RECOMMENDATIONS MATTER TO US!

Li Xiang, Piyush Pritam Sahoo, Rhea Chandra, Malvania Smeet Saunil

TEAM ACCURO

Contents

1. Project Overview

1.1 Introduction and Background

We live today, in what could be best described as the age of consumerism, where, what the consumer increasingly looks for, is information to distinguish between products. With this rising need for expert opinion and recommendations, crowd-sourced review sites have brought forth one of the most disruptive business forces of modern age. Since Yelp was launched in 2005, it has been helping customers stay away from bad decisions while steering towards good experiences via a 5-star rating scale and written text reviews. With its vast database of reviews, ratings and general information, Yelp not only makes decision making for its millions of users much easier but also makes its reviewed businesses more profitable by increasing store visits and site traffic.

The Yelp Dataset Challenge provides data on ratings for several businesses across 4 countries and 10 cities to give students an opportunity to explore and apply analytics techniques to design a model that improves the pace and efficiency of Yelp's recommendation systems. Using the dataset provided for existing businesses, we aim to identify the main attributes of a business that make it a high performer (highly rated) on Yelp. Since restaurants form a large chunk of the businesses reviewed on Yelp, we decided to build a model specifically to advice new restaurateurs on how to become their customers' favourite food destination.

With Yelp's increasing popularity in the United States, businesses are starting to care more and more about their ratings as "an extra half star rating causes restaurants to sell out 19 percentage points more frequently". This profound effect of Yelp ratings on the success of a business makes our analysis even more crucial and relevant for new restaurant owners. Why do some businesses rank higher than others? Do customers give ratings purely based on food quality, does ambience triumph over service or do geographic locations of businesses affect the rating pattern of customers? Or is the old adage "location, location, location" indeed an important factor for the success of a business on Yelp? Through our project we hope to analyse such questions and thereby be able to advice restaurant owners on what factors to look out for.

1.2 Review of Similar Work

1) Visualizing Yelp Ratings: Interactive Analysis and Comparison of Businesses:

The aim of the study is to aid businesses to compare performances (Yelp ratings) with other similar businesses based on location, category, and other relevant attributes.

The visualization focuses on three main parts:

a) Distribution of ratings: A bar chart showing the frequency of each star rating (1 through 5) for a single business.

b) Number of useful votes vs. star rating A scatter plot showing every review for a given business, with the x-position representing the "useful" votes received and y-position representing the for the business.
c) Ratings over time: This chart was the same as Chart 2, but with the date of the review on the x-axis

The final product is designed as an interactive display, allowing users to select a business of interest and indicate the radius in miles to filter the businesses for comparison. We will use this as a base and help expand on some of its shortcomings in terms of usability and UI. We will further supplement this with analysis of our own using other statistical methods to help derive meaning from the dataset.

2) <u>Your Neighbors Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction</u>

This study focuses on the influence of geographical location on user ratings of a business assuming that a user's rating is determined by both the intrinsic characteristics of the business as well as the extrinsic characteristics of its geographical neighbors.

The authors use two kinds of latent factors to model a business: one for its intrinsic characteristics and the other for its extrinsic characteristics (which encodes the neighborhood influence of this business to its geographical neighbors).

The study shows that by incorporating geographical neighborhood influences, much lower prediction error is achieved than the state-of-the-art models including Biased MF, SVD++, and Social MF. The prediction error is further reduced by incorporating influences from business category and review content.

We can look to extend our analysis by looking at geographical neighbourhood as an additional factor (that is not mentioned in the dataset) to reduce the variance observed in the data and improve the predictive power of the model.

3) Spatial and Social Frictions in the City: Evidence from Yelp

This paper highlights the effect of spatial and social frictions on consumer choices within New York City. Evidence from the paper suggests that factors such as travel time, difference in demographic features etc. tend to influence consumer choice when deciding what restaurant to go to.

"Everything is related to everything else, but near things are more related than distant things" (Tobler 1970).

1.3 Motivation for the Project

Our personal interest in the topic has motivated us to choose this as our area of research. When planning trips abroad, we explore sites like HostelWorld and TripAdvisor that make planning trips a lot faster and easier; not only is this helpful to customers planning trips but also to the businesses that have been given honest ratings. Since the team consisted students from a Management university, our motivation when choosing this project was more business focused. Our perspective on recommendations was more catered towards how a business can improve its standing on Yelp, and thereby improve its turnover through more visits by customers.

