Text Analytics and Association Rules Models

Homework Assignment #1

March 23, 2018

**Team #2**

Rakesh Anne

Swaraj Machiraju

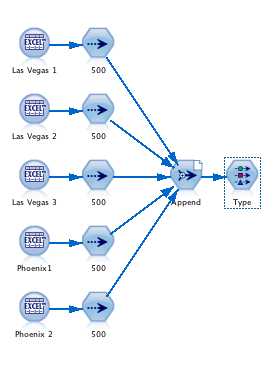
Srilekha Lava

Prasanthi Lingamallu

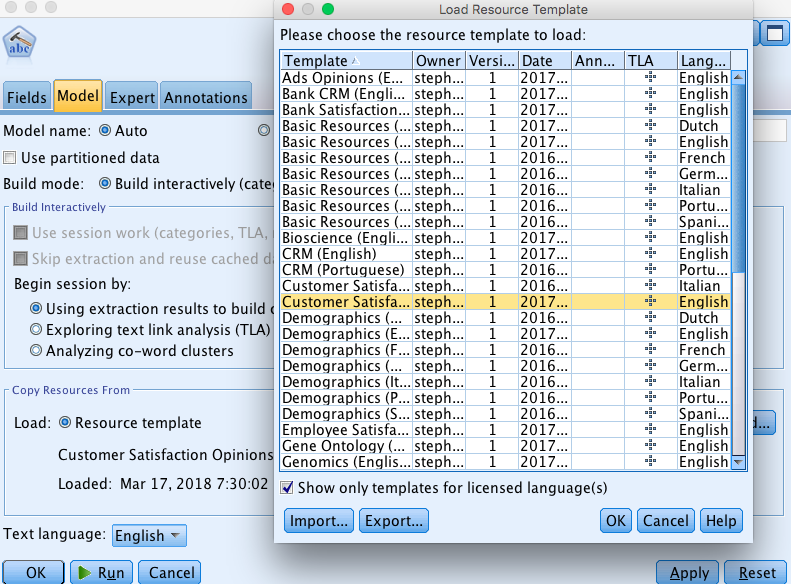
Remigiusz Szupryczynski

Stephanie Winiarski

To start in building a functioning Text Analytics model, one must first complete any appropriate pre-modeling operations. In the case of the three tasks related to building and validating Text Analytics and Association Rules models for Yelp Restaurant Reviews, this was opening the five data files to review the provided information and determine that no cleaning or manipulation was needed prior to loading into IBM SPSS Modeler. However, it was noted that each file contained a high volume of data, which was concerning from a time and computer limitations (memory and capacity) perspective. As a result, it was decided that the test model would first be built using a sampling of the data before attempting to run the model with the full data set.

 When creating the sample population to test, five Excel source nodes were added to the canvas to load the five distinct data files. The Source nodes were each connected to an individual Sample node, each of which was edited to include a sample of the first 500 records. The five samples were combined using an Append Node, which was then connected to a Type Node. In opening up the Type Node, the “Read Values” button was used to allow for the system to select the appropriate choice for each field. The choices selected by the “Read Values” button appeared reasonable, and as such, no changes were made to these default values.

Given the nature of the task, the Text Mining Node was selected and added to the canvas. Once on the canvas and connected to the Type Node, the Text Mining node then needed to be edited. The primary edit needed was to upload a different library underneath the Model Tab. Various libraries were tested, including Product Satisfaction and Hotel Satisfaction (English), but ultimately it was determined that Custom Satisfaction Opinions (English) provided the best results for the task at hand.

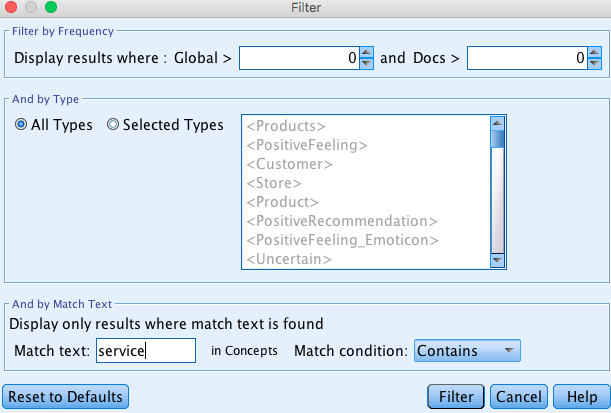


All other pre-selected fields within the node were reviewed, and ultimately determined to be appropriate for the task at hand. As such, options like Build interactively and Using extraction results to build categories were used.

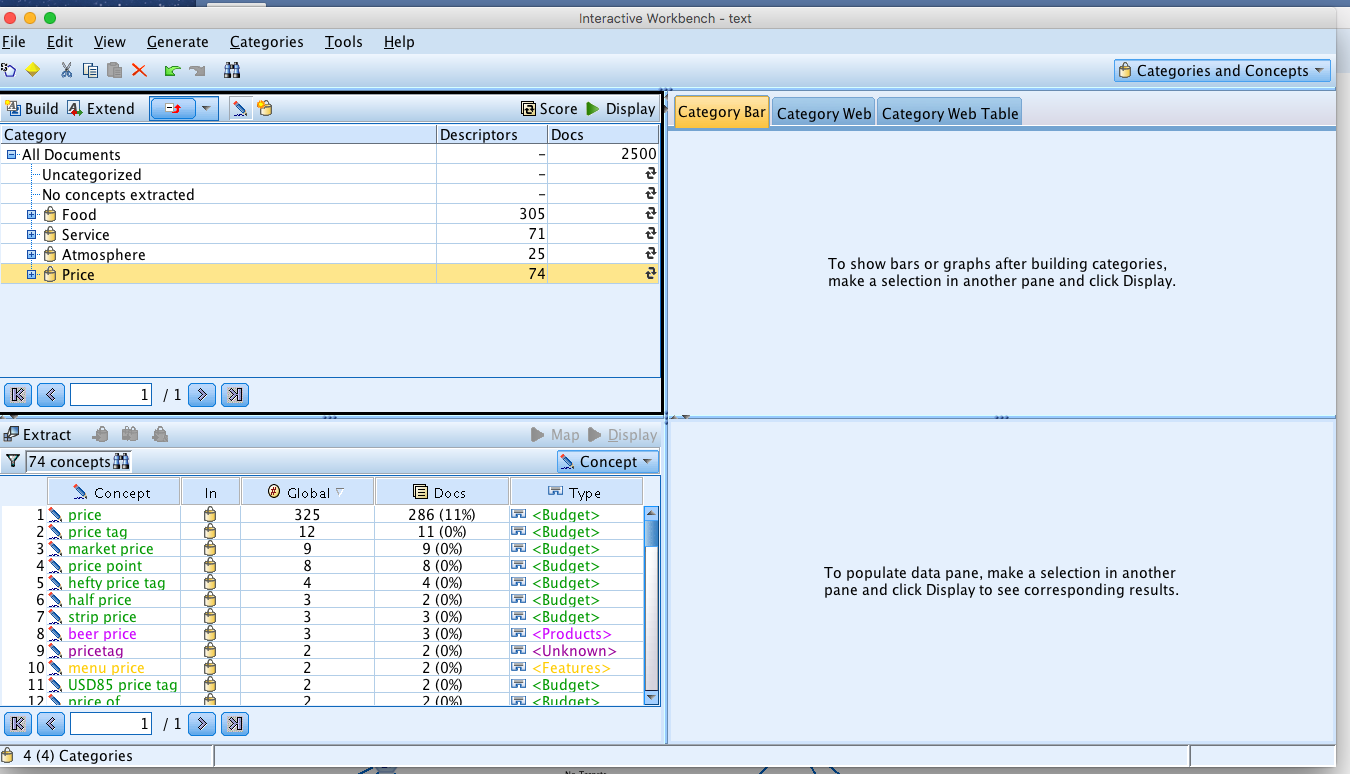
With all the pre-work completed, the stream was run through the Text Analysis node, producing a golden nugget.

