# **Exploratory Analysis：**

## **Basic Statistical Analysis and Data Cleaning Insights**:

### Data Cleaning and feature engineering are essential steps prior to any statistical testing or data analytics in a data science project. This section illustrates the procedures and steps taken for data cleaning and feature transformation including basis statistics analysis, dealing with missing data, outlier detection and feature categorization.

### **Basic Statistics Analysis:**

There are overall four aspects in the restaurant dataset. The first dimension is the restaurant basic information, including number of reviews, ratings, category of the results and so on. The second dimension is the internal attribute of the restaurants, like WIFI option, Ambience, and Alcohol Availability. The third dimension is information about surrounding facilities near the restaurants, like number of schools, shopping malls or bus stops. The last dimension consists of the demographical information of the restaurants nearby, including the proportion of white people, unemployment rate, education levels and so on. In summary, there are altogether 5133 rows of data, extracting from 7 major cities and metropolitan areas, and 59 dimensions representing different aspects of the restaurants.

Among the 79 features, there are 56 numerical features, and 23 categorical features. Table1 shows the physical meaning and types of each columns in the data.

Table 1 Dataset Feature List



Table 2 shows a snapshot of basic statistics of numerical columns, including distinctive count, mean, min, max and quantiles. Table 3 shows the mode of some categorical features in the dataset. As shown in the following tables, there are missing values in some columns like ‘bank’, ’bar’. Therefore, dealing with the missing values will be the first step for data cleaning.

Table 2 Basic Statistics of Numerical Columns

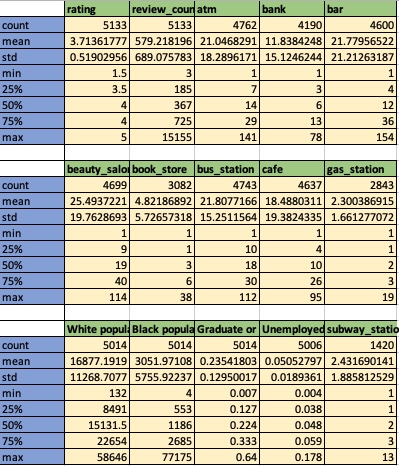
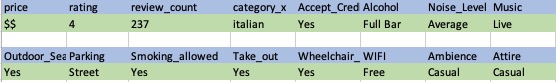
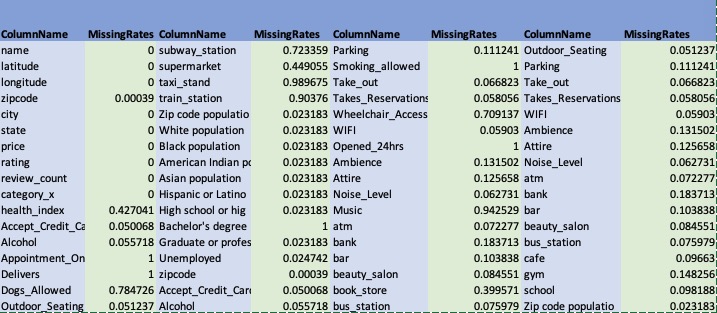


Table 3 Mode of Categorical Features



### **Handling with Missing Values**

Table 4 Missing Rate for Columns



### As shown in table 4, the missing rate varies significantly over different features. Therefore, different missing value handling techniques need to be applied. The missing values are mainly due to information missing in the Yelp website and Google Map, as not all restaurant has all the features above. In general, three different techniques have been applied to deal with NA values in this dataset.

### First, for all features with more than 40% of missing rate are dropped off, as there will never be a fair imputation method to fill in the blanks. Second, for majority of the remaining features, imputation by median/mode to fill in the blanks. The rationale behind is that variance in some of the numerical columns are large, imputation by median will introduce less variance towards the dataset. Thirdly, for missing data in columns like ‘LowPrice’, the values can be inferred by the price range column in the dataset.

### **Outlier Detection**

Although the dimensions in this dataset all have physical meanings and extreme values happens occasionally, outlier detection and handling are still needed as outliers will create barriers for obtaining high accuracy machine learning models. Three outlier detection methods are applied to detect extreme values in this dataset.

#### Z-Score Based Methods

The most common outlier detection method is to make use of box plot and z-score to flag out any value that is far away from the population mean in a univariate manner. Figure 1 illustrates the boxplot for 5 numerical features in the dataset, where there are some high spikes exist in those columns. A z-score threshold of 3 are applied to flag out any point far away from its mean.