We believe that our topic of analysis is crucial for the following reasons:

1) It will make the redirection of customers to high quality restaurants much easier and more efficient.
2) It can encourage low quality restaurants to improve in response to insights about customer demand.
3) The rapid proliferation of users trusting online review sites and and incorporating them in their everyday lives makes this an important avenue for future research.
4) Prospective restaurant openers (or restaurant chain extenders) can intelligently decide the location based on the proximity factor to other restaurants around them.

1.4 Project Scope and Methodology

"How to dominate the restaurant scene in a city?"

Primary requirements (for "restaurants" and one city only):

Step 1: Descriptive Analysis - Analysing Restaurants specifically for what differentiates High performers, low performers and Hit or Miss restaurants. The analysis will further be segmented into for example region, review count, operating hours, etc. For each of the 3 segments mentioned, the following analysis will be done:

A. Clustering to analyse business profiles that characterize the market. Explore various algorithms and evaluate each of the algorithms to decide which works best for the dataset.
B. Time series analysis of whether any major trends have emerged in restaurants by region – further decipher the does and don'ts for success

Step 2: Key factors identification for prescriptive analysis (feature extraction) for new restaurants by region, in order to succeed. Regression will be used to identify the most important factors and the model will be validated so that we can analyse how good the model is. This will constitute the explanatory regression exercise.

As an extension, we will also attempt to predict the rating for new restaurants, thereby informing existing restaurants of potential competition from new openings.

Step 3: Spatial Lag regression model. This section will focus on Geospatial Analysis to examine the effect of location of a business on its rating. The goal of this will be to modify the regression model in Step 2 by adding the geospatial components as additional variables to the model. This section will explore the three spatial regression models and use the model that best fits the dataset:

- Checking for Spatial Autocorrelation: Spatial dependencies existence will be checked using Moran's I (or any other spatial autocorrelation index) to see if they are significant.

- Spatial lag model (for regression) will be used if the dependent variable, the business rating, is spatially auto correlated i.e. the ratings of businesses in one location are correlated to the ratings in nearby locations. Spatial proximity will be defined using an n x n matrix and various weight matrices (to test validity) will be used in the estimation of spatial regression.
- Spatial cross-regressive model will be used in place of Spatial lag model if the independent variables (or business attributes) in the regression are spatially auto correlated.
- Spatial error model will be used if the residuals of the OLS regression are spatially auto-correlated.
- Results from the exercise will be interpreted to recommend salient features of regions to describe to businesses typical characteristics of similarly rated, close-by restaurants.

Step 4: Build a visualization tool for client for continual updates on business strategy. Focus will be to build a robust tool that helps the client actively visualize all insights developed during the project.

Secondary requirements:

Expand and recreate the analysis for all other cities.

This analysis will be recreated to include other kinds of businesses eg. Bars, Salons, etc. For some businesses, new methods of analysis such as latent factorization will be employed (especially for those with minimal information on attributes)

Spatial Lag Regression model will be replicated to other business categories.

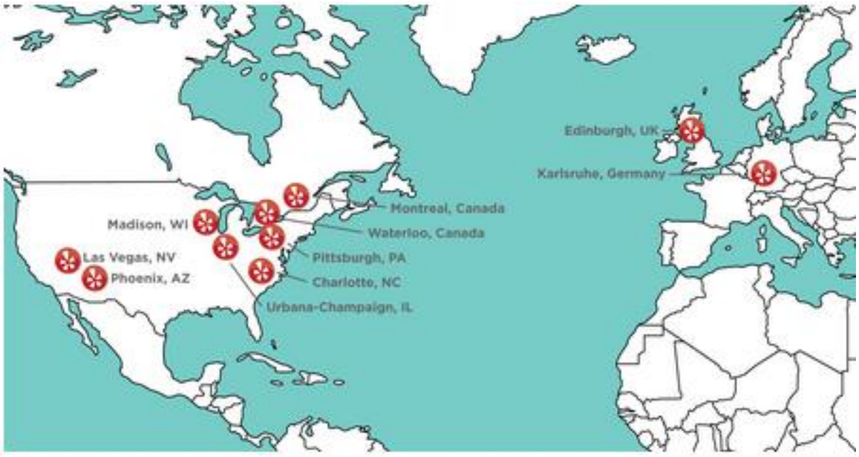
Future research:

Evaluating the importance of review ratings for restaurants – Are they effective to improve ratings? Do restaurants that utilize recommended changes succeed?

