CS 838 (Spring 2017): Data Science Project Report - Stage 5 (Group 12)

Deepanker Aggarwal
deepanker@cs.wisc.edu

Saket Saurabh
ssaurabh@cs.wisc.edu

Vishnu Lokhande
lokhande@cs.wisc.edu

1 Objective

—- Fill in the details here —-

2 Statistics of Table E

- The schema of table is as follows (name, address, zipcode, cuisine, price, violation_code, critical_flag, grade, median_household_income, median_real_estate_value, population_density, cost_of_living, population, neighborhood, borough).

- ### 2.1 Attribute definitions

  —- Fill in the details here of what each attribute means —-

- The number of tuples in the Table E are $5,560$.

- Some sample tuples from table E are as follows:

  1. (juniper,"237 w 35th st,",10001,bars american,$$,06C,Critical,A, 81671,650200,35350,157.4,21966,Chelsea and Clinton,Manhattan)
  2. (mission chinese food,"171 e broadway, ,",10002,desserts bars chinese,$$,10F,Not Critical,A,33218,535600,93461,168.9,82191, Lower East Side, Manhattan)
  3. (okinii,"216 thompson st, ,",10012,japanese bars,$$,10B,Not Critical,A, 86594,1000001,80873,164.7,26145, Greenwich Village and Soho,Manhattan)
  4. (the fitz,"fitzpatrick manhattan, 687 lexington ave,",10022,bars american irish,$$,10B,Not Critical,A,109019,866100,67873,158.2,29618, Gramercy Park and Murray Hill,Manhattan)

5. (ny sweet spot cafe,"2376 coney island ave,",11223,middle eastern european,$$,06C,Critical,A,41328,613000,35968,167.4,74606,Southern Brooklyn,Brooklyn)

- Apart from the attribute 'cuisine', no other attribute in the table has missing values.

3 Data Analysis Tasks

We undertook the following categories of data analysis tasks on the Table E.

- OLAP Analysis
- Classification

We describe the process and the outcome of these tasks in the following sections.

4 Data Analysis I: OLAP

4.1 Questions that we wanted to answer:

Using OLAP analysis, we wanted to answer the following questions for our NYC Restaurant dataset as represented by Table E:

1. How many restaurants are located in each borough and neighborhood of New York City?
2. How many NYC restaurants are there for each kind of price type as reported at Yelp?
3. How many NYC restaurants are there for each kind of Yelp rating?
4. How many NYC restaurants are there for each grade of health violation?
5. How many NYC restaurants have received 'critical' health violation status?
6. Which borough and neighborhoods of New York City have restaurants with highest number of 'critical', grade 'A' health violations?
7. Which Yelp ratings correspond to the highest number of 'critical', grade 'A' health violations?
8. How many 'critical', grade 'A' health violations exist for even expensive restaurants in New York City?

4.2 OLAP Data Analysis Process

We had to do a couple of interesting things before we could perform OLAP analysis on our dataset as represented by Table E. We loaded the dataset in a Python package called 'cubes', which provides a ROLAP server backed by SQLite relational database. We introduced a concept hierarchy for area of restaurant in New York City- borough > neighborhood > zipcode. We defined a set of dimensions across which we could perform aggregation, roll-up, drill-down, slicing and dicing. The following subsections define our OLAP methodology in detail.

4.2.1 Introducing Concept Hierarchy for Area

The table E produced at the end of Stage 4 just had the zipcode information for each restaurant. Therefore, in this stage, we added two more new attributes: borough and neighborhood, and created a concept hierarchy for area. This borough and neighborhood data for each zipcode was obtained by web scraping this information from this url: `https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm` and afterwards, we joined this scraped data with the zipcode available in Table E.

The resulting concept hierarchy can be represented as: borough > neighborhood > zipcode. This implies that each zipcode belongs to a larger neighborhood, which in turn is a part of larger borough. In fact, New York City has 5 boroughs (Bronx, Brooklyn, Manhattan, Queens, Staten Island) and 42 neighborhoods and 180 zipcodes.

4.2.2 Dimensions and Fact in our OLAP analysis

Based on the data in the Table E, we created these dimensions for our OLAP cube: area, price, rating, and critical_flag. The dimension 'area' represents a concept hierarchy as explained above. The restaurant data represents the fact table with the number of restaurants as one of the measures of the fact table.