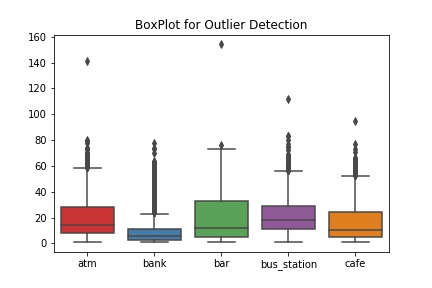


Figure 1 Outlier Detection using Box Plot

#### Local Outlier Factor(LOF):

Local outlier factor is an outlier detector for finding anomalous data point by measuring the local deviation of a given data point with respect to its neighborhood. Figure 2 illustrates the results for applying LOF methods on the 5 features shown above in a multi-dimensional way. The following plot is plotted using PCA decomposition, and the radius of the cycle denotes the outlier score. The larger the radius of the cycle, the higher chance the data point is an outlier.

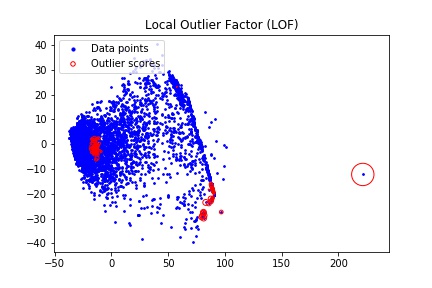


Figure 2 Outlier Detection using LOF

#### Isolation Forest:

Employing decision tree-based technique, isolation forest detects the outliers by assuming that outliers are rare and different from the main population, and therefore it is easier to be spitted out using shallow decision trees. Figure 3 illustrates the results of isolation forest using same feature stated above, the dot in red are treated as outliers.



Figure 3 Isolation Forest Outlier Detection

Data point will be treated as outlier if all three of the methods repost certain data point as outlier. In this case, there is one common data point being flagged by all three methods, and it will be removed from the analysis.

### **Binning Features**

As shown in table2, some of the features will have a large standard deviation, and the distribution is highly skewed. Data binning or categorization is a useful method to deal with such situation. In this dataset, one important data issue is that the review counts are highly variant, and it adds difficulty in the classification task. Figure 4 shows an effective binning method in deal with high variance review count column. The binning strategy here is to make sure the frequency in each category are comparable, in order to facilitate later data analytics.

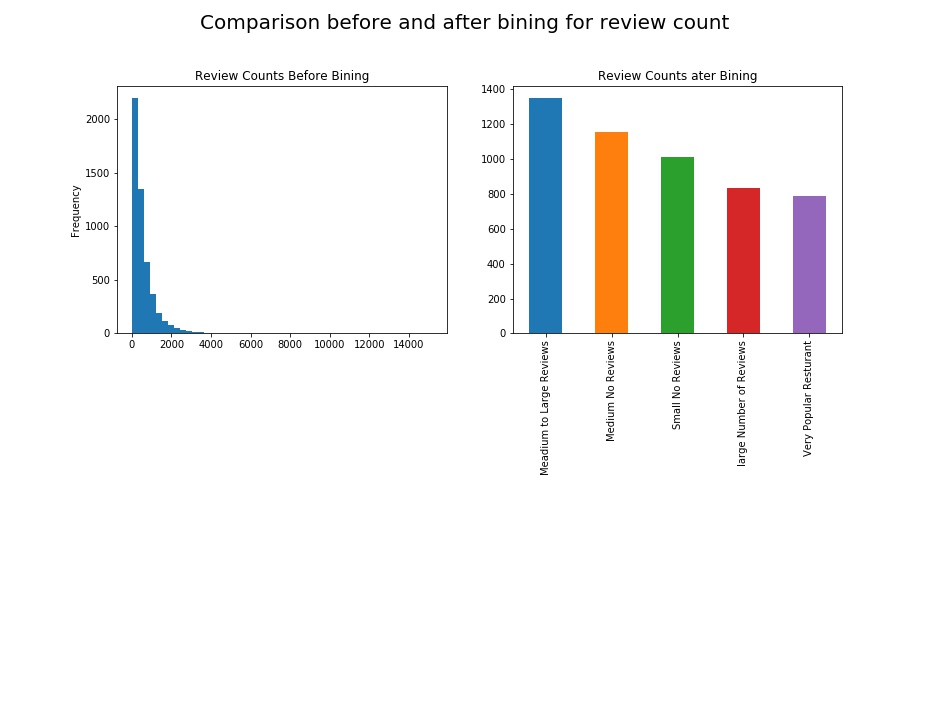


Figure 4 Data Binning Example

### **Additional Feature Engineering**

In additional to the stated data cleaning and feature generation method, there several other steps being taken to transform the data features to more user-friendly format. For example, any Boolean column with value ‘Yes’ or ‘No’ will be filled by ‘1’ and ‘0’ so that it can be used directly in any numerical analysis. In addition, instead of using absolute population number, the ratio of each race has been calculated against the total population within a neighborhood.