With the execution run completed, the output displayed a list of extracted concepts, but no categories. In order to complete the task, it become obvious that it was necessary to manually build the four categories specified - food, service, atmosphere, and price. To build a category, “Build” was clicked in the top left corner of the Interactive Workbench window. This opened a secondary window, which allowed the “Match text” field to be filled with each of the four categories.

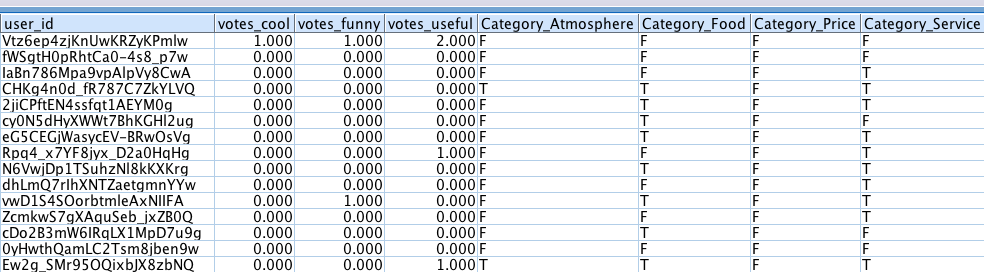
Of course, this resulted in having four empty categories that now needed filling. To fill each category, the concepts displaying in the Extract window below were filtered by selecting the Filter icon and using the Match text field again to display only results where the match text was found in the extracted concepts (in this case for food, price, atmosphere, and service).

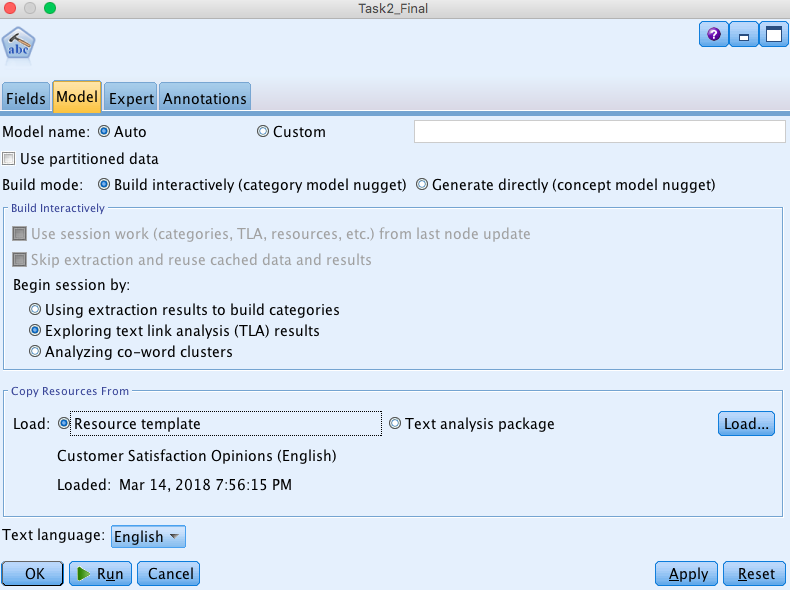


Filtered concepts were then added to the four created categories, which meant the node now had all the necessary components to be successfully run (as shown in the screenshot below). The golden nugget icon was clicked to run the stream.

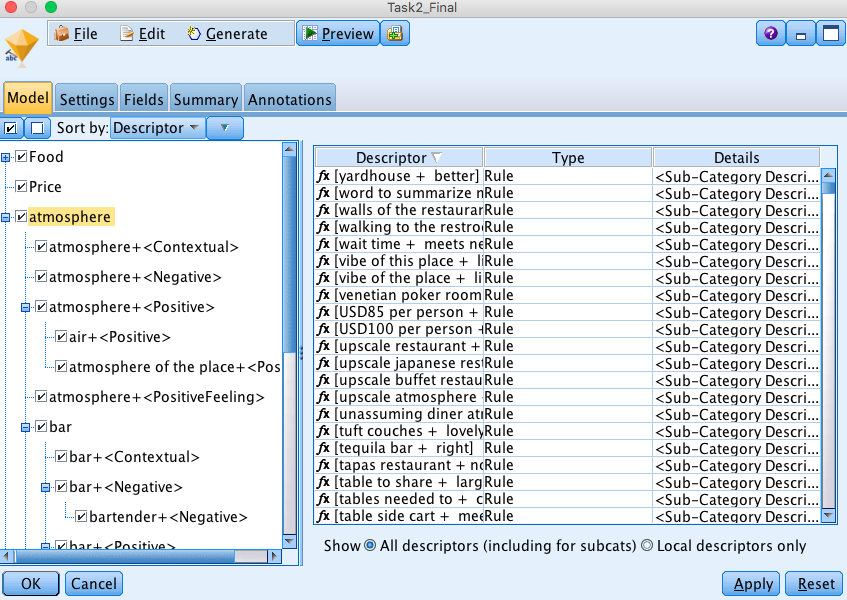


After running the stream, the anticipated golden nugget (labeled Task1\_Final) appeared on the canvas as anticipated. To confirm the needed output was generated within this golden nugget, a Table node was connected and the stream run through this new output node. The execution of the Task 1 stream was proven successful upon opening the Table and seeing four new columns in the Table representing each of the build categories. These new columns were populated with a T (True) or F (False), with True indicating that the category was found in the text data (and False signifying the category was not contained in the text).

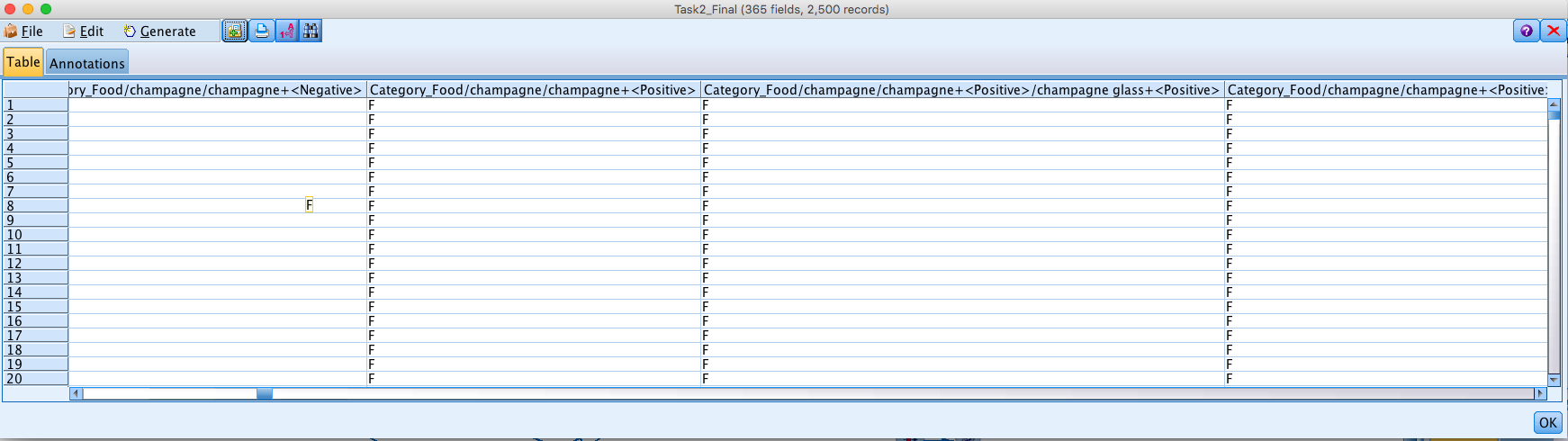
****

The goal of the second task was to extract linguistic “concepts” from the review text to determine the sentiment of the text reviews and at a high level determine whether the reviews were positive or negative towards the restaurants. To achieve this goal, a new stream was built using a new Text Mining node. However, unlike in Task 1 where the Text Mining node was run “using extraction results to build categories”, the Model field was edited and “exploring text link analysis (TLA) results” was selected. With the libraries already having been tested for Task 1, the Customer Satisfaction Opinions (English) library was used again, and the new stream was run. 