4.2.3 Summary of Schema used in our OLAP analysis

Figure 1 summarizes the Star-Schema of our OLAP model.

4.2.4 A Note on using the Python 'cubes' package for OLAP

We used the Python 'cubes' package to perform our OLAP analysis. The 'cubes' package provides a default ROLAP server backed by a SQLite database that was sufficient for our given dataset size. The 'cubes' package takes as input a JSON file that describes the OLAP logical model. (We have attached this JSON file in the Appendix section of this report.) The logical model described by the JSON file is independent of the underlying physical model. In our case, the underlying physical model is defined by a single relational table in which we load the Table E. The 'cubes' package defines a nice set of abstractions on top of the logical model that offers standard operations like drill-down, slicing, dicing, etc. For example, a browser represents a view of the cube, a point cut represents a slice along a dimension and one can invoke the aggregate() method on the browser object to perform various kinds of drill-down.

4.3 Outcomes of OLAP Analysis

In this section, we report the results of various OLAP analysis that we performed.

- Roll-up on number of restaurants in each NYC Borough
- Drill-down on number of restaurants in each NYC Neighborhood
- Roll-up on number of restaurants with each type of price
- Roll-up on number of restaurants with each type of rating
- Roll-up on number of restaurants with critical health violation type
- Roll-up on number of restaurants with each health violation grade
- Slice and dice on health violation by area
- Slice and dice on health violation by restaurant rating
- Slice and dice on health violation by restaurant price

5 Data Analysis II: Classification

The attributes from the table E which have been used to solve the classification questions are (zipcode, price, violation_code, critical_flag, grade, median_household_income, median_real_estate_value, population_density, cost_of_living, population, neighborhood, borough).

5.1 Questions that we wanted to answer:

1. Can we predict the price of the restaurant from the other attributes used for classification?
2. Can we predict the critical flag of the restaurant based on the other attributes used for classification?
3. Can we predict which borough the restuarant belongs to based on the other attributes used for classification?

5.2 Classification Process

5.3 Preprocessing

We analyzed the dataset and found no missing values. As some attributes had string values, we transformed them to integer values and also analyzed how data is distributed in Figure 1. We find that price attribute is skewed for $ and $$, grade is also skewed. Before classifying the data, we also standardize it, such that mean of the data is 0 and variance is 1.

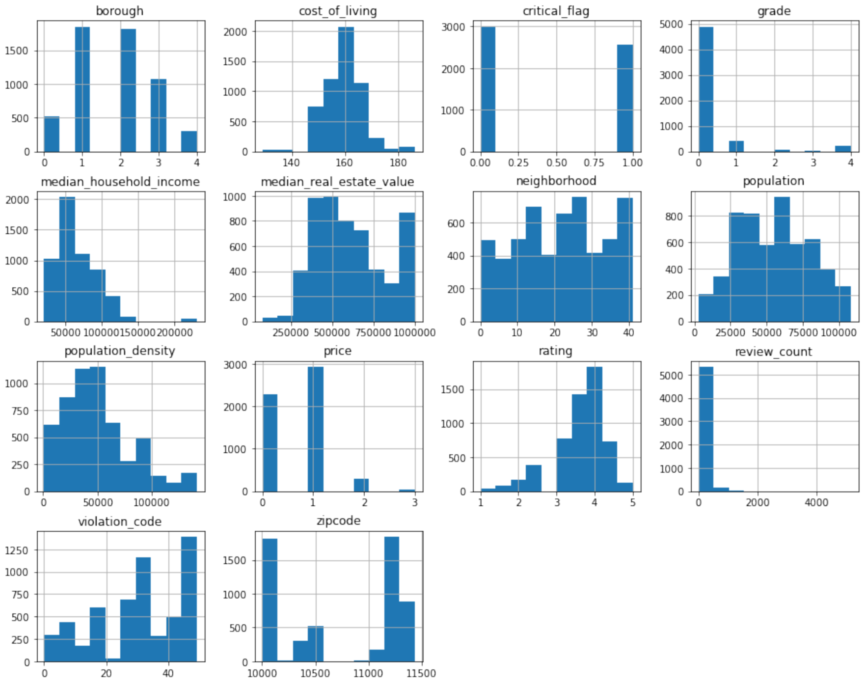

Figure 1: Convolution Operation Example

5.4 Prediction

For each ques we try below classifiers:-

- Logistic Regression
- Linear Discriminant Analysis
- K-Nearest Neighbors
- Decision Tree Classifier
- Naïve Bayes

