PROJECT REPORT FOR BACHELOR THESIS PROJECT
(VII SEMESTER)

LOCATION ANALYTICS USING FACEBOOK PAGE DATA

SUBMITTED BY:
ASHUTOSH YADAV
13IM30022

UNDER THE GUIDANCE OF
PROF. J.K. JHA

DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY

JULY-NOVEMBER 2016

AUTHOR'S DECLARATION

I certify that :

- *The work contained in this report has been done by me under the guidance of my supervisors.*
- *The work has not submitted to any other Institute for awarding any degree or diploma.*
- *I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.*
- *Whenever I have used materials (data, theoretical analysis, figure and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.*

Place: IIT Kharagpur

ASHUTOSH YADAV
13IM30022

Date:

Department of Industrial and Systems Engineering

Indian Institute of Technology Kharagpur – 721 302

CERTIFICATE

*This is to certify that the thesis entitled "**Location Analytics using Facebook Page Data**", submitted by **Ashutosh Yadav (13IM30022)** to Indian Institute of Technology Kharagpur, is a record of bonafide project work carried out by him under the concerned supervisor.*

Dr. J. K. Jha
Associate Professor
Department of Industrial and Systems Engineering

CERTIFICATE OF EXAMINATION

*This is to certify that we have examined the thesis entitled "**Location Analytics using Facebook Page Data**", submitted by **Ashutosh Yadav (13IM30022)**, and hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in partial fulfilment of his Bachelor of Technology in Industrial Engineering for which it has been submitted. This approval does not necessarily endorse or accept every statement made, opinion expressed or conclusions drawn, as resolved in the thesis, it only signifies the acceptance of the thesis for the purpose for which it is submitted.*

Date: ***----------***

Place: IIT Kharagpur ***External Examiner***

1. INTRODUCTION

Location is a crucial factor of retail success, as 94% of retail sales are still transacted in physical stores. To increase the chance of success for their stores, business owners require not only the knowledge of where their potential customers are, but also of their surrounding competitors and complementary businesses. From the property owner's standpoint, it is also important to assess the potential success values of their property locations so as to determine the appropriate businesses to lease the locations to and for the right amounts. However, assessing and picking a store location is a cumbersome task for both business and property owners. To carry out the above tasks well, many factors need to be taken into account, each of which requires gathering and analyzing the relevant data. Traditionally, business and property owners conduct surveys to assess the value of store locations. Such surveys, however, are costly and do not scale up well. With fast changing environments (e.g., neighborhood rental, local population size, composition, etc.) and emergence of new business locations, one also needs to continuously re-evaluate the value of store locations.

Fortunately, in the era of social media and mobile apps, we have an abundance of online user-generated data, which capture both activities of users in social media as well as offline activities at physical locations. Facebook is one of the world's largest social media platforms, with more than 1 billion active users everyday. From the business standpoint, the massive availability of user, location, and other behavioral data in Facebook is attractive, and has changed the way people do businesses. For instance, many small/medium business owners are now setting up Facebook Pages to:

(i) allow customers to find their businesses on Facebook
(ii) connect with customers via "likes" and "check-ins"
(iii) reach out to more customers through advertising their business pages on Facebook
(iv) conduct analytics of their pages to get a deeper understanding of their customers and marketing activities.

Consumers are also adapting both their online and offline behaviors to the introduction of Facebook Pages for businesses. Other than "liking" businesses on their Facebook Pages, they can do a "check-in" whenever they physically visit the respective business stores. Facebook Pages have turned many offline signals into online behavior that can be analyzed for business insights. In particular, features such as "likes" and "check-ins" can be used as indicators of popularity, and by extension, success. Similarly, Instagram, Twitter, and Foursquare also have variants of these quantitative signals that can be retrieved from their geotagged photos, tweets, and tips. These data allow us to study the dynamics of brick-and-mortar stores and discover meaningful patterns and insights that will help retail and property owners make better decisions.

