Author: Aakanksha Baid

Yelp Dataset Challenge – Round 13
Predicting star rating class for restaurants

Introduction

With proliferation of ecommerce and online businesses, consumers can provide instant feedback on products and services. Yelp is the biggest online platform connecting people with local businesses. Yelp helps people share and rate their experiences and it is the leading local guide for word-of-mouth publicity on everything from boutiques and mechanics to restaurants and dentists. With over 180 million unique monthly visitors, Yelp is increasingly being used by businesses for advertising. A well-constructed profile with high star ratings can be the reason prospective customers choosing one business over another. According to a study by Nielsen, 98% of Yelp users have made a purchase from a high-rated business they found on Yelp, with around 90% doing so in a week. So, it is important to understand what drives these high star ratings.

Star ratings are accompanied by user reviews, which are often mismatched with the star rating. Knowing and altering factors that contribute to high ratings can lead to higher earned media and growth. Studies have shown that customers are likely to spend 31% more on a business with excellent ratings. Harvard Business School researcher, Michael Luca, found that a 1-star rating improvement on Yelp translated to anywhere from a 5 to 9 percent swing on revenues. So that will be a win-win-win situation for Yelp, listed businesses and customers.

This analysis provides a feasible model to predict the star rating class based on business attributes and other factors like location, photos etc. As business attributes used in the model can be controlled by the business, knowing which one's influence star ratings can help businesses pull the correct levers to raise their ratings. In addition to helping raise review ratings, alteration of attributes would provide the average customer with a positive experience, more relevant information and trustworthy reviews which can attract new customers through network effect and earned media. If Yelp can help the business influence its rating accurately and gain trust among its users then it can become the go-to place for users to search for business reviews, like Amazon does for product reviews.

Yelp's major source of revenue is advertising which requires an active and engaged user base that responds to the business ads. Restaurants with higher star ratings are less likely to advertise with Yelp as their high star rating provides earned media. Using this model, Yelp can partner with low star rating businesses and provide them with a playbook of attributes that can improve their star ratings. Yelp can also provide partnership to newly listed businesses to help them achieve higher star ratings. Both these partnerships can be monetized by Yelp. As average star ratings for businesses increase, there will be less differentiation with existing high star rated businesses and these businesses would have to sponsor with Yelp to get user attention which will ultimately benefit Yelp's ROI.

Yelp data on business and photos was used to predict the star rating class for restaurants. This analysis included only business specified attributes and excluded the actual text reviews. However, the algorithm was still able to predict star rating class with 77% accuracy and ROC AUC of 84% using classification algorithm Random Forest.

Author: Aakanksha Baid

Problem Definition and Algorithm

Task Definition

The objective is to implement a model that predicts Yelp star rating class of restaurants based on restaurant attributes and other factors like location, photos etc. This analysis' hypothesis is that a combination of restaurant's attributes, location, reviews and photos are indicators of quality and can be used to predict the star rating class (high/low). The star rating class is defined as high for stars greater than equal to 3.5 and low for stars below 3.5.

If a model that calculates business rating based on business features listed on Yelp can be created, then the model can be used to significantly predict the ratings another establishment will receive, because the features of a business have great impact on how a business is rated.

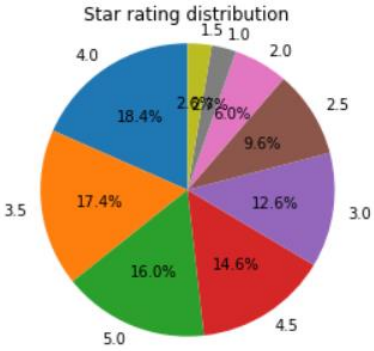
The data was downloaded and converted to JSON files for analysis in Python. The major steps involved in the analysis are as follows –

1. Data pre-processing: Process of cleaning raw data that includes treating missing, noisy and inconsistent data to prepare data for modelling.
2. Exploratory data analysis (EDA): Process of finding initial insights from data through summaries and visualizations to describe variables and their distributions. This step helps in understanding the data and identifying the best algorithm to use.
3. Model Creation: Train the data to make predictions and assess model results on test data.
4. Feature importance: Identify the important attributes that predict target variable.

