<u>Research Question</u>: Use Multinomial Logistic Regressions to predict star ratings for restaurants in Arizona and identify the most important features that contribute to better predictive accuracy.

Introduction

In this report we compare the performance of two different multinomial logistic regression models in predicting star ratings for restaurant businesses in Arizona. We then use the best model to identify the most important features that contribute to higher predictive accuracy of the ratings category.

Methodology

After the initial data cleaning and pre-processing phase, we end up with a total of 21 features. We know that the large number of features and their interdependencies may introduce overfitting and may hinder the performance of our classification model. When multicollinearity occurs among variables, the estimated logistic regression coefficients may be inaccurate, thus reducing the predictive power of the model.

To address the issue of multicollinearity we have used two different approaches:

- **Approach 1:** We use correlation analysis to select a subset of feature based on the Spearman correlation coefficient for numerical variables and Cramer's V coefficient for categorical variables.
    - We first identify pairs of highly correlated features using a threshold of 0.5.
    - For each pair we select the feature that has the highest correlation with the target variable, to use as a regressor in our classification model.
- **Approach 2:** We use PCA to map our set of correlated variables onto a new set of completely uncorrelated variables (called principal components).
    - Principal components are ranked in order of importance based on the fraction of the variance of the original data that they can explain.
    - We select the principal components **that explain at least 95% of the variance** of the original dataset to use as regressors in our model.

In conclusion, we end up with two multinomial LR models to predict star ratings, that are structurally similar, however their feature space is different.

We select the best performing model relying on the out-of-sample accuracy and the class-based comparisons of the F1 and AUC scores. The selected model is used for the interpretation of results.

Data Pre-processing

The data comes from Yelp Dataset Challenge. In this project, we focus on restaurants in Arizona. To increase the robustness of our analysis we select only the restaurants that are currently active, as denoted by the 'Open' column in the dataset. This filtering criteria allows us to reduce the sample size from **6817** to **4925** restaurants.

Another important part of the data pre-processing phase is feature engineering. Using domain knowledge and the information contained in the original features, we construct five additional variables. All categorical features in our analysis are encoded and are therefore represented either

as binary or nominal variables. We end up with a list of **5 numerical features and 16 categorical features** which are encoded for the purpose of building our predictive model.

For further information on the data pre-processing refer to **Appendix 1.**

Exploratory data analysis

We make the following important observations based on the Exploratory Data Analysis. See **Appendix 2** for further information.

- **Distribution of ratings:** We observe that there is enough variation in the ratings category for prediction purposes. The two most frequent ratings are 3.5 and 4.0 accounting for 30% and 35% of our sample. While the 5-star rating accounts only for 1% of our sample. Also, it is worth noting there are no restaurants in our filtered sample that have a 1-star rating.
- **We analyse the pairwise relationship between numerical features in our dataset.**
    - Based on an initial visual inspection of the histograms, our variables do not seem to be normally distributed. We also check this observation quantitatively by employing the **Shapiro-Wilk test** for normality.[1]
    - The scatter plots do not support linearity between the pairs of numerical variables in our dataset.

These observations suggest that Pearson correlation coefficient might not be appropriate for our analysis, therefore we use Spearman's coefficient to understand the correlation structure of the numerical features in our dataset.

Multicollinearity and Dimensionality reduction

When there is a multi-collinearity among explanatory variables, the estimation of the logistic regression coefficients may lead to invalid statistical inference. We have addressed this issue using two different approaches:

Approach 1: Correlation analysis

Correlation of numerical features

As explained in the previous section, the bivariate relationships between our numerical features are not linear, moreover these variables are not normally distributed.

This suggests that the Spearman rank correlation may be more appropriate because: i) it does not carry any assumptions about the underlying distribution of the data and ii) it captures non-linear dependencies. Moreover, Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. These properties make it a suitable candidate for our correlation analysis of numerical features.[2] See **Appendix 3** for further details on the calculations and the correlation matrix.

[1] Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)"
[2] Page, E. B. (1963). "Ordered hypotheses for multiple treatments: A significance test for linear ranks"

We observe the that the pair of **(review count; tips)** with a correlation of 0.88, is highly correlated based on our 0.5 threshold. In this pair, review count is the feature with the highest correlation with the target variable i.e., star-ratings. Therefore, based on Approach 1 we will drop the variable 'tips' and keep 'review count' as a regressor in our multinomial LR model.

Correlation of categorical features

However, Spearman correlation is not appropriate for categorical variables. Multiple measures of association exist that quantify the mutual dependency between two categorical variables. In our analysis we use the Cramer's V test. [3] Cramer's V is used as a post-test to determine strengths of association between two variables after chi-square has determined significance. Chi-square says that there is a significant relationship between variables, but it does not say just how significant and important this is. Cramer's V is a post-test to give this additional information. See **Appendix 3** for further details.