2. OBJECTIVE

If we were to open our own cafe, would we not want to effortlessly identify the most suitable location to set up our shop? Choosing an optimal physical location is a critical decision for numerous businesses, as many factors contribute to the final choice of the location. Here, we seek to address the issue by investigating the use of publicly available Facebook Pages data which include user "check-ins", types of business, and business locations to evaluate a user-selected physical location with respect to a type of business. Using a dataset of 6323 food businesses in Bangalore, we conduct analysis of several key factors including business categories, locations, and neighboring businesses. From these factors, we extract a set of relevant features and develop a robust predictive model to estimate the popularity of a business location. Our experiments have shown that the popularity of neighboring business contributes the key features to perform accurate prediction.

In this work, we make use of data collected from Facebook Pages to answer important research questions such as:

"Where should an owner set up his physical retail store at, so as to optimize the store's popularity?"
and
"What are the important factors influencing a store's popularity?"

To this end, we propose a location analytics framework that operates on top of Facebook Pages data. The centerpiece of this framework is the following prediction task: Given a target location that a business/property owner wants to hypothetically set his/her store at, how can we extract the relevant data of businesses within the vicinity of the target location and use them to estimate the popularity of the target location in terms of number of 'check-ins' ?

3. METHODOLOGY

The first point of business for this project was to harvest data from Facebook pages using the Facebook Graph Api. For achieving this purpose we built a Data Harvesting bot that scrapes various facebook pages and collects the relevant data needed for the project.

Harvested Data

	A	B	C	D	E	F	G
5054	yju's JEE and NEET	1	4782	12.92578	77.62025	Education	Education
5055	ake Me Learn	1	15782	12.92829	77.58287	Education	Education
5056	sc.IT Bacheler Science Information Technology	36	271	13.0839	80.27	Education	Admissions Training
5057	ECS - Global Education Consultancy Services	3	109959	12.91812	77.58671	Education	Educational Consultant
5058	elight Academy of Education	67	6206	12.9648156514	77.5880206879	Education	Tutoring
5059	ilicon City Academy Of Secondary Education	345	662	12.88097576	77.57882045	School	School
5060	est Fitness Education	0	475	12.90921	77.64943	Consulting/Business Service	Physical Fitness
5061	rinity Education Services	4	2681	13.03162	77.64415	Professional Service	Professional Service
5062	outh east asian Education Trust	344	1308	12.91947	77.59771	Education	Educational Organizatior
5063	ajajinagar College of Education	2	248	12.9201899	77.5837097	Education	Education
5064	he Vision Education Group	1	1896	13.0190496	77.6490784	Education	Education
5065	oots Education	0	3523	12.9266657421	77.6089775562	Consulting/Business Service	Consulting/Business Ser
5066	obs & education	0	879	12.9156199	77.5737228	Education	Business Service
5067	MAGE Creative Education Yelahanka	56	766	12.9833	77.5833	Education	Arts & Marketing
5068	tudius Education Private Limited	1	2301	12.92447	77.63175	Education	Exchange Program
5069	.S.S. Academy of Technical Education	10	35	12.9020916667	77.5047361111	University	Technical Institute
5070	oorna Prajna Education Centre	203	1190	13.0066214	77.5809061	School	Junior High School
5071	istance Education (Correspondence Courses)	0	540	12.97163	77.63581	Education	Education
5072	oung Innovators Education Services Pvt Ltd	1	5077	12.9665975878	77.535367012	Education	Professional Service
5073	dmissiongyan Education Consultancy	2	1584	12.9365301	77.6148605	Education	Educational Consultant
5074	ollegesniper	0	2111	13.017122732	77.6540064812	Education	Education
5075	aunchPad	79	100202	12.99610529	77.53964304	Education	Education
5076	uperprofs	57	208524	12.99646423	77.66864261	Education	Education
5077	y Mathews Educare Pvt Ltd	0	22713	12.91732	77.61765	Education	Education
5078	kyfi Labs	5	15281	12.95696	77.70117	Education	Education
5079	ivil Simplified	133	120923	12.9738802632	77.7015317734	Education	Education
5080	yju's AIPMT	0	4865	12.91858	77.63142	Education	Education
5081	earson Schools India	268	12566	12.9833	77.5833	Education	Education
5082	FAI - Teaching For Artistic Innovation	30	217782	12.9817801	77.6000595	Education	Education
5083	al Bahadur Sastri College of Education Bangalore	0	1	12.99562	77.57941	Local Business	Local Business
5084	ditya Bangalore Institute For Pharmacy Education and	0	4	12.96698	77.5872879	Local Business	Local Business
5085	V Education Society School Bangalore	0	0	12.92452	77.59352	Local Business	Local Business
5086	Poorna Prajna Education Centre	Sadashivanagar	Bangalore"	0	0	13.01472	
5087	onfluence Educational Services Pvt Ltd Bangalore	1	3	12.9321999	77.583582	Local Business	Local Business
5088	Tata Unisys Limited Education	TULEC"	0	0	12.92868	77.58277	Local Business
5089	l Ameen college of education bangalore	0	0	12.8820599	77.673	Local Business	College & University
5090	Infotech Computer Education	Bangalore"	0	1	13.07614	77.55673	Local Business
5091	bhinava Bharath Education Career Guidence Trust Ba	0	7	12.8811314705	77.5820654843	Organization	Educational Organizatior
5092	angalore North Education Society	5	0	13.00473	77.54167	Local Business	College & University

