Assignment 3 – Alan Kessler, PREDICT 410 Section 57

Introduction

This analysis looks at customer ratings of a restaurant. The restaurant is rated on a variety of characteristics and these ratings are positively correlated with each other. The analysis considers ways to account for this correlation such as Factor Analysis and Principal Component Analysis (PCA) when advising the restaurant owner on how to maximize the restaurant's overall score.

Proposal

The following analysis shows that characteristics of the food including taste, temperature, and freshness are the most important indicators of the overall score. This indicates that investing in improvements related to food quality will generate the best return on investment. Additionally, service-related characteristics like waiting time and friendliness are more often rated lower than other categories indicating room for improvement.

Analysis

Based on the correlation matrix shown in Figure 1, several variables in the data are highly correlated with each other. This could result in multicollinearity concerns when building a regression model to predict overall score and skew interpretation of variable importance. These variables are correlated within groups of three that will be referred to as food, service, and locale. This characteristic of the data lends itself to using PCA and Factor Analysis as those techniques can reduce the overall dimensionality of the data.

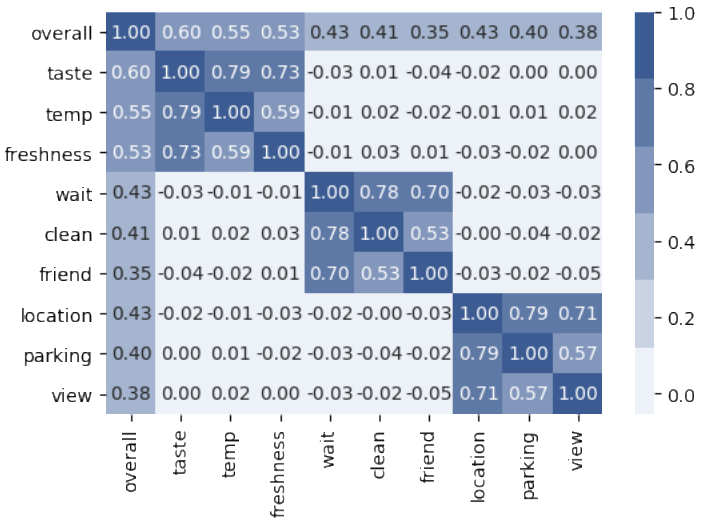

Figure 1 – Correlation Matrix

PCA indicates that the first three principal components explain 78% of the variance as shown in Figure 2. This is also demonstrated in the eigenvalues as the first three have much larger eigenvalues than the remaining components. An analysis of the correlation of the first three principal components show that the procedure is successful in creating orthogonal vectors.

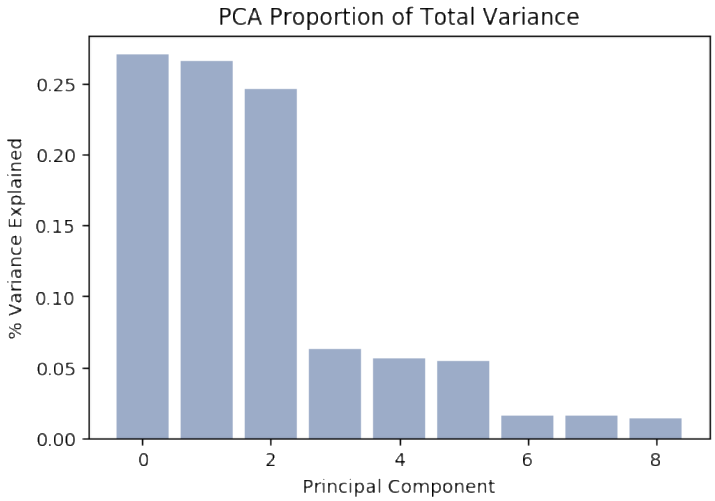

Figure 2 – Proportion of Total Variance Explained by Each PC

The loadings of the first three principal components show that the groupings do not correspond to the correlation matrix shown in Figure 1. The second and third components have significant loads to both the service and locale groups. Figure 3 shows this relationship which makes the interpretation of the principal components more difficult.

When factor analysis is applied, the groups are more defined. Now each factor corresponds to only one of food, service, or locale. Each of the un-rotated factors is perfectly correlated to one of the variables. Rotating the factors can result in a more meaningful set of factors that correlate more closely with the underlying variables.