Out of the highly correlated categorical pairs *(listed in Appendix 3)* we will only keep the variable 'Cuisine'.

Testing Statistical significance of correlation coefficients

We quantitatively establish random permutation tests to check whether correlation coefficients are indeed statistically significant or whether they are a consequence of random fluctuations. We observe that the coefficients associated with the highly correlated numerical and categorical pairs are statistically significant. See **Appendix 3** for further information. This test provides a robustness check to our feature selection process.

A summary of the final subset of features that will be used as regressors in the first multinomial LR model is provided in **Table 3, Appendix 3.**

Approach 2: PCA

The scope of this analysis is to map a set of correlated variables onto a new set of completely uncorrelated variables (called principal components). Principal components can be ranked in order of importance based on the fraction of the variance of the original data that they can explain. In our analysis we will select the principal components **that explain at least 95% of the variance of the original dataset**. See **Appendix 3** for further details.

To correctly apply PCA we need to standardise our numerical features to mean zero and unit variance. This standardisation requirement and the fact that it assumes linear relationship among variables, suggests that PCA is not a very suitable method for dimensionality reduction of categorical variables. [4]

Acknowledging the limitations of the PCA and since most of the categorical variables in our analysis have a relatively low correlation with the target variable, our approach is to use PCA only on the set of numerical features and on a subset of the top 5 categorical features that that have the highest correlation with the target variable.

A summary of the set of variables that will be used in the PCA is provided in **Table 4, Appendix 3.**

[3] Baak et al (2019). "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristic."

[4] Nguyen and Holmes (2019): 'Ten quick tips for effective dimensionality reduction'

In our analysis we select the principal components that explain **at least 95% of the variance** of the original dataset. This amounts to **2 PCs**. We then use uncorrelated PC instead of original correlated variables to regress the target variable in our second multinomial LR model. [5]

Multinomial Logistic Regression

To predict the star ratings of the restaurants in AZ, we perform standard multinomial logistic regressions for both models using ”mnrfit” in MATLAB. Multinomial logistic regression predicts the probabilities of the dependent variable falling into each of the star-rating categories and classifies the prediction into the class with the highest probability. Our analysis is as follows:

- First, **we split the sample into the train set and test set.** In order for all ratings to be represented with the same proportions as in the whole dataset, we assign a fraction of 70% of the businesses in each rating category to the training set, and a fraction of 30% to the test set.
- **We calibrate the multinomial logistic regression** model in the training set. The model's structure is: $\log\frac{\mathrm{p_j}}{\mathrm{p_K}} = \beta_{\mathrm{jo}} + \beta_{\mathrm{j1}}\mathrm{x_{i1}} + \cdots \beta_{\mathrm{jM}}\mathrm{x_{iM}} = \boldsymbol{\beta_j} \cdot \boldsymbol{x_i}$, where class K is kept fixed as a pivot class (in our model class K corresponds to the 5-star rating category). The model's solution is a list of K-1 vectors of coefficients $\boldsymbol{\beta_j} = (\beta_0, \beta_1 \dots \beta_M)$.
- Once the coefficients are estimated we can **compute the in-sample log likelihood** for Model 1 and Model 2, as the sum of log probabilities of observing the train data if they were generated by the model in use. The predicted probabilities for each class given our set of independent variables and estimated coefficients, are computed as follows:

$$\mathrm{P}(class\ j \neq K) = \frac{\exp\left(\boldsymbol{\beta_j} \cdot \boldsymbol{x_i}\right)}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{\beta_k} \cdot \boldsymbol{x_i})}$$

$$\mathrm{P}(class\ K) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{\beta_k} \cdot \boldsymbol{x_i})}$$

  - We then **evaluate the out of sample accuracy** for each model. A not very rigorous possibility is to consider the predicted star rating for each business in the test set, as the one to which the model assigns the highest probability and compute the out-of-sample fraction of objects assigned to the correct class.
- We then perform **class-based comparisons** to identify which classes are better predicted by the competing models through the F1 and AUC scores.
- **We select the model that has the highest out-of-sample accuracy** and the best overall performance for different classes. We use this model to interpret results and identify the most important features that contribute to better predictive accuracy of the ratings.

