Capstone Project Yelp Business

Memed

22 novembre 2015

PREDICTING RATING OF A BUSINESS

Abstract

Predicting user preferences, for items such as for commercial products, movies, and businesses, is an important and well-studied problem in recommendation systems. In this project, we investigate potential factors that may affect business performance on Yelp. We use features available in the Yelp Business dataset, especially Business Attributes. After preprocessing the data to handle missing values, we ran two selections techniques(Anova, Tree-Based), to evaluate which features might have the greatest importance. After that we performed regression models (linear regression). Thus, in order to improve business performance, a future step would be to conduct additional analysis of review text to determine new features that might help the business to achieve a higher number of positive reviews.

INTRODUCTION

Yelp, founded in 2004, is a multinational corporation that publishes crowd-sourced online reviews on local businesses. As of 2014, Yelp.com had 57 million reviews and 132 million monthly visitors [1]. A portion of their large dataset is available on the Yelp Dataset Challenge homepage, which includes data on 42,153 businesses, 252,898 users, and 1,125,458 reviews from the cities of Phoenix, Las Vegas, Madison, Waterloo, and Edinburgh [2]. For businesses, the dataset includes business name, neighborhood, city, state, latitude and longitude, review rating, number of reviews as "review_count", and categories such as "Ambiane Romantic". The goal of this work is to analyze what factors may affect the performance of a business on Yelp. Specifically, we wanted to investigate the effects of business attributes on the business rating.

For this purpose, we used two methods of features selection to determine which features best predict business performance, as represented by star rating, and finally, we implement regression models (linear regression),and evaluate prediction models.

Methods and Data

DATA PROCESSING

The dataset we used is available on the Yelp Dataset Challenge homepage. We limited ourselves to businesses, which included 10964 business records, with 78 variables. We imported these json files into R arrays using mongodb for our analysis.

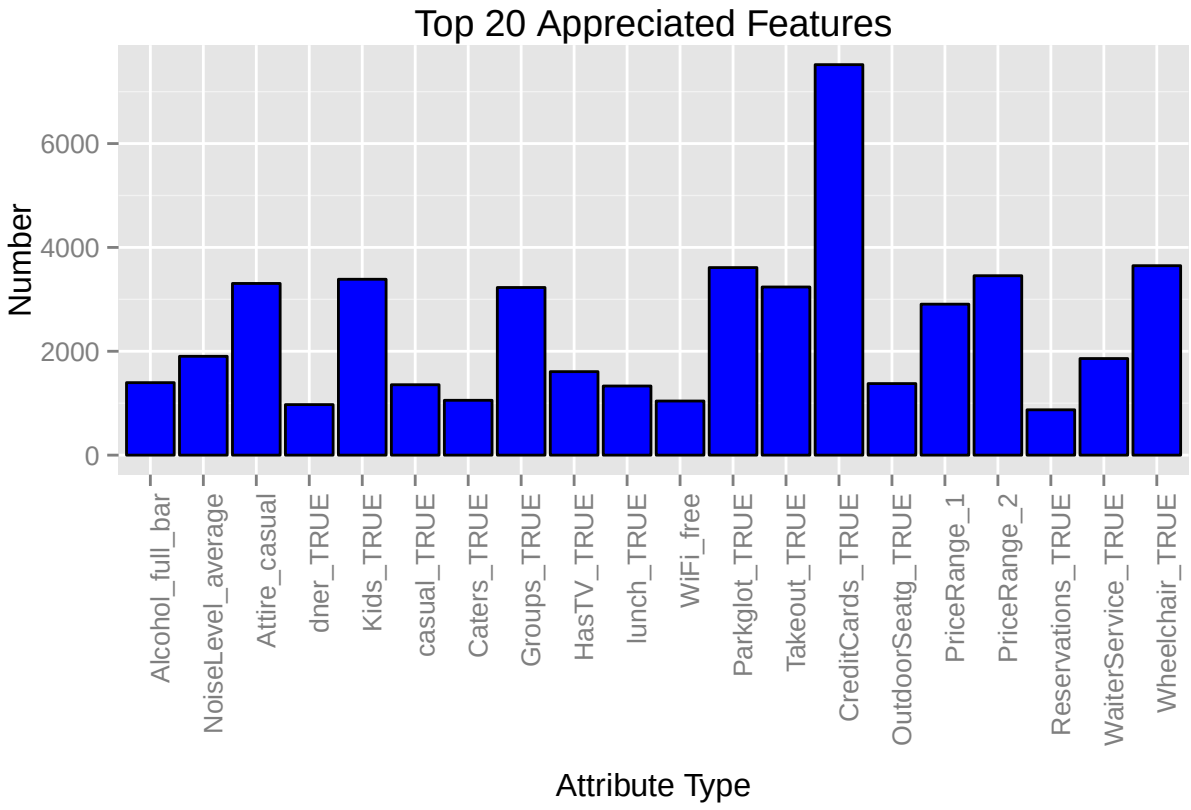
We limited ourselves to a subset of yelp businesses (subData)), which included 10964 records, with 78 variables. To handle missing data, we Change to "NR" (No respond), and Add another Factor Level "NR".In subData, stars is numeric, and the other 77 variables are factors.