## **Predictive Analysis:**

### **Hypothesis Generation and Proposed Methods:**

Based on the exploratory data analysis and heuristics, several hypothesis has been brought out to extract more insights from the dataset:

First, two of the most significant attributes in the restaurant data set are review counts and price. The former is a great indicator of the popularity of the restaurant, and the latter is an important metric for customer to rate the restaurant. Therefore, the first hypothesis being brought up is whether the review count distributions are the sample among different price groups. The null hypothesis is that the distribution of review counts are the sample among all 4 price groups, namely ‘$’,’$$’,’$$$’ and ‘$$$$’ To verify the hypothesis, Analysis of Variance(ANOVA) method will be performed. Since there multiple levels in the price features. T-test will also be applied to further support the hypothesis.

Second, intuitively, popularity and ratings of the restaurants should have a strong association. Popularity can be measured by review counts. The second hypothesis is that review count and ratings of the restaurant have a linear relationship. This can be tested using a linear regression model.

The third hypothesis states as the good rating restaurants, moderate rating restaurants and good restaurants can be well-separated using the features listed in the dataset. In other words, the hypothesis stats that there is a clear decision boundary between different classes of the restaurants. Testing this hypothesis is the processing of building a multi-class classification model. To verify the hypothesis, logistics regression and other data driven machine learning models will be applied.

### **Class Label Generation:**

One of the most important tasks prior to any supervised classification task is to make sure the data is properly labeled. The objective in this dataset is to predict whether a restaurant are good, and the most direct metric is the rating provide by Yelp. However, one issue with this label is that number of reviews will potentially influence this rating. To compensate for the bias introduced by review count, a new fusion metric has been introduced. The new rating is calculated by treating the original rating minus 0.5 as base score and it will be penalized by the review count z-score within its original base score group. For example, if a restaurant has a rating of 4, its base score will be 3.5 and the final score is calculated by 0.5 times the ratio of the review count of this restaurant to the maximum review count of restaurants whose rating are also 4.

After the final score has been calculated, the class label is generated by binning the final score. Three classes have been generated, namely 0, 1 and 2, they represent poor rating restaurant, moderate rating restaurant and good rating restaurant respectively. The objective of the classification task is to correctly classify each restaurant to the appropriate class using supervised classification techniques. Table 5 shows the class distribution in the class label.

﻿

Table 5 Class Label Distribution

|  |  |  |
| --- | --- | --- |
| Class Label | Count | Physical Meaning |
| 0 | 2482 | Poor Rating |
| 1 | 2151 | Moderate Rating |
| 2 | 499 | Good Rating |

### **Parametric Statistical Methods:**

#### T-Test and ANOVA:

#### ANOVA test is used to test the first hypothesis, which states that there is no significant deference on the number of reviews(review\_counts) and price range. Figure 5 illustrated a two way ANOVA test results. And Figure 6 shows the Quantile-Quantile Plot of theoritical quantiles and sample quantiles in the test.

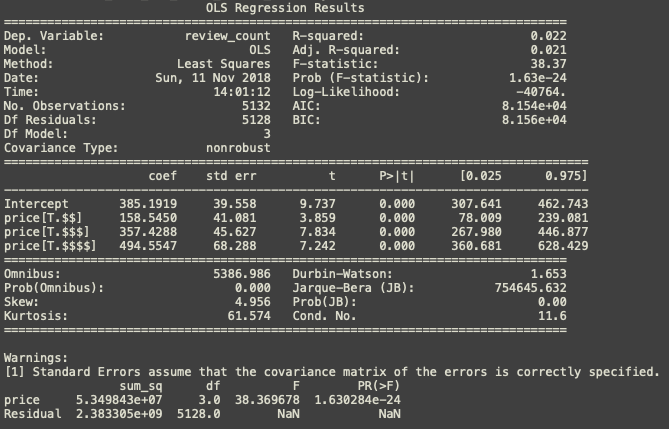


Figure 5 ANOVA Test Result



Figure 6 ANOVA Q-Q Plot

As illustrated above, there is an significant number of sum of square error, a F score of 38.37, and a near 0 p-value. All those statistics shows strong evidence against the null hypothesis, therefore, the hypothesis is not valid. There are significant differences in review counts among different price groups.

To further verify the results, t-test is conducted among pair-wisely among different price groups. Figure 7 shows an example of price group ‘$’ against other price groups. Again, this is an evidence against the proposed hypothesis.

