PRELIMINARY ANALYSIS OF RESTAURANT RECOMMENDATION SYSTEM

Sukraat Ahluwalia; Guangze Jin; Shih-Ting Huang
Rochester Institute of Technology, Rochester, New York, USA
[sxa4430, gj9530, sh3964]@rit.edu

Motivation

- We love exploring new restaurants and this is an interesting project to perform preliminary analysis of restaurant recommendation system.
- We practiced database and data mining techniques using real-world Yelp open dataset.

Methods

- Building relational DBMS database from Yelp dataset
- Building predicting models for target variable (e.g. goodForKids) using decision tree, random forest, and SVM
- Performing cluster analysis of user attributes

Building Yelp Relational DBMS

Business Entity Set

- Extract business type = [Restaurant, Food, or Diners]
- 65,028 observations
- Attributes: goodForGroup, goodForKids, wheelchair, noiseLevel, alcohol, takeOut, reservation, delivery, hasTV, wifi, priceLevel, creditCard etc.

User Entity Set

- 1,183,362 observations, 9 attributes

Review Entity Set

- Systematic sampled (every 10th) from raw data
- 473,690 observations, 7 attributes

Checkin Entity Set

- One attribute per hour, 24*7 variables, and 135,148 observations

Tip Entity Set

- 1,028,802 observations, 4 attributes

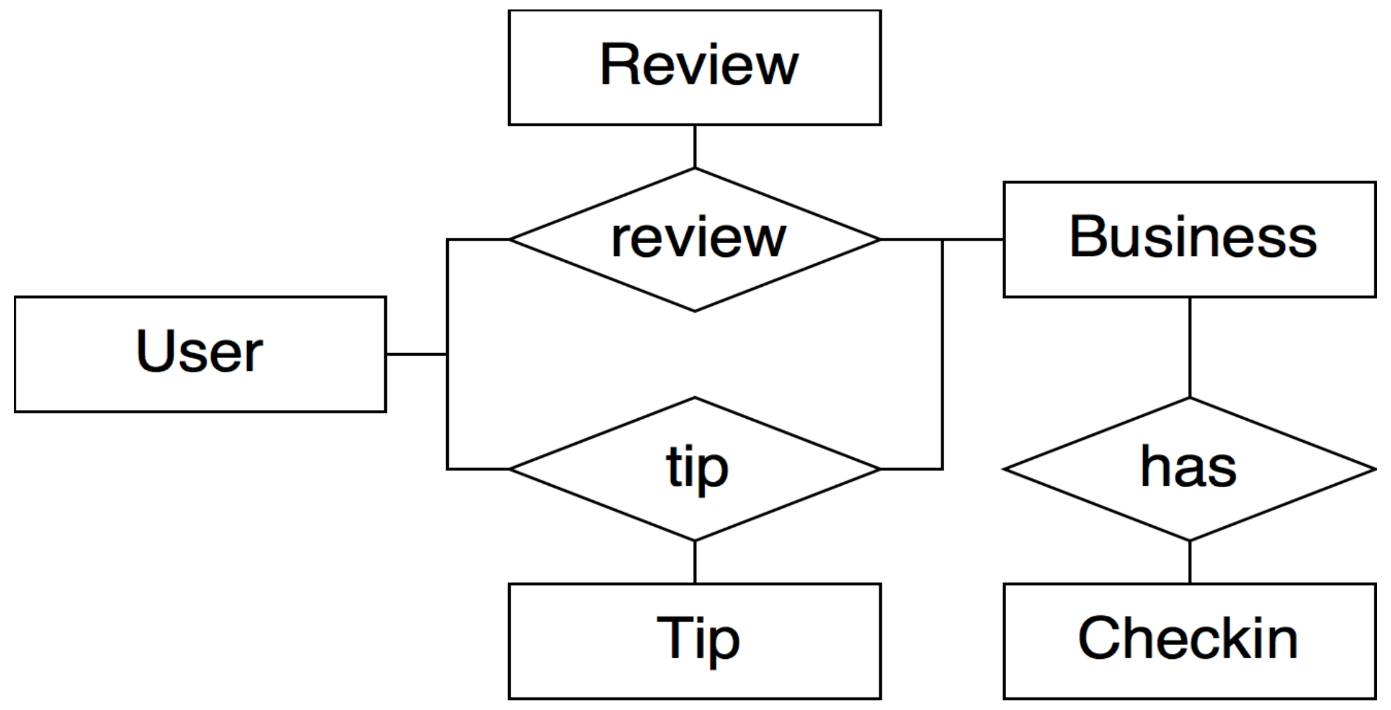
Review
review
Business
User
tip
has
Tip
Checkin

Figure 1. Entity-Relationship diagram of yelp database.

