

# **DATA SCIENCE BASED APPROACH FOR BUSINESS LOCATION SUITABILITY RECOMMENDATIONS**

J. I. Chinthaka

209318G

Degree of Master of Science

Department of Computer Science and Engineering  
Faculty of Engineering

University of Moratuwa  
Sri Lanka

March 2024

# **DATA SCIENCE BASED APPROACH FOR BUSINESS LOCATION SUITABILITY RECOMMENDATIONS**

J. I. Chinthaka

209318G

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree  
Degree of Master of Science

Department of Computer Science and Engineering  
Faculty of Engineering

University of Moratuwa  
Sri Lanka

March 2024



## DECLARATION

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

2024-07-16

The above candidate has carried out research for the PhD/MPhil/Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: [Dr. A.S. Perera](#)

Signature of the Supervisor:

Date: [2024-07-17](#)

## **ACKNOWLEDGEMENT**

This research project on “DATA SCIENCE BASED APPROACH FOR BUSINESS LOCATION SUITABILITY RECOMMENDATIONS” would not have been possible without the support of many people.

First, I would like to thank my supervisor, Dr. Shehan Perera, Senior Lecturer, Department of Computer Science and Engineering, whose expertise was invaluable in formulating the research literature and methodology. Your insightful feedback pushed me to choose the right direction and successfully complete my dissertation.

I would also like to thank my project proposal evaluator, Dr. Dulani Meedeniya, Senior Lecturer, Department of Computer Science and Engineering, for her valuable guidance throughout my studies. You provided me with useful guidance that helped to sharpen my thinking and to bring my work to a higher level.

Especially I would like to thank Prof. Uthayasanker Thayasivam, Head of Department, Department of Computer Science and Engineering for his assistance and coordination to conduct the research without any issues during the final year.

Finally, I wish to thank the academic and non-academic staff of the Department of Computer Science and Engineering and colleagues for the support and encouragement given.

## ABSTRACT

While existing research primarily focuses on optimizing established businesses, it overlooks a critical group: aspiring entrepreneurs seeking to establish ventures in their own locales. Instead of relocating, businesses often pivot based on region-specific attractiveness. Rather than relocating businesses to more attractive areas, it is better to determine the intrinsic value of each geographical location. Such a methodology will explore additional factors impacting business success and have the potential to significantly enhance urban planning, policy-making, and resource allocation strategies, thereby fostering a more conducive environment for economic growth and development. However, selecting a high-demand business idea for the current location involves navigating various physical, economic, social, and environmental factors, underscoring the complexity of entrepreneurship in today's landscape.

In the past decades, the rapid increase of smartphones and enhanced location-based applications has united individuals on platforms like Yelp, Trip Advisor, Foursquare, and Zomato, facilitating the sharing of experiences across different locations. These platforms, known as location-based social networking (LBSN) platforms, are crucial for business owners seeking to understand customer interests through reviews and visitation patterns. Similar to finding the proper location and time for businesses, we can enhance business category selection mechanisms for a given location using data from LBSN platforms.

By analyzing the Yelp Dataset, we aim to establish a methodology that accurately assesses the suitability of different business categories for specific locations. To achieve this, we first identify key factors influencing business success and filter them based on their availability in the Yelp dataset. Our methodology prioritizes the Size Index aspect of the given area. Finally, we developed a recommendation model that predicts the order of suitable business categories, ranking them from highest to lowest suitability, with one model notably achieving an accuracy of 77.97% while testing the current success of the existing businesses.