Results and Interpretability

We summarise the results of our in-sample log-likelihood and the out-of-sample accuracy for both models in the table below:

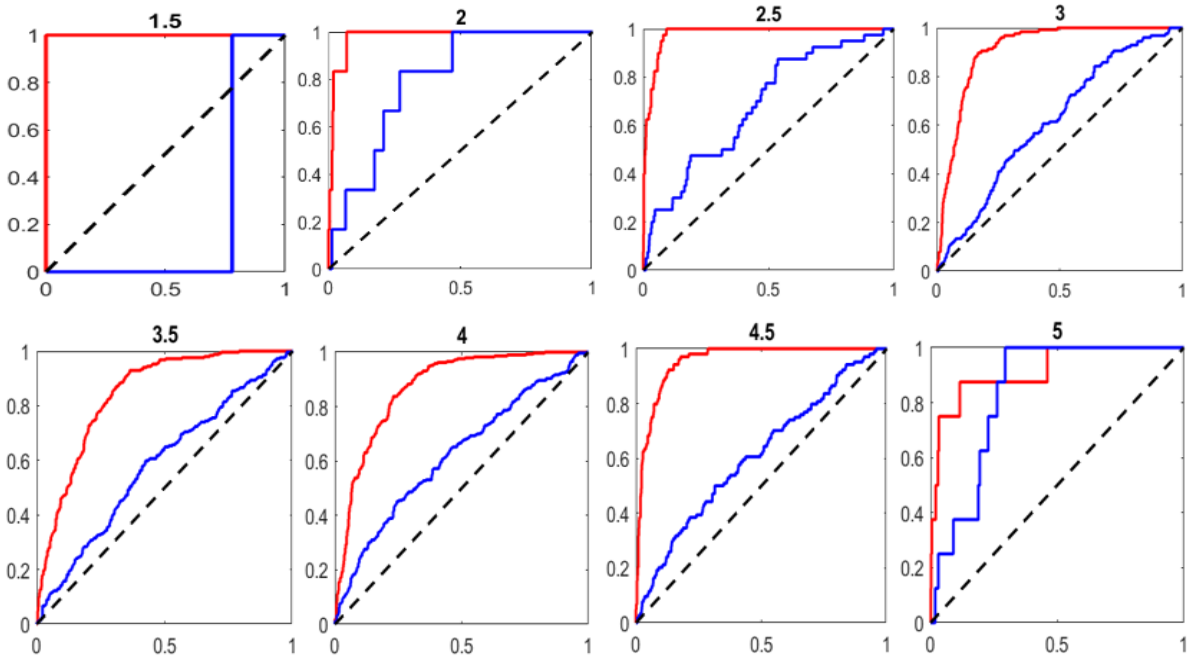
[5] Aguilera et.al. (2006): 'Using principal components for estimating logistic regression with high-dimensional multicollinear data'

Table 1: In-sample log likelihood and out of sample accuracy

	Model 1 (Correlation analysis)	Model 2 (PCA)
In-sample log likelihood	-1.51E+03	-2.73E+03
Out of sample accuracy	0.64	0.36

These two global indicators suggest **that Model 1 has the best performance**. This implies that the joint probability conditional on the training dataset (i.e., the probability of observing the data in the train set), is higher when using Model 1. Additionally, **64%** of the out-of-sample restaurants are assigned to the correct class. We have also assessed the performance locally, using **class-based comparisons** to identify which classes are better predicted by the competing models. Based on the AUC score we observe that **Model 1 consistently outperforms Model 2** across all classes. This is also supported when looking at the F1 scores for each class.[6]

Figure 1: Class based comparisons of Model 1 (in red) and Model 2 (in blue) based on the AUC Score



Note: There are no restaurants in our filtered sample that have a 1-star rating.

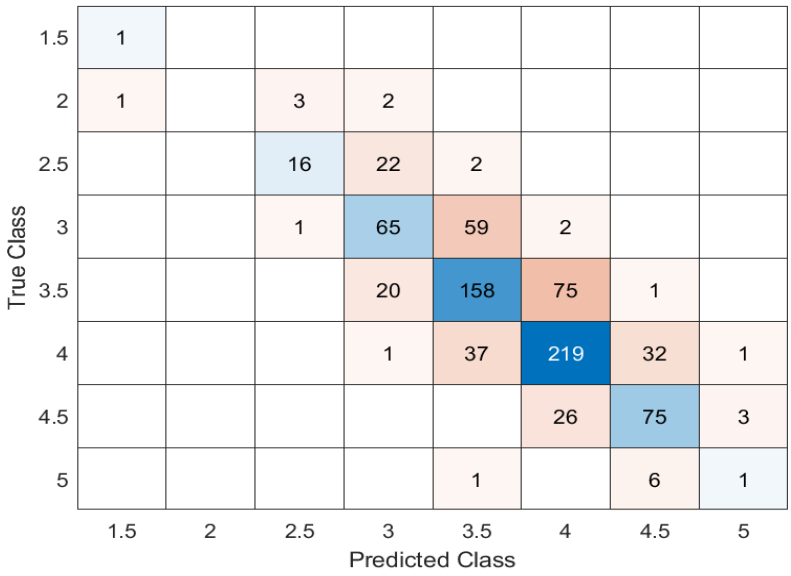
[6] We observe some classes with F1 score of zero. This should be interpreted that across our testing sample no observation has the highest probability corresponding to that class. This does not imply that AUC is zero because each class will have an assigned probability and by construction when creating the AUC plot, observations will be assigned to that class for sufficiently low thresholds.

