MKTG 5220: BIG DATA AND STRATEGIC MARKETING

ASSIGNMENT 1

April 14, 2017

SUBMITTED BY:
Kunja Dutta
MAYANK GUPA
DEBALEENA SAHA
SAMARPITA CHAKRABARTY

University of connecticut, School of business
SPRING' 17

Contents

I. Statement of the Problem

In this report, we have based our analysis on two primary goals:

Specific Task 1 : To predict the star ratings of restaurants accompanying the reviews provided by customers on Yelp, based on an analysis of the review texts.

Specific Task 2 : To find the most used words by reviewers in their language, in order to recommend an approach to provide these terms as word options to be selected in the review column if applicable, to encourage more reviews.

II. Abstract

In this day of e-Commerce, a large extent of a customer’s shopping choice is largely based on reviews by other customers. Such is also the case for Yelp, where the customer reviews have a high potential of influencing the choice of restaurant for other customers. It is often observed that the star ratings are misaligned with the content of the reviews. In such a case, it is important to understand how the star ratings are related to the semantic content of the reviews. It would also be useful to identify which are the most common terms occurring in the review content. These words can be given in the form of checklists or select options for the customers as a substitute for detailed reviews or as an addition to their reviews. This would be more effective for the restaurant managers than the review comments which are often ignored by managers because of their sheer volume and unstructured content. Hence we aim to analyse the review contents, extract the important and relevant concepts and base the rest of our analysis and modeling on these concepts, which is covered in a later part in this report.

III. Background

Almost seven out of ten people read online reviews before they buy something, according to a recent report. Online reviews are one of the most important parts of any business today. From gaining local organic search rankings to becoming word-of-mouth, online reviews create branding. They play a very important role in spreading information and influencing user decision.

There's only a small link between the average user rating of an item and the actual quality of an item. In other words, just because a product has an average rating of five stars doesn't mean it's a great product. Nor does an average rating of one star mean a product is sub-par.

A user may only read a limited number of reviews before coming to a decision. An important aspect to the success of a rating and reviews site such as yelp is to identify which reviews to promote as being useful.

Yelp is the #1 review site on the web and immensely powerful, making it a mighty friend or a bad enemy. Yelp lets consumers (Yelpers) be the jury. Yelp trusts consumers to leave authentic reviews about their experiences. It also focuses on long-form reviews rather than short one-liners, providing the kind of in-depth, thorough review wary customers crave. Sometimes "good coffee" just isn't good enough.

IV. METHODOLOGY

There were two datasets provided for this specific task: 1) Average Star Rating - Contained average rating of the restaurants along with user text reviews, 2) Individual Star Rating - Contained individual star rating of each of the user reviews along with their text reviews. Team had explored both datasets and brainstormed to find meaningful inference and how both datasets can be used in the prediction. As Individual Star Rating had all the information that Average Star Rating dataset had including the individual star ratings corresponding to each user, team had concluded that it would be better to proceed modeling with Individual Star Rating dataset to predict the stars of the review rather than Average Star Rating.

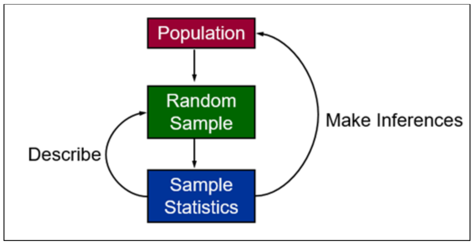
IBM SPSS Modeler and JMP Pro 13 has been used to predict stars for review comments from the data set yelp_reviews.csv. The Team followed a Sample-Explore-Modify-Model-Assess (SEMMA) approach which included 5 steps

1. Data Sampling
2. Data Exploration
3. Data Modification
4. Data Modeling
5. Data Assessment

1. Data Sampling

The yelp_review.csv i has 2,685,066 rows which is our total population. Out of this, a random sample of 50,000 was extracted and stratified with stars. There were two reasons to sample

- The initial data set was huge.
- Sample statistics of the random sample make it easier to explore the data and analyze it to make interferences of the population. This sample statistics will help to describe the random sample better. This is why sampling the data was important.



The data was imported to JMP Pro 13. From the option **subset**, a sample of 50,000 records was selected and stratified.

2. Data Exploration

In JMP Pro 13, we have explored the data using various techniques.

There was no missing value in the data set.