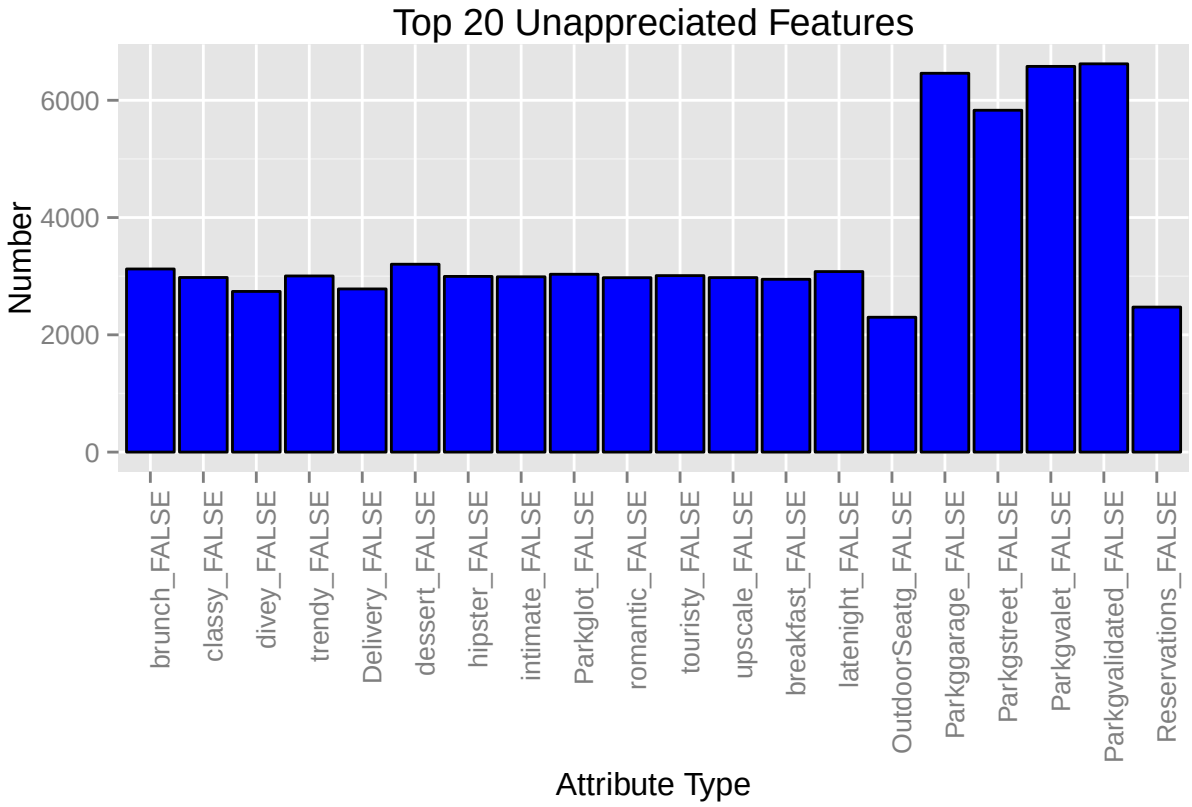
```
## Warning: package 'vcd' was built under R version 3.2.2
```

```
## Loading required package: grid
```