Table 2: Class based comparisons of Model 1 and Model 2 based on the AUC Score

	F1-scores (Model 1)	F1-scores (Model 2)
1.5 rating	0.67	0.00
2 rating	0.00	0.00
2.5 rating	0.54	0.00
3 rating	0.50	0.17
3.5 rating	0.62	0.39
4 rating	0.72	0.50
4.5 rating	0.70	0.05
5 rating	0.29	0.00

In our setting, even though a prediction is wrong, we also want to know how much deviated the prediction is: for a restaurant with a 4.5-star rating, classifying it as a 1.5- star restaurant is a totally different story from classifying it as a 4-star restaurant, even though both predictions are wrong. For this we use the confusion matrix. From the confusion matrix, we can see that even when the prediction is wrong, our predicted value is still clustered around the true value. This implies that **even if our prediction is not 100% accurate**, **Model 1 still provides a good estimate of the true value**.

Figure 2: Confusion matrix for the best performing model (Model 1)



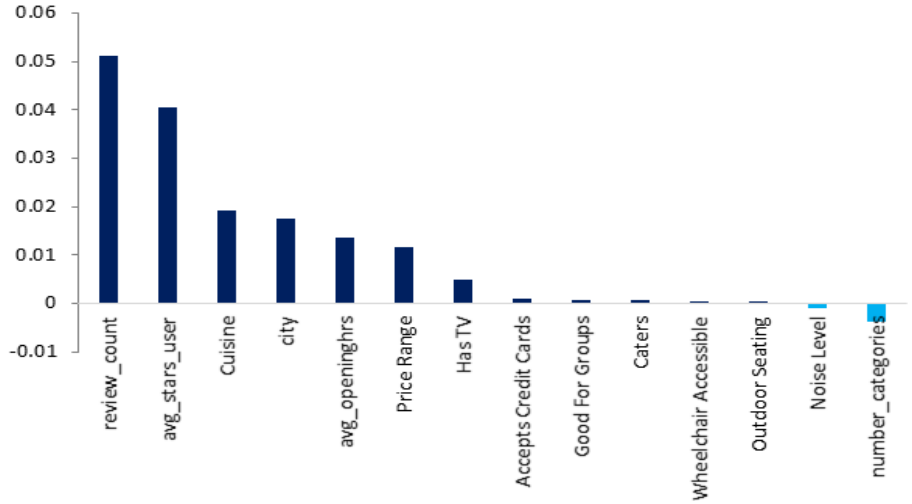
Evidence suggests that Model 1 performs significantly better, and it can be considered a reliable predictor of the star-ratings category. On the other hand, Model 2 that relies on PCA falls behind, potentially because in this analysis we are disregarding a considerable number of categorical features that might explain part of the variance of the target variable.

Finally, we use the best performing model for the feature importance analysis. Our feature importance analysis is based on the permutation feature importance. The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly

shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature.[7]

We observe that the 5 most important features that contribute to better predictive accuracy of the star-ratings are: review counts, average ratings per user, type of Cuisine, City, and the average opening hours.

Figure 3: Permutation feature importance (Model 1)



[7] https://scikit-learn.org/stable/modules/permutation_importance.html

Appendix 1: Data pre-processing

The data comes from Yelp Dataset Challenge. In this project, we focus on restaurants in Arizona. To obtain the relevant sample for our analysis we use the 'Category' column to extract all the businesses located in Arizona that have the 'restaurant' tag in their categorical description. In total there are **6817** restaurants in the state of Arizona. To increase the robustness of our analysis we select only the restaurants that are currently active, as denoted by the 'Open' column in the dataset. This filtering criteria allows us to reduce the sample size to **4925** restaurants.