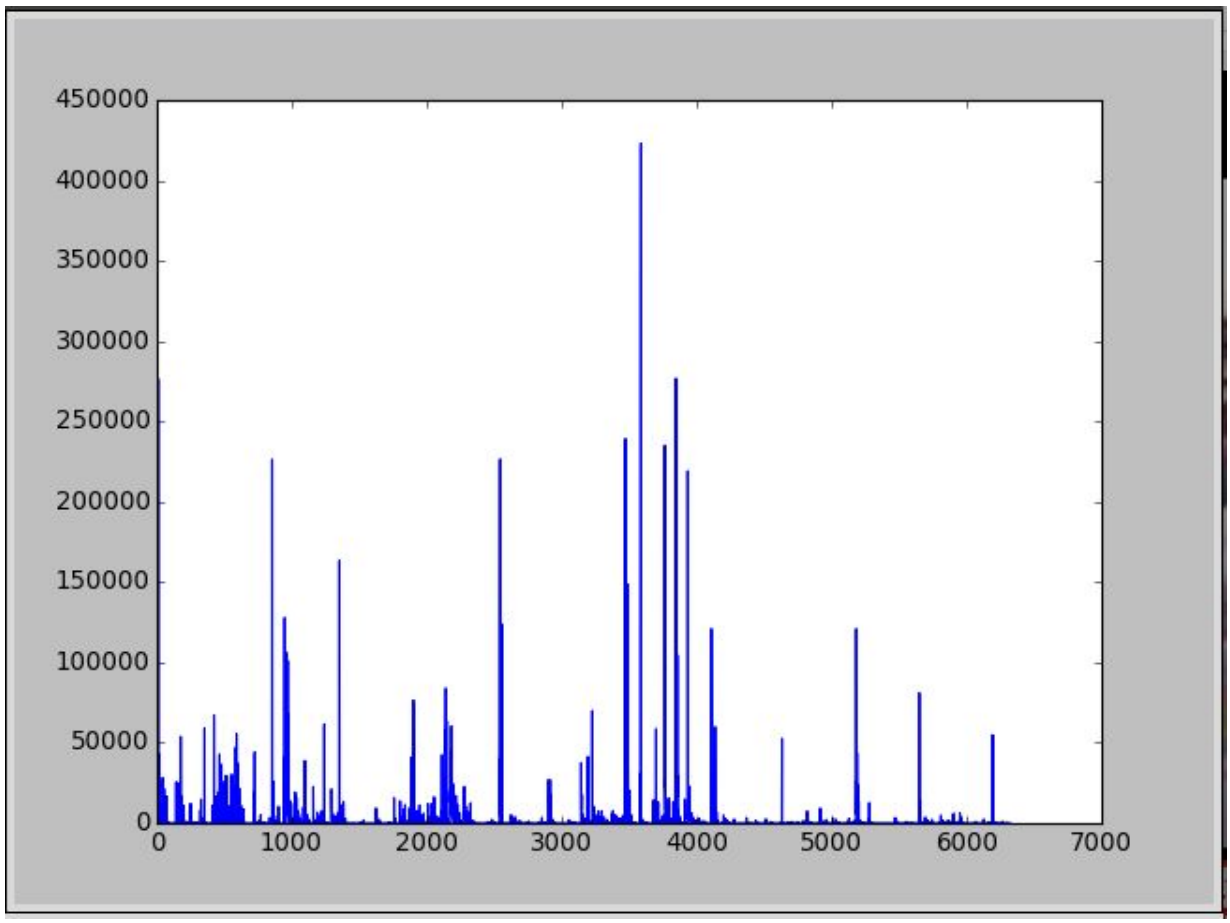
In this project, we are focusing our studies on food-related businesses found in the Facebook Pages of Bangalore. The food-related businesses were defined based on a manually-curated list that consists of 83 food-related categories of business, such as those containing the words "restaurant", "pub", "bar", etc. In particular, we consider food-related businesses in Singapore Facebook Pages that explicitly specify latitude-longitude coordinates, and these coordinates must be within the physical boundaries of Singapore. Using Facebook's Graph API, we obtained a total of 19939 business profiles within Bangalore boundaries, of which we categorically filtered 6323 (31.7%) profiles that are food-related. All business data were analyzed in aggregate, and no personally-identifiable information was used.