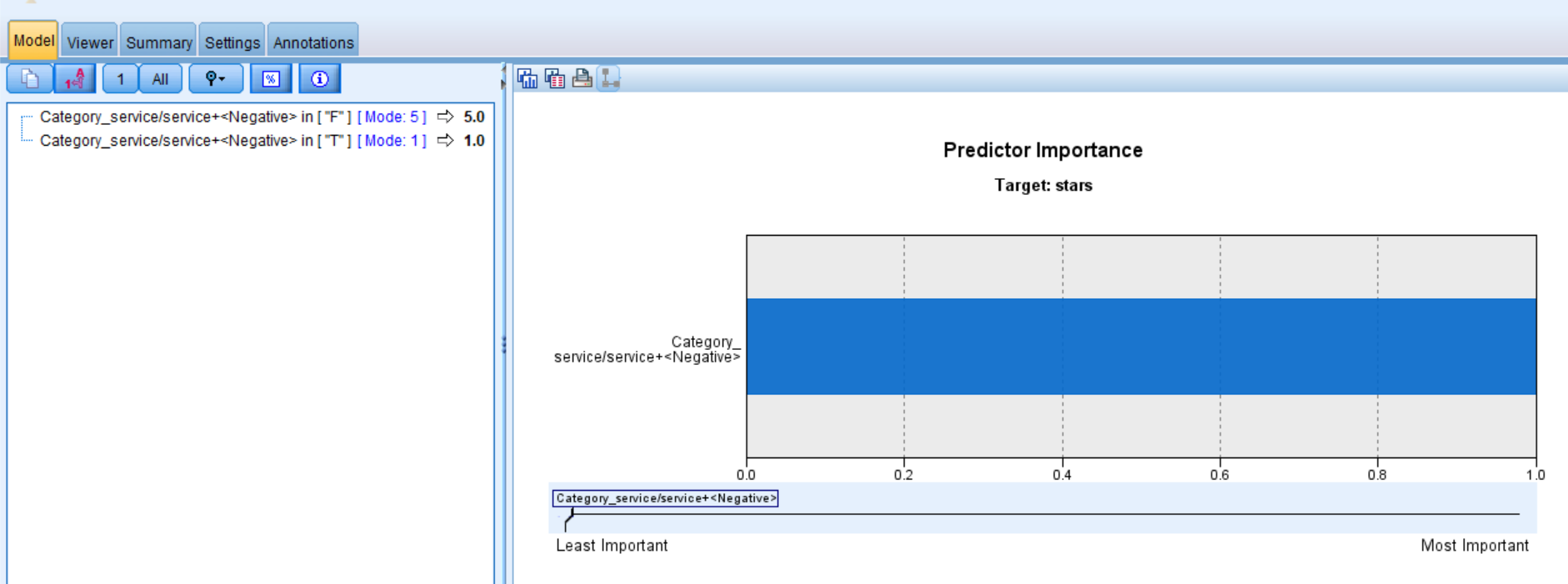
The golden nugget output revealed a lengthy list of category attributes, which ranged from a specific food product like “crab” to a generic term like “experience”. However, because the task goal was to derive the embedded sentiment corresponding only to the four provided attributes, this meant that each category attribute produced by the model needed to be individually reviewed and then bucketed into either food, service, atmosphere, or price. The act of moving category attributes into one of the higher-level categories was a simple drag and drop within the model window, but the more difficult piece to this process was interpreting the ambiguous category attributes and placing them in the correct bucket. Once this was complete, positive and negative inferences tied to the extracted linguistic concepts were recognized in the data.



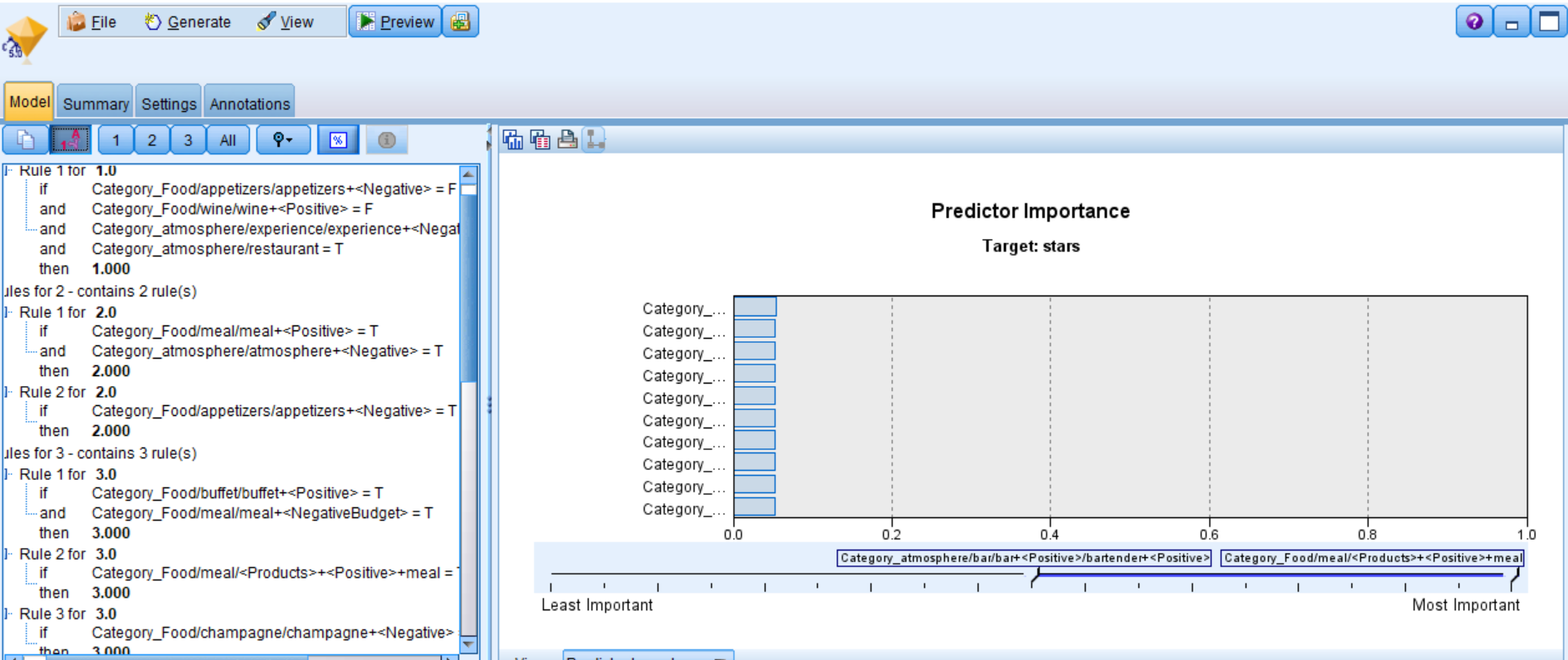
As a final step to complete Task 2, a Table node was again called upon to confirm that the linguistic concepts were successfully extracted from the review text. Like with Task 1, if the built stream was successful, these concepts would be added to the table as new columns to the right of the existing data and each column would be populated with a T or F flag. Again, these flags signified whether or not the linguistic concept was found within the review. As shown below, the Table added for Task 2 did contain the new linguistic concept data, thus achieving the goal.