Another important part of the data pre-processing stage is feature engineering. Using domain knowledge and the information contained in the original features, we construct the following additional variables:

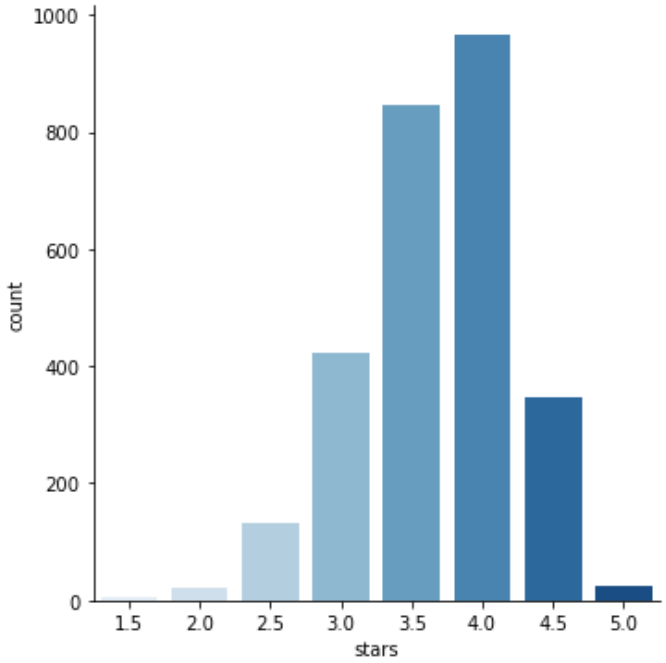
- **Average opening hours**: defined as the difference between closing and opening time, averaged over the days of the week.
- **Number of categories:** defined as the number of services that the business offers. It is obtained by counting the types of services listed in the 'Category' column for each business. Other than the" restaurant" tag, this column also provides additional information on the diversity of cuisines/services offered such as: Chinese, Korean, fast food, breakfast, barbeque etc. To better reflect this information in our analysis, we create a numerical feature 'numer_categories' to quantitatively depict the number of different services, and a categorical feature 'Cuisine' which corresponds to a dummy variable for each category.
- **Number of tips:** defined as the number of tips received by each individual business. This information is extracted from the 'tips dataset' by counting the total number of tips received by restaurant businesses in AZ.
- **User average ratings:** defined as the average rating that a user gives to all the businesses that he/she reviews.[8] For instance, if a user gives a rating of 3.5 to a restaurant in our sample, and on the other hand the user has given an average of 3 to all other businesses that he/she has rated, then in relative terms the 3.5 rating given to the restaurant in our sample is considered rather high.
- **Categorical variables obtained from the 'attributes' column:** The attributes column in the business dataset contains useful holistic information on additional features that could potentially contribute to higher star ratings. For instance, this column denotes whether the restaurant offers: Take-out, delivery services, Free Wi-Fi, outdoor settings etc. To reflect this information in our analysis, we create categorical variables for each attribute. We keep only the variables that have less than 20% rate of Null values. Out of the filtered columns, we further exclude all the observations that have null values, thus reducing the number of restaurants in our sample to **2768**.

Appendix 2: Exploratory data analysis

Distribution of star ratings:

[8] For each business ID we have used information from the 'reviews' dataset to extract all the users that have reviewed that particular business. We then use the information in the 'users' dataset to estimate the average star rating that a particular use gives to other business that he/she reviews. We then aggregate this information to obtain an average score for each business.

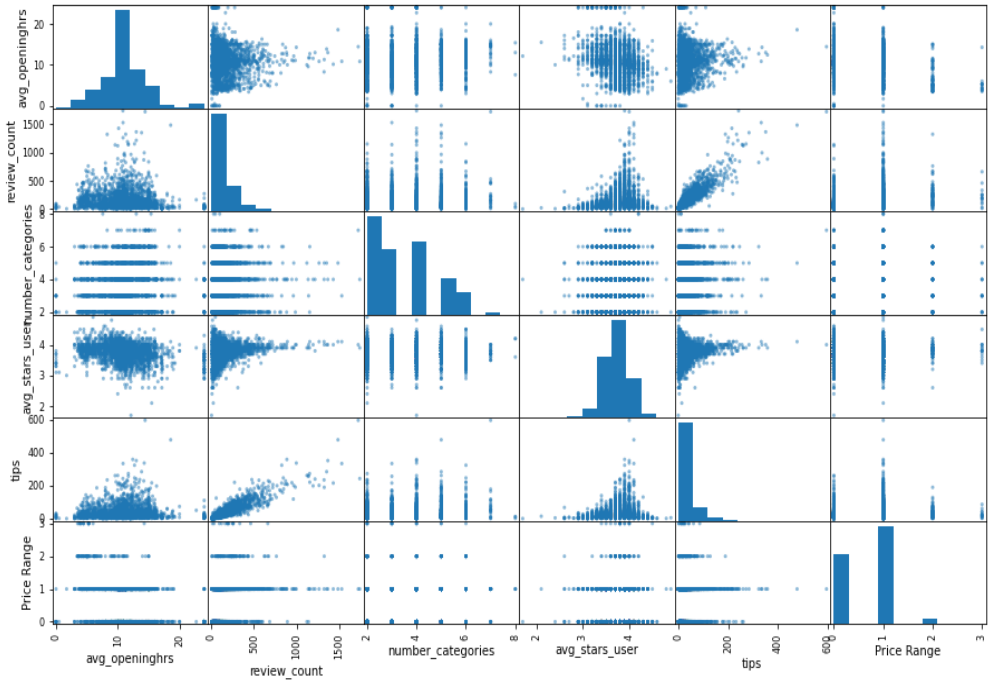
Figure 4: Distribution of star ratings for restaurants in Arizona



