"Spotting Optimal Location For a New Business"

Kapil Kumar (CS13M025) and G K Sudharshan (CS13M050)

April 30, 2014

Abstract

There are lots of start up businesses who are confused among several locations to commence their business. Traditionally, such problems are solved using the demographics, revenue earned and human flow statistics of all similar businessmen in given region. This approach is very slow and doesn't consider the consumer's perspective. With the emergence of social networking services like check-ins, we can additionally add user's flavor in prediction of optimal geospatial location for a business. We propose to build a machine learning platform using the location based social networking data to identify the best location among different locations suggested.*Firstly we have decided to focus on finding optimal location for the restaurants.*

Part I

Dataset Used:

Dataset selection:

We have experimented with following two datasets for the project:

- **Foursquare Dataset**
- **Yelp Dataset**

It is difficult to get the desired data, as data we intend to use involves private information of the user. Finally we decided to use a data set available from FourSquare and Yelp and try to extract all possible features from the data set. In future we are planning to use Twitter's streaming API to extract publicly available check-ins of people at various restaurants.

Filtering the Foursquare Dataset:

The dataset contains the information of the check-ins of a particular user at a particular time at a particular venue all over the world. We decided to use the data from New York as it has high number of venues on foursquare. We filtered the user data and the venues in area of $100km^2$ with Geo-Coordinates of New York as centre. There are around 9300 venues, the *data is unlabeled so we* considered Top 1% of venues based on check-ins as a *common places* on the assumption that the traffic on *common places* like airports, railway stations is constant throughout the day and we verified the same in Google Maps, whereas the other places like restaurants will not have such a high volume of check-ins.

Filtering the Yelp Dataset:

The dataset available from the Yelp contains data from Phoenix, Arizona, United States. The dataset was available in JSON format. We converted the data to desired form such that we can extract the maximum number of features as mentioned in Part II of the document. The data set has around 15000 venues out of which 8600 are restaurants. We also have check-in information for 11000 venues.

Part II
Defining Features

We measure the features of the surrounding area by analyzing the set $\{ p \in P : dist(p, l) < r \}$ of places that lie in a disk of *radius r* around *desired location l*. The function *dist* denotes the geographic distance between two places and P is the set of venues in New York or Phoenix depending upon the dataset.

Input features for the Supervised Learning Model:

1 Density: It measures the number of neighbors (venues) around the place, we assess to what extent the popularity of a place depends on the number of venues in the same area.

$$density = |\{\text{p}\epsilon\text{P} : dist(p, l) < \text{r}\}|$$

-

2 Crowdedness: This measures the number check-ins around the venue. The intuition behind this is that more the number of people around a venue more is the chances of getting customers. $c(p)$ represents the number of check-ins at venue p.

$$Area\ Popularity = \sum_{dist(p,l)<r} c(p)$$

3 Competitiveness: The amount of competition is measured in terms of ratings in the surrounding radius of the venue. Competition in the context of restaurants can have either a positive or a negative effect. One would expect that, for instance, placing a bar in an area populated by nightlife spots would be rewarded as there is already an ecosystem of related services and a crowd of people being attracted to that area. However, being surrounded by competitors may also mean that customers will be shared across multiple venues.It considers $R(p)$ as number of check-ins at venue p.

3.1 Average Rating: It is the measure of the average of the ratings given to the surrounding venues by the user.

$$Average\ rating = \frac{\sum_{dist(p,l)<r} R(p)}{\sum_{dist(p,l)<r} 1}$$

3.2 Count competetive venues: It is the measure of number of venues with ratings greater than 4.5. This metric tries to capture, if a venue is surrounded by good venues how it affects the ratings of the venue.

$$Count\ Competetive\ venues = |\{\text{p}\epsilon\text{P} : dist(p,l) < \text{r}\&R(p) > 4.0\}|$$

3.3 Count weak surrounding venue: It is the measure of venues with ratings less than 3. Similar to above metric, this tries to explore that if a venue is surrounded by bad venues how it effects the ratings of the venue.

$$Count\ Weak\ venues = |\{\text{p}\epsilon\text{P} : dist(p,l) < \text{r}\&R(p) < 2.5\}|$$

4 Locality Influence: It captures influence on the venues by the presence of common places like Railway Stations, Airports etc. around the venue. Here C corresponds to the set of all common venues considered.