The descriptive statistics of numerical variable are shown in the table below:

SumRat	Median	Mean	Min_Star	Max_Star	stddev	cv	skewn	kurto
38681.5	3.5	3.53	1	5	0.88	0.25	-0.4	-0.11
Let us p	lot data	for the	20 feature	s that are	highly ap	preciat	ed and t	hose unappreciated for business**





With 7520 positive appréciations, businesses that accept CreditCards are the most appreciated, followed by Wheelchair and Parkglot.
By contrast, businesses that do not have Parkgvalidated , Parkgvalet, and Parkggarage , are not appreciated.

METHODS

After this summary study of features,to determine which features might be most useful in predicting business performance, we will select a limited number of features that will enable us to train our model. To this end, we explored two methods for future selections (Anova, Tree-Based), and compile selected features of both methods.

A. Anova Feature Selection

We used the Anova F-value as the scoring function for the feature set, then selected features with the highest score since the F-values for these features were significantly higher than the rest.

B. Tree-Based Feature Selection We used a tree-based estimator to compute feature importances, we choosed a coefficient of complexity cp= 0.0025371 which represent 6 nodes, and we used the default "gini" criterion to evaluate splits and selected those features with the highest importance score.

After compiling Anova and Tree-Based, there are 38 other predictor variables in the data set which may play bigger role to determination of business rating:

We can now split the cleaned training set into a pure training data set (70%) and a test data set (30%). We will use the test data set to conduct cross validation in future steps.

```
## Loading required package: lattice
```

The new training dataset contains 7676 observations while the testing data set contains 3288 observations.

Stars are numeric and continuous, we will use Linear Regression

- Y = a + b1 * x1 + b2 * x2 + b3 * x3 +.... bk * xk
- Y = Continuous output value (stars)
- x1,x2,... xk : Features of the input.
- a, b1, b2..bk : Weights of the features.
- Stars = F(BusinessAttributes)
- Stars = a + b1 * (wifi) + b2 * (drive_thru)... etc

Results

After performing an analysis of variance model in the training dataset, the p-value of Model is very small we found we removed 14 variables with a p-value below 5%:

```
## Warning in predict.lm(reg, testData): prediction from a rank-deficient fit
## may be misleading
```

```
## [1] 0.04710145
```

```
## Warning in predict.lm(reg, testData): prediction from a rank-deficient fit
## may be misleading
```

```
## [1] 0.07575758
```

The Peirces skill score (PSS) is 0.08, the global error rate (TG) is 0.05, the root mean squared error (RMSE) is 0.9494524 and this is our final model.

Hurting and Helpfull Attributes

Attributes	Hurting Features	Attributes	Helpfull Features
djTRUE	-0.36	intimateTRUE	0.42
PriceRange4	-0.31	Attiredressy	0.31
NoiseLevelvery_loud	-0.27	ParkgstreetTRUE	0.23
DriveThruTRUE	-0.21	CreditCardsFALSE	0.22
PriceRange3	-0.20	diveyTRUE	0.22
AppointmentFALSE	-0.16	TakeoutFALSE	0.18
CatersFALSE	-0.12	Alcoholbeer_and_wine	0.13

Discussion

Advices to improve your Business

Don't:

- Have price Range between 2 and 4.
- Accept only Credit Cards.
- Have a Drive-Thru.
- Paid WiFi.

- Take out.
- Have a noisy environment

Do:

- Take Appointments.
- Being well dressy and casual.
- Play smooth/ambient background music.
- have a parking in the street.
- Dog friendly
- offer a catering service