Location Profiling for Retail-Site Recommendation Using Machine Learning Approach

Choo-Yee Ting(✉) and Mang Yu Jie

Faculty of Computing and Informatics, Multimedia University, 63000 Cyberjaya, Selangor, Malaysia
cyting@mmu.edu.my

Abstract. Retail site selection is a critical stage for a new retailer since it helps them to decide which locations have the best chance of delivering a good return on investment. Most of the new retailers will face problems while selecting a retail site for new business. Work presented in this paper will focus on predictive modelling by using the geographical variables and demographic variables. Besides, an analytical dataset will be constructed that generated by several algorithm and the five different feature selection will be performed on the analytical dataset to increase the efficiency of models. There are six classification models were developed in this project, which is Random Forest Classifier, XGBoost Classifier, Logistic Regression, Naive Bayes Classifier and Decision Tree Classifier. Besides, a deep learning classification models will be developed in this project, which is Multi-layer Perceptron Classifier. Accuracy, Precision, Recall, and F1-Score are used to evaluate the performance of classification models in this project. Among all models that constructed by using different features of several feature selection, XGBoost Classifier has the highest accuracy, which is around 94%.

Keywords: Retail-Site-Recommendation · Feature Selection · Machine Learning Model

1 Introduction

Retail, is the sales of items or services by a corporation to a customer for personal consumption, it may be a way of earning profit by selling products or services through various channels of distribution and merchants satisfy demand that has been identified by a supply chain [1]. According to [2], retail has been divided into three main categories: (i) Convenience Goods, such as groceries and daily provisions; (ii) Shopping or Comparison Goods, which refer to relatively more expensive items purchased at less regular intervals; and (iii) Specialty Goods, which are other items that appeal to higher-income customers. Moreover, retail can be split into small-scale and large-scale establishments. Single-propriety stores and non-store operators such as hawkers, peddlers, and market stalls are examples of small-scale retailers. Superstore, discount store, department store, supermarket, hypermarket, and shopping center are examples of large-scale retailers [3].

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 48–67, 2022.
https://doi.org/10.2991/978-94-6463-094-7_5

Recently, the retail industries in Malaysia have been impacted by the COVID-19 pandemic. COVID-19 is a worldwide pandemic that has spread to 190 countries, with an increasing death toll [4]. As a result of the COVID-19 pandemic, many retailers or small and medium-enterprise (SMEs) are struggling to stay afloat in the aftermath of the global COVID-19 outbreak and Malaysia's MCO, which has caused revenue losses of 50% or more for several SMEs [5].

Moreover, Global GDP is forecast to fall by 2.1%, while GDP in developing countries is expected to fall by 2.5% and GDP in high-income countries is expected to fall by 1.9%. As a result of their extensive trade integration and the direct impact on tourism, countries in East Asia and Pacific (EAP) such as Cambodia (3.2%), Singapore (2.1%), Hong Kong Special Administrative Region of China (2.3%), Thailand (3%), Vietnam (2.7%), and Malaysia (2.1%) are expected to suffer the greatest GDP losses under a global pandemic scenario [6].

By implementing Machine Learning in retail site selection, it will help retailers to determine which are the possible locations that have the best opportunity to deliver a good return on investment [7]. Location is the most important role in business, a strategic location will attract the customers by making products or services convenient, which has a significant impact on market share and profitability [8]. Site selection considers a wide range of ecological, social, and monetary variables that can often be restricted in site selection [9]. The best site selection will effectively utilise the city's assets while also providing high network proficiency [10]. Thus, the project aims at using location profile, demographic, and footfall information as input to machine learning models and its output would be the optimal recommendation for retail given a particular location.

The objective of this paper was to construct an analytical dataset for retail-site recommendation. Secondly, this paper aimed to employ machine learning models for optimal recommendation of retail to site matching. Lastly, this paper also aimed to develop a web-based app as front-end to the machine learning models.

2 Background Studies

2.1 Variables Related to Retail Site Selection

The features is an important factor to select a retail site. There are a lot of essential variables that can the site of selection. Ting et al. [8] used many features to select a retail site such as Point of Interest (POI), nearby property or neighbourhood, population and education data. The features can be splitted into geographical features and demographic features.

Geographical features is the variables which around the retail site, such as POI, nearby property or neighbourhood and transportation. All of the location attractors are referred to the features of location. Erbıyık et al. [11] said that providing a convenient location for customers is important. Ting et al. [8] do an experiment by applying the food and beverage data to the computer algorithm to predict the retail business in an area. In the paper, the result shown that it has the highly mean accuracy. Therefore, there is highly possible that existing strong dependencies between the retail business in an area [8]. Wang et al. [7] used the geographical features including POI, Yellow Pages (YP), iProperty, economy and education data, the author also used demographic features in the

paper. The author used geographical features and demographic features to construct an analytical dataset. The POI and YP data were categorised into many location features, such as Chinese restaurants, fashion stores, bank and post office.