4.1 Distance: Average distance from the prominent or common places within certain radius. It is observed that closer the common place more is the influence on number of check-ins in the venues.

$$Average\ distance = \frac{\sum_{dist(p,l)<r\ \&\ c\epsilon C} dist(p,l)}{\sum_{dist(p,l)<r\ \&\ c\epsilon C} 1}$$

4.2 Weighted distance: Weighted average of distance from prominent or common places within certain radius. In this case we weight the average distance based on the check-ins in the individual common place, it tries to capture the phenomenon that closer the venue to the common place with most check-ins more is the influence.

$$Average\ distance = \frac{\sum_{dist(p,l)<r\ \&\ c\epsilon C} dist(p,l).c(l)}{\sum_{dist(p,l)<r\ \&\ c\epsilon C} c(l)}$$

4.3 Count common places: Number of common places in certain radius.

$$Count\ common\ venues = \sum_{dist(p,l)<r\ \&\ c\epsilon C} 1$$

5 User transitions: It tries to track the users consecutive transitions, this assumes that increased mobility between places in the area can increase the number of random visitors towards the target place, we measure the density of transitions between the venues inside the area.

5.1 Transition density : Number of user transitions inside region by overall transition. Transition is the movement of user from one venue to another venue.

$$Transition\ Density = |\{m,n\epsilon T : dist(m,l) < \text{r}\ \&\ dist(n,l) < r\}|$$

5.2 Incoming transitions : We also define a feature to account for the in-coming flow of external user traffic towards the area of the place in question. *How many people are visiting the place from far away places?* We define it as the number of user transitions from outside area by overall transition within region.

$$Transition\ Density = |\{m,n\epsilon T : dist(m,l) > \text{r}\ \&\ dist(n,l) < r\}|$$

Outputs for the Supervised Learning Model:

Following are the scores that we are using for prediction:

1 Ratings:

Ratings are the average ratings that will be given by the users visiting. The value of ratings lies between 1 to 5. Eventually this particular output parameter was not considered as there were lots of noise on this particular parameter. For e.g we tried to correlate it with number of checkins , we observed that for some sample of data of sufficient size, the rating is positvely correlated to the number of check-ins but for other random sample of equivalent size it is negatively correlated.

2 Number of Check-ins :

Counts the number of check-ins that will happen at a particular venue during a specific period of time. We found that the check-ins are more robust score compared to the other parameters.

Part III Extracting Features

We extracted the features mentioned in Part II of the document. We used *the radius of 1 km* around the venue under consideration to measure the various features. In order to check the usability of features for the supervised model we considered *Pearson Correlation Coefficient,* which is covariance of two variables divided by the product of their standard deviations.

Table 1 shows the correlation coefficient of *Attraction Coefficient for Foursquare* with different features considered for top venues based on number of check-ins. We observe that there is a drastic drop when considered top 2000 venues, the reason for this is may be the absence of the venue information. In this case we are considering different categories of venues like Restaurants and Gyms together while analyzing.

In case of Yelp dataset as show in Table 2, we observe that there is not much difference in correlation for Top 200 and Top 2000 venues. This is because we have label for venues i.e we know venues that are restaurants.

Table 1: Correlation coefficient of Attraction Coefficient for Foursquare with different features for top venues based on number of check-ins

Correlation with Attraction Coefficient	Crowdedness	Number of Neighboring venues	Average Inverse Distance	Number of Common Venues	Average Ratings	Number of Competitive venues	Number of Weak Competitive Venues
Top 200	0.2938	0.4379	0.4727	0.3147	0.4361	0.4092	0.4160
Top 2000	0.0891	0.0875	0.0846	0.0122	0.0915	0.0839	0.0717

Table 2: Correlation coefficient of number of Check-ins for Yelp for top venues based on Attraction Coefficient

Correlation with Number of check-ins	Crowdedness	Number of Neighboring venues	Number of Common Venues	Average Ratings	Number of Competitive venues	Number of Weak Competitive Venues
Top 200	0.7287	0.6802	0.3676	0.2659	0.6156	0.4025
Top 2000	0.5717	0.4970	0.2878	0.2389	0.4620	0.2316

All above features extracted are checked for the correlation with the different scores, but we observed Yelp dataset to perform well for number of check-ins, while Foursquare dataset perform better for the Attraction Coefficient.