The first step in starting Task 3 was to add a Type node that was then connected to the text analytics model nugget. Within the Type node, the “Stars” variable was changed to “Ordinal” type and was set to “Target”. Three nodes were selected to build models - C&R Tree, C5.0 and Association Rule Mining nodes. Initially, the C&R Tree model was built by setting “Stars” as target and the 42 concepts from each category (food, service, price and atmosphere) as Predictors. The model produced only one rule.



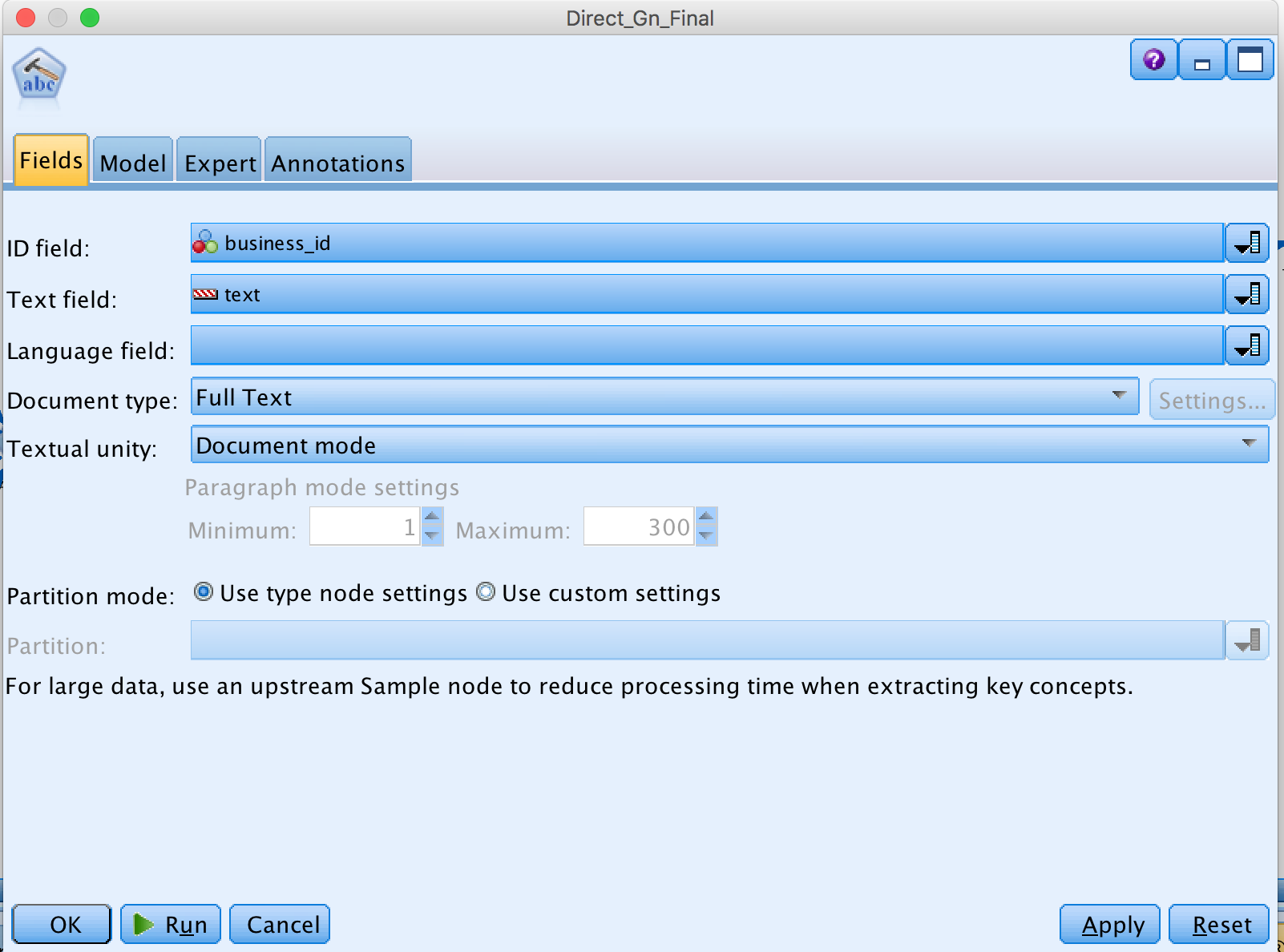
The C5.0 model was built using Custom Field Assignments where again “Stars” was set as the Target and the 76 concepts from each category were set as Predictors. The output type was formatted as rule set as opposed to decision tree. The model produced more intuitive rules for each star rating.



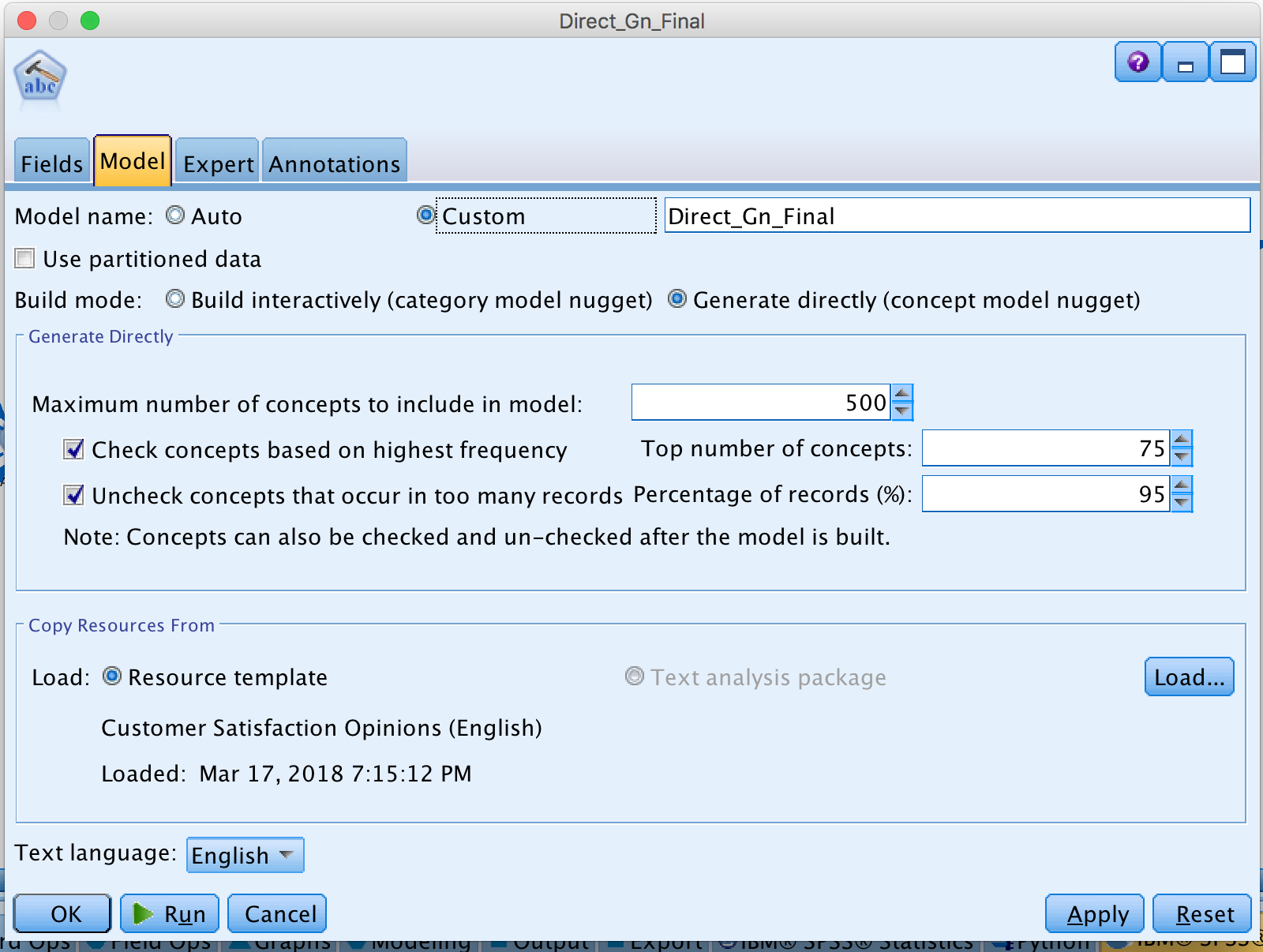
These rules included:

1. Food is good, but the experience is unpleasant = Star Rating 1
2. A portion of the meal like appetizers is not great and the atmosphere is unpleasant – Star Rating 2
3. Food is costly = Star Rating 3
4. Food is good, but costly, atmosphere is unpleasant, and service is good = Star Rating 4
5. Appetizers are not satisfactory = Star Rating 5

To complete further testing for Task 3, the 5 spreadsheets were appended with a sample of 500, and a Type node was used to validate that the data was loaded properly using the preview button. From the type node, a Text Mining node was connected (named “Direct\_Gn\_Final”). Within the Field tab, the text field and the business id were selected as shown below.



Within the Model tab, “Generate directly” was selected, the maximum number of concepts for the model (500) and the top number of concepts (75) were determined, and the decision to ignore concepts that occur 95% of the time was made. Like with previous tasks, the “Customer Satisfaction Opinions (English)” dictionary was used.



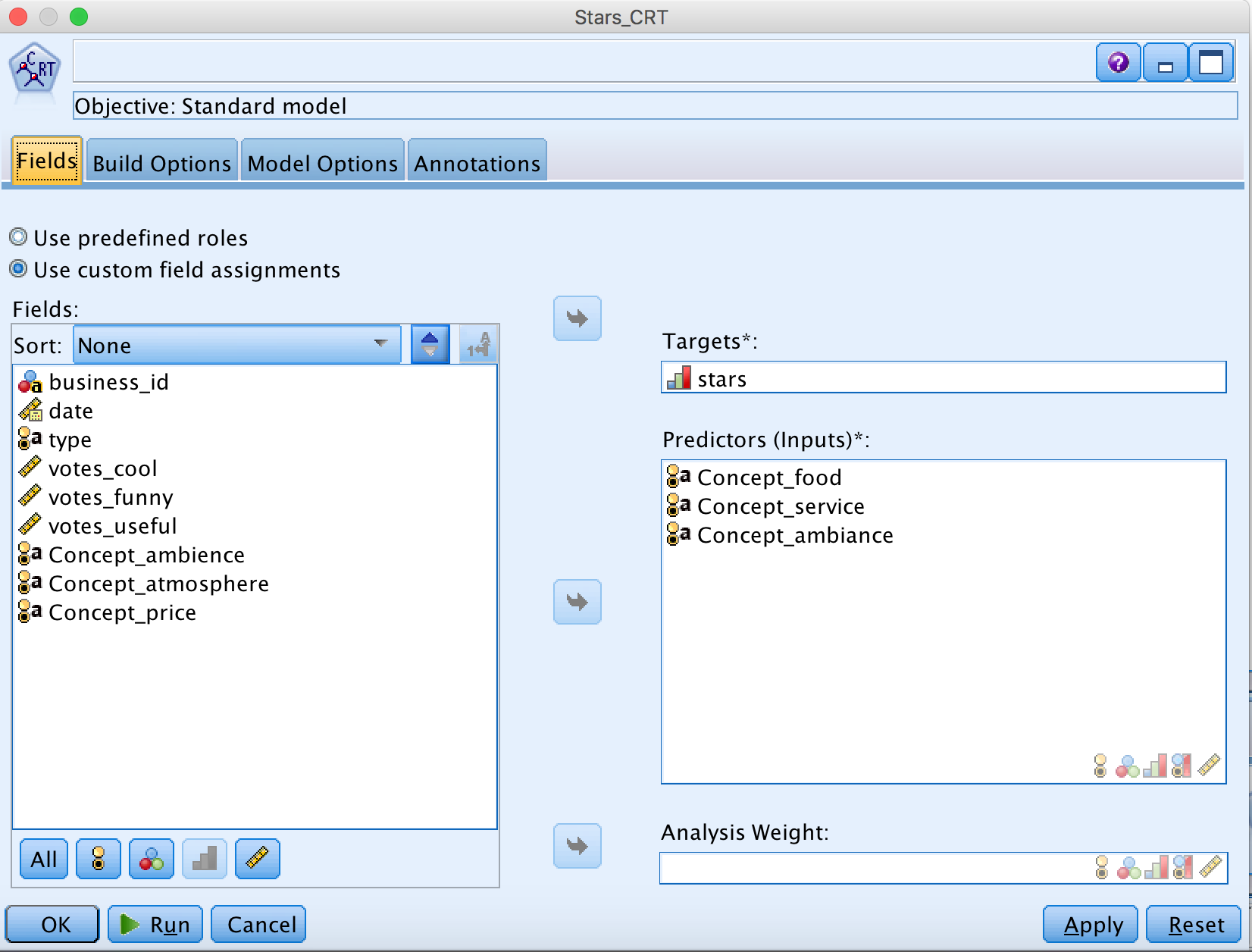
The model produced a corresponding golden nugget, titled “Direct\_Gn\_Final”, which was then edited to add 6 concepts for the model builds. The concepts chosen were the four previously provided (food, price, service, atmosphere), and ambience and ambiance. The additional two concepts were added once it was discovered that without them, no rules were being generated.

As a quick vaidation, this golden nugget was connected to a table to ensure that the newly added 6 concepts werte showing up with F (False) / T (True).