Number of extreme values were very minimal.

Stars are recognized as the **target variable** and **text** is taken as the **predictor variable**.

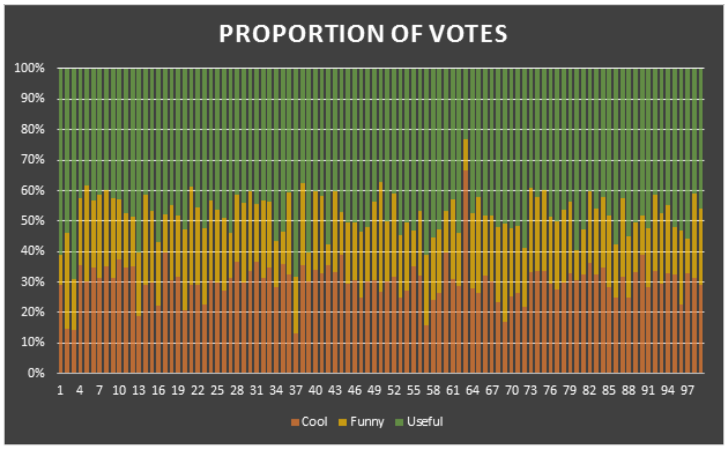
The Team tried to explore the data with the help of distributions, correlations, multivariate analysis, and graphs.

There was a weak correlation between the variables. There was not much influence of the votes on the stars.

Next, the impact of the IDs and type was analyzed.

Graphs with variables against dates did not show any pattern.

Stacked column graph of the Cool, Funny and Useful vote for each business_id showed some interesting comparison.



Votes vs Business_ids

These steps assisted in getting a sense out of data and perform data treatment moving forward.

3. Data Modification

Although the data quality was good, some minor modifications had to be carried out in order to finetune the data set and analyze it better.

On trying to recode few variables, it did not make any significant impact. So we went ahead without recoding.

No new variable was required to be derived.

The Team eliminated few variables such as business_id, user_id, date, type, review_id. This is because the IDs, date, and type were not making much difference on the output.

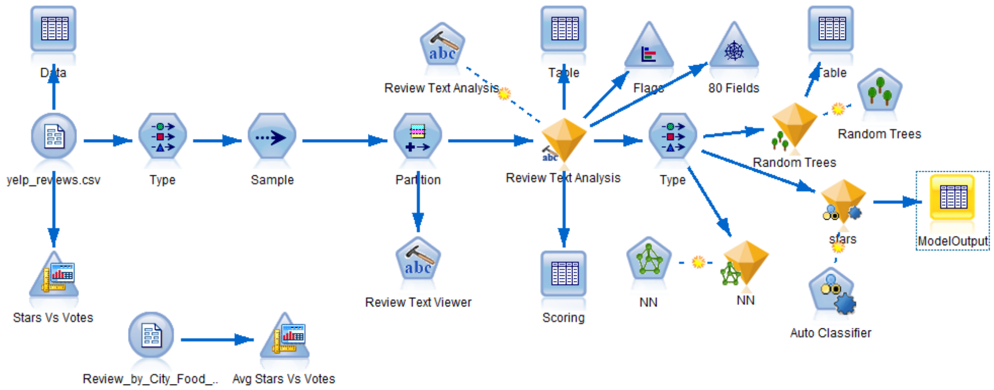
Data set is now ready to be modeled to find the best outcome.

Further steps were performed in IBM SPSS Modeler.

4. Data Modeling

The yelp_review.csv file was imported through var file to IBM SPSS Modeler.

The below figure shows the entire model built in SPSS modeler:



Model built in IBM SPSS Modeler

Following steps were performed to build the model after importing dataset:

- Var. File node yelp_reviews.csv was created which contained the data from yelp_review.csv.
    - Data node Data was added to the Var. File node to visualize the data.

- Type node was added to the Var. File node. This defined the variable types. Stars is defined as the target variable. Predictor variable is text.