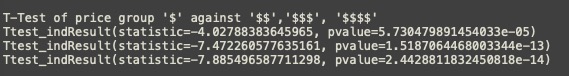


Figure 7 Pair-wise T-test Example

#### Linear Regression

Linear regression is one of the most effective way in examining the liner relationship between two variables. To verify the second hypothesis, a linear model has been fitted into the data and figure 8 shows the results of the linear regression model. The R-score obtained by this model is 0.16, which suggest that the liner relationship between this two variables is weak.



Figure 8 Linear regression Model Result

### **Logistics Regression Classifier and other Data Driven Predictive Models**

In next portion, six data driven predictive models are applied to test the third hypothesis. And all six methods utilize the same set of training and testing data. K-fold cross validation method has been used to test the robustness of the model, and ROC-AUC plot and confusion matrix for each model will be generated separately.

#### Logistic Regression

Logistics Regression is a widely used statistic model which utilize logistic function to model binary/multiclass dependent variables, in this case, the restaurant classes. Figure 9 and 10 illustrated the Receiver Operating Characteristic (ROC) curve for all class label and the confusion matrix heat map

#### 

Figure ROC for Logistic Regression

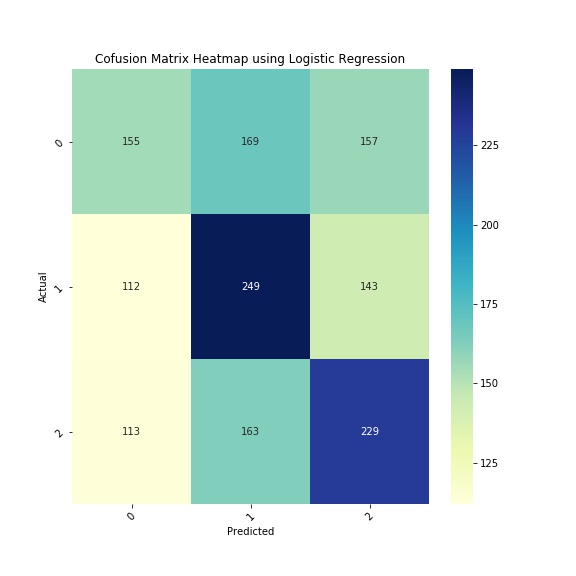


Figure Confusion Matrix Heatmap for Logistic Regression

#### Decision Tree Classifier

### Decision tree-based classifier is one of the most-easy-to-understand classification techniques. The properties that the results of a decision tree classifier can be easily interpreted has make it popular. Figure 11 and 12 illustrated the Receiver Operating Characteristic (ROC) curve for all class label and the confusion matrix heat map.

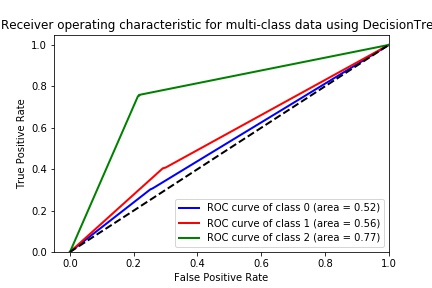


Figure 11 ROC Curve for Decision Tree



Figure 12 Confusion Matrix Heatmap for Decision Tree

#### Naïve Bayes Classifier

### Naïve Bayes classifier is a probabilistic classifier based on Bayes theorem by assume the data feature are independent. Gaussian Naïve Bayes are uses here to deal with the continuous values in this data set. Feature 13 and 14 illustrated the ROC curve for all class label and the confusion matrix heat map.