From the 6323 food-related businesses, we retrieved a total of 690 unique categorical labels (as standardized by Facebook) from the attribute "category list", which represents the subcategories of a business. These categories contain not only food-related labels (i.e., "bakery", "bar", "cafe", "coffee shop"), but also non-food labels such as "movie theatre", "shopping mall", and "train station." The existence of non-food related labels within food businesses is Facebook's way of allowing business owners to choose more than one categorical label for their business profile.

Each business profile has a location attribute that contains the physical address and latitude-longitude coordinates. Knowing the location of every business allows us to calculate the neighborhood of a selected business through the spatial distribution of other businesses around the vicinity. Specifically, we consider the set $P\ l = \{p|dist(p, l) \leq r\}$ of places p that lie in a radius r around a target location l. We can then create a two-dimensional distance matrix containing the distance between every pair of business. For efficiency, we only consider a maximum radius of 1km (i.e., $r \leq 1km$). This allows us to quickly retrieve the k nearest neighbors of any location.

Facebook provides two possible indicators for a business page's popularity: "check-in" and "like". The "check-in" metric is common in location-based social media like Facebook and Foursquare. Meanwhile, the "like" metric is more unique to Facebook, allowing users to express their recommendation/support for an entity. A "check-in" is the action of registering one's physical presence, and the total number of "check-ins" received by a business gives us a rough estimate of how popular and well-received it is. In contrast, the number of "likes" literally reflects an online vote for the business. Intuitively, therefore, a "check-in" should be a more suitable measure of a physical store's popularity, as it indicates a physical presence. Furthermore, "check-in" can be repeated, i.e., a user could "check-in" to a place on Monday and do so again on Tuesday. By contrast, "likes" cannot be done repeatedly—it is a one-time event.

