Kevin Roberts

Professor Guha

Data Science

3/9/2017

Restaurant and Consumer Data Data Set

Recommendation systems have been an interesting topic within the data science community. Using relatively big data, one can predict a top-n list of objects according to the consumers’ preferences. According to the Restaurant and Consumer Data Data Set (RCDDS) donated to UCI Machine Learning Repository in August of 2012, it is the goal to produce a list of recommended restaurants to consumers. This data set is relatively new and large containing 130 restaurants located in Mexico, 138 consumers, and over 47 attributes. The data includes nine csv files in total; four of which contain information regarding restaurant data, three of which contain information regarding consumer data, and one of which contains data on user restaurant recommendations. There are a wide range of data types ranging from numeric, nominal, and ordinal, with majority being nomial type. All the data has been imported into jupyter notebook, cleaned, information analyzed for merging of files, and visualized with basic box, bar, and scatter plots. Within the current research, discussion of how these processes took place and what information was gained will be the primary precept.

The RCDDS is a relatively clean data set. Throughout the cleaning process, all null data points were filled out with question marks to represent missing values. These values were disregarded when the files were used for further visual analysis. As for column, certain rows were cleaned and disregarded from the data completely. This included the fax,url and country column in the geoplaces2.csv file. 130 of 130 data points were missing in the fax column, thus, no inferences could be made, 116 of 130 data points were missing in the url column, therefore, not much influential data could come from such a small amount of data, and finally, all data points in the country column were listed as Mexico, hence, there was no variation to deduce information. It is important to get rid of unneeded data in a dataset as information just becomes messy and if used, will only hinder your results. Data preparation is not just a first step, but must be repeated multiple times over the course of analysis as new problems come to light or new data is collected, therefore, there will be more steps to clean the RCDDS in the future when finding other interesting patterns and analyzing for k-nearest neighbors. [1]

Once majority of cleaning was completed, it was safe to analyze and look for patterns in the data. After running some python scripts, information stated consumers were born between 1930 and 1994 with the average being 1984.7. The birth years are slightly skewed, because the median varies by almost five years at 1989. We found out that 74% of people prefer to use cash for payment, followed by 12.43% debit/bank cards and 9.61% VISA. This is a very sizeable portion of the data favoring cash. Not surprisingly, consumers prefer Mexican cuisine the most, however, again, a very large portion of data at 29.39%. This may not seem like much at first, but, when looking at all the cuisine options (103) and comparing the Mexican cuisine to the second highest (3.33%), it is evident consumers preference for Mexican cuisine overpowers all others. Restaurant cuisine aligns closely with the consumer cuisine preferences. Majority of restaurants overwhelmingly offer Mexican cuisine, followed by International then American. Consumers had an overall, slightly above average rating for restaurants. The average rating gave a 1.2 on a 0-2 scale. Rating, food rating, and service ratings all were similar among consumers, however, service rating varied more with a lower mean and wider middle 50%. Further analysis of consumers broke down their budget (high, medium, low). When graphing this information in a bar chart, majority of consumers had a medium budget, followed by low, unsure, then, lastly, high.

When merging files to find correlations between attributes, interesting patterns were made. First, data wrangling was done to push this information to better compare and graph data. Once this was done, comparative algorithms were done to separated budget into low, medium, and high data frames. With this, the goal was to see if consumers with varying budgets had differing cuisine prefers. However, this was not the case. All consumers still prefered Mexican cuisine overwhelmingly number one followed by several varying cuisines. There was no evident follow up restaurant as many others had below ten consumers preferring each cuisine. With smaller data like this, one needs to limit experimentation and stick to simple models; There are many options, but one should stick to constrained modeling, smoothing and quantification of uncertainty. [2] If more data were collected and not as many consumers overwhelmingly preferred mexican cuisine, this would not be the case. When determining if budget has an effect on how consumers prefer to pay, there was the same result. Budget had little to no effect on how consumers prefer their cuisine or how they prefer to pay. Other merging of file information was collected such as how high consumer budget and low consumer budget effects dress preference.

Throughout the data cleaning, wrangling, and processing steps, getting use to the python syntax was an issue. This involved learning through trial and error. Also, wrangling the data caused problems. Importing the restaurant summary file caused issues as a line in the file is unreadable or too large. Learning to put all data into Data Frames made algorithms more efficient as well as visualizing this information simpler and easier to see. Python provides tons of libraries for large data that made this process very helpful.

Further analysis will include identifying more interesting patterns in consumer behavior such as what cuisines are enjoyed by consumers versus their ratings given to restaurants with those cuisines, as well as identifying any interesting patterns in credit card usage by consumers. Also, using inference and machine learning technique analysis will include determining which of the consumer attributes are most influential when comparing users to their most favorable restaurant, and using k-nearest neighbors in an n-dimensional graph to identify the top-n generated list according to consumers.

Overall, the RCDDS is a very interesting data set. Majority of the data shows consumers favor cash and mexican cuisine, with no variation dependant on budget. All users and restaurants are from Mexico which can lead to some bias data, so conclusions can only be applied to that specific region. There are a wide range of restaurants and consumers listed, ranging in age and cuisines. Cleaning the data is a very important step that will continue to happen throughout research as many more patterns can still be looked into. Further analysis will include learning from these data patterns, finding relevant attributes, and generating a top-n list for future consumers with a collaborative filtering recommendation system.

Github link: <https://github.com/kevinroberts94/DataScience>

**References**

1. Wickham, Hadley. *Tidy Data*. *Journal of Statistical Software*. American Statistical Association, n.d. Web.
2. Deeb, Ahmed El. "What to Do with "small" Data?" *Rants on Machine Learning*. N.p., n.d. Web.