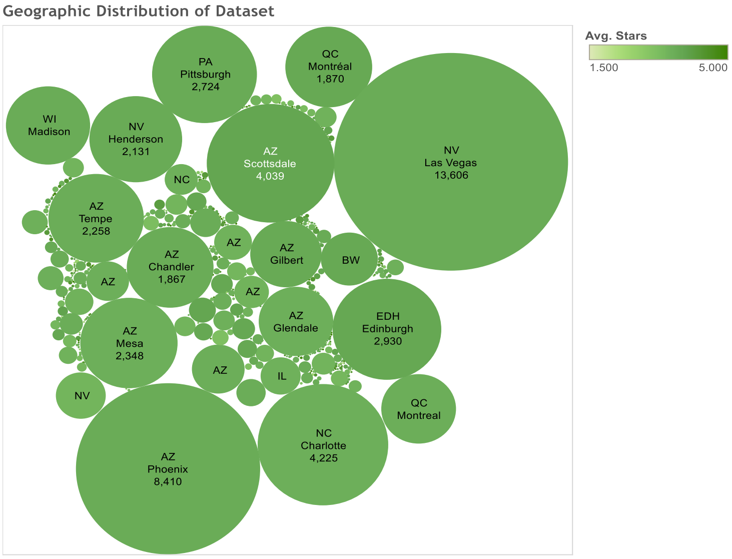
Can the ratings and reviews of local experts be assimilated in feature extraction to help improve the predictability of ratings success? We realize that people are social entities and can be heavily influenced by reviews from local experts in their criticism on Yelp. Future research in this area can enrich our analysis for a business as well.

1.5 Preliminary Findings and Methodology Implications

The data is spread over 10 cities with the highest number of data from Las Vegas and Phoenix. 45% of the businesses received ratings of 3.5 or 4. 28% received a score of 4.5 or 5. The weekly opening hours of businesses with 3.5, 4, or 4 stars is higher than the others. For Restaurants, Bars, Hotels, and Fast Food Restaurants, the majority number of businesses receive an average of 3.5 or 4 stars. For Hair Salons, Travel Services and Gymnasiums, majority of the businesses received higher rating of 4.5 or 5. For Shopping, Banks, Books and Music, majority of the businesses received 3.5 or 4 stars. These findings suggest that the deviation we may observe for predictions may not be very massive, and the results could prove to not be as insightful and "neat" as imagined as this stage.



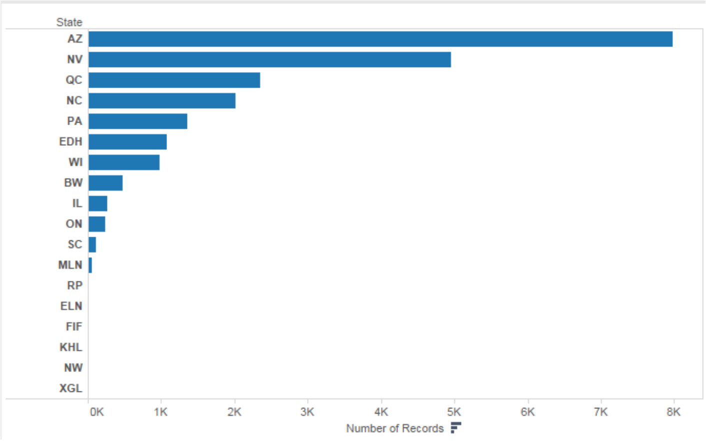
Upon carefully analysing the dataset, we found the following:



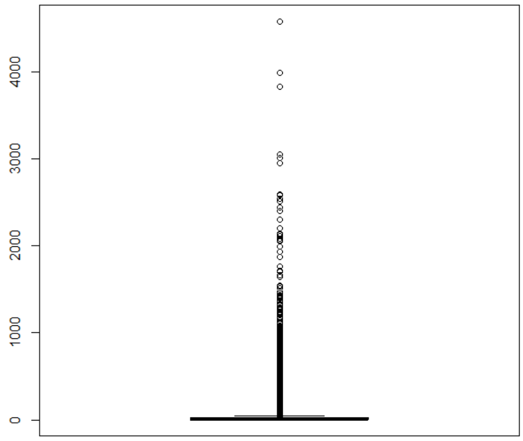
State, City and sum of Number of Records. Color shows average of Stars. Size shows sum of Number of Records. The marks are labeled by State, City and sum of Number of Records.

This suggested that the total number of data points in the business dataset were small, and even smaller if we broke it down by region.