Part IV Evaluation

We compared the actual score and predicted score using correlation methods. We tested our model for following correlation methods

1. Pearson's correlation coefficient (PCC): $\frac{cov(X,Y)}{\sqrt{var(X)var(Y))}}$
2. Spearman's rank correlation coefficient (SRCC): A nonparametric (distribution-free) rank statistic proposed by Spearman in 1904 as a measure of the strength of the associations between two variables (Lehmann and D'Abrera 1998). The Spearman rank correlation coefficient can be used to give an R-estimate, and is a measure of monotone association that is used when the distribution of the data make Pearson's correlation coefficient undesirable or misleading. The value is calculated as P(rank(X),rank(Y)).
3. Kendall tau rank correlation coefficient (KRCC) :Let (x1, y1), (x2, y2), . . . , (xn, yn) be a set of observations of the joint random variables X and Y respectively,such that all the values of (xi) and (yi) are unique. Any pair of observations (xi, yi) and (xj, yj) are said to be concordant if the ranks for both elements agree: that is, if both xi $>$ xj and yi $>$ yj or if both xi $<$ xj and yi $<$ yj. They are said to be discordant, if xi $>$ xj and yi $<$ yj or if xi $<$ xj and yi $>$ yj. If xi $=$ xj or yi $=$ yj, the pair is neither concordant nor discordant.The value is calculated as

$$\frac{number of concordant pairs - number of discordant pairs}{all - possible pairs}$$

Random cross-validation with 100 random validation set is considered and reports average values across experiment.

Part V Results

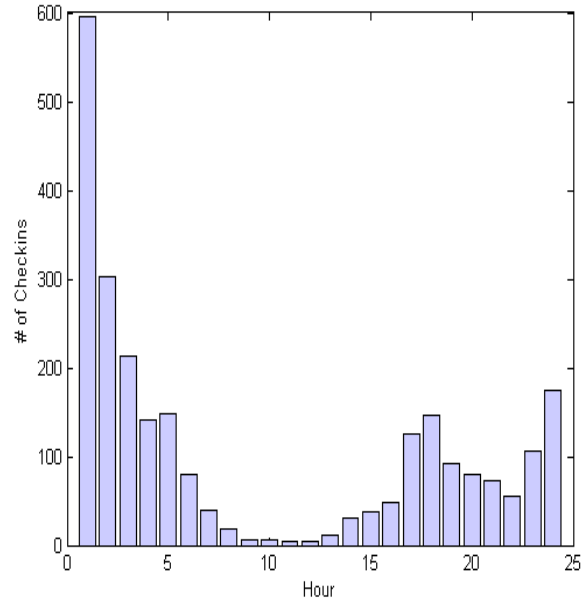
Foursquare Dataset

In case of foursquare dataset, we observed that the final results obtained were not so credible, we obtained Pearson correlation coefficient of 0.42 for top 200 venues. We believed that the bad results were because we assumed that all venues considered are of similar type. Hence we pruned the dataset to contain only restaurants, with following assumption, venues that are more

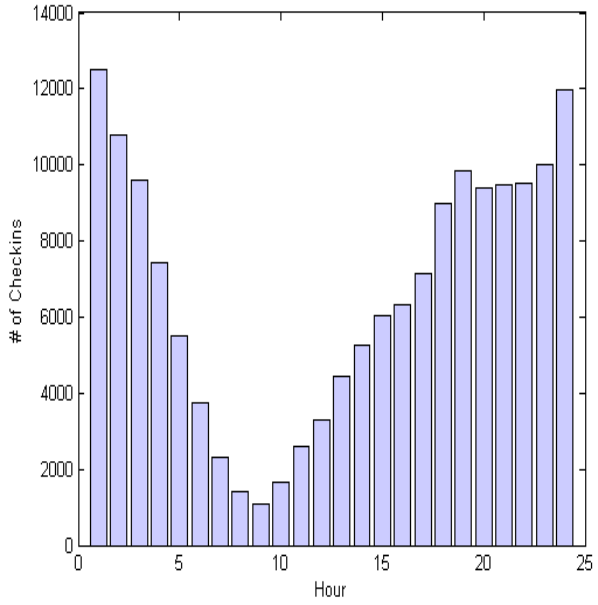
frequently visited during peak lunch and dining hours are restaurants. So we clustered Top 200 venues based on check-ins at hour of day using a simple K means clustering with 2 cluster and distance measure as correlation between features vectors (that compares the checkins of any venue at any hour of day). Then we eliminated cluster based on the average checkins at hour of day in each cluster.