Demographic features which is the population in the area, the education data, the economy data and POI customer number. These are the data that most of the researchers using by selecting the retail site. It can be used on predicting the popularity of an area and find out the optimal retail store. Xu et al. [12] train different machine learning models to predict a retail site. The author used the geographical features and demographic features which is POI and POI customer number to predict the customer of the retail site, the retail site with the higher value means that the retail site has the highly potential to open a new store. Therefore, the demographic features are very important for retail site selection. Moreover, [13] used the nearby existing business data which is from social media network to predict the "check-in" score for new retail site. In the paper, the author focused on the food-related business, the data has been categorised into restaurant, bar, pub and others. The author analyzed the visitor number of nearby business and train the machine learning algorithm to predict the potential popularity of the retail site.

Table 1 shows the existing works on selecting the geographical and demographic features for retail site selection.

2.2 Feature Selection

Recently, the amount of data available has exploded in terms of both sample size and dimensionality in many machine learning applications [26]. These data usually have a high number of dimensions, which makes data analysis and decision-making difficult [27]. Therefore, utilization of the large-scale data is very important in the machine learning applications. It is possible that learning algorithms will become relatively sluggish and even degenerate in their performance of learning tasks if there are too many noisy, redundant, and irrelevant dimensions present. The majority of researchers will use feature selection on the data in order to increase the effectiveness of the learning algorithm. This issue may be effectively handled with the help of feature selection, which is a dimensionality reduction approach that removes unnecessary and superfluous data. It can speed up computation, enhance learning accuracy, and make the learning model or data more intelligible [27]. Feature selection methods can be classified into three general types which is filter, wrapper and embedded methods according to the various search strategies [25–27].

Table 2 shows the comparison of methods used in feature selection.

2.3 Site Selection Techniques

In site selection, different approaches will perform differently. Therefore, the researchers have to determine the optimal approach as each approach has its set of advantages and disadvantages [25]. The researchers will choose the different approaches based on the different situation. Most of the researchers using five types of techniques to perform retail site selection, which is Machine.

Table 1. Summary of Variables Used in Existing Retail Site Selection Works

	Geographical			Demographic			
Authors	Point of Interest (POI)	Nearby Property or Neighbourhood	Transportation	Population	Education	Economy	POI Customer Number
[14]	✓		✓				
[18]	✓	✓		✓	✓	✓	
[8]	✓	✓		✓	✓		
[13]	✓						✓
[23]	✓	✓		✓	✓	✓	
[24]	✓	✓	✓				
[12]	✓						✓
[15]			✓	✓			
[19]	✓		✓	✓		✓	
[20]	✓	✓	✓				
[16]			✓		✓		
[25]	✓	✓		✓	✓	✓	
[17]	✓		✓	✓			
[21]			✓	✓		✓	
[7]	✓	✓	✓			✓	
[22]	✓	✓	✓				

Learning, Deep Learning, Analytic Hierarchy Process (AHP), Geographic Information System (GIS) and Technique for Order Preference by Similarity to Ideal Solution (TOPSIS).

Machine Learning is a learning algorithm which can do the prediction with a large amount of data. Machine learning model will calculate the weights of different factors and compute the output of its prediction [35]. Besides that, Deep learning is a subfield of machine learning that works with artificial neural networks, which are algorithms inspired by the structure and function of the brain. Deep learning algorithms are built in a hierarchy of increasing complexity and abstraction while traditional machine learning are linear. Moreover, AHP is used to determine the weights of each criteria and compare two alternatives pairwise, it will select optimal site for a specific objective from a range of potential options [25]. Furthermore, Geospatial data describes both the location and the attributes of spatial features, and geospatial data can be captured, stored, queried, analysed, and displayed using Geographic Information System (GIS) [36]. As a visual inspection approach, Geographic Information System overlays various data on a map and allows for analysis to be performed on it [25]. Lastly, TOPSIS is a multiple criteria method

Table 2. Comparison of Methods Used in Feature Selection

	Filter					Wrapper				Embedded				
Author	Mutual Information	Chi-squared Statistics	Information Gain	Relief	Correlation based	Genetic Algorithms	Recursive Feature Elimination	Boruta	Hill-climbing Searching	Lasso	Elastic Net	Random Forest	Artificial Neural Network	Support Vector Machine
[8]			✓				✓					✓		
[29]		✓												
[28]												✓	✓	
[30]					✓							✓		✓
[31]	✓	✓	✓											
[25]	✓			✓		✓	✓	✓	✓			✓		
[32]												✓		
[23]								✓	✓			✓		
[24]												✓		
[33]					✓		✓							

Table 3. Comparison of Techniques Used i Site Selection

	Site Selection Techniques				
Author	Machine Learning	Deep Learning	AHP	GIS	TOPSIS
[25]	✓				
[34]					✓
[38]	✓				
[14]	✓			✓	
[39]			✓	✓	
[7]	✓	✓			
[40]			✓		
[24]	✓				
[41]				✓	
[42]			✓	✓	
[12]	✓				

that uses simultaneous minimization of distance from an ideal point and maximum of distance from a nadir point to look for solutions from a finite set of alternatives [37].

