Analysis

2.1 Data Collection

This analysis will be performed on the combination of two data sets: One containing business information and their corresponding Yelp reviews, the other source being 2010 Census data which will provide demographic information to the businesses. The business information was downloaded directly from Yelp (Yelp Inc., n.d.) and contains multiple data tables: tips, businesses, check-ins, reviews, and users. Each of these data tables contains a large number of observations. The Census data was obtained using their public API and contains the number of establishments, employment during the week of March 12, first quarter payroll, and annual payroll for all industries by 5-digit ZIP Code as of the last census in 2010 (U.S. Census Bureau, n.d.). Census data was only collected for zip codes we analyzed to reduce the request size.

The Yelp dataset contains five tables of data. The business data table contains records for over 200,000 businesses with data such as geographical information about the business, open hours, various attributes that vary depending on the business category and provided information, and the star rating. The check-ins data table contains a comma-separated string of check-ins a business has received for businesses that have received at least one check-in. The users data table includes 1,968,703 observations and 23 variables, with information on Yelp users that left reviews for the businesses such as the number of reviews they have made, how long they have been Yelping since, and how many times one of their reviews has been marked as being a certain criterion, e.g. funny or useful. The review data includes 8,021,122 observations over 9 variables, including the review text, the date, star rating, user id, business id, and how many votes the review has received. Finally, the tips data has 1,320,761 observations over 5 variables that are business id, user id, compliment count, date, and tip text.

The Yelp dataset contains a large amount of information, both structured and unstructured, this is quite beneficial as it opens multiple channels for analysis. It is also being actively updated from time to time with new information and records. Although the analysis we have performed is at the business level we also received the information necessary to do a customer-level analysis which could be utilized to target market specific customers in the future. The Census data we used is from 2010 however within the next year the 2020 census will be completed and the model will be able to be updated with the more current information. Further, although in this analysis we only utilized the business patterns API there are many more endpoints containing more information that we could add to the models.

The data used in this project does have some limitations. The Yelp data includes a few common attributes amongst the businesses, meaning most are unusable for large scale data analysis. Also, the Yelp dataset only covers observations from 10 metropolitan areas, potentially limiting the information and insights that can be concluded from our model. There is a discrepancy between the review count in multiple tables and the actual number of reviews in the reviews table due to the review count being the total number of reviews during data collection, while only the reviews that were recommended by Yelp were included in the reviews data table during collection. Businesses were only included when they had at least three reviews older than two weeks at the time of collection, this will lead to some businesses being excluded from the data set. Although the census data endpoint we used in our analysis was aggregated on the zip code level the majority of endpoints for the census data API are aggregated by Zip Code Tabulation Areas (ZCTAs) which are generalized areal representations of United States Postal Service (USPS) ZIP Code service areas created by the U.S. Census Bureau (U.S. Census Bureau, 2020). Furthermore, not all zip codes have data available from the endpoint, some return partial results whilst others return no data at all.