Figure 1: This figure shows average checkins at hour of a day at each clusters. We considered right (as more frequent check-ins are observed) cluster for our regression model and it yields a 10% improvement in correlation.

(a) Cluster 1



(b) Cluster 2

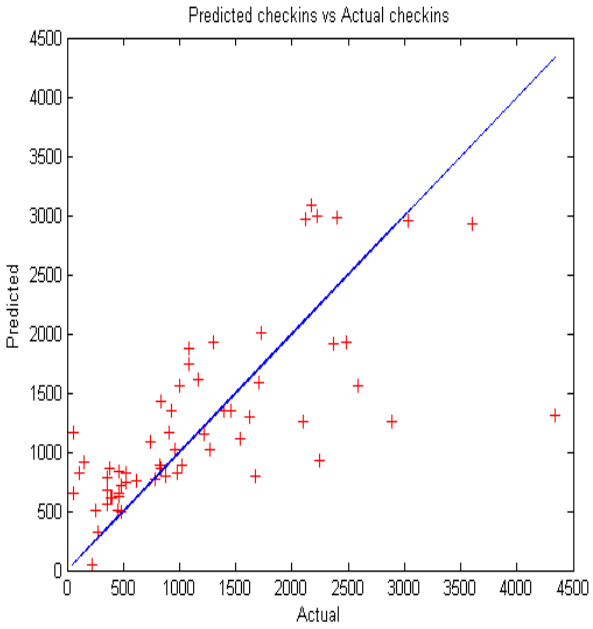


Yelp Dataset

Below figures shows the results of the *Linear regression model* built using the above features. The Figure 2 shows scatter of predicted and actual scores, the diagonal line represents the ideal scatter that is expected . The Table 3 shows the performance of the simple regression model that was built using extracted features on complete data using different regression models.

Figure 2: Scatter plot for predicted check-ins vs actual check-ins for Yelp dataset

(a) Building for Top 500 Venues for Yelp dataset



(b) Building for Top 2000 Venues for the Yelp dataset

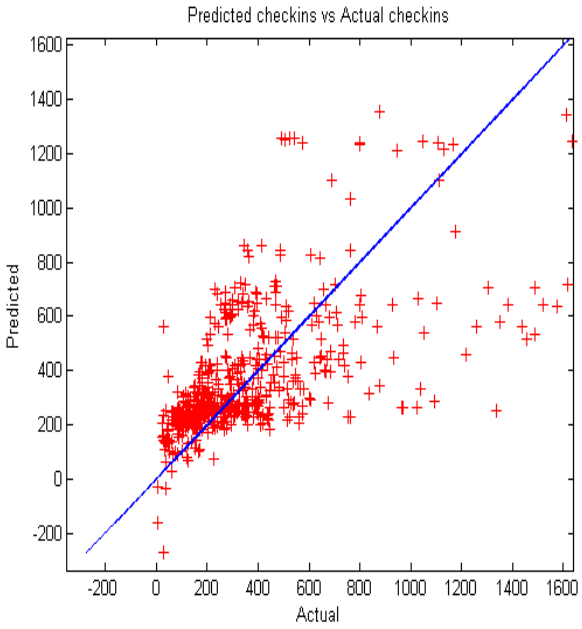


Table 3: Correlation Coefficient of Predicted and Actual Scores using multiple regression model

	Linear Regression	SVR Gaussian Kernel	Regression Tree
PCC	0.5218	0.4700	0.4676
SRCC	0.6512	0.6155	0.5438
KRCC	0.4710	0.6051	0.3854

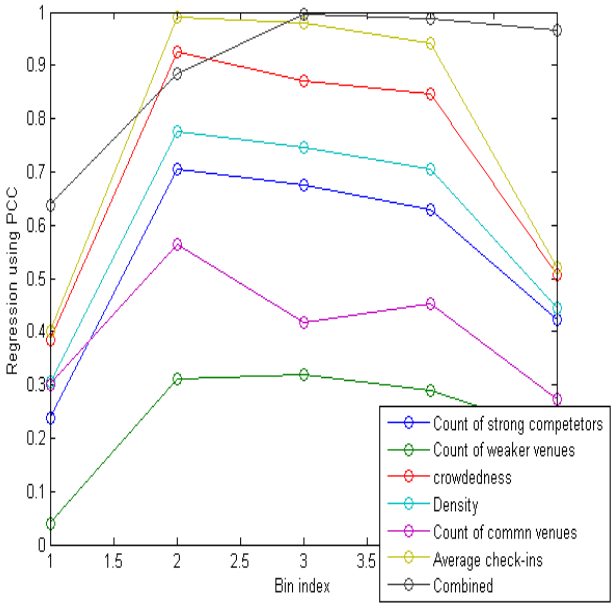
Part VI Binning:

We observed the model perform not so well for complete 8000 venues but performs considerably well on top 2000 venues and 2000 venues. So we sorted the the venues based on number of check-ins and divided these data sets into equifrequent bins and built different models for each bins.