Frequency Distribution of the Number of 'Check-ins'



FrameWork

Our location analytics framework, consists of two phases: training and prediction. The training phase involves extracting a set of features from the existing business profiles and feeding them to a machine learning algorithm in order to a predictive model for the "check-in" count. In turn, the prediction phase involves extracting features from a target business profile and invoking the trained predictive model to generate a (predicted) "check-in" count for that profile. The modules in our proposed framework consist of three main types: (i) input profiles, (ii) feature extractor, and (iii) predictive model. We shall describe each module type in turn.

The input profile represents a physical business, and is used in both training and prediction phases. An input profile contains several attributes of a business, namely:

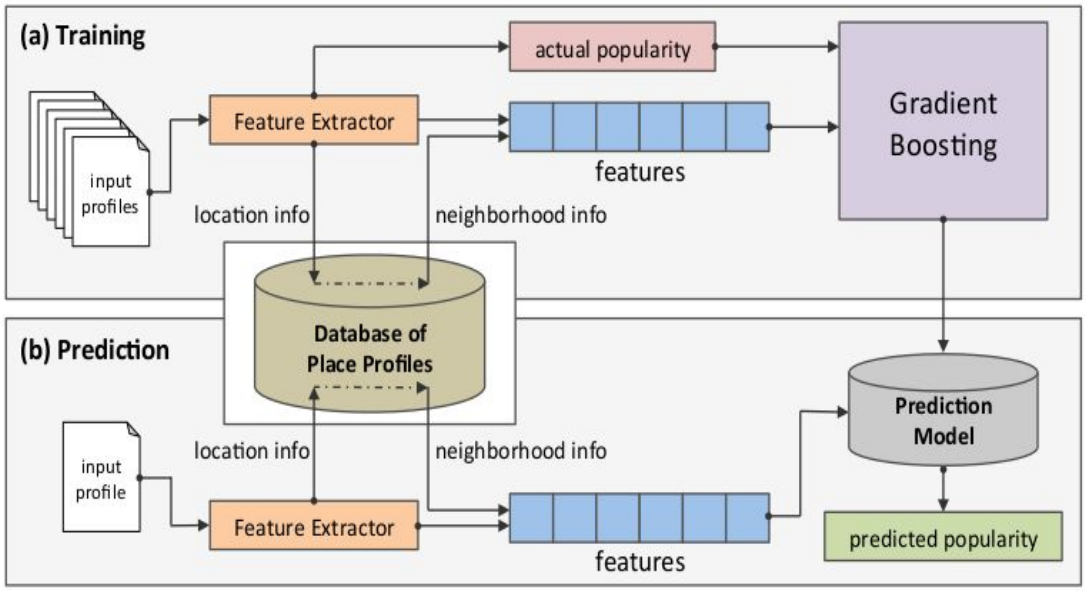
• The lat-long coordinate of the business. This is used in both training and prediction phases.

• The business categories. Examples include "bar", "cafe", "dining", "train station", etc.. This information is also used in both training and prediction phases.

• The "check-in" counts. In the training phase, we feed the actual "check-in" counts of the existing businesses as the target variable of our algorithm. In the prediction phase, the "check-in" counts of the queried locations are assumed to be unknown and our algorithm is supposed to predict them, whether it is for an existing or a new location.

All input profiles of the existing businesses are stored in a database of place profiles. Using this database, we can extract a set of features for a given business, which include features derived from its own (input) profile as well as features from its neighbors (computed based on a range of radius).

Location Analytics FrameWork



Feature Extraction

The feature extraction module serves to construct a feature vector representing a particular business. In this work, we divide our feature vector into six chunks, which represent different aspects of a target business. The first two chunks are associated with categorical data, while the remaining four are about hotspots (i.e., location and "check-ins") data.