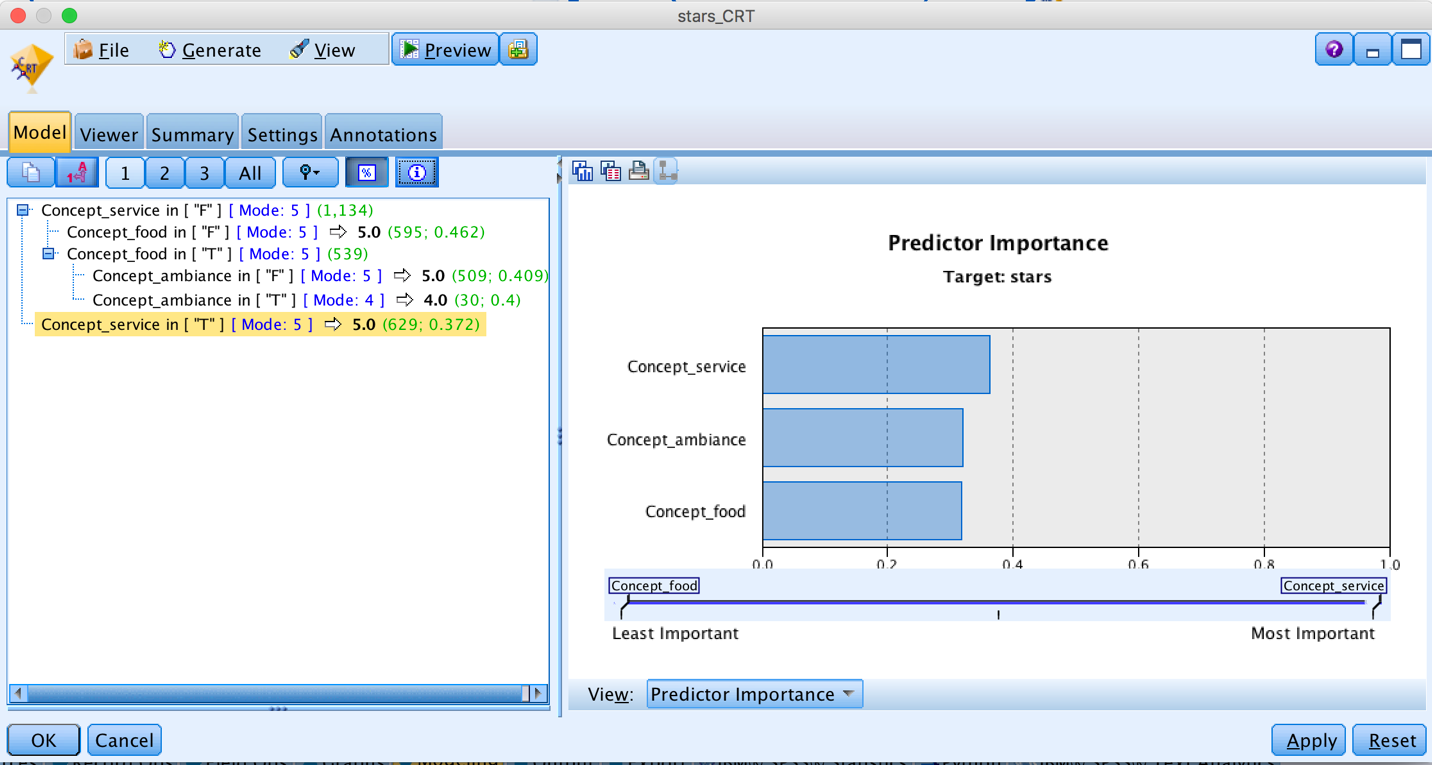
Once confirmed, the “Direct\_Gn\_Final” nugget was connected to a Type node (“Stars\_Type”), and the mearsurement of stars was changed from continuous (default) to ordinal and read values was used. The Type node served as a connection to the model nodes C&R Tree, C 5.0, and Association Rules:

* **C & R Tree**: Within the model nugget, selected custom field assignment (as opposed to using predefined roles) and stars as target. The inputs used were food, service and ambience concepts. No other settings were changed, and the model was run and named “Stars\_CRT”.

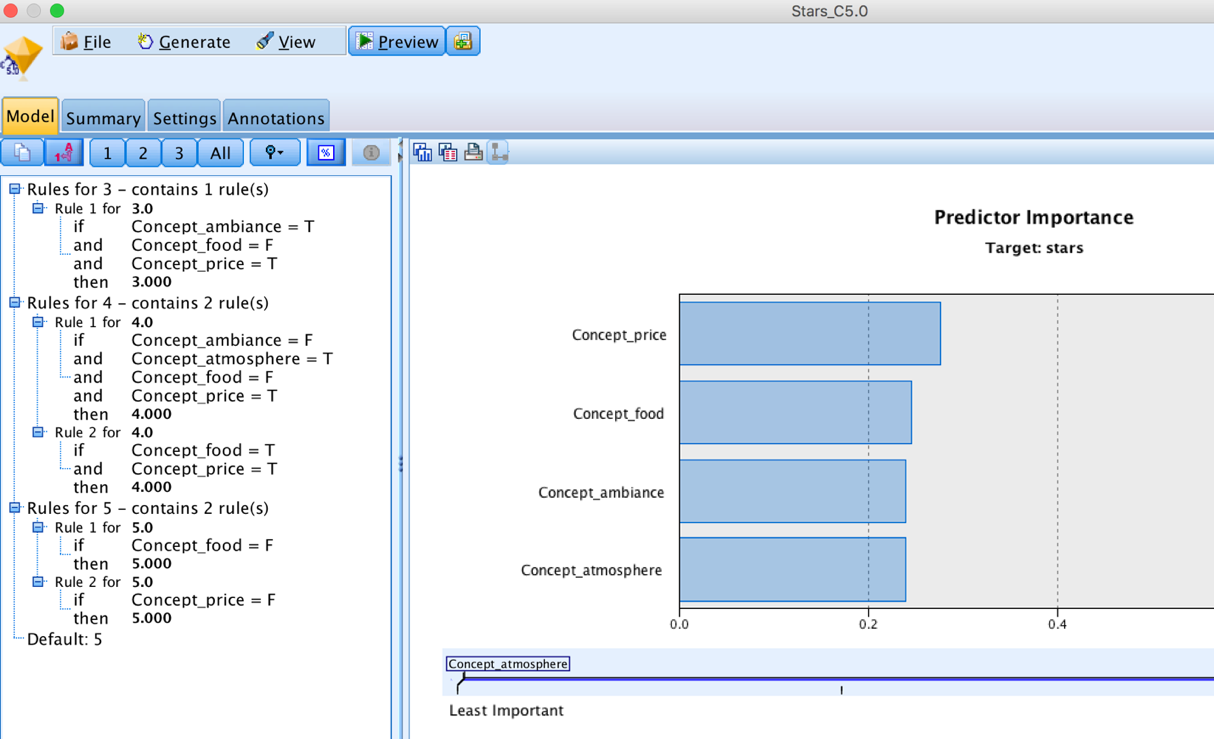
*Please note: Price was note included, as this input did not produce any rules.*



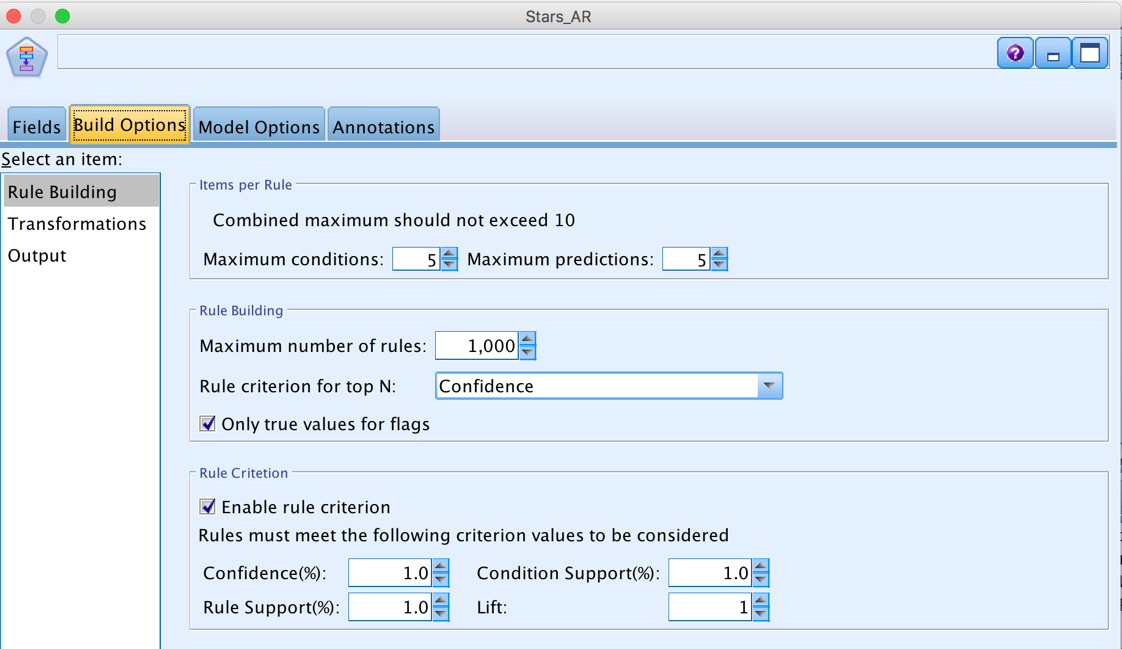
Using the C & R Tree, a set of rules were produced (see below). The predictor importance shows that service (0.36) is slightly more important to customers than ambiance and food (both at 0.32) when they are determining a star rating.