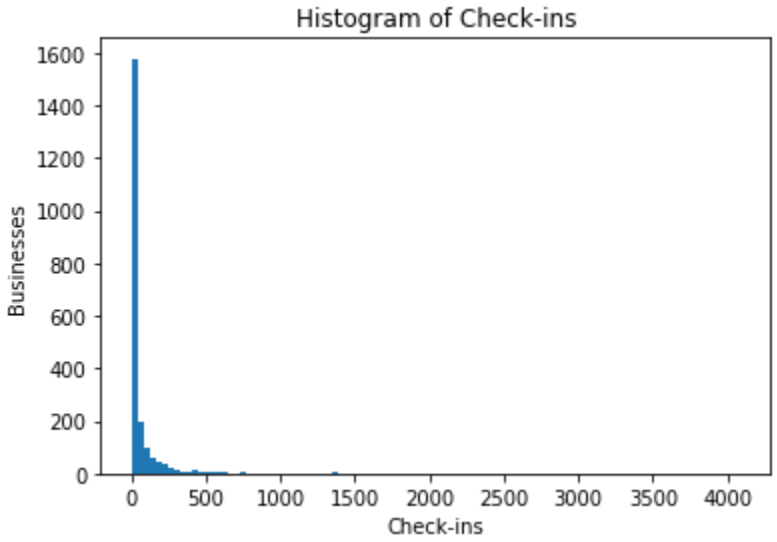

Figure 1: Histogram of the number of Check-ins across businesses in Wisconsin

2.2 Data Preparation

Out of the five tables provided by Yelp, we only used three of them, the business, check-ins, and reviews tables. The reviews table had no missing values but did contain columns we would not use for analysis and so were removed to reduce the size of the data: useful, funny, cool, and date. The updated reviews data frame was then uploaded to a local NoSQL database, MongoDB. We reduced the scope of the data to only include locations in the USA with more than 1000 businesses, this reduced the states included to only: AZ, IL, NC, NV, OH, PA, SC, and WI; reducing the number of businesses from 209,393 to 153,499. For each zip code in the business

data, census data was retrieved. This contained some missing values as not all return values were complete, these values were imputed with the median. Further, there were not records for all zip codes, the missing zip codes were also imputed with the median. The two cleaned data tables, reviews, and businesses were uploaded to a NoSQL database, MongoDB.

Without some processing the unstructured review text is unusable in the analysis, so we used the VADER sentiment analysis tools in the Natural Language Toolkit (NLTK) Python package (Hutto & Gilbert, 2014). Before running the sentiment analysis we did some cleaning over the text. Due to time limitations, the scope was reduced to approximately 2000 businesses within Wisconsin. First, punctuation and numbers were removed, making all letters lowercase, removing any English stopword, and finally stemming the words to their word stems. VADER sentiment analysis was then run over all of the reviews cleaned text and the four sentiment scores returned were added: negative, neutral, positive, and the compound score. For each business, the average sentiment score across the reviews for that business was then added to the business document in the database for each sentiment score.

The final step in preprocessing was to create the response variable, which is the absolute difference between the normalized star rating and the normalized number of check-ins for each business.

Figure 2: Scatter plot of the number of check-ins by Star rating for businesses in Wisconsin

2.3 Descriptive Analytics

The first interesting observation in the data to be modeled is the high frequency of businesses with no check-ins and the existence of several outliers going into thousands of check-ins, which can be seen in Figure 1. Secondly, as seen in Figure 2, when looking at the number of check-ins a business has against the star rating of the business we see the data is approximately normally

distributed except for some outliers, that is left-skewed and centered at a star rating of 4.0. Interestingly, businesses with 5.0 star ratings have a relatively low number of check-ins as compared to businesses with a star rating between 2.5 and 4.5. Finally, in Figure 3, we have plotted check-ins against the target variable, grouped by the star rating. We observe the data is stratified by star rating with the businesses with a star rating of 5.0 having the largest target distance. Further, businesses with a star rating of 1.5 have the smallest target distance. Finally, businesses with a star rating of 4.0 have some businesses with a high number of check-ins that cause the strata to become non-linear.

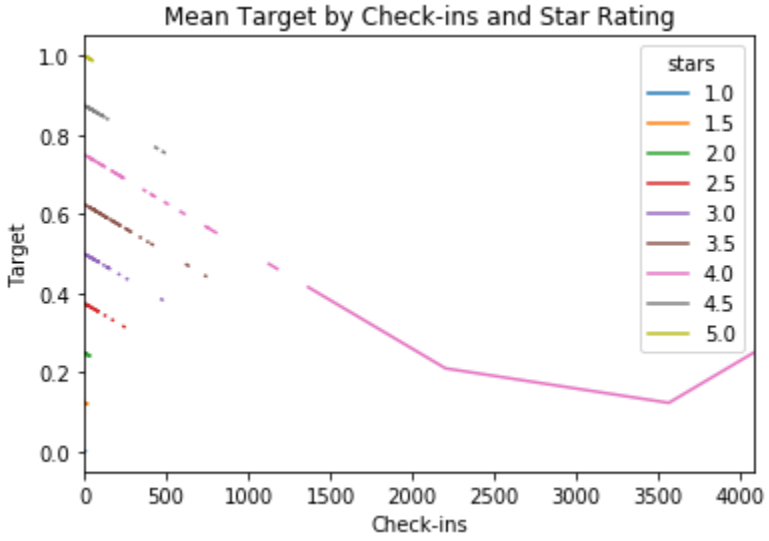

Figure 3: The mean of the target variable by the number of check-ins stratified by the star rating for businesses in Wisconsin