Data Pre-processing and EDA

Business data contains attributes of 192,609 businesses. These attributes included location, hours of operation, number of reviews, open status, categories in which the business was listed like Insurance, Dim Sums, Restaurants etc. and attributes like ambience, takes reservations etc. Only stars and number of reviews was numeric, rest are all categorical variables.

After dropping irrelevant variables, data summaries revealed that categories and attributes have missing values. To investigate possible reasons for this missing data, only current open businesses (82%) were included in the analysis. There are 158,525 open businesses in the data which have received 5,626,949 reviews. Next, univariate analysis of stars and reviews showed that stars is negatively skewed. This means that more values are concentrated on the right side of the distribution making the left tail of the distribution longer. Median stars is 3.5 or 50% of open businesses have less than 3.5 stars. Only



Author: Aakanksha Baid

30% of open businesses get the high star ratings of 4.5 and 5.

Examining reviews distribution shows that each business has at least 3 reviews. Reviews is positively skewed having maximum at 8,348, median at 9 while average reviews is 35.5. The high standard deviation of 116.6 implies that reviews has high outliers.

On performing bivariate analysis of reviews and stars, scatterplot shows weak correlation (0.04). This is due to the presence of outlier reviews which alters the linear relationship between these 2 variables. 4-star rated businesses receive the highest median reviews per business of over 14.

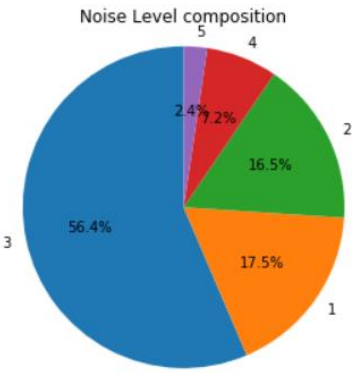
A quick inspection of top 20 most reviewed businesses shows that restaurants located in Las Vegas are the highest reviewed category of businesses. Las Vegas being a prime tourist destination receives lot of attention and high star ratings from Yelpers which drives new customers leading to even more reviews. E.g. Mon Ami Gabi which is a French restaurant in Las Vegas, has the highest number of reviews of 8,348.

As restaurants category received such high proportion of reviews, they became the focal point of this analysis and data was filtered to keep only restaurants giving 42,237 restaurants.

Next, restaurant attributes were split and those having more than 40% missing values were removed from data as their correct imputation would not be possible due to the high variability among restaurants. Some of the excluded attributes included having happy hour, open 24 hours etc. This brought down restaurant attributes from 39 to 17. Also, dropping restaurants with more than 40% missing values left 35,105 restaurants for analysis.

For data munging, these attributes were converted to binary values 0 and 1 and missing values were also converted to 0 denoting absence of the attribute. Ambience which had low missing values but 9 elements like touristy, casual, trendy etc. with 2 levels of true and false each was still excluded due to high cardinality (9*2 = 18). Two attributes “Noise Level” and “Restaurant Price” were kept as ordinal categorical variables as a natural order is present in them. Over 50% of restaurants are average noisy and medium priced ($$).

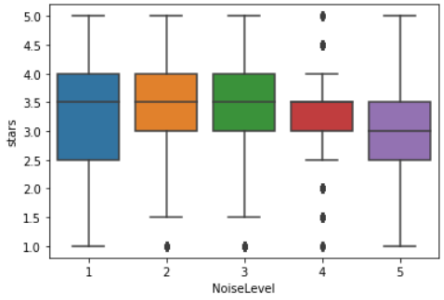
Author: Aakanksha Baid



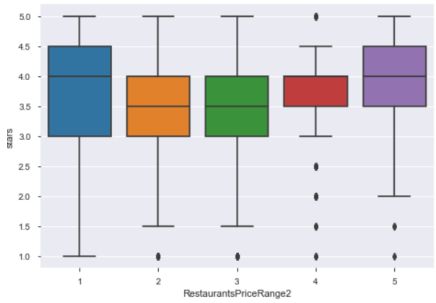
1 - None or missing noise 2 – Quiet 3 – Average 4 – Loud 5 -Very Loud	1 - None or missing price 2 – price range 1 ($) 3 - price range 2 ($$) 4 – price range 3 ($$$) 5 – price range 4 ($$$$)

Proceeding with testing out the hypotheses if certain attributes impact stars.