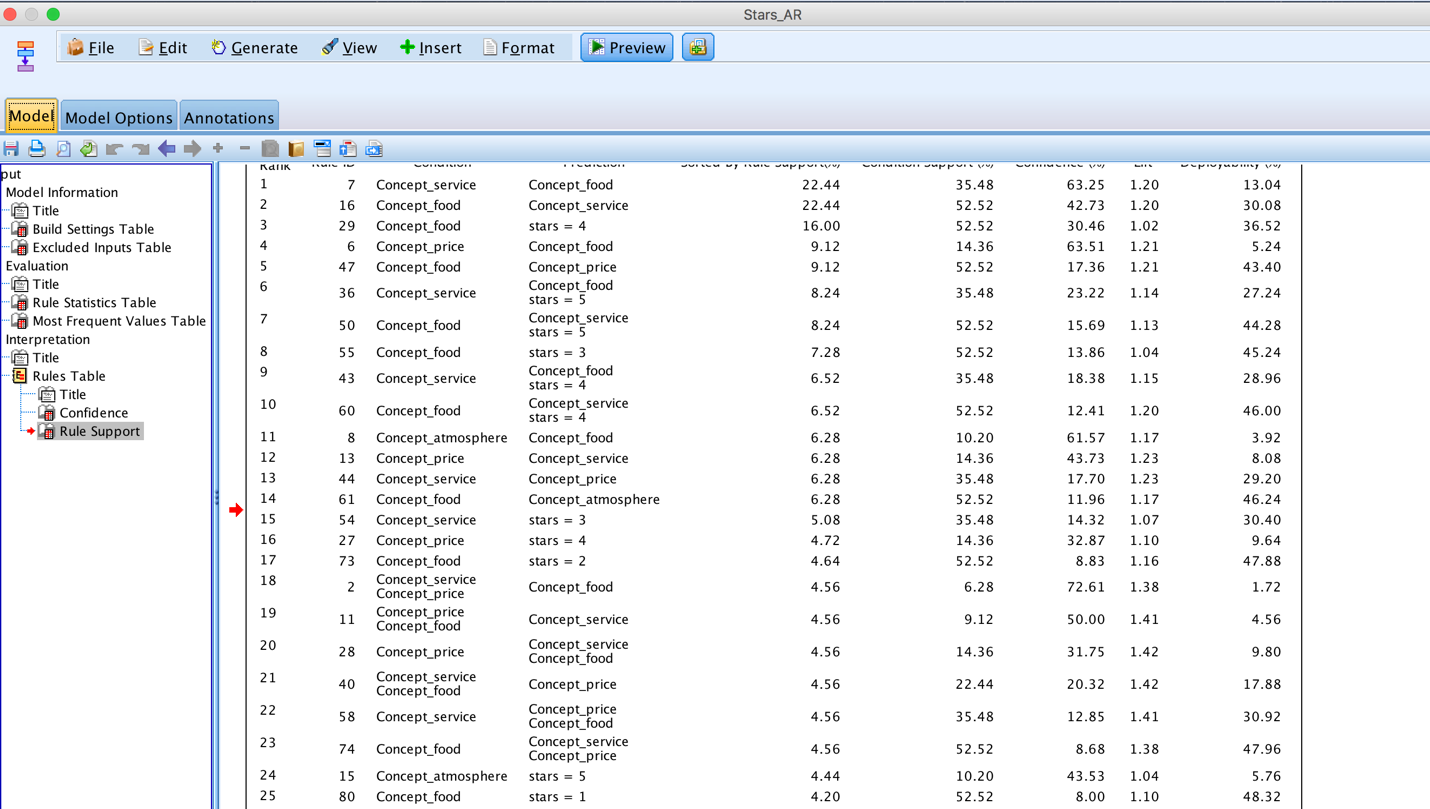
* **C 5.0 Model**: Within this second model nugget, Custom field assignment was again used. Stars were selected as the target and the inputs were the six concepts. On the Model tab, Rule Set was selected in lieu of Decision Tree. This model can be seen on the stream as “Stars\_C5.0”.



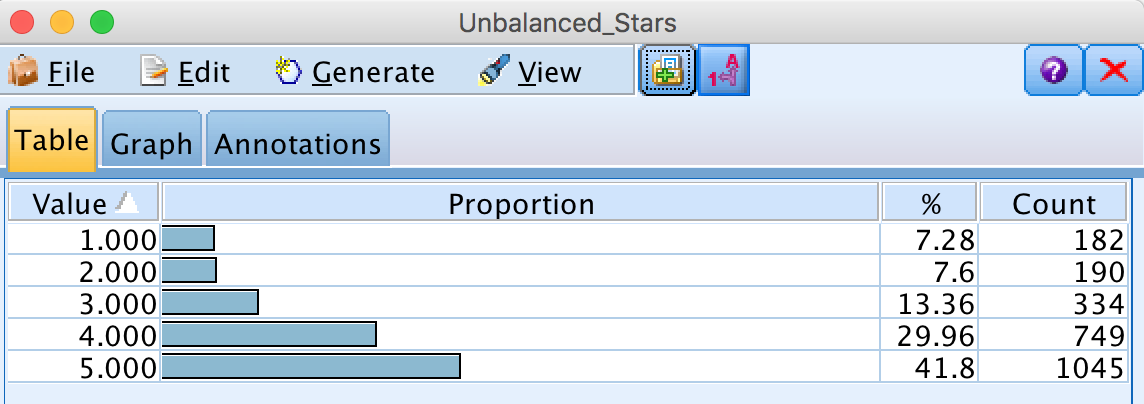
* **Association Rules**: When using Associaion Rules as a third model for testing, similar selections were made like with other models. Again custom field assignment was selected, and stars were identified as the target while the six concepts were identified as inputs. The thresholds for the model were modified to the levels shown below, and the model was run.



The Association Rules model generated a number of rules, as seen below.