3 Model Building & Evaluation

After data collection, preparation and preliminary analysis was completed. We were finally set to build, evaluate and choose a model that best predicted our continuous target variable. In order to test the performance of our models, we must create a process to evaluate each of the models efficacy. To accomplish this, we have subset our data into a "Training Dataset" and a "Validation Dataset". Once these were created, we constructed models on the training dataset and then observed the models performance using the validation dataset. Once all models were created and validated we were able to directly compare them and choose the model that best fit our dataset. Table 1 below displays the eight variables included in the analysis.

Table 1

Variables Used in Analysis

Variable Name

review_count
zbp_employees
zbp_establishments
zbp_annual_payroll
ss_compound
ss_neg
ss_neu
ss_pos

Model 1: Multiple Linear Regression

The first model we constructed was a linear regression model. A linear regression is a statistical approach to find the linear relationship between variables. Because our data contains multiple (eight independent variables) our model will be classified as a multiple linear regression model. Due to our relatively low amount of variables and the fact that we are using this model for predictive purposes, not explanatory, we will train our model using all independent variables. The summary statistics of our full multiple linear regression model built on the training data can be found in *Table 2* below.

Table 2

Model Summary

Variable	Coef	P-Value
Review Count	-.0006	0.0
ZBP Employees	~0	0.343
ZPB Establishments	~0	0.32
ZBP Annual Payroll	~0	0.214
SS Compound	0.2930	0.0
SS Negative	-1.1276	0.0
SS Neutral	0.3172	0.0
SS Positive	1.385	0.0

The full multiple linear regression model built on all eight available predictors had a F-statistic of 0.00 with an associated p-value of 0.00 which means that at least one of the eight predictors is statistically significant in predicting the target variable. Likewise, the full model has a very large R-squared value of 0.929. This means that approximately 93% of the variance in our target variable is explained by the eight independent variables.

When the full multiple linear regression model was applied to the validation dataset, there was a statistically significant decrease in predictive capability. This means that the model was overfit. The R-squared value of the model using the validation data fell from 0.929 to 0.496. This means that the multiple linear regression model will likely experience some difficulties when trying to predict new data. The limitations of this model can best be understood through analysing its MSE (Mean Squared Error) which is a good proxy for how well the model fits, where a value closer to zero is desirable. Our model when analysed on the validation has a MSE of 0.031. We will later be able to compare the models MSE to other models to determine which model best fits the data.

Model 2: Gradient Boosted Regression Tree

The next model that was constructed was a gradient boosted regression tree. This is an additive model that builds in a stage-wise fashion which allows for the optimization of a specified loss function. For each stage of the model, a regression tree is fit on the negative gradient of the loss function. In doing so each regression tree reinforces the examples that the previous regression tree had difficulty with and misclassified.

As a baseline, a gradient boosted regression tree with the default settings was run on the training data. Using the default settings, the model returned a MSE of 0.0315 and a R-squared value of 0.4386. Using the default settings alone, the model was unable to outperform the multiple linear regression model. In an attempt to improve the model's predictive capabilities we cross validated the model over a specified tune grid. The tunegrid built models using the following characteristics and allowed us to pick the optimal combination of model parameters. Our tunegrid built models that contained bootstrap staples of 100, 200, 300, 400 and 500. Bootstrap samples is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement (Brownlee, 2019).