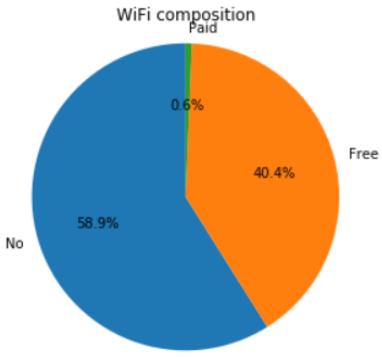
Noise Level - Quiet and average noisy restaurants have identical stars distribution.

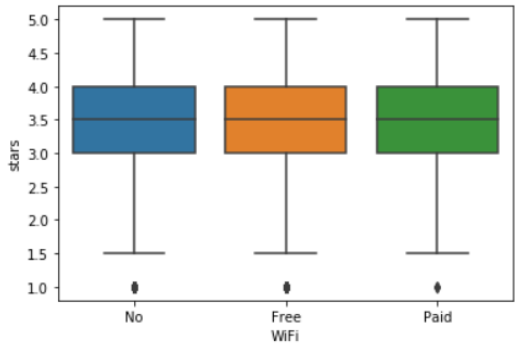


Price - Restaurants in price range categories 2 and 3 have no variation in star ratings. Most expensive restaurants (price range category 5) have the highest median star rating of 4.



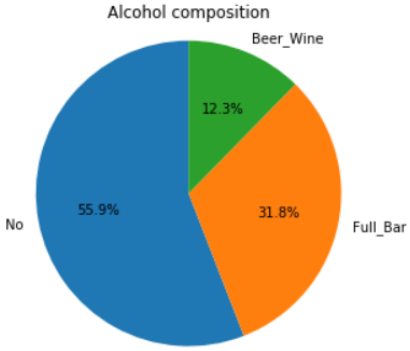
Wi-Fi - Boxplot shows restaurants having no, free or paid Wi-Fi have identical stars distribution. However, outliers are present in all these 3 types of Wi-Fi and around 60% restaurants have no Wi-Fi.

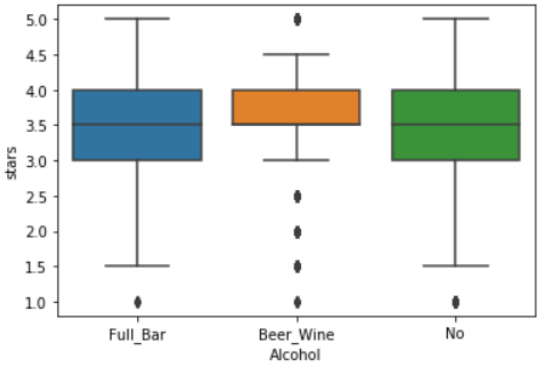




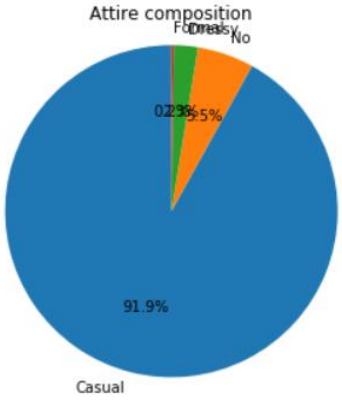
Author: Aakanksha Baid

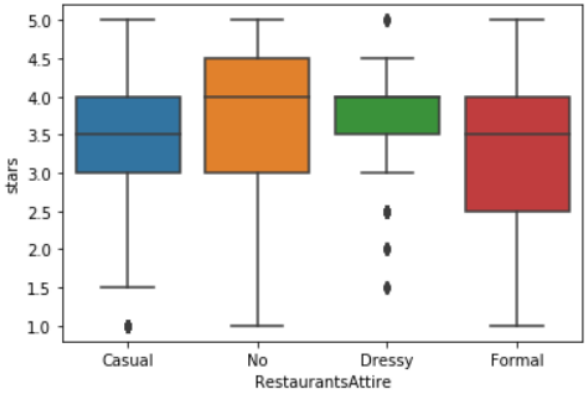
Alcohol - Surprisingly, full-bar restaurants and those serving no alcohol have similar stars distribution. But outliers are present in all 3 alcohol types. More than 50% restaurants serve no alcohol.