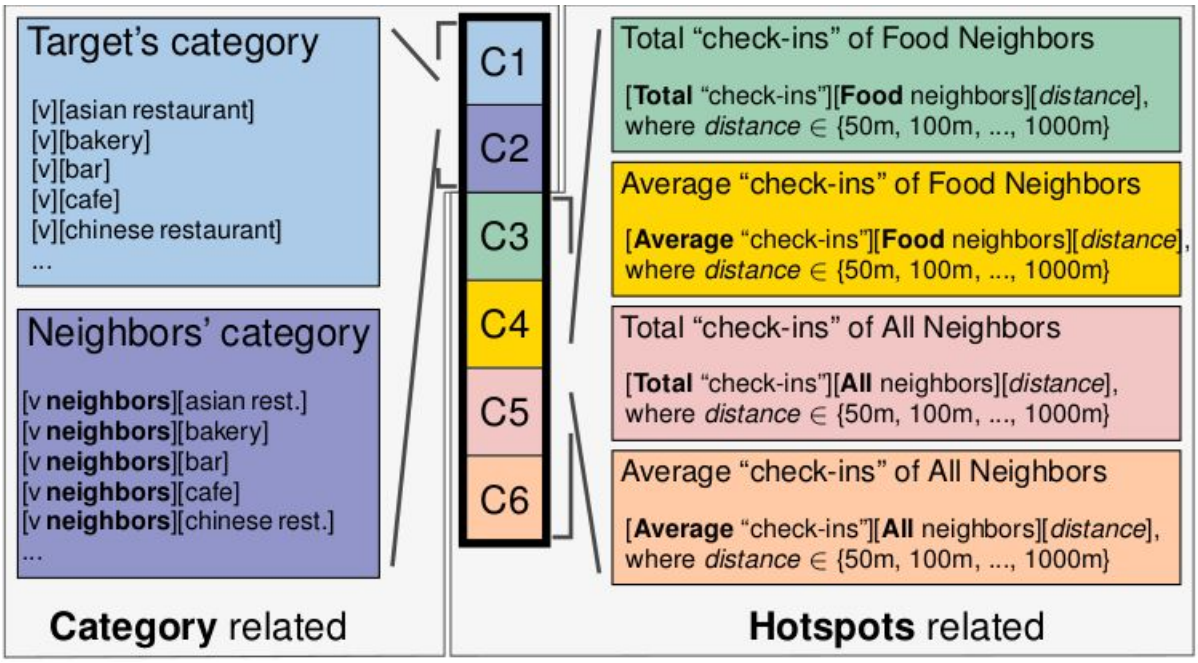
Chunk C-1 : The categories of the target business. This chunk is represented using a binary feature vector. For example, a categorical variable with four possible values: "A", "B", "C", and "D" is encoded using four binary features: [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1], respectively. To represent multiple categories, we simply use "0" and "1" to indicate the absence and presence of each category label respectively. In other words, we use a one-vs-all scheme where we convert multi-class labels to binary labels (i.e., belong or does not belong to the class).

Chunk C-2 : The categories of the target business' neighbors. We first select from our database of place profiles the neighboring food businesses within r meters from the target business, after which we extract and sum up the category feature vectors of the neighbors. To

define category neighborhood, we use r = 200 meters, which we found to give optimal performance.

Chunks C-3 and C-4 : Food-related hotspots. The two chunks are related in that both only use food-related neighbors. In other words, they exclude neighbors that have no relevance to food, such as clothing and electronic stores. For each chunk, we are interested in "hotspots", which are circular areas with the profile in the center and each area is quantified by the "popularity" of stores within it. We define 20 hotspots around the profile whereby each hotspot is demarcated by a maximum distance of r meters, of which r {50, 100, 150, . . . , 1000}. Finally, the only difference between C-3 and C-4 is in how "popularity" is defined; the former computes the (natural) logarithm of the total "check-ins" within a hotspot, while the latter computes the logarithm of the average "check-ins". It must be noted that the total and average "check-ins" include only the "check-in" counts of the neighbors and not the count of the target business itself (which is assumed to be unknown). Also, the purpose of applying logarithmic transformation to the "check-in" counts is to reduce the skewness in the counts distribution (i.e., most businesses have small "check-ins" counts, but there is a handful number of businesses with unusually large "check-ins"). In other words, applying logarithm transformation would allow us to mitigate the impact of (unusually) high "check-ins" for popular businesses.