We visualise the pairwise relationship between numerical features in our dataset. This visualisation helps us understand the interdependencies between variables as well as the distributional properties of our data. We make two important observations:

- Based on an initial visual inspection of the histograms, our variables do not seem to be normally distributed. We also check this observation quantitatively by employing the **Shapiro-Wilk test** for normality. This method tests the null hypothesis that the sample comes from a normally distributed population. We reject the null hypothesis at the 5% confidence level for all the variables, which substantiates the conclusion that the numerical features in our dataset are not normally distributed.
- Also, the scatter plots do not support linearity between the pairs of numerical variables in our dataset.

Figure 5: Pairwise relationship between the numerical features in our sample



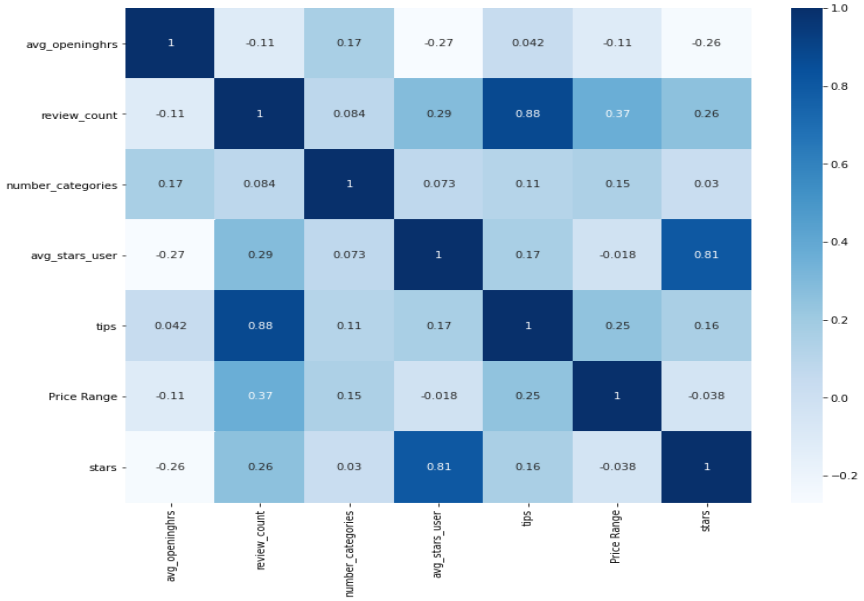
Appendix 3: Multicollinearity and dimensionality reduction

Spearman correlation

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables. It is appropriate for both continuous and ordinal variables and it is measured as follows: $\rho_S = \frac{Cov(Rank_x, Rank_y)}{\sigma_{Rank_x}\sigma_{Rank_y}}$

This results in the following correlation matrix:

Figure 6: Correlation matrix of numerical features using Spearman correlation coefficient