Attire - Boxplots revealed that median rating of restaurants with no dress code is highest at 4 and outliers are present in Dressy and Casual attire type. 92% restaurants have a casual dress type.

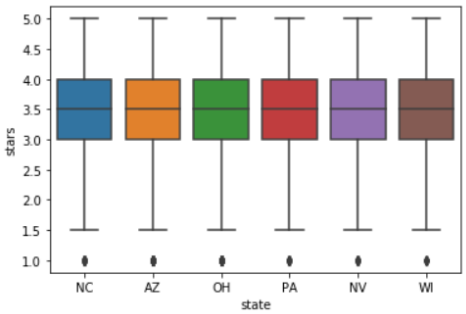


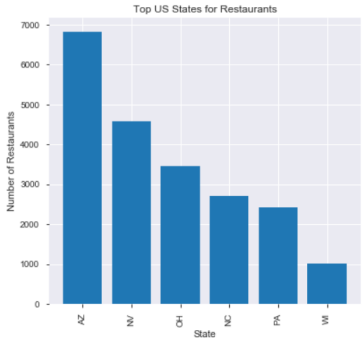


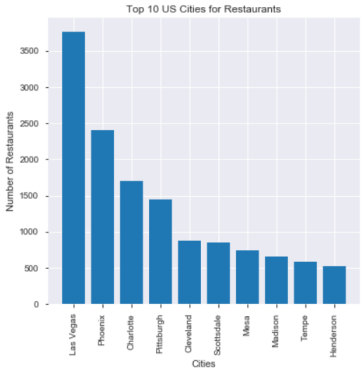
State - A frequency distribution of states showed some high counts for states outside US. As food is affected by a country's culture and topography, the data was filtered for U.S. states only leaving 21,640 restaurants. Subway, McDonald's and Taco Bell have the highest presence in US cities. McDonald's, Hash House A Go Go and Chipotle Mexican Grill are the top American chains receiving the highest reviews.

As states varied highly by the number of restaurants, restaurants were dropped in those states where total restaurants were less than 500. This left 21,016 U.S. restaurants in the data. A boxplot of states on stars shows no variation in stars distribution due to location. But, outliers are present in all states. Arizona has the maximum number of restaurants followed by Nevada. Arizona may be having multiple cities with high density of restaurants. In terms of citites, Las Vegas has the maximum restaurants which can explain the high reviews of restaurants like Mon Ami Gabi.