Chunks C-5 and C-6 : All (food + non-food) hotspots. These chunks are similar to C-3 and C-4 , respectively. The only difference is that, instead of solely using food-related neighbors, chunks C-5 and C-6 use food and non-food neighbors together. The non-food neighbors include bookshops, transportation facilities like bus and train stations, furniture stores, universities, etc. We include non-food hotspots so as to capture the complementary (non-food) businesses within the neighborhood of a target business.



PREDICTIVE MODEL

In order to learn the association between the extracted features and "check-in" scores of a given business. we trained various supervised regression models and then compared the results for selecting the most suitable one the outcome of which is shown in the results section.

EVALUATION METRICS

To measure how accurate our predicted "check-in" scores differ from the actual (observed) scores, we use two popular regression quality metrics: mean-squared logarithmic error (MSLE) and mean absolute logarithmic error (MALE). The MSLE and MALE metrics are respectively defined as:

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2$$

$$MALE = \frac{1}{n}\sum_{i=1}^{n}|\log(p_i + 1) - \log(a_i + 1)|$$

where n is the number of samples in the test set, p is the predicted "check-ins", and a is the actual "check-ins". The MSLE metric measures the averaged squared errors, which gives a higher penalty to large

logarithmic differences | log(p i + 1) − log(a i + 1)|. On the other hand, the MALE metric measures the averaged absolute errors, whereby all the individual differences are weighted equally. To assess the performance of our predictive model, we perform a 10-fold cross-validation procedure whereby the dataset is randomly partitioned into 10 equal sized subsamples. A single subsample is retained as the validation data for testing our models, while the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. We then report the averaged performance. Finally, to test for the statistical significance of our results, we utilize the independent two-sample t-test. In particular, we look at the p-value of the t-test involving two performance vectors, at a significance level of 0.01. If the p-value is less than 0.01, we can conclude the performance difference is statistically significant.

RESULTS AND ANALYSIS

For predicting the number of checkins based upon our location input profile we created a number of machine learning models using different algorithms and compared the performance of these algorithms for selecting the most suitable one using different criteria of Regression analysis. The criteria used for the selection of these models are MAE where MAE stands for the Mean Absolute Error and is calculated using the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

Another criteria used for comparing the regression models is the RMSE or the Root Square Mean Error the value of which is calculated using the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(f_i - y_i)^2}{n}}$$

Another criteria for comparing the performance of the models that can be used is obtained using RMSE is R-Square which is calculated as follows:

$$\textit{R-Square} = 1 - \sqrt{\frac{\sum_{i=1}^{n} (f_i - y_i)^2}{n}}$$

Model	MAE	RMSE	RSquare
Decision Tree	0.00757297012521	0.00110840664368	0.998891593356
Gradient Boosting	0.304888562165	0.0262190695387	0.973780930461
KN Neighbours	0.912908655424	0.983375969341	0.0166240306585
Linear Regression	2.31457479302	0.791377052873	0.208622947127
LogisticRegression	0.127461358361	0.0643547561705	0.935645243829
RandomForest	0.0371440281563	0.0163303917154	0.983669608285

Based on these criteria for selection of the models we made the following observations:

- *It was found that the lowest level of R-Square was found for the KN Neighbours algorithm which means that the KN Neighbours has the least accuracy and maximum errors which was quite an expected result as KN Neighbours algorithm is mostly used only classification problems rather than for Predictive Regression Problems.*
- *Linear Regression algorithm also resulted in significantly small value of R-Square which implies that number of 'check-ins' is not a linear model and hence cannot be predicted accurately using Linear Regression*
- *Decision Tree algorithm resulted in a very high value of R-Square which was nearly equal to 1 which suggested overfitting of the data in the Decision Tree model. Upon further investigation of this theory by dividing the data into training and testing datasets it was found to be true as it resulted into a small value of R-Square for the testing dataset as compared to the Training dataset. As a*

result the Decision Tree algorithm was not found suitable for predicting the number of 'check-ins'.

