Fun with Yelp

In this document I am going to inspect a free data resource from Yelp in order to find interesting questions that can be further analyzed.

It should be noted that this document represents the very first draft. The model used in this analysis is just an initial run without further parameter or feature tuning.

Questions

The analysis consists of two questions that are related to both the fashion and the travel industry:

- **Question 1**: How good are the overall reviews of fashion and travel businesses compared to each other and to other businesses?
- **Question 2**: Is it possible to predict the star rating of a business by using the business attributes (Credit cards excepted, Parking spots provided,etc) and which attributes are critical for the rating?

Fashion and Travel - how good do they perform?

In a first step I analyzed the category column provided in the Yelp business dataset. To identify the relevant business categories I split the category strings into single words and created a frequency table (**Table 1**).

Most businesses are tagged as "services" which is a very unspecific identifier and can relate to a variety of businesses. **Fashion** and **Travel** are on 28 and 34 respectively. This position seems reasonable given the usual cityscape where restaurants, food stores and small shopping business are dominating.

Table 1: Yelp Category Count

word	n
services	59414
restaurants	54621
food	39256
shopping	28562
home	25877
spas	20287
bars	19289
beauty	18854
medical	16772
health	15414
hair	13205
event	12487
nightlife	12154
local	12118
repair	11385
automotive	11052
planning	11049
salons	11046
stores	10957
american	10578
auto	9416
life	8699
arts	8498
hotels	8419
active	8257
estate	7394
real	7222
fashion	7172
pet	6733
tea	6657
sandwiches	6345
fast	6280
traditional	6273
travel	6124
coffee	6080

Comparison of Yelp KPI's

To answer the first question I averaged some of the most important KPI's on Yelp for the relevant categories. In **Table 2** we can see that fashion and travel businesses have a lower star rating on average compared to all other categories. However, the fashion businesses are very close to the group average.

In terms of review frequencies the picture is different. The fashion businesses are clearly below average while the travel businesses are above. One possible explanation could be the closer relation of customers and employees in travel shops. However, this finding leaves room for further analysis.

The average starting date on Yelp compared over the three categories shows no significant differences. However, the average Check-Ins are strongly correlated to the number of reviews. Travel business have again on average more check-Ins compared to fashion and the group average. One reason for that might be the close relation between a Check-In and a review.

Table 2: This is the table caption

business	average_stars	average_review_count	average_start_date	average_checkins	count
fashion	3.541181	12.85861	2012-01-03	1.721251	7030
other	3.649344	30.31507	2012-10-01	1.973846	161811
travel	3.259343	46.31994	2011-12-21	2.751275	5726

Predicting Star Ratings

For the second question my first task was to look at the Yelp attribute data to check whether those features can be used to predict the star rating of a business. It turned out that most of those attributes were null for most of the businesses. However, some of the attributes had a null ratio of 90% and better (up to 75%). In the end I used 15 attributes which had at least 1000 none null entries. As an additional feature I used the average Check-Ins.

I applied an XGBoost model in combination with the well-known R package 'caret'. The decision to use XGBoost was mainly driven by the execution speed but also by the good applicability to almost all machine learning tasks with very good results.

Table 3: XGBoost Model Results

RMSE	Rsquared	MAE	RMSESD	RsquaredSD
0.9790353	0.0466839	0.7878964	0.0022475	0.0014105

The results show that the initial model can already predict star ratings with an MAE of 0.78 stars. This is of course far from being an acceptable outcome for the business. However, given the lag of additional reliable features and without any parameter tuning, this first result is at least a good start.

Variable Importance

Table 4: Variable Importance of Star Rating Prediction

	Overall
avg_checkins	100.0000000
DogsAllowed	98.0250321
BikeParking	54.4008090
BusinessAcceptsCreditCards	52.3687039
BusinessParking_garage	38.6787127
BusinessParking_valet	16.8258366
GoodForKids	13.1163178
BusinessParking_validated	11.9803192
Alcohol	3.8825194
BusinessParking_street	3.7538583
HappyHour	2.5216760
WheelchairAccessible	1.6508721
GoodForMeal_dinner	1.3579977
Caters	1.1946592
BestNights_monday	0.9155425
GoodForMeal_breakfast	0.0000000

The second part of the question was which attributes are relevant for a star rating. In **Table 4** you can see the most important business attributes in descending order.

The most important feature is the average Check-In which is actually not a business attribute. Check-Ins probably occur more likely when the business also has a good star rating. However, the actual causal relationship between Check-Ins and the star rating leaves again room for further analysis.

The most important attributes are:

- whether or not dogs are allowed
- bike parking is possible
- credit card acceptance
- the possibility to park the car

It can be doubted that the allowance for dogs in a fashion or travel shop would have significant positive influence on the star rating. However, credit card usage and parking possibilities are most likely not counter-productive in this context. But again, this is only a first draft and more analysis needs to be done to answer those questions properly.