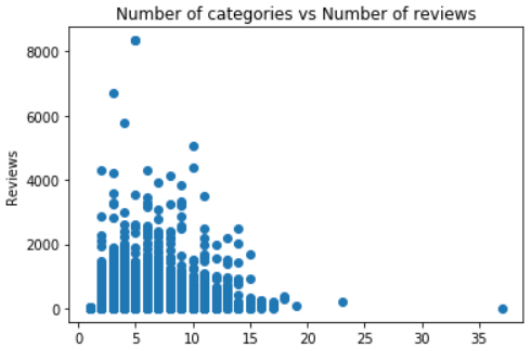
Author: Aakanksha Baid

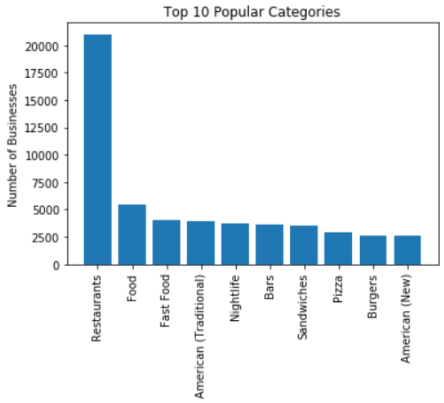




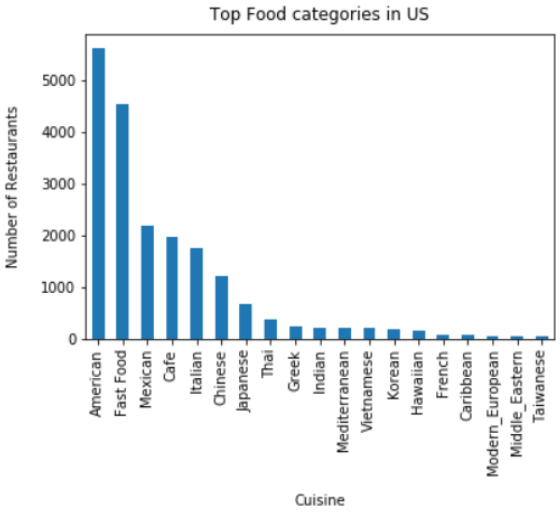


Categories - A restaurant is labelled under several categories. In total there are 510 unique categories paired with restaurants. Among these, the top popular categories are Food, Fast Food and American (Traditional). For example, Taco Bell has categories - Restaurants, Breakfast & Brunch, Mexican, Tacos, Tex-Mex and Fast Food. To test the hypothesis if more categories lead to more reviews, a scatterplot of number of categories vs number of reviews revealed no strong correlation (0.15). Assigning more categories to a restaurant may help in user search but does not impact number of reviews. There are outlier restaurants assigned over 35 categories but still receiving single-digit reviews.



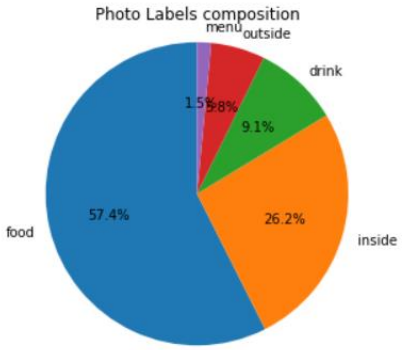


As some of the above categories seem unrelated, imputing cuisines from these categories and dropping restaurants with missing/ unpopular cuisines left 19,656 restaurants in data. Among these, over 5000 US restaurants serve American cuisine.



Author: Aakanksha Baid

Next, examining Photos dataset -
Photo data included photo related information like photo captions and labels etc. for 30,488 unique businesses. Since multiple records of same business was found having different photo labels, businesses were grouped by total photos. As the photo caption variable has over 50% missing values, it was dropped. Photo label dummies were created with binary 0/1 form. Photo data contains 57% Food labelled photos for businesses. This data was merged with the U.S. restaurants data and missing photo labels were made 0.

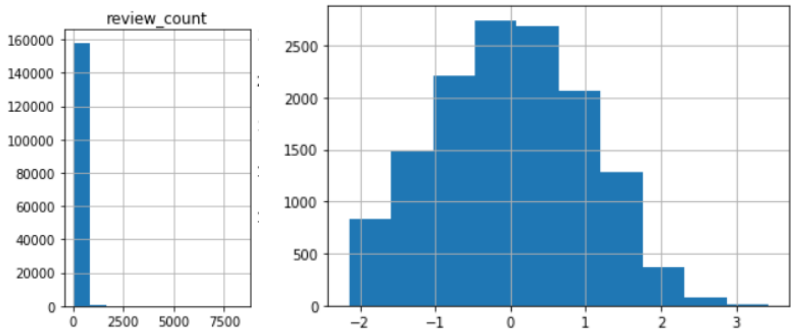


Correlation - A correlation heatmap revealed negative correlation among alcohol types and among attire types while positive correlation among good for meal types and photo label types. To detect presence of multicollinearity, variance inflation factor (VIF) was calculated. VIF estimates how much the variance of a coefficient is "inflated" because of linear dependence with other predictors.

$VIF = \frac{1}{1-R^2}$, where R^2 is the multiple regression of a predictor on other predictors.