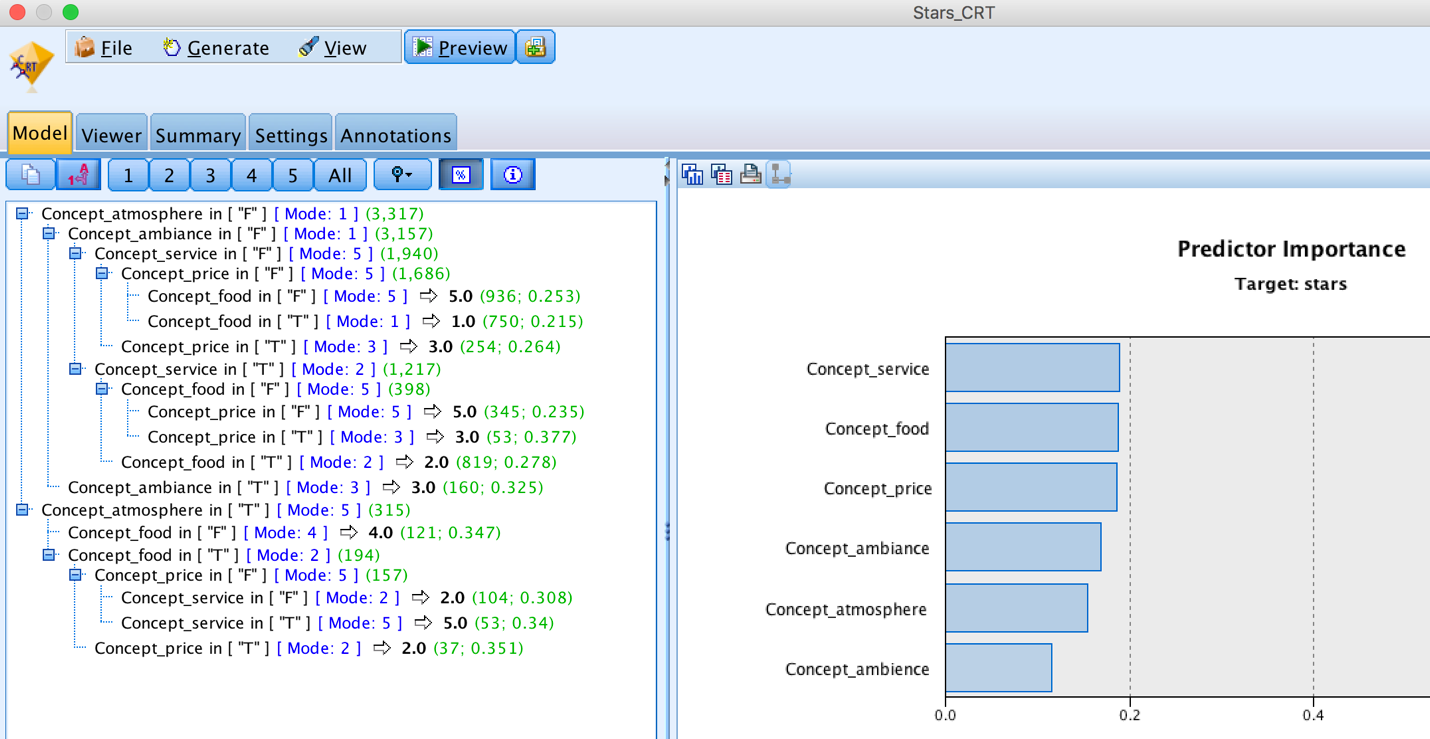


Unfortunately, during this first run the Association Rules model did not generate rules for all five star ratings. It was suspected that star ratings were missing due the data samples selected not representing a balance of the overall data available. This suspicion was confirmed upon checking the distribution of the sample data, as 5 star ratings represented a much larger proportion than other ratings.

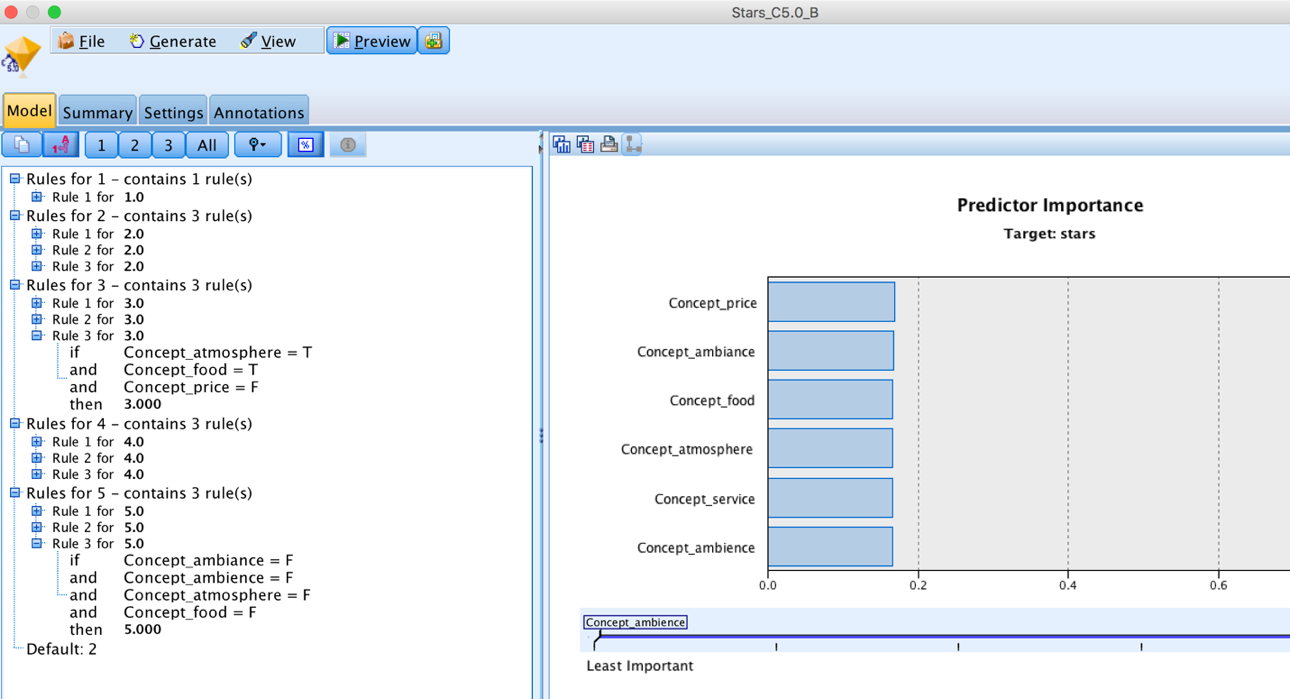


A Balance node was added to equalize the distribution of the five star ratings represented in the sample data. Adding this node allowed for each model to produce better rules.

The balanced data samples produced the following final business rules in the C&R Tree and C 5.0 modeling nodes.



A sample rule from this new set is that if the atmosphere is positive and the food is delicious (positive), but the price is considered expensive (negative), a predicted rating would be for 3 stars.



**Executive Summary**

All three specific tasks were successfully completed using IBM SPSS Modeler software. The first task required that category attributes be identified (i.e., features - provided as “food,” “service,” “atmosphere,” and “price”) to develop Text Analytics models. For this task, a data sample of 500 lines was used from each of 5 data populations for final data modeling and analysis, which enabled the software to run the data streams more efficiently while still generating meaningful results, as expected by the requirements. IBM SPSS software capabilities allowed for several different data file sources to be merged into one data stream, making all analysis easy to navigate and to read through the generated graphical outputs. Several different data modeling approaches were considered to determine if run simulations would generate expected results. Once a data model generated favorable results, the final task one deliverable was verified by adding a Table to the source, where each category including food, service, atmosphere and price would be marked with True or False output.

The goal of the second task was to extract linguistic “concepts” from the data to determine which reviews were positive or negative towards each restaurant. The Customer satisfaction opinions data library was used to generate best results. No other data modifications were made using other third party software (such as Excel) to manipulate the original data for use within the IBM software. After building four categories (one for each provided attribute), every concept read through Text Analysis was individually reviewed and placed into either the food, service, atmosphere or price category. Following this review process and a rerun of the data, positive and negative inferences were linked to newly extracted linguistic concepts as generated. To verify the final task two deliverable, a table was added to already created data stream, where each category would display true or false output. The true or false outputs signified whether or not the linguistic concept was found within the review.

The third and final task was to extract business insights using the associative rules generated from the models and understanding the highest predictors for star ratings provided by customers. As expected, when food in combination with price or service was bad, customers provided a bad rating of 1. At the other extreme, a star rating of 5 was provided when food was good in combination with either good service or reasonable prices. The true insight was finding that when food was really good in combination with good service and/or atmosphere, customers rated overall experience high despite price being high. Therefore, mangers focused on profitability should strive to maintain or improve the quality and taste of their food, while also working to provide good service, as customers don’t mind paying more for an overall good experience.

**Appendix:**

The Final SPSS model is as follows –