Cramer's V test

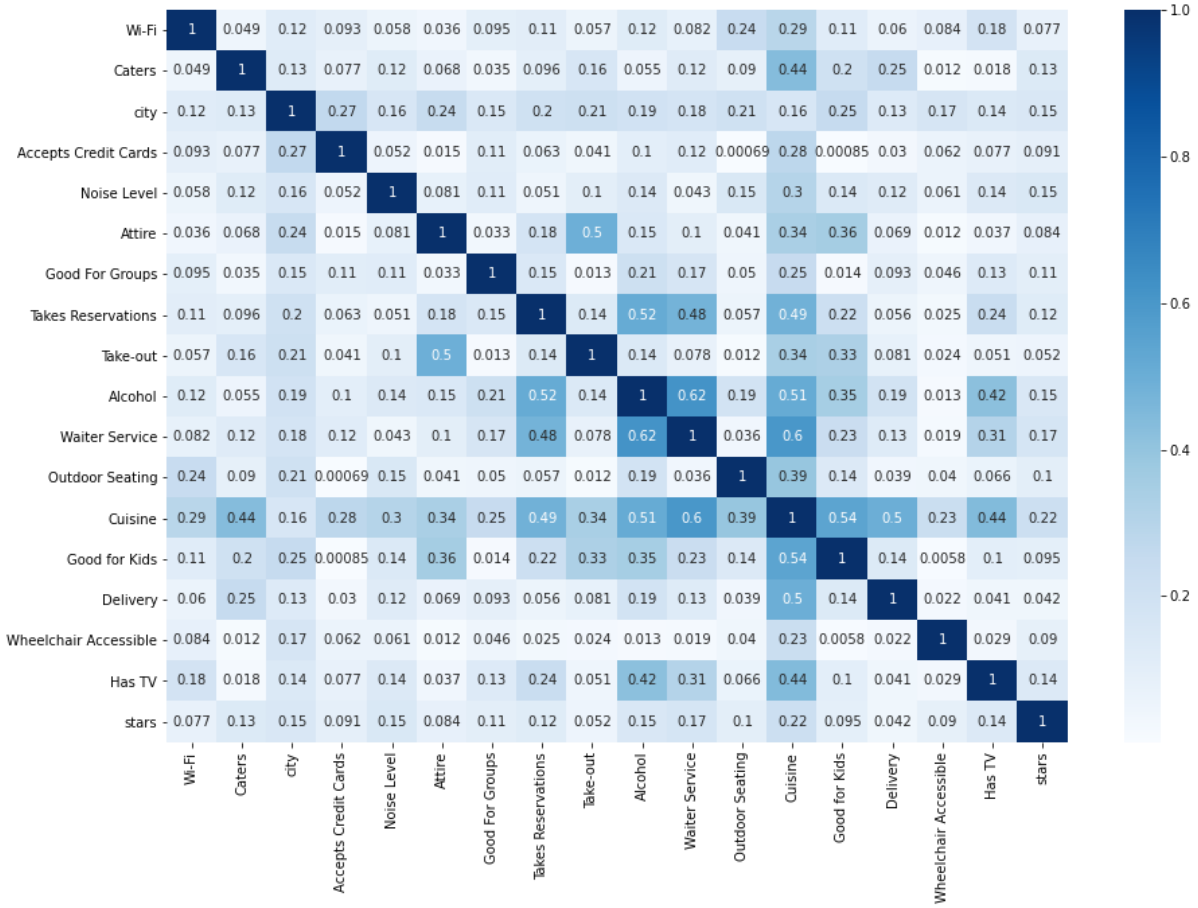
Cramer's V is calculated as follows:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

Where χ^2 refers to Pearson chi-square statistics. A p-value close to zero means that our variables are very unlikely to be completely un-associated in the population. k is the number of rows/columns in the contingency table.

This results in the following correlation matrix:

Figure 7: Correlation matrix of categorical features using Cramer's V correlation coefficient



We have identified the following pairs of highly correlated categorical features based on a 0.5 threshold:

- (Take reservations; Alcohol) with a correlation of 0.52. 'Alcohol' is the feature with the highest correlation with the target variable.
- (Alcohol; Waiter Service) with a correlation of 0.6. 'Waiter Service' is the feature with the highest correlation with the target variable.
- (Cuisine; Waiter Service) with a correlation of 0.6. 'Cuisine' is the feature with the highest correlation with the target variable.
- (Cuisine; Alcohol) with a correlation of 0.51. 'Cuisine' is the feature with the highest correlation with the target variable.

- (Cuisine; Good for Kids) with a correlation of 0.54. ‘Cuisine’ is the feature with the highest correlation with the target variable.
- (Cuisine; Delivery) with a correlation of 0.5. ‘Cuisine’ is the feature with the highest correlation with the target variable.

Testing Statistical significance of correlations

The random permutation test consists of randomly reshuffling the sequence of variables and recomputing correlations many times. By doing this, we want to identify the residual correlation that is associated with the distributional properties of the variables as opposed to random fluctuations. Repeating this process many times (i.e., for each pair of variables we are repeating this process 500 times), helps us establish a confidence interval that we can use as a benchmark for our correlation coefficient. We observe that all the correlation coefficients in our analysis are statistically significant, except for pair number 4 which corresponds to (review counts; average opening hours). However, this does not affect our analysis as this pair is not highly correlated. As far as categorical features are concerned, we observe a higher number of correlation coefficients that are not statistically significant. However, they are not associated with the highly correlated pairs identified above therefore they do not impact our feature selection process.