A VIF > 10 represent critical levels of multicollinearity where the coefficients are poorly estimated. The 5 dummy variables Casual attire, Noise level, Take Out, Fast Food and American along with price fall under this category. As VIF can be high for dummy variables when reference category used has small frequency which was the case with this data, only the multicollinear price predictor was dropped before splitting data into train and test.

As reviews and photos have right skewed distribution, a power transformation using Box Cox transformer was fit to the training data and test data was transformed accordingly. Below figure displaying before and after transformed reviews from the training data makes reviews distribution Gaussian-like.



Before transformation After transformation

As data was imbalanced (64% of data is high rating class or 1), an attempt at up sampling the low star rating score (0) using the SMOTE algorithm (Synthetic Minority Oversampling Technique) was also tried. However, after over-sampling the training data and using the power transformation to reduce skewness in the data, the mean absolute error (MAE) after modeling increased so the original imbalanced data was used.

Author: Aakanksha Baid

Model Creation

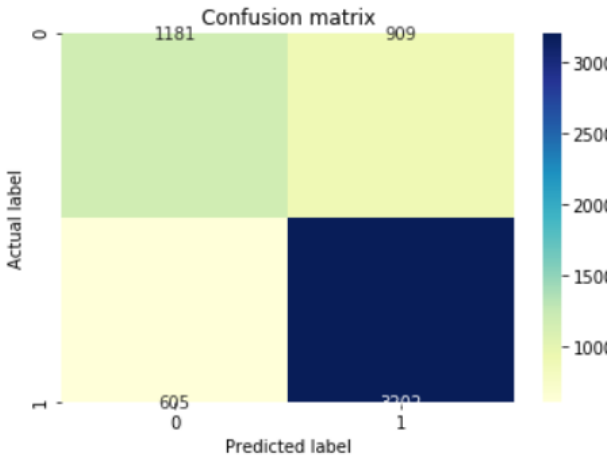
Supervised machine learning algorithms were deployed because the target variable (star rating class) is already known. Since the star ratings are not continuous, classification results are expected to perform better than regression. To simplify the problem and the implementations of algorithms, the target variable stars was converted to binary target star rating class by binning 2 categories – high and low rating, 1 for stars above 3 and 0 for stars 3 or below. The majority star class is 1 (64% of restaurants) giving baseline for model 64%, or the classification algorithm needs to predict with at least 64% accuracy.

The data contains 57 features and one target (low or high rating class). The target is derived from the training set, and its performance is measured on the test set. Two classification algorithms were used -

a) Logistic regression: Logistic regression assumes that the dependent variable is a category and the fitted line predicts the probability that the dependent variable will have a specific value (0 or 1). Logistic regression was fitted on all 57 features.

Model Results

Mean Absolute Error (MAE)	**Accuracy**
It describes the *typical* magnitude of the residuals $\frac{1}{n}\sum \lvert y_{True} - y_{Predicted} \rvert = 0.25$	$\frac{True\ postives+True\ negatives}{All\ Outcomes}$ = 0.75 74% cases are correctly identified.
Precision	**Recall (true positive rate or sensitivity)**
$\frac{True\ postives}{True\ positives+False\ positives}$ = 0.78 78% of positive cases are correctly classified from all predicted positive cases	$\frac{True\ postives}{True\ positives+False\ negatives}$ = 0.84 84% of positive cases are correctly identified from all the actual positive cases.



Confusion matrix - The diagonal elements in the confusion matrix represent number of cases for which predicted class is correct while off-diagonal elements are incorrectly classified.