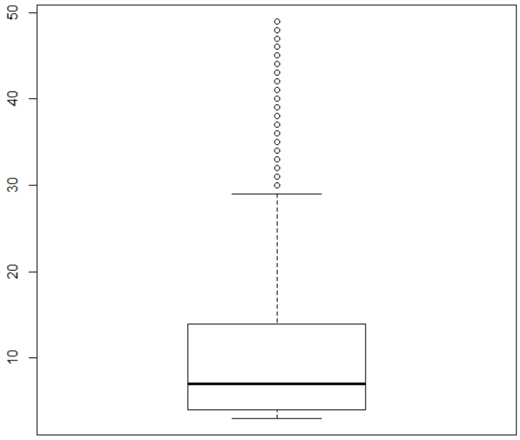
Breaking the analysis down further by restaurant (as one of the business categories), we found the following:



Given that the size of the data set per region is small, and the total number of regions given are also limited, our analysis has to be on a very micro level. The methodology hence must be robust and scalable.



A simple box-plot of the review count reveals a drastic distribution in the dataset for business reviews. In order to keep our analysis relevant and credible, we will hence seek to moderate the business dataset according to places with higher review counts.

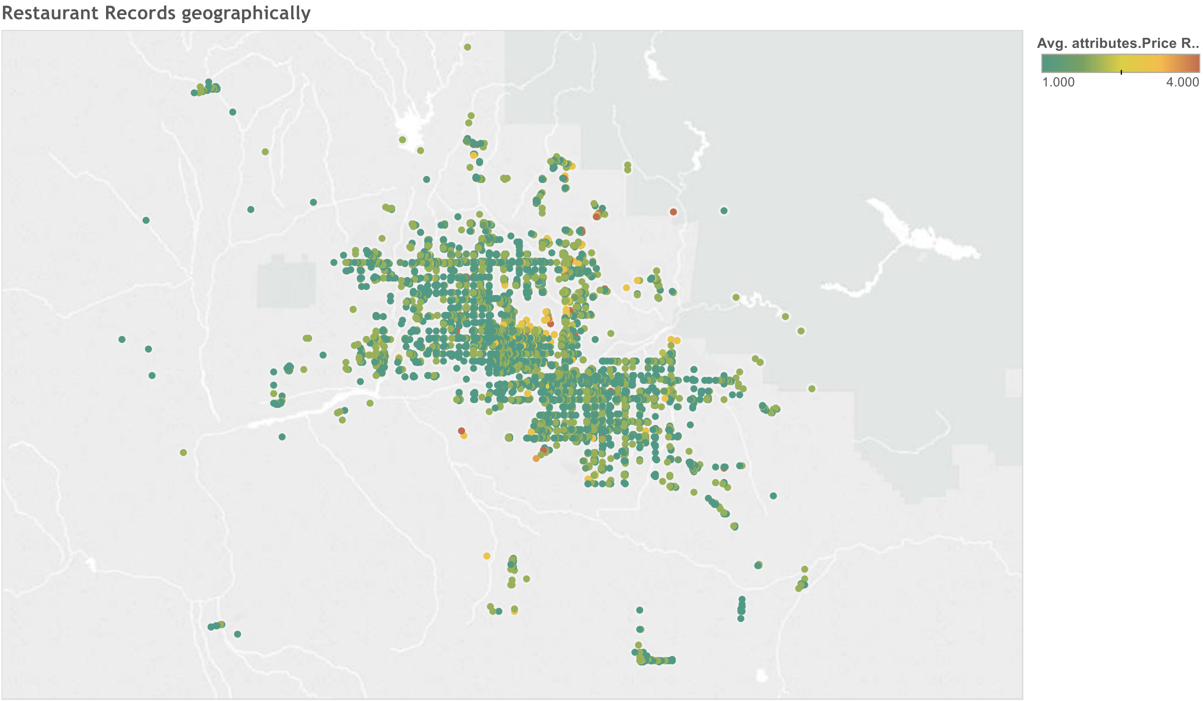


Filtering review count by less than 50, we saw that the distribution was even more concentrated towards a lower review count. The ratings for high performers must hence be moderated in some way (by either splitting the dataset or weighing by review count) as per the review count for a location. The same logic would follow for Low Performers and the Hit or Miss category.

What will also be interesting for businesses to learn is how the Hit or Miss restaurants behave. These are typically restaurants with a huge deviation in ratings, indicating a great divide in preferences among users. It would be interesting to see the characteristics of such a restaurants since businesses could then look to augment its service offering depending on the segment that likes it.