**Keywords:** Location-based Social Networking (LBSN), Business category selection, Yelp Dataset, Entrepreneurship, Urban planning

# TABLE OF CONTENTS

|   |      |
|---|------|
| DECLARATION .....                               | i    |
| ACKNOWLEDGEMENT .....                           | ii   |
| ABSTRACT.....                                   | iii  |
| TABLE OF CONTENTS .....                         | iv   |
| TABLE OF FIGURES .....                          | vii  |
| TABLE OF TABLES.....                            | viii |
| 1 CHAPTER 1 .....                               | 1    |
| INTRODUCTION .....                              | 1    |
| 1.1 Demand For The Locations.....               | 1    |
| 1.2 Advancement In Smartphones .....            | 2    |
| 1.3 Location-Based Social Network (LBSN).....   | 2    |
| 1.4 Availability Of Location-Related Data ..... | 3    |
| 1.5 Research Problem .....                      | 4    |
| 1.6 Research Objectives .....                   | 5    |
| 2 CHAPTER 2 .....                               | 6    |
| LITERATURE REVIEW .....                         | 6    |
| 2.1 Location Prediction .....                   | 6    |
| 2.2 Business Prediction For A Location .....    | 13   |
| 2.3 Social Commerce .....                       | 15   |
| 2.4 Yelp Reviews .....                          | 19   |
| 2.5 Other Works With Yelp .....                 | 21   |
| 2.6 Point-Of-Interests (POI).....               | 23   |
| 3 CHAPTER 3 .....                               | 24   |
| PROPOSED METHOD .....                           | 24   |
| 3.1 Data Summary.....                           | 25   |
| 3.1.1 Business .....                            | 27   |
| 3.1.2 Review .....                              | 28   |
| 3.1.3 User .....                                | 29   |

|       |   |    |
|-------|---|----|
| 3.1.4 | Checkin .....   | 31 |
| 3.1.5 | Tips.....   | 32 |
| 3.2   | Data Pre-Processing And Preparation .....             | 32 |
| 3.2.1 | Remove Null Values .....                              | 33 |
| 3.2.2 | Frequency Mining For Business Categories .....        | 33 |
| 3.2.3 | Remove Businesses Not In Above Mined Categories ..... | 37 |
| 3.2.4 | Cluster Business Locations .....                      | 38 |
| 3.3   | Successor Metric .....                                | 42 |
| 3.3.1 | Candidate Metrics .....                               | 43 |
| 3.3.2 | Exploring Metrics .....                               | 43 |
| 3.3.3 | Final Successor Metric .....                          | 45 |
| 3.4   | Aspects Of Reasons .....                              | 46 |
| 3.4.1 | Size Index.....                                       | 47 |
| 3.4.2 | Accessibility Index.....                              | 49 |
| 3.4.3 | Evaluate Aspects Of Reasons .....                     | 53 |
| 3.5   | Model Construction.....                               | 55 |
| 3.5.1 | Models.....   | 55 |
| 3.5.2 | Evaluate Existing Models .....                        | 60 |
| 3.6   | Validation.....                                       | 61 |
| 3.6.1 | Evaluation Metrics .....                              | 61 |
| 3.6.2 | Cross-Validation .....                                | 67 |
| 3.7   | Results .....   | 69 |
| 3.7.1 | Average Values .....                                  | 69 |
| 3.7.2 | Standard Deviations .....                             | 69 |
| 3.7.3 | Box Plots: .....                                      | 70 |
| 3.7.4 | Conclusion .....                                      | 71 |
| 3.8   | Location Suitability Recommendation Model .....       | 72 |
| 3.8.1 | Introduction .....                                    | 72 |
| 3.8.2 | Methodology .....                                     | 73 |
| 3.8.3 | Evaluation .....                                      | 74 |
| 4     | CHAPTER 4 .....                                       | 83 |