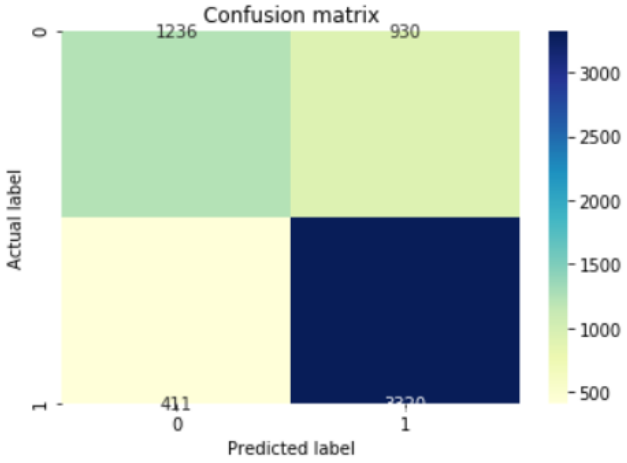
Author: Aakanksha Baid

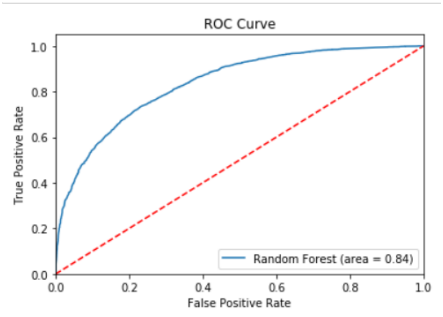
Evaluation Metrics – As target classes are imbalanced, accuracy is not an appropriate model performance metric. ROC AUC and MAE will be used as evaluation metrics.

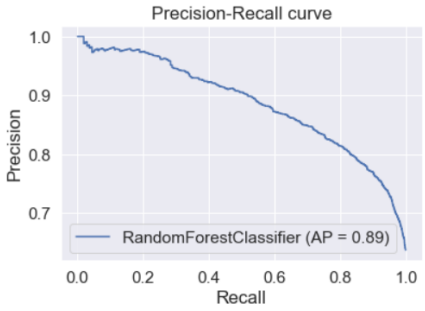
The receiver operating characteristic (ROC) curve which illustrates the diagnostic ability of a binary classifier system has area under the curve (AUC) as 0.79. ROC is a probability curve and AUC represent degree or measure of separability or discrimination. Higher AUC implies better correct predictions of the model (0's as 0 and 1's as 1). So logistic regression is giving average result as assumptions are violated in this data and still has some classification errors.

b) Random Forest – This algorithm was deployed to improve model results as it has relaxed assumption, provides better predictive performance, low overfitting, and easy interpretability. Compared to the logistic regression model performance, the random forest model's

- Accuracy increased to 0.77 from 0.75
- Precision remained the same at 0.78
- Recall increased to 0.89 from 0.84
- MAE reduced to 0.23 from 0.25
- ROC AUC increased to 0.84 from 0.79







Precision Recall curve - Average precision is the area under precision recall curve and is 0.89. It is the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

AP = $\sum_n (R_n - R_{n-1}) * P_n$, where R_n and P_n are the precision and recall at the nth threshold.

Thus, random forest performs better at classification and is used to find important features that affect star rating class. Reviews, number of categories, food photos and noise levels are the top 4 predictors.

Feature Selection -
In the Random Forest model, 57 features were used, not all of which provide strong predictive capability to the model and can, thus, be removed through feature selection, which is advantageous because –

Author: Aakanksha Baid

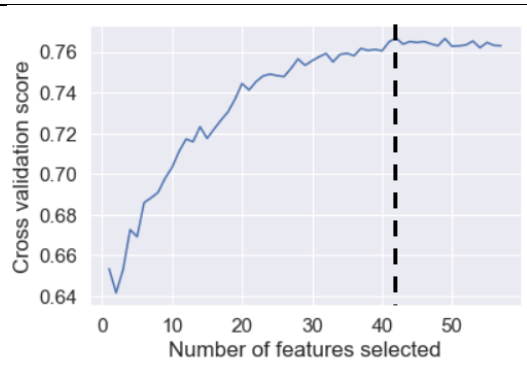
- It helps to avoid overfitting as it removes redundant data
- it may increase prediction performance, as learning will concentrate only on meaningful data
- it reduces execution time and is memory-efficient as there is less data to process

As variables are correlated, univariate selection method like chi-square test is not used for feature selection as the results are reliable only for independent variables.