Predicting Models for goodForKids

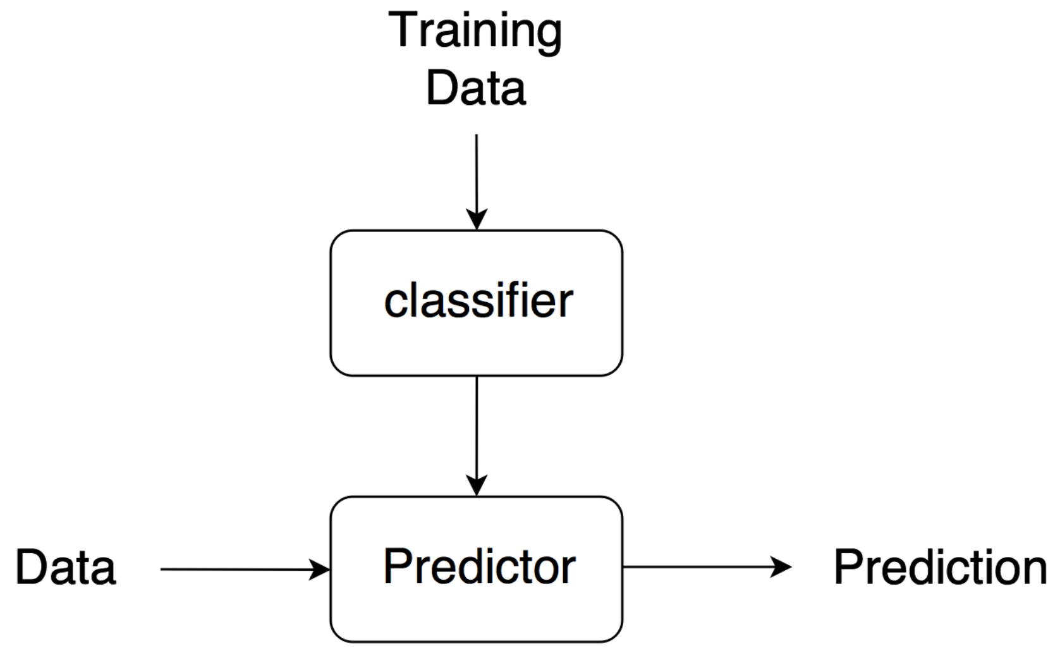
Training Data
classifier
Data
Predictor
Prediction

Figure 2. Attribute Predictor flowchart

Date Preparation

- Created dummy variables for multiple category variables

Variable Selection

- Correlation analysis for selecting important variables (Figure 3)

Model Establishment

- Split dataset into three sets:
  - training set (70%)
  - validation set (15%)
  - test set (15%)