|                 |    |
|-----------------|----|
| CONCLUSION..... | 83 |
| REFERENCES..... | 85 |

## TABLE OF FIGURES

| <b>Figure</b> | <b>Description</b>   | <b>Page</b> |
|---------------|--|-------------|
| Figure 1.1:   | Introduction to this study                                     | 3           |
| Figure 2.1 :  | Example restaurant vicinity graph [1]                          | 8           |
| Figure 2.2:   | Examples of the reverse nearest neighbour (RNN) [6]            | 12          |
| Figure 2.3:   | Examples of social elements in Yelp.com review [11]            | 16          |
| Figure 2.4:   | An example of entity linking within the Yelp [13]              | 18          |
| Figure 2.5:   | Example of user reviews on Yelp [12]                           | 19          |
| Figure 2.6:   | Relationships between check-in activities and users[50]        | 23          |
| Figure 3.1:   | Wireframe for the proposed methodology                         | 25          |
| Figure 3.2:   | Summary of dataset attributes                                  | 26          |
| Figure 3.3:   | Top 50 categories  | 33          |
| Figure 3.4:   | Steps in Apriori Algorithm                                     | 35          |
| Figure 3.5:   | Business distribution of final categories                      | 37          |
| Figure 3.6:   | Yelp business location distribution in the world map           | 38          |
| Figure 3.7:   | Example illustration for after and before K-Mean clustering    | 39          |
| Figure 3.8:   | Steps of the K-means clustering algorithm                      | 40          |
| Figure 3.9:   | Elbow curve for the business location distribution             | 41          |
| Figure 3.10:  | Cluster centers for $K = 6$                                    | 41          |
| Figure 3.11:  | Pairwise relationships between Candidate Metrics               | 44          |
| Figure 3.12:  | Map of neighbourhood business for a given business in 1Km area | 48          |
| Figure 3.13:  | A Delaunay triangulation in the plane with circumcircles shown | 51          |
| Figure 3.14:  | Delaunay graph for the selected cluster                        | 51          |
| Figure 3.15:  | Delaunay graph for the selected business                       | 52          |
| Figure 3.16:  | Possible hyperplanes for given data points                     | 56          |
| Figure 3.17:  | Hyperplanes in 2D and 3D feature space                         | 56          |
| Figure 3.18:  | Converting Non - Linear to Linear SVR in 2D Hyperplane         | 57          |
| Figure 3.19:  | An example for Decision Tree Model                             | 57          |
| Figure 3.20:  | Decision Tree evolution in Random Forest Regression            | 58          |
| Figure 3.21:  | Illustration of Multi-Layer Perceptron network                 | 59          |
| Figure 3.22:  | Illustration of sample Convolution Neural Network              | 60          |
| Figure 3.23:  | Illustration of sample Recurrent Neural Networks               | 60          |
| Figure 3.24:  | A graphical representation of cross-validation                 | 67          |
| Figure 3.25:  | Detailed methodology structure                                 | 69          |
| Figure 3.26:  | Box Plot results   | 71          |
| Figure 3.27:  | Confusion Matrix sample  | 76          |
| Figure 3.28:  | Three ROC curve samples  | 78          |
| Figure 3.29:  | Area Under the Curve AUC - ROC Curve                           | 79          |
| Figure 3.30 : | Resulted confusion matrix                                      | 80          |
| Figure 3.31:  | Resulted ROC   | 81          |

## TABLE OF TABLES

| <b>Table</b> | <b>Description</b>   | <b>Page</b> |
|--------------|--|-------------|
| Table 1.1:   | Existing research problem summary                                    | 4           |
| Table 2.1:   | The distance between each customer and each existing facility [6]    | 13          |
| Table 2.2:   | The distance between geographic objects after choosing p1 and p2 [6] | 13          |
| Table 3.1:   | Details of each type of dataset                                      | 26          |
| Table 3.2:   | Descriptions of data in Yelp Business dataset                        | 27          |
| Table 3.3:   | Descriptions of data in Yelp Review dataset                          | 28          |
| Table 3.4:   | Descriptions of data in Yelp User dataset                            | 29          |
| Table 3.5:   | Descriptions of data in Yelp Check-in dataset                        | 31          |
| Table 3.6:   | Descriptions of data in Yelp Tips dataset                            | 32          |
| Table 3.7:   | Important insights related to business categories                    | 34          |
| Table 3.8:   | Selected parameters to mine using Apriori                            | 36          |
| Table 3.9:   | Example set of mined rules   | 36          |
| Table 3.10:  | Number of businesses in each cluster labels                          | 42          |
| Table 3.11:  | Sample of filtered candidate metrics                                 | 43          |
| Table 3.12:  | Correlation between Candidate Metrics                                | 44          |
| Table 3.13 : | Model error results  | 70          |
| Table 3.14 : | Results of recommendation models                                     | 81          |