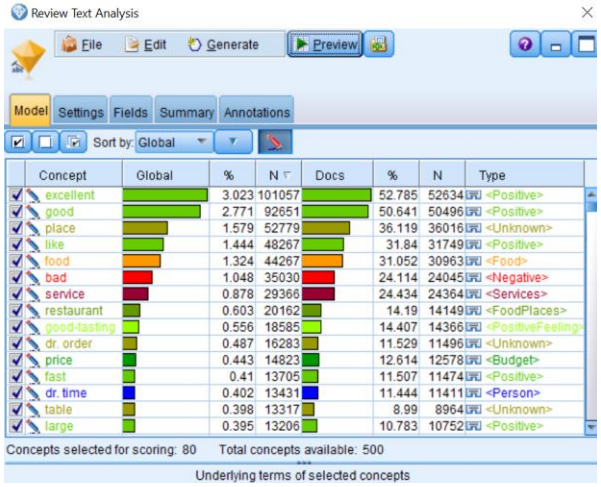
Rest of the variables were assigned types based on the understanding of the data using best judgment.

This node has inputs for text analysis.

- Sample node was added to the Type node to take out 50,000 of the population as a random sample.

- Partition node was added to the Sample node. A training set of 40%, a validation set of 30% and a test set of 30% were assigned.

- Review Text Analysis node was added to the Partition node whose output was generated as ,. The Text Analysis node analyzes the text field- review text to find concepts, categories from the text. It scores the concepts and assigns term weight to each of the concepts it finds. It also assigns the Type of the concepts and categorizes them into positive and negative categories. While generating the text model, concept model nugget setting was selected to select the concepts and generate the model directly by the SPSS Modeler itself. Concepts with global percentage below 0.174 were deselected by the Modeler itself. Global percentage, global count, doc percentage, doc count and type (positive or negative) of each concept is shown here. Sort by: global was selected. Scoring mode Concepts: Fields, Field values: Flags, True value: T, False value: F, Field name: extension Concept_, Add as: Prefix were selected and, Accommodate punctuation errors was enabled. Text field: Text, Text field represents: Actual text and Text language: English were selected.



Sort by: Global

	Concept	Global	%	N	Docs	%	N	Type
☑	excellent		3.023	101057		52.785	52634	<Positive>
☑	good		2.771	92651		50.641	50496	<Positive>
☑	place		1.579	52779		36.119	36016	<Unknown>
☑	like		1.444	48267		31.84	31749	<Positive>
☑	food		1.324	44267		31.052	30963	<Food>
☑	bad		1.048	35030		24.114	24045	<Negative>
☑	service		0.878	29366		24.434	24364	<Services>
☑	restaurant		0.603	20162		14.19	14149	<FoodPlaces>
☑	good-tasting		0.556	18585		14.407	14366	<PositiveFeeling>
☑	dr. order		0.487	16283		11.529	11496	<Unknown>
☑	price		0.443	14823		12.614	12578	<Budget>
☑	fast		0.41	13705		11.507	11474	<Positive>
☑	dr. time		0.402	13431		11.444	11411	<Person>
☑	table		0.398	13317		8.99	8964	<Unknown>
☑	large		0.395	13206		10.783	10752	<Positive>

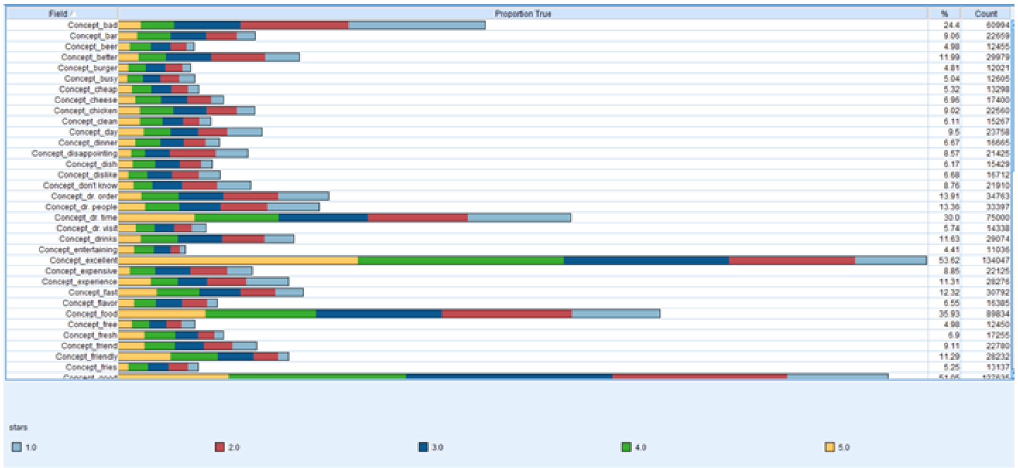
Concepts selected for scoring: 80 Total concepts available: 500

Underlying terms of selected concepts

Concepts displayed in Review Text Analysis

Review text analysis is the output node from text analysis which displays the concepts and its distribution.