In doing so we can ensure that each tree is likely built on different data which can help reduce the variance of the overall gradient boosted regression tree. The tunegrid also built models that contained a learning rate of .02, .04, .06, .08, .10 and .12. The learning rate shrinks the contribution of each additional tree by the specified learning rate. Thus, there is a tradeoff between the learning rate and the number of trees built. Lastly the tunegrid built models that contained alpha-quantiles of .1, .2 and .3. Each model was built using a loss function "hubar"

which is a combination of least squares regression and least absolute deviation. A value of k = 10 was chosen for the cross-validation, meaning the data set was split into 10 parts and the process was run ten times.

Using cross validation, the best combination of tune grid parameters contained an alpha off .2, loss function “hubar”, random state of 1137 and a bootstrap sample size of 200. The cross validated gradient boosted tree returned a less effective model with an associated R-squared of 0.42728 and MSE of 0.03217.

Model 3: Random Forest

In addition, a regression random forest was constructed. A regression random forest model is an ensemble technique that allows algorithms to combine predictions in order to predict more accurately than an individual model (Chakure, 2019). In this case, because the response variable is not binary (1 or 0), a regression random forest is necessary. The algorithm aggregates numerous decision trees with the training data and outputs the class that is the mean prediction of the individual trees for the random forest.

Initially, a random forest with the default settings and a random state of 1337 was run on the training data. These settings outputted a mean-squared-error of .03196 and an R-squared of .43099 on the validation data. There was room for improvement, so the random forest was cross-validated with bootstrap sampling. Bootstrapping means random sampling with replacement, which will reduce the variance of the model. The max_depth parameter refers to the maximum depth of each tree. In this case, the max_depth range was 3 to 11. Max_features refers to the number of features considered when splitting a node, set to a range of 1 to 8. Min_samples_leaf, or the minimum number of samples required on each leaf node was set to a range of 3 to 10. The min_samples_split parameter, or the minimum number of samples required to split an internal leaf node, was set to a range of 20, 120, and 20. The last parameter that was adjusted for the random forest is n_estimators, which is the number of trees in the forest of the model. N_estimators were set for the algorithm to test a range of 300, 700, and 100. A value of k = 10 was chosen for the cross-validation, meaning the data set was split into 10 parts and the process was run ten times.

The random forest model with the best MSE and R2 values included the maximum depth of each tree to be 9, the number of features considered when splitting a node to be 2, the minimum number of samples required at each leaf node to be 3, the minimum number of splits at each internal leaf node of 20, and the number of trees in the forest equaled 300. The cross-validated random forest improved the mean squared error to .03037 and the R2 to .45926 on the validation data. This means that approximately 46% of the variance in our target variable is explained by the eight independent variables for the cross-validated random forest model.

Model Selection

After building the models, it was necessary to identify the machine learning algorithm that is best-suited for predicting the target variable on new data. So, the different models built were compared on key criteria (MSE and R-Squared), and the best-performing one was selected. Of the 13 unique models built, the multiple linear regression and the random forest with cross validation performed the best. In order to choose between these models, their R-Squared and MSE will be evaluated and compared.

The multiple linear regression model had a slightly higher R-Squared variable than the random forest model with values of 0.49601 and 0.45926 respectively. This means that the variance in the response variable is explained by approximately 4% greater using the independent variables in a linear regression model than a random forest model. However, the MSE of the random forest model is slightly lower than the MSE of the multiple linear regression model with values of 0.03037 and 0.03128 respectively. This means that the average difference between the predicted values and the actual values are slightly higher in the linear regression model. Due to the fact that both models have similar R-Squared values and MSE they would both likely perform very similarly in practice. For our purpose, the ease of implementation and explanation for Yelp! are both important criteria. Thus, the multiple linear regression model which is easier to explain than a random forest (an ensemble model) will be the model chosen. See *Table 3* below for details on additional models that were created and their metrics.