Analysis of other such variables has led us to conclude that variables such as opening hours and ratings for business categories are vastly different, which would support our analysis methodology of dividing the dataset into regions and categories before analysis.

Significantly large missing values for attribute data could be good and bad. Good in a sense that classification of business for clustering or classification will become easy. Bad in that the reliability of data may be circumspect, since you would typically see multiple attributes for a business especially in Yelp, which would give the business better results of being recommended to a user.



Map based on Longitude and Latitude. Color shows average of attributes.Price Range. The data is filtered on Categories, which has multiple members selected.

The above image shows the spatial distribution of restaurants in Phoenix, Arizona. The color intensity from green to red indicates increasing price range. As can be seen and expected, more expensive restaurants tend to aggregate towards the centre of the city. This might indicate that certain network effects as a result of location may translate into an effect on Yelp ratings. As a result of this simple exercise, we believe we should be able to include the effect of spatial lag in our regression model, thereby improving the credibility of the results.

1.6 Limitations and Assumptions

In doing our analysis, we have overall concluded below some of the major limitations we can foresee from this project:

Limitations	**Assumptions**
Limited data points on businesses and cities	Project methodology will be scalable for looking at regional trends
Limited action-ability of insights since companies may not care about Yelp ratings	Project findings will help set priorities for improvement for business owners
Businesses attribute may not be completely accurate	Assuming that data has been updated as accurately as possible
Defining business categories	Assuming business tags under categories are comprehensive for the competitive set

Future projects can further seek to mitigate some of these by adopting larger datasets and actually partnering with a real business to assess the impact of the recommendations in terms of a profitability analysis to recommend the best solutions.

1.7 Risks and Mitigation

Risk Assessment Metric:

	Likelihood			
Impact		Low	Medium	High
	Low	C	C	B
	Medium	C	B	A
	High	B	A	A

Risks	**Level**	**Mitigation**
Insufficient statistical knowledge	B	Consult with supervisor and online course materials
Lack of actionable business insights	A	Continuous literature search on meaningfulness of insights for businesses according to each city
Dashboard UI design may not be intuitive or extensive	A	User testing and consistent updates with the supervisor

2. Project Execution

2.1 Work Scope

Through this project we are hoping to build to an interactive dashboard as a solution to the ratings and recommendations system Dataset Challenge by Yelp. Some research methods and machine learning techniques we would like to look into are:

- Cultural Trends

- Seasonal Trends
- Location Mining
- Change-points analysis
- Hierarchical and Non-Hierarchical Clustering
- Classification analysis
- Explanatory & Predictive Regression analysis
- Spatial Lag Regression Analysis

2.2 Deliverables

- Project Proposal
- Mid-term presentation
- Mid-term report
- Final presentation
- Final report
- Project poster
- Visualizations of findings and insights hosted on Tableau
- Wiki page