- Table node was added to the Review Text Analysis node to visualize the data after analyzing the text field.
- A distribution node called Flags was added to the Review Text Analysis node to see the frequency of the true concepts. Based on this representation, concepts with global percentage less than 10 were eliminated. Distribution of concepts is given below. The stars are indicated by the colors in each bar. The color legends for stars are mentioned at the bottom of the below figure.



Field	Proportion True	%	Count
Concept_bad		24.4	60994
Concept_bar		9.06	22659
Concept_beer		4.98	12455
Concept_better		11.99	29979
Concept_burger		4.81	12021
Concept_busy		5.04	12605
Concept_cheap		5.32	13298
Concept_cheese		6.96	17400
Concept_chicken		9.02	22560
Concept_clean		6.11	15267
Concept_day		9.5	23758
Concept_dinner		6.67	16665
Concept_disappointing		8.57	21425
Concept_dish		6.17	15429
Concept_dislike		6.68	16712
Concept_don't know		8.76	21910
Concept_dr. order		13.91	34763
Concept_dr. people		13.36	33397
Concept_dr. time		30.0	75000
Concept_dr. visit		5.74	14338
Concept_drinks		11.63	29074
Concept_entertaining		4.41	11036
Concept_excellent		53.62	134047
Concept_expensive		8.85	22125
Concept_experience		11.31	28276
Concept_fast		12.32	30792
Concept_flavor		6.55	16385
Concept_food		35.93	89834
Concept_free		4.98	12450
Concept_fresh		6.9	17255
Concept_friend		9.11	22780
Concept_friendly		11.29	28232
Concept_fries		5.25	13137

Distribution of Concepts

- Another table called Scoring Scoring was added to to the Review Text Analysis node visualize the score of the concepts given by SPSS.

- Type node Type was added to the Review Text Analysis node. This node has inputs for the predictive model to predict the stars based on the concepts created by review text analysis.

- A web node called Fields 80 Fields was added to the Review Text Analysis node to visualize the pattern of 80 most linked concepts.

Next, the best modeling technique was selected through Auto Classifier by adding an Auto Classifier node Auto Classifier to Type Node. After executing the model the output node stars was derived. Based on the data type, the models that tried were C5, Logistic regression. Decision List, Bayesian Network, Neural Network, Discriminant, KNN Algorithm, LSVM, Random Trees, Tree-AS, CHAID, Quest, and C&R tree (classification and regression tree).

Next, to visualize the Model output an output node ModelOutput was added to the Auto Classifier Model output.

Results of each of the above-stated model were evaluated. The factors based on which the model selection was done are

- Accuracy of the model : higher value is better
- Misclassification rate of the model : lower value is better

Based on the above factors, random trees and neural network was found as the best model.Model accuracy rate obtained for both is around 40%.

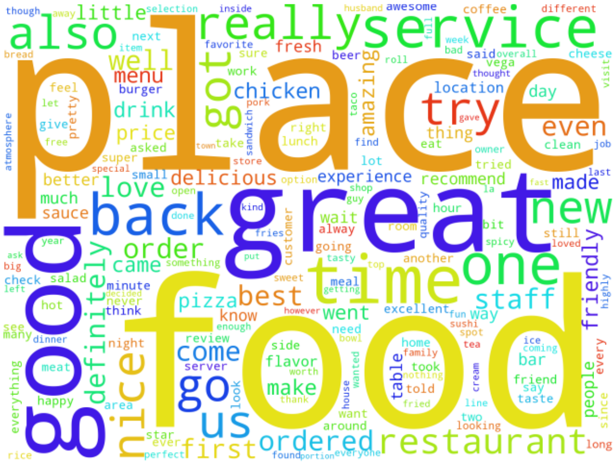
To decide on the best between these team had further done individual modelling of both of these and compared the results. To do that, a Random Trees model node Random Trees was added to the Type node whose output was Random Trees . Table node Table was added to the random tree to see the model output.

For, Neural Network results, N-N node (neural network node) NN was added to the Type node whose output was generated as NN .

Neural Network also has given similar results like random trees, but to keep the model simpler team has chosen Random Trees as the best model with 0.402 model accuracy and 0.594 misclassification rate.