Predicted Label	Precision(Mean)	Recall(Mean)	F1(Mean)
Logistic Regression	0.954497	0.898707	0.925463
Linear Discriminant Analysis	0.940495	0.932362	0.936340
K-Nearest Neighbors	0.954768	0.893945	0.922586
Decision Tree Classifier	0.940495	0.932362	0.936340
Naïve Bayes	0.945024	0.932362	0.938578
SVM	0.948239	0.932362	0.940171

- SVM

We do 5-fold Cross Validation, then choose the best classifier on the basis of accuracy as in the questions we are trying to answer has labels with equal importance. Then we report Precision, Recall and F1-Scores for the best classifier.

6 Classification Results

We report the results for each question below.

6.1 Question 1

Can we Predict the price of the restaurant from the other attributes used for classification?
After cross-validation, we found........ to perform best as seen in Figure ??. Precision, recall and F-1 scores of each label for the classifier for ...... classifier are in Figure

Predicted Label	**Precision(Mean)**	**Recall(Mean)**	**F1(Mean)**
Logistic Regression	0.954497	0.898707	0.925463
Linear Discriminant Analysis	0.940495	0.932362	0.936340
K-Nearest Neighbors	0.954768	0.893945	0.922586
Decision Tree Classifier	0.940495	0.932362	0.936340
Naïve Bayes	0.945024	0.932362	0.938578
SVM	0.948239	0.932362	0.940171

subsection for each question paste results and box plots

7 Learnings and Conclusion

—- Fill in the details here —-

8 Future proposals

—- Fill in the details here —-

9 Appendix

```
##################################
#--- olap_model.json ---#
##################################
# This is the logical model for the OLAP cube that was fed into
# the Python 'cubes' package for analysis.

{
    "dimensions": [
       {
         "name": "area",
         "levels": [
              {
                  "name": "zipcode",
                  "label": "ZipCode"
              },
              {
                  "name": "neighborhood",
                  "label": "Neighborhood"
              },
              {
                  "name": "borough",
                  "label": "Borough"
              }
           ],
         "hierarchies": [
              {
                  "name": "area_hierarchy",
                  "levels": ["borough", "neighborhood", "zipcode"]
              }
         ]
       },
       {
         "name": "price",
         "levels": [
              {
                  "name": "price",
                  "label": "Price"
              }
         ]
       },
       {
         "name": "rating",
         "levels": [
              {
                  "name": "rating",
                  "label": "Rating"
              }
```

```
      ]
    },
    {
      "name": "critical_flag",
      "levels": [
           {
               "name": "critical_flag",
               "label": "Critical Health Violation Flag"
           }
      ]
    },
    {
      "name": "grade",
      "levels": [
           {
               "name": "grade",
               "label": "Health Violation Grade"
           }
      ]
    }
],
"cubes": [
    {
       "name": "nyc_restaurants",
       "dimensions": ["area", "price", "rating", "critical_flag",
           "grade"],
       "measures": [
               {"name":"median_household_income", "label":"Household
                   Income"},
               {"name":"median_real_estate_value", "label":"Real
                   Estate Value"},
               {"name":"population_density", "label":"Population
                   Density"},
               {"name":"cost_of_living", "label":"Cost of Living"},
               {"name":"population", "label":"Population"}
             ],
       "aggregates": [
               {
                  "name": "household_income_avg",
                  "function": "avg",
                  "measure": "median_household_income"
               },
               {
                  "name": "real_estate_value_avg",
                  "function": "avg",
                  "measure": "median_real_estate_value"
               },
               {
                  "name": "cost_of_living_avg",
                  "function": "avg",
                      "measure": "cost_of_living"
                  },
                  {
                      "name": "population_avg",
                      "function": "avg",
                      "measure": "population"
                  },
                  {
                      "name": "num_restaurants",
                      "function": "count"
                  }
              ],
           "mappings": {
                        "area.zipcode": "zipcode",
                        "area.neighborhood": "neighborhood",
                        "area.borough": "borough"
                       }
       }
    ]
}
```