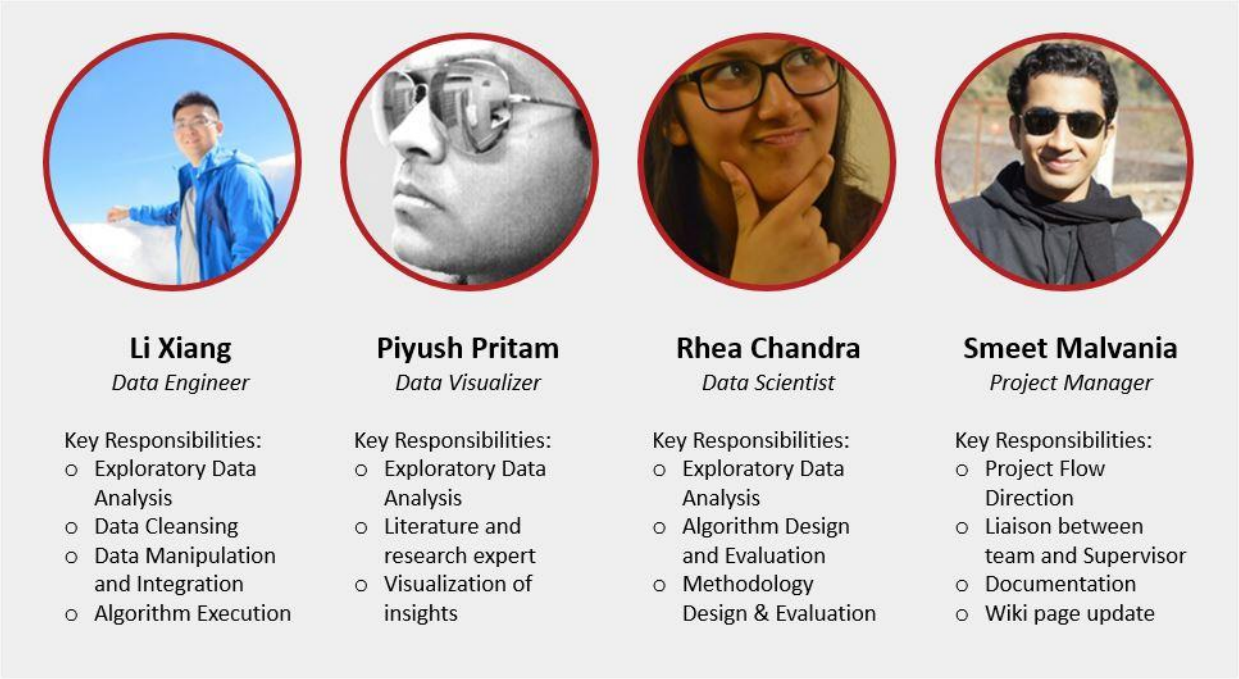
2.3 Tools

Tableau for final visualizations and overall dashboard, Python/R for EDA and data preparation (and manipulation of datasets if required), SAS EG for data mining (if needed).

2.4 Project Timeline

Task	I/C	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13
Research														
Project scope exploration	All	🟥												
Tools Familiarity	Rhea, Piyush, Smeet	🟥	🟥	🟥										
Exploratory Data Analysis	Li Xiang, Piyush		🟥	🟥										
Proposal Development	Rhea, Smeet		🟥	🟥										
Project Proposal														
Proposal Document	All		🟥	🟥										
Wiki Update	Smeet		🟥	🟥										
Step 1: Descriptive Analysis														
Data Preparation	Li Xiang			🟥										
Advanced Data exploration	Li Xiang, Piyush			🟥	🟥									
Clustering Business Profiles	Rhea, Smeet			🟥	🟥									
Trend Analysis	All			🟥	🟥									
Step 2: Feature Extraction														
Model Development	Rhea, Piyush, Li Xiang				🟥									
Model Evaluation	All				🟥	🟥								
Predictive model assessment	All					🟥								
Spacial autocorrelation assessment	All					🟥								
Midterm Report														
Interim Report Preparation	All						🟥							
Interim Presentation Preparation	All						🟥							
Wiki Update	Smeet						🟥							
Step 3: Spatial Lag Regression														
Spatial Lag model development	Rhea, Li Xiang							🟥						
Alternative model testing and eval	Piyush							🟥	🟥					
Salient region interpretation	All								🟥					
Step 4: Data Visualization														
Tool development	Piyush, Li Xiang									🟥	🟥			
User Testing	Rhea, Smeet										🟥			
UI Improvement	Li Xiang, Smeet										🟥			
Step 5: Final Presentation														
Final Report	All											🟥	🟥	
Final Presentation	All											🟥	🟥	
Wiki Page	Smeet												🟥	
Poster	Smeet												🟥	

3. Roles & Responsibilities



Li Xiang
Data Engineer

Key Responsibilities:
- Exploratory Data Analysis
- Data Cleansing
- Data Manipulation and Integration
- Algorithm Execution

Piyush Pritam
Data Visualizer

Key Responsibilities:
- Exploratory Data Analysis
- Literature and research expert
- Visualization of insights

Rhea Chandra
Data Scientist

Key Responsibilities:
- Exploratory Data Analysis
- Algorithm Design and Evaluation
- Methodology Design & Evaluation

Smeet Malvania
Project Manager

Key Responsibilities:
- Project Flow Direction
- Liaison between team and Supervisor
- Documentation
- Wiki page update

4. References

1) Predicts the ratings of the business based on the review text provided by the user.
http://arxiv.org/abs/1401.0864

2) Is there a correlation between the business' ratings and the neighbours ratings?
http://dl.acm.org/citation.cfm?id=2557526

3) Spatial and Social Frictions in the City: Evidence from Yelp.
http://faculty.chicagobooth.edu/jonathan.dingel/research/DavisDingelMonrasMorales.pdf

3) M. Anderson and J. Magruder. "Learning from the Crowd." The Economic Journal. 5 October, 2011.

4) The Yelp Dataset Challenge: http://www.yelp.com.sg/dataset_challenge