The importance of the predictors and the top decision rule for ‘stars’ helped us achieve the results for **Specific Task 1**.

For **Specific Task 2**, team has prepared a word cloud to find the most used typical words, using Python programming, from yelp_review.csv data set.



Word cloud of most used words in reviews

The word cloud is the visual representation of the typically used words. The font size of the words shows the importance of the words used in the reviews. The bigger font size depicts the higher frequency of occurrence of the words in the reviews. This word cloud indicated the following words as most used in order: place, food, great, good, back, time, service, one, really, also, nice, got, try.

5. Data Assessment

The Confusion Matrix for the model gave a clear picture of the actual and predicted values of the stars which indicated mostly a low proportion correct. The word cloud helped us identify the most used words in the review. These outputs helped us achieve the predicted star ratings and the most used words accurately.

V. Result

The model built for Specific Task 1 indicated that the star rating given by the reviewers is different from the ratings predicted by the review text.

The word cloud built successfully for Specific Task 2 helped us identify some of the most used words in the language used by the reviewers which can be used to create easier review templates and encourage more customers to review.

VI. Conclusion and Recommendation

Desired outcomes for Specific Task 1 and Specific Task 2 have been achieved.

Ratings based on review comments is essential since the predictive model indicates a difference between the ratings provided and the ratings based on the reviews. Also, words such as **great/good/bad place**, great/good/ bad **food**, will/will not **come** back, more/less **time** to serve, great/good/bad **service**, must **try**, can be provided in the review column for a reviewer to select them if they think that holds good for the particular product.

Recommendations for utilizing the review comments better would be

- Ratings/stars based on review comments should be included to highlight the difference between the true ratings/stars based on words and feelings of the reviewers and the impulsive ratings/stars given at the moment by the reviewers. This rating/star plays an important role while taking any decision by the customer in today's market, so the true 'star' need to be displayed.
- Customer's use a certain language to communicate among themselves in review forums. It is important to identify that language so that marketing for a product can be done in the same language to connect easier with the customer.
- Ease of review with an option to just select words that best describes a product will encourage more reviewers to provide a feedback. The option of words that would be provided to the reviewers can be decided from the most used words by the reviewers.
- Votes option under reviews should be made more attractive to catch the attention of people who are lazy to review. If they agree to a certain comment, they can just vote for it and help other buyers.

VII. Executive Summary

Yelp reviews and ratings are an important source of information for consumers to make informed decisions about a venue. It is also an important source of feedback from customers for restaurant managers to understand the level of customer satisfaction. Yelp dataset has much more information than some similar rating-based dataset, such as Netflix dataset. The goal of our project is to analyse the review comments from the Yelp dataset and predict the ratings based on review texts to understand the true ratings of the restaurants based on the feelings of the reviewer. Another important task that we achieved was to find the most used words by reviewers in their language from the reviews and provide them as word options to be selected in the review column, to encourage more effective reviews. This information might come useful while designing social media campaigns and online advertising that proactively seek to leverage favorable consumer reviews as part of a digital marketing strategy.