- Decision Tree Model
- Random Forest Model
- SVM Model

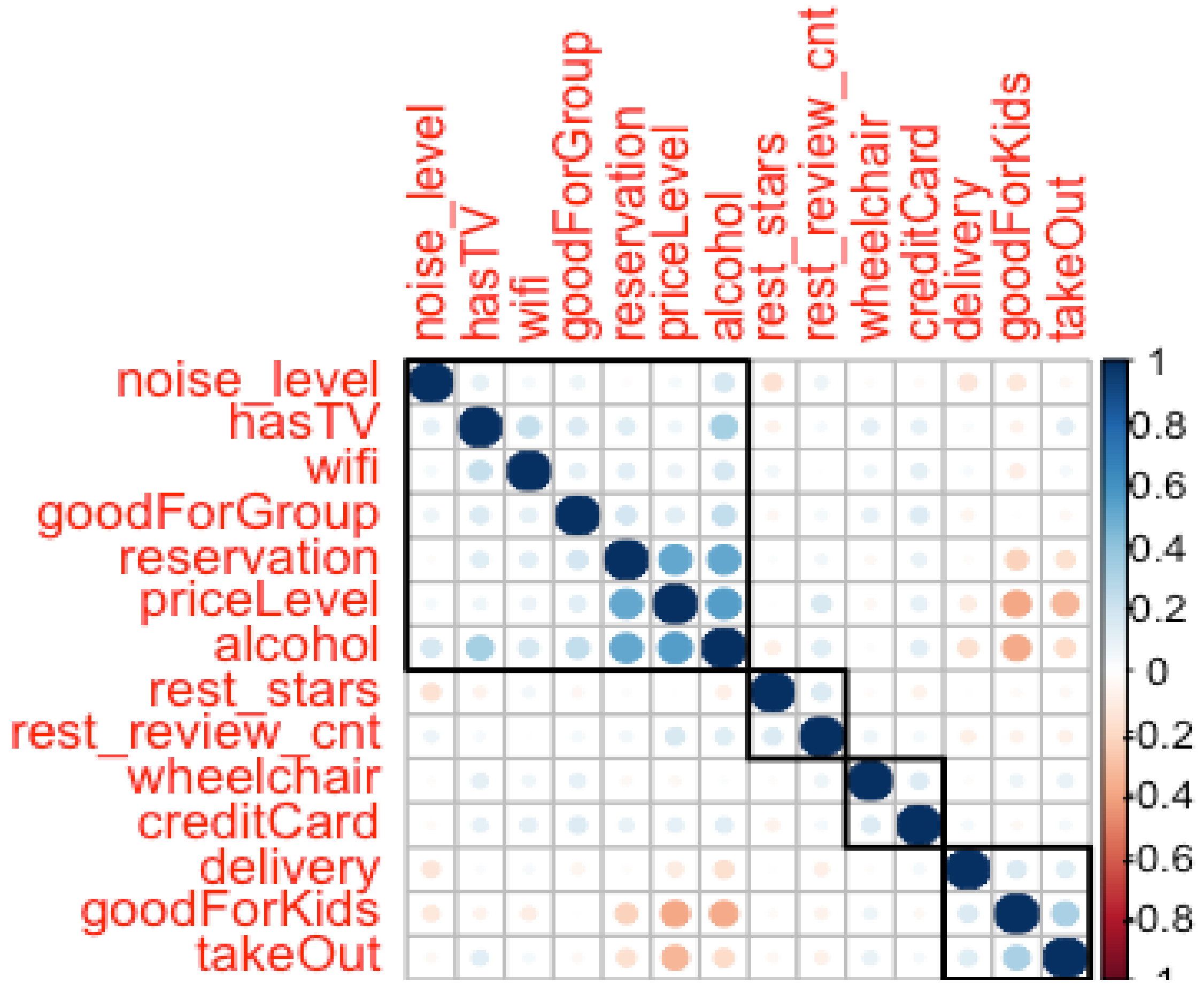

Figure 3. Variable correlation using Pearson

Results

- Built all three models with all variables

Table 1. Overall errors with all variables

	Full	Training	Validation	Test
Decision Tree	14.5%	14.6%	14.3%	13.9%
Random Forest	13.5%	13.2%	13.8%	14.6%
SVM	13.7%	13.6%	13.2%	14.7%

- Excluded variables with lower correlation and built models

Table 2. Overall errors with relatively important variables

	Full	Training	Validation	Test
Decision Tree	14.5%	14.6%	14.3%	13.9%
Random Forest	12.9%	12.8%	13.9%	12.6%
SVM	12.8%	12.8%	13.5%	12.4%

Cluster Analysis of User Attributes

- Collaborative filtering based: finding other users whose rating patterns match the targeted user
- Group similar entities based on a set of attributes
- KMeans clustering algorithm:

$$argmin_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

- Use K=7 as the number of clusters for the system.