The correlation of actual and predicted venues is above 90 for nine bins. The one of the bins for which the performance was low, was the one which had least attraction coefficients.

So the experiment was quite successful *assuming that the venue for which we are testing has went to the correct bin. Now the problem reduces to allocating the correct bin to the venue.*

Figure 3: Regression performance after binning the entire dataset into multiple equi-frequent bins



Part VII Defining the Bins On basis of Normalized checkins:

After seeing good results on bining we tried to formalize the notion of bining. For this we used new parameter normalized checkin.

Normalized check-in at venue v is defined as follows

$$N(v) = \frac{NumberofCheck-insAtPlace}{AverageNumberofCheck-insInArea}$$

The reason choosing the normalized check-ins was that it gives a intution whether the venue is performing good are bad. For example if the Normalized check-in is sufficiently large then the venue is in some-sense performing better than the average while if it too low then the venue is not able to attract the users to the restaurants.So on the basis of how well a venue performs in its neighborhood we divided venues into four categories based on normalized checkin values:

- Less than 0.5.
- 0.5 to 1.
- 1 to 5.
- Above 5.

Table 4: Regression accuracy on binning as per normalized check-ins and after classfication using nearest neighbor

Bin Model	No Classifier	Nearest Neighbor (5)
N(v)<0.5	**PCC**: 0.6987 **SRCC** : 0.5552 **KRCC** : 0.4629	**PCC**: 0.6354 **SRCC** : 0.5542 **KRCC** : 0.4091
0.5<=N(v)<1	**PCC**: 0.9958 **SRCC** : 0.9531 **KRCC** : 0.9258	**PCC**: 0.9573 **SRCC** : 0.9216 **KRCC** : 0.7667
1<=N(v)<5	**PCC**: 0.9686 **SRCC** : 0.9535 **KRCC** : 0.8895	**PCC** : 0.8526 **SRCC**: 0.8308 **KRCC** : 0.7674
N(v)>=5	**PCC**: 0. 9686 **SRCC** : 0.9564 **KRCC** : 0.8505	**PCC** : 0.8081 **SRCC** : 0.7544 **KRCC** : 0.7497

At this point we know that if we can allocate bin to a particular test venue then we can predict the number of checkins for that venue with a high accuracy. We used nearest neigbor approach for classifying the to which particular bin would the given test example belong based on voting scheme with respect to nearest neighbors. The below table depicts the result of regression model after applying nearest neigbor classification with 5 nearest neighborhood.

The Nearest neighbour approach that we considered performs a better result than Support vector machine. While experimentation we observed for sufficient large neighbors the performance of model is not so good. The result we obtained is not perfect as the accuracy of classication was found out to be around 69% .

Part VIII Conclusion and Road Ahead:

The results above show that there are definitely some signals in the dataset that can be used for the task of identifying a best location for the restaurant.We observed the accuracy to be above 90% if we can classify venue to correct bin. We observe that best results were shown on Top 50% of data when sorted in increasing order compared to bottom 50%. All the features considered affects the prediction of number of chekins, i.e popularity of a place can be clearly explained using the spatial features.We plan to build different model for the different category of venues which would boost the models performance.We may also explore how we can learn different models on the basis of the neighbourhood of the venue.We would also use the Information of the types of restaurants like whether a Italian restaurant will be more or less influenced by neighboring Italian Restaurant. We may also look for public check-in information available from Twitter to get more venues and better results. We would like to study the impact of our features on various retail store chain like Pizza Hut and get some insight from it.

References

[1] *'Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement',Dmytro Karamshuk,IMT Lucca, Italy*

[2] *Foursquare Dataset "https://archive.org/details/201309_foursquare_dataset_umn"*

[3] Yelp Dataset *"https://www.yelp.com/dataset_challenge/dataset"*