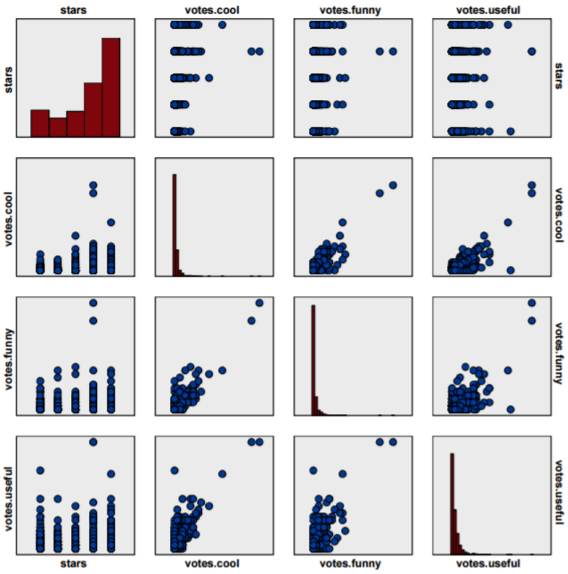
VIII. References

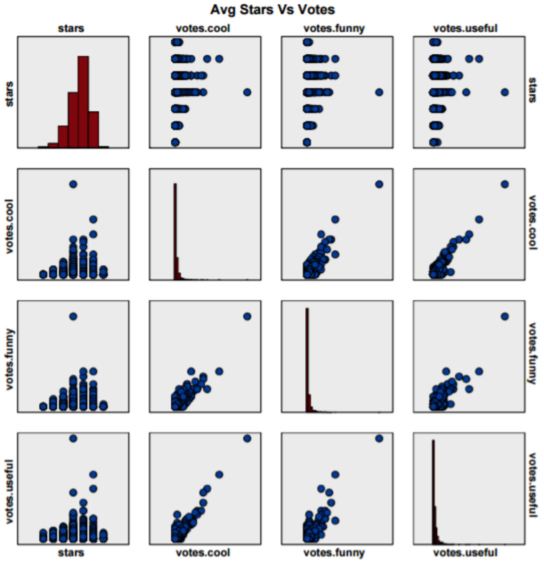
1. “When Shopping Online, Can You Trust the Reviews?” - Elizabeth Holmes
2. “When 4.3 Stars Is Average:The Internet’s Grade-Inflation Problem”-Geoffrey A. Fowler

IX. Appendix

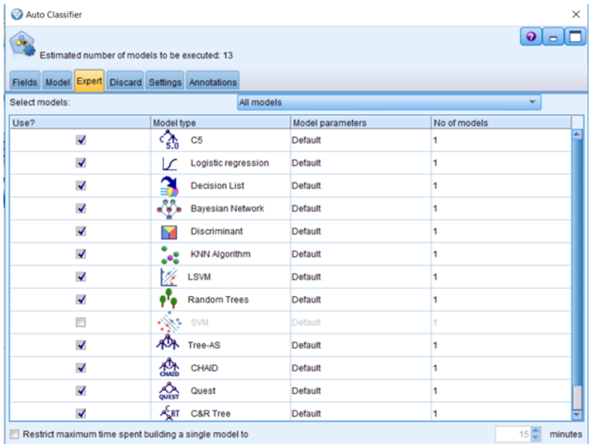
1.Weak correlation between votes and stars

2. Correlation between average stars and votes





3. Models tried with Auto Classifiers to find the best outcome in IBM SPSS Modeler



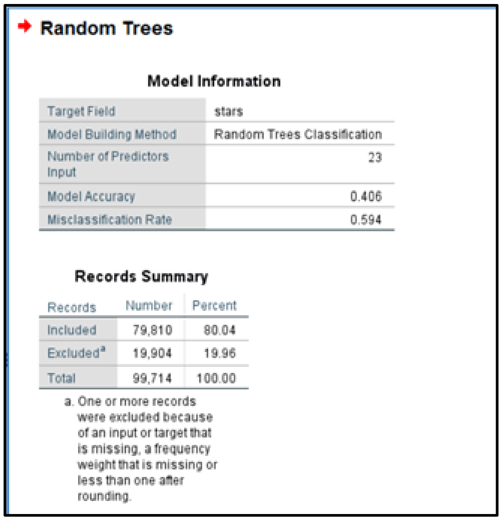
Estimated number of models to be executed: 13

Select models: All models

Use?	Model type	Model parameters	No of models
☑	C5	Default	1
☑	Logistic regression	Default	1
☑	Decision List	Default	1
☑	Bayesian Network	Default	1
☑	Discriminant	Default	1
☑	KNN Algorithm	Default	1
☑	LSVM	Default	1
☑	Random Trees	Default	1
□	SVM	Default	1
☑	Tree-AS	Default	1
☑	CHAID	Default	1
☑	Quest	Default	1
☑	C&R Tree	Default	1

□ Restrict maximum time spent building a single model to 15 minutes

4. Random Trees Model Information



Random Trees

Model Information

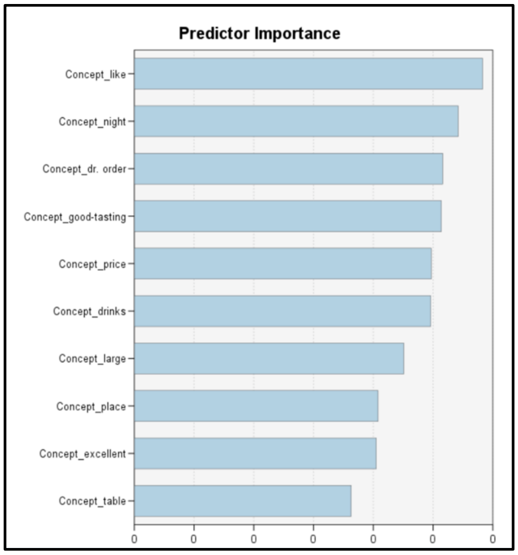
Target Field	stars
Model Building Method	Random Trees Classification
Number of Predictors Input	23
Model Accuracy	0.406
Misclassification Rate	0.594