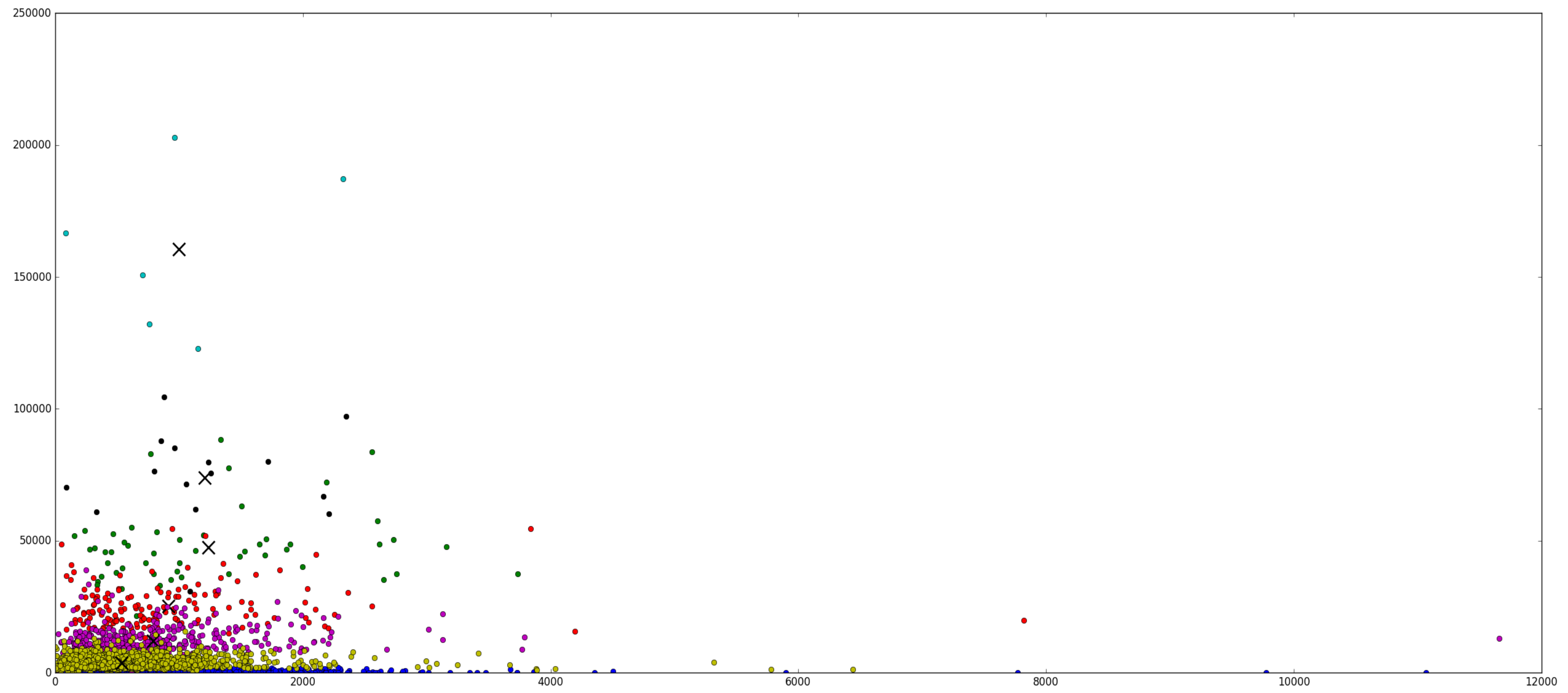

Figure 4. Cluster plots at K=7

Concluding Remarks

- Used Python to build the restaurant database in SQLite (from .json files to .db database).
- Used R to perform correlation analysis of variables.
- Excluded variables with lower correlation to reduce error rate of the predicting models for goodForKids.
- K=7 is the most suitable cluster number for clustering analysis of user attributes.
- Our preliminary findings can contribute to the restaurant recommendation system in the future.

References

[1] 2017. Yelp Open Dataset. (2017). https://www.yelp.com/dataset/

[2] Daniar Asanov. 2011. Algorithms and Methods in Recommender Systems. Berlin Institute of Technology.

[3] Aarshay Jain. 2016. Quick Guide to Build a Recommendation Engine in Python. (2016). https://www.analyticsvidhya.com/blog/2016/06/quick-guide-build-recommendation-engine-python/

[4] Ashish Kumar. 2017. Recommendation Engine - Content-Based Filtering & Collaborative Filtering. (2017). https://www.linkedin.com/pulse/recommendation-engine-content-based-filtering-ashish-kumar