Figure 8: Testing statistical significance of Spearman correlation for numerical variables using random permutations method

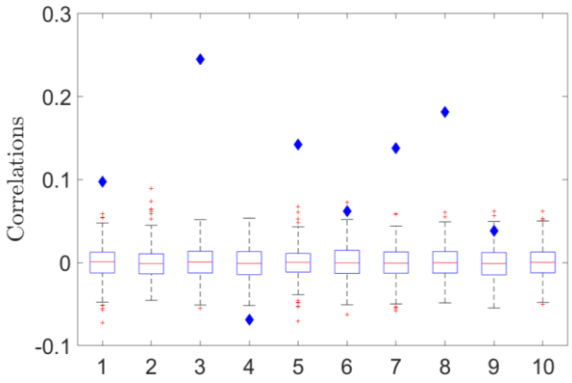


Figure 9: Testing statistical significance of Cramer’s V correlation for categorical variables using random permutations method

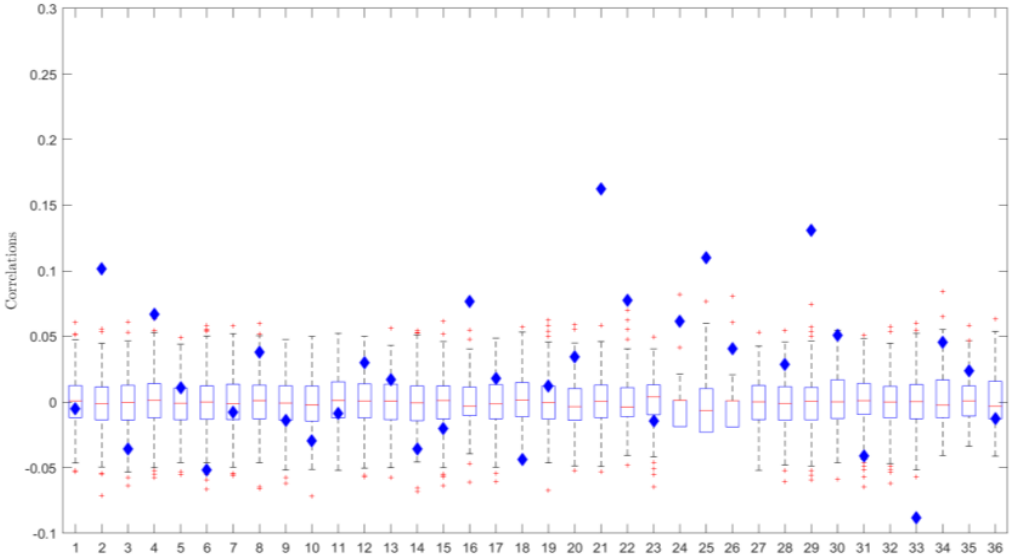


Table 3: Selected features for the multinomial LR based on Approach 1

Target variable	Independent variables (numerical)	Independent variables (categorical)
Star-ratings for restaurants in Arizona	Review count; Number of categories; Average star-ratings per user; Average opening hours	City; Cuisine; Wi-Fi; Caters; Accepts Credit Cards; Price Range Noise Level; Good for Groups; Take out; Outdoor Seating; Wheelchair accessibility; Has TV; Price Range

PCA

The goal of the PCA is to reduce the number of **m** variables to a smaller number of **p** uncorrelated variables which are linear functions of the variables in the original dataset. Principal components are mathematically defined as:

$$e_{it} = \frac{1}{\sqrt{\lambda_i}} \sum_{k=1}^{N} v_{ik}\, x_{kt}$$

Where $v_i = \left(v_{i1}, v_{i_2} \ldots v_{iN}\right)$ are the normalised eigenvectors, $0 \leq \lambda_N \leq \cdots \leq \lambda_1$ are the eigenvalues of the correlation matrix and x_k are the original variables standardised to mean zero and unit variance.

Whenever a few eigenvalues dominate, we can approximate the original variables in terms of a few PCs and use them to gain intuition:

$$x_{it} = \sum_{k=1}^{N} \sqrt{\lambda_k}\, v_{ki} e_{kt}$$

The presence of large eigenvalues means that the corresponding PCs account for a large fraction of the overall variance in the data.

Table 4: Set of variables that will be subject to PCA

	Numerical features	Top categorical variables (hot-encoded)
PCA	• Review count; • Tips; • Number of categories; • Average star-ratings per user; • Average opening hours ;	• Cuisine; • City; • Noise Level; • Alcohol; • Waiter Service;