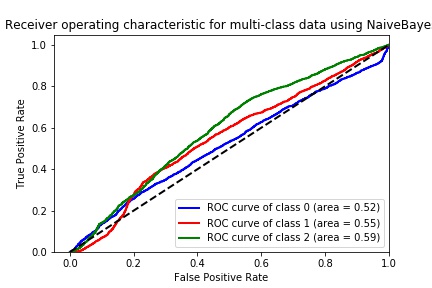


Figure 13 ROC curve for Naive Bayes Classifier

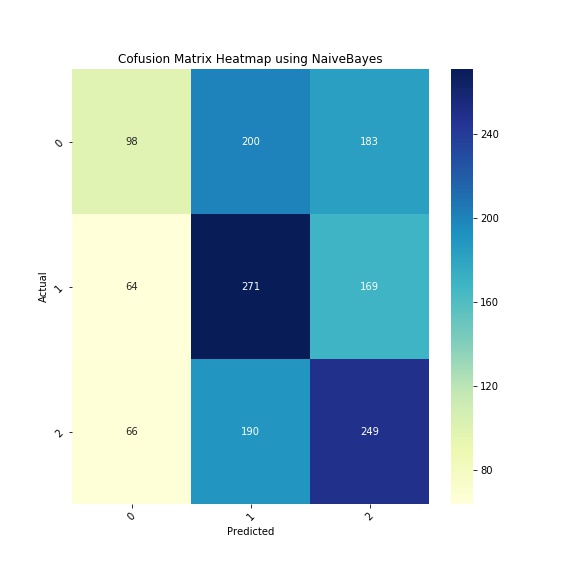


Figure 14 Confusion Matrix Heatmap of Naive Bayes

#### K-Nearest-Neighborhood (KNN) Classifier

As a lazy learner, KNN classifier is an instance-based learning technique by applying majority vote principle in its neighborhood data point, while the neighborhood is obtained the pairwise distance measure. Feature 15 and 16 illustrates the ROC curve for all class label and the confusion matrix heat map.

### 

Figure 15 ROC for KNN

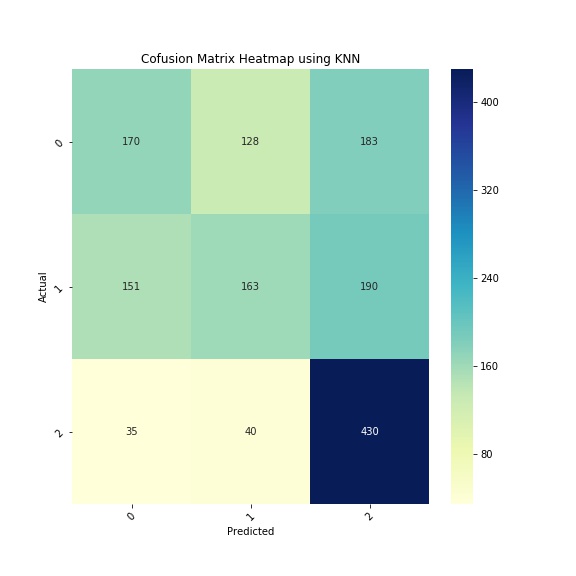


Figure 16 Confusion Matrix Heatmap for KNN

#### Support Vector Machine (SVM)

Support Vector Machine is a popular supervised technique to deal with non-liner classification or regression problems. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In this case, a radial basis function kernel is applied. Figure 17 and 18 illustrates the ROC curve for all class label and the confusion matrix heat map.

### 

Figure 17 ROC for SVM

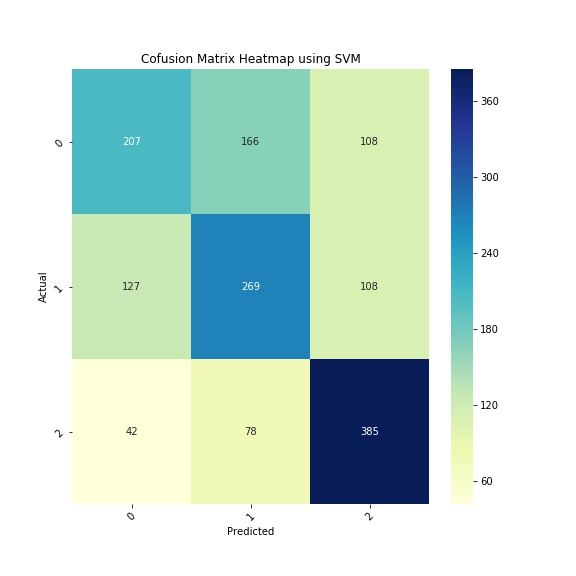


Figure 18 Confusion Matrix Heatmap for SVM

#### Random Forest

Employing bagging principle, random forest is an ensemble-tree based machine learning technique in classification and regression problems. Figure 19 and 20 illustrates the ROC curve for all class label and the confusion matrix heat map.

### 

Figure 19 ROC for random Forest

### 

Figure 20 Confusion Matrix Heatmap for Random Forest

### **Results Comparison and Discussion**

Based on the model results above, there are several observations we can infer from the ROC curves and confusion matrix:

1. The data features are not independent, this is the primary reason why Naïve Bayes classier generated near random results.
2. The problem is not a linear-separable problem, therefore SVM or decision tree-based methods achieved relatively better classification accuracy.
3. Class 0 and Class 1 are not as distinctive as Class 2, none of the classier achieved good separation between Class 0 and Class 1. This suggests that the class label generation may be biased.
4. Out of the 5 classifiers, Random Forest achieved the best AUC score for Class 2. A AUC of 0.92 suggest that the dataset does have predictive power for the rating of restaurant based on the information given.