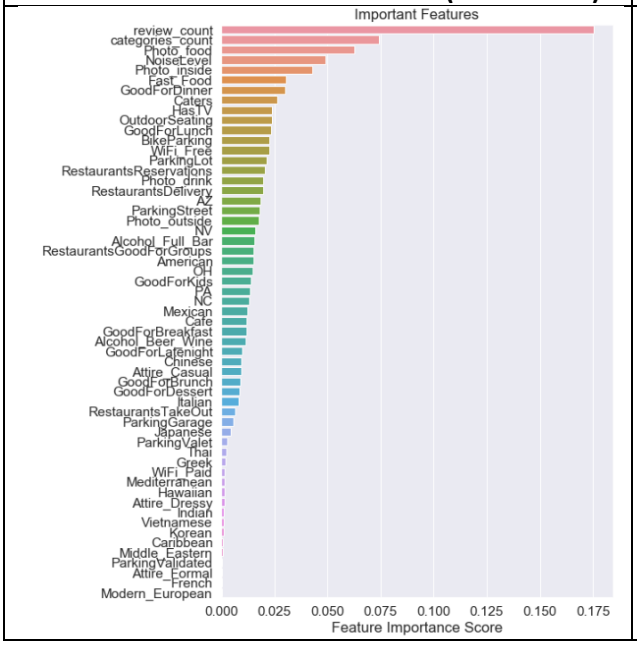
RFE for feature selection - Using recursive feature elimination (RFE), a feature pruning technique that fits the specified model and removes the weakest features (lowest standardized coefficient) iteratively until the specified number of features is reached. To find the optimal number of features and detect overfitting, cross-validation is used with RFE where the fitting is accompanied by testing, using a training set and test set chosen according to a folding parameter to score different feature subsets and select the best scoring collection of features. 42 features are selected according to F-1 score while retaining the accuracy of 0.77.

F-1 score is the harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases than accuracy in the case of imbalanced classes. Harmonic Mean penalizes the extreme values.

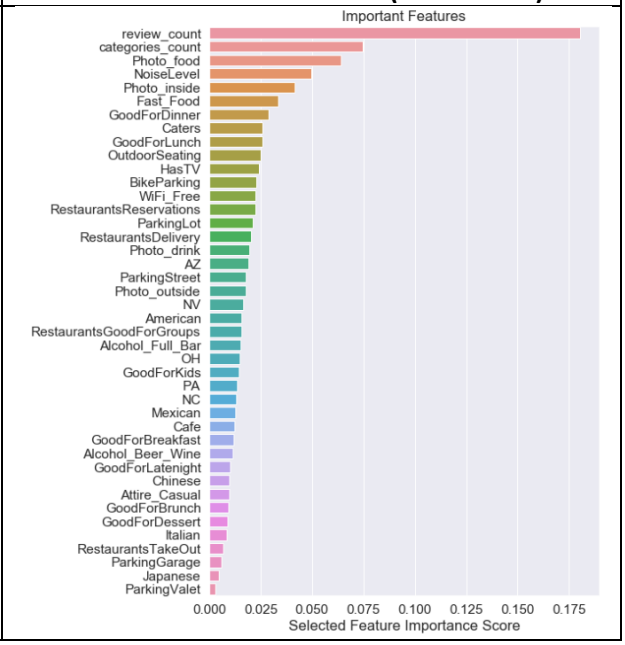
F-1 score = $2 * \frac{Precision * Recall}{Precision + Recall}$



Random Forest Feature Selection (57 features)



RFE-CV Feature Selection (42 features)



Author: Aakanksha Baid

Conclusion

Using Random Forest classification, 77% accuracy along with 84% ROC area is achieved to predict star rating class (high and low) for US restaurants. 42 features are selected by recursive feature elimination with cross-validation technique.

Future work may involve analysis of Yelp's other datasets like reviews, user etc. to understand if other features increase the predictive power of star rating class. Other complex classification algorithms can also be explored to improve model performance.

Yelp can use this model to advise restaurants (as a paid service or as part of its business analytics suite) on what factors lead to high star ratings. The restaurants can then focus on specifically improving/ targeting those attributes to drive up their ratings for local customer growth.

References

https://towardsdatascience.com/feature-selection-in-python-recursive-feature-elimination-19f1c39b8d15

http://www.stat.cmu.edu/~larry/=sml/forests.pdf

https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f