Table 3

Model Evaluation

Model Type	R-Squared	MSE
Lasso Model	.01175	.05505
Lasso with CV	.39705	.03386
Elastic Model	.01339	.05541
Elastic with CV	.43979	.03146
ADABoost Model	0.39782	.03382
ADABoost with CV	.42915	.03206
Gradient Boosted Tree	.43865	.03152
Gradient Boosted with CV	.42728	.03217
Decision Tree	.39660	.03389

Decision Tree with CV	.36732	.03555
Multiple Linear Regression	**.49601**	.03128
Random Forest	.43099	.03196
Random Forest with CV	.45926	**.03037**

4 Conclusion

4.1 The Outcome

The analysis set out to bridge the gap for businesses who have a disparity in their number of check-ins and their star ratings using data from Yelp! and the US Census. Sentiment analysis on the reviews of businesses were included as predictors of the regression models. The normalized difference between the two factors was created as the response variable and eight other predictors were used in the analysis. The regression models were cross-validated and evaluated on the validation data and the model with the best R-squared was selected, a multiple linear regression model. Businesses on Yelp! can use this information to better understand their business operations and create solutions to improve these factors. Specifically, the businesses included in the analysis from Wisconsin can use this solution to brainstorm ways to improve their operations and marketing, which could be associated with an increase in check-ins. Additionally, the businesses can improve customer service or ambience which may be associated with improved star-ratings. Additionally, by now knowing that the two variables that have the biggest impact on the difference between star ratings and review score are the positive and negative ratings of reviews left by customers, businesses can then focus more on what the reviews specifically say about their operations that they could improve on.

4.2 Implementation

Yelp! businesses who have a high discrepancy between ratings and check-ins will be the ones who can be positively impacted by this solution. Yelp! can share this information and solution with the businesses involved. As mentioned, the businesses can then focus more on what the reviews say and will be able to identify what is causing the discrepancy. For example, let's say a business owner with a low-star rating but high number of check-ins sees this solution and then pays more attention to their reviews. If reviews are consistently mentioning poor service, the business owner can recognize how this may lead to a low star rating. Then, the business owner can talk to the staff about top-notch customer service and re-establish training guidelines in order to in the future have reviews say that the service was excellent on their Yelp! page. This could in turn be associated with an increase in their star-ratings. In addition, Yelp! can use our multiple linear regression to predict and quantify the impact that their changes could have on the gap between ratings and check-ins. For instance, if a business is identified as having a low rating but high check-ins and is revamping their customer service. Yelp! can estimate the impact this will

have on the sentiment score of their reviews and use our model to predict the impact that this change will have on the overall target value. This information can be used to direct business leaders on how to effectively run their organizations.

4.3 Limitations & Moving Forward

A major limitation of this project is the scale of this solution, which was limited due to time constraints and processing power. So, in order to apply this solution to more businesses, rather than just a single state, it will need to be a long-term project. Luckily, this analysis is reproducible, meaning that the same process could be applied to a larger dataset of different states or regions. Furthermore, the Yelp data included only a few common attributes amongst the businesses, meaning most are unusable for large scale data analysis. Moving forward, it will be necessary to find more complete data in order to include more variables in the analysis, which could give businesses more insights. In addition, further analysis can be done on the solution by doing a text analysis for each business' reviews, and to see which words or phrases are the most common for both positive and negative reviews. In future analyses, the 2020 Census data can easily replace the 2010 Census data to give more updated information to the businesses. Lastly, this analysis was completed at a business level. In the future, a customer-level analysis could be utilized to target specific customers in the future, and the data to complete this readily available.

5 References

Brownlee, J. (2019, August 8). *A Gentle Introduction to the Bootstrap Method*. Machine Learning Mastery. Retrieved November, 2020, from https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

Chakure, A. (2019, June 29). *Random Forest Regression*. Medium. Retrieved November, 2020, from https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Hutto, C.J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI, June 2014

U.S. Census Bureau. (n.d.). *2010 Zip Code Business Patterns*. Retrieved November 22, 2020, from https://api.census.gov/data/2010/zbp.html

U.S. Census Bureau. (2020, August 26). *ZIP Code Tabulation Areas (ZCTAs)*. Retrieved November 28, 2020, from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

Yelp Inc. (n.d.). *Yelp Open Dataset*. Yelp Dataset. Retrieved October 13, 2020, from https://www.yelp.com/dataset

6 Appendix