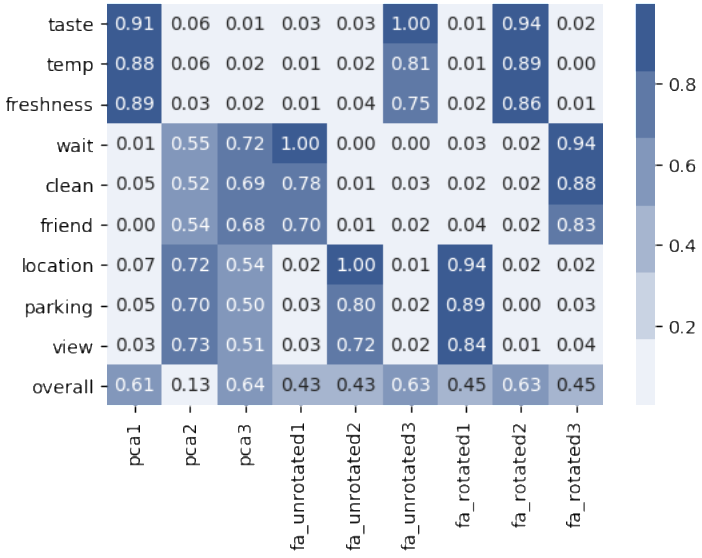

Figure 3 – Absolute Value of Correlation between Variables and Factors/Components

Figure 4 shows the fit statistics of the six models used to predict overall score. The additional model uses rotated factors from the “fa-kit” Python module. The full model has the best fit to the data and all of the terms in the model are statistically significant. The second model, containing one variable of each type, is much simpler to calculate and there is much greater confidence in

the coefficient estimates while the R-squared is only marginally impacted. The third model containing all principal components results in the same fit statistics as the first, which is intuitive because it includes all of the same information as the first model. The fourth model containing only the first three principal components improves on the fit of the second model but has a lower R-squared than the full model. However, the coefficient estimates for the fourth model have a smaller confidence interval, providing greater confidence in the model specification. The fifth model using factor analysis has fit statistics more closely related to Model 2. The varimax rotated factors used in Model 6 result in the best fit of the reduced-dimensionality models.

Model	R-Squared	AIC
Model 1: Full Regression	0.822	379.9
Model 2: Subset of taste, wait, and location	0.760	665.9
Model 3: Full PCA Model	0.822	379.9
Model 4: First 3 Principal Components	0.798	492.0
Model 5: 3 Un-rotated Factors	0.769	627.4
Model 6: 3 Varimax Rotated Factors	0.817	397.0

Figure 4 – Model Fit Statistics

In analyzing the variance inflation factors (VIF), the full model has VIFs all under five (excluding the intercept term), indicating that multicollinearity may not be a large concern for generating predictions of overall score. However, the correlation does impact the ability to attribute variable importance to any single variable.

Model 2 would be my preferred model for implementation because it has VIFs very close to one resulting in stable coefficients, has a high R-squared, and is the simplest to interpret and explain to a client without a statistical background. For all of the other models, the VIFs are equal or close to one.

The strong grouping provided by the rotated factor analysis can be used to comment on the relative importance of the various categories. Based on the standardized coefficients, the most important category for predicting overall score is food. Relative to that category, the other two categories are each 75% as important. The sum of the coefficients from the full model by group almost identical to that of Model 2 indicating that most variance picked up in the simplified model. Because the data is all on the same scale, these coefficients provide variable importance information similar to that from the factor analysis.

In addition to looking at which variables are important it is also good to look at areas that have room for improvement. Figure 5 shows that service, while less important than the food category, has somewhat lower scores and could be an area for improvement. However, no variable is significantly different. If one category is rated very low, perhaps a smaller investment would be needed for a correspondingly larger improvement. For example, even if food is the most important, improving a score from 9 to 10 may have a smaller return on investment than improving a service score from 1 to 5.

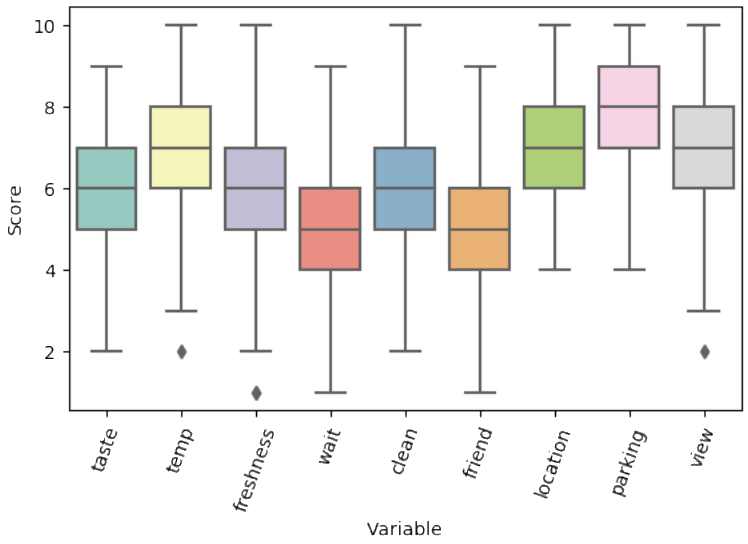

Figure 5 – Box-plot of Survey Variables

Conclusion

Different techniques are used to reduce dimensionality in restaurant rating data. The methods differ in their interpretability and ease of implementation. Ultimately, these techniques allow for a greater ability to gain insights about the business problem and create more robust predictive models. In this case, it is clear food characteristics are related and are the most important indicators of the restaurant's overall score.