- *Logistic Regression resulted into a significant high value of R-Square suggesting that the data follows an inverse exponential relationship. However the value was smaller that the values of obtained for Gradient Boosting and Random Forest Algorithms because of which it was not chosen as the final predictive model.*
- *Both the Gradient Boosting and the Random Forest algorithm resulted in high values of R-Square which were not as high as Decision Tree model showing that there is no overfitting in this model and both can be used for the predictive model. However Random Forest algorithm was chosen as the final predictive model as it has lower value of mean absolute error.*

IDENTIFIED FACTORS

Based on our analysis the following factors were identified that affect the popularity of a physical retail store and how they affect their popularity:

- ***Neighbouring Competitive businesses:***

Neighbouring competitive businesses affect the popularity of the physical restore in a negative fashion as they tend to eat into the possible customer base. The closer the competitive business is to the the more is the impact. However there is a local effect which results into a boost in the business if a significant number of stores start operating in each others proximity resulting into the formation of a hub.

- ***Neighbouring Complementary businesses:***

Neighbouring Complementary businesses affect the popularity of the physical store in a positive fashion as they tend to have the same set of customers and the customers visiting one of them is more likely to visit the same store at the same time hence both these complementary businesses help increase the business of each other .

- ***Proximity to Places of Importance and Tourism:***

The proximity of the physical store to a place of significance helps boost the popularity of the physical store as this has the same effect on the popularity of the place as do the complementary businesses. They help increase the popularity of the store by bringing in

more customers who were initially visiting the place of significance or the tourist spot.

CONCLUSION AND FUTURE WORK

In this work, we investigate whether businesses can benefit from other (popular) businesses within its vicinity. Our results show not only a positive correlation between the popularity of a target business and its neighbors, but also the critical importance of the "hotspot" features: the nearer a target location is to a popular place with larger "check-ins", the more successful it would be. This finding conforms with our intuition. But more importantly, it demonstrates that ubiquitous online data (such as Facebook Pages) can be used to gauge the socio economical values. We also show how our predictive model can be used to accurately estimate the "check-in" score of a particular location, allowing us to identify the best locations that would bring popularity, and by extension, success. Despite the promising potentials of our approach, there remains room for improvement. For instance, our current work has not taken into account the temporal aspects of the business popularity, such as modeling the trend of the "check-in" scores over time. Further quantitative and qualitative studies may also be needed in the future to compare our work with other location-based services such as Foursquare. To facilitate more comprehensive location analytics, we can extend our approach by building a two-level location recommendation system, whereby we first (coarsely) recommend a city district and then pinpoint (multiple) promising locations within that district. As we include more data, such as non-food categories and auxiliary data that reflect the human flow of different areas of an urban city, we will be able to further improve on our current model and findings. To address all these, we plan to develop a new spatiotemporal predictive model that integrates a richer set of residential, demographics, and other social media data.

REFRENCES

- *J. Chang and E. Sun. Location3: How users share and respond to location-based data on social networking sites. In Proceedings of the International AAAI Conference on Weblogs and Social Media, pages 74–80, 2011.*
- *H. Chen, R. H. L. Chiang, and V. C. Storey. Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4):1165–1188, 2012.*
- *N. Cohen. Business location decision-making and the cities: Bringing companies back. Technical report, Brookings Institution Center on Urban and Metropolitan Policy, 2000.*
- *ESRI. Revealing the "where" of business intelligence using location analytics.*
- *A. Natekin and A. Knoll. Gradient boosting machines: A tutorial. Frontiers in Neurorobotics, 7(21):1–21, 2013.*
- *Facebook Graph API reference. https://developers.facebook.com/docs/graph-api/reference/page, 2016.*