Records Summary

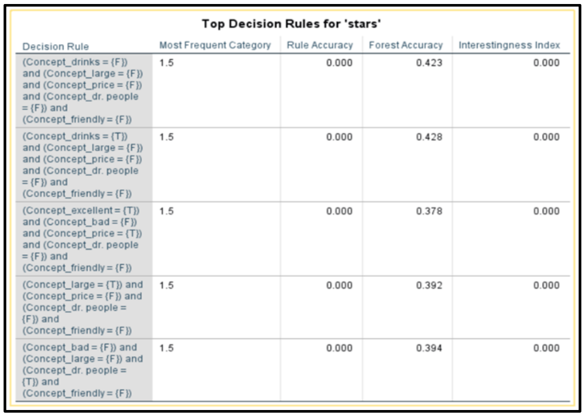
Records	Number	Percent
Included	79,810	80.04
Excluded[a]	19,904	19.96
Total	99,714	100.00

a. One or more records were excluded because of an input or target that is missing, a frequency weight that is missing or less than one after rounding.

5. Random Trees Predictor Importance



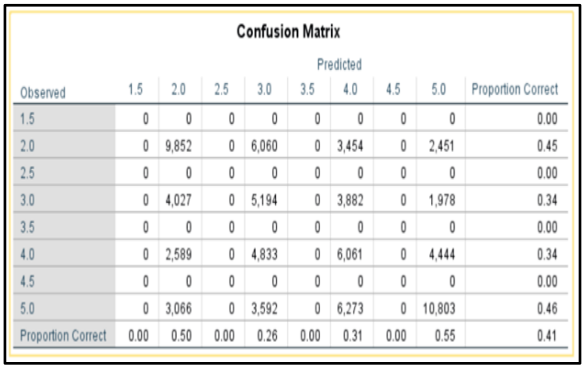
6. Random Trees Top Decision Rules for ‘stars’



Top Decision Rules for 'stars'

Decision Rule	Most Frequent Category	Rule Accuracy	Forest Accuracy	Interestingness Index
(Concept_drinks = {F}) and (Concept_large = {F}) and (Concept_price = {F}) and (Concept_dr. people = {F}) and (Concept_friendly = {F})	1.5	0.000	0.423	0.000
(Concept_drinks = {T}) and (Concept_large = {F}) and (Concept_price = {F}) and (Concept_dr. people = {F}) and (Concept_friendly = {F})	1.5	0.000	0.428	0.000
(Concept_excellent = {T}) and (Concept_bad = {F}) and (Concept_price = {T}) and (Concept_dr. people = {F}) and (Concept_friendly = {F})	1.5	0.000	0.378	0.000
(Concept_large = {T}) and (Concept_price = {F}) and (Concept_dr. people = {F}) and (Concept_friendly = {F})	1.5	0.000	0.392	0.000
(Concept_bad = {F}) and (Concept_large = {F}) and (Concept_dr. people = {T}) and (Concept_friendly = {F})	1.5	0.000	0.394	0.000

7. Random Trees Confusion Matrix



Confusion Matrix

	Predicted								
Observed	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	Proportion Correct
1.5	0	0	0	0	0	0	0	0	0.00
2.0	0	9,852	0	6,060	0	3,454	0	2,451	0.45
2.5	0	0	0	0	0	0	0	0	0.00
3.0	0	4,027	0	5,194	0	3,882	0	1,978	0.34
3.5	0	0	0	0	0	0	0	0	0.00
4.0	0	2,589	0	4,833	0	6,061	0	4,444	0.34
4.5	0	0	0	0	0	0	0	0	0.00
5.0	0	3,066	0	3,592	0	6,273	0	10,803	0.46
Proportion Correct	0.00	0.50	0.00	0.26	0.00	0.31	0.00	0.55	0.41