Table 3 shows the existing works on site selection.

3 Analytical Dataset

This section discusses the dataset and its source, all of the data will be used to construct an analytical dataset to perform the prediction in this paper.

3.1 Source of Data

3.1.1 Point of Interest Data (D_{poi})

Point of Interest (POI) data is a dataset that consist a list of retail businesses in Malaysia which is prior to 2022. The data were obtained from Yellow Pages which is an online business directory. The POI dataset consists of 443, 041 rows with address, category, latitude, longitude, retail name and State of retail.

3.1.2 Base Data (D_{base})

Base data is the subset of POI data that consist lists of retail business of interest, such as restaurants and cafes. The information provided by base dataset are address, name of retail business, latitude and longitude.

3.1.3 Residential Property Data (D_{ppt})

The residential property data is a list of properties which obtained from Brickz and the data is from the Value and Property Services Department (JPPH). The residential property data is ranged from 2000 to 2019 and it consists of 491, 721 rows with residence name, type and tenure of property, price of the property, latitude and longitude.

3.1.4 Population Data (D_{pop})

The population data is a list of population in Malaysia and its source of population data is Humanitarian Data Exchange (HDX). The information of population provided by dataset are the total density population, men population, women population, women of reproductive age population (ages 15–49), elderly population (ages 60+), children population (ages 0–5), and youth population (ages 15–24).

3.1.5 Geospatial Data

The Geospatial data is a list of geographical details of Malaysia. The dataset is mainly used to carry out the geocoding and reverse geocoding. It will show all of the information in Malaysia from State to District and the coordinates of each State and District.

3.1.6 Competitor Data (D_{cpt})

The competitor data is a dataset which is same with the POI data that consist a list of retail businesses in Malaysia. It consists the address, category, latitude, longitude, retail name and State of retail.

Table 4 shows the overview of data of this paper.

Table 4. Overview of Data

Dataset	Description	Dimension
Base Data (D_{base})	The retail business data, which include different categories of retail data	1714, 4
Point of Interest Data (D_{poi})	The nearby location features	443041, 6
Competitor Data (D_{cpt})	The nearby competitors	443041, 6
Residential Property Data (D_{ppt})	The nearby residence and property	491721, 6
Population Data (D_{pop})	The population of men, women, women of reproductive age, children, youth and elderly	3855823, 8
Geospatial Data	The geographical details of Malaysia	144, 17

3.2 Data Preprocessing

To ensure that the result can be a good performance, the data preprocessing is needed. Data preprocessing can transform the raw data into a format which is usable and efficient. There are many meaningless and missing data in a dataset. Therefore, the data cleaning is needed to handle the missing data and noisy data. There are some missing values of coordinates in D_{poi} and D_{ppt} have been dropped since the coordinates are important to identify the location in this paper. Besides, the reverse geocoding has been implemented in the D_{base} by using the Geospatial Data, the City and State of the coordinates will be generated by implementing reverse geocode. Lastly, there are some unimportant features in D_{ppt} has been filtered out.

3.3 Construction of Analytical Dataset

Algorithm 1 shows the steps to construct the analytical dataset. In this algorithm, the input data is D_{base}, D_{poi}, D_{ppt}, D_{pop} while the output will be the $D_{analytical}$ which is the analytical data. D_{base} is the base data to merge the data with others data. Firstly, the ***get-nearby-poi*** function is to get the nearby POI from the base data from the base data. By passing the base data latitude and longitude, a nearby POI within 2.0 km will be generated and the top 5 POI of all of the POI will be selected. Besides, a nearby POI category within 2.0 km will be generated and the top 5 POI category of all the POI category will be selected by using ***get-nearby-category*** function. After that, the nearby property will be generated by using ***get-nearby-property*** function. The top 3 nearby neighbourhoods will be generated by passing the latitude and longitude of base data and all of the details of the nearby neighbourhoods will be shown. Moreover, the ***get-nearby-population*** function is used to calculate the population density. The population density of each neighbourhood will be generated through this function. Furthermore, the nearby competitor within 2.0 km will be selected from the POI data. Lastly, D_{base}, D_{nearby_poi}, D_{nearby_ppt}, D_{nearby_pop} and D_{nearby_cpt} will be merged based on the latitude and longitude and the output will be analytical dataset. In this project, all of the analytical dataset will be appended together and construct as an analytical dataset.

Algorithm 1 generate-Analytical-Dataset

Require: D_{base_1}, D_{base_2}, D_{base_3}, D_{poi}, D_{ppt}, D_{pop}
Ensure: $D_{analytical}$
1: **for** $data$ in $(D_{base\_1}, D_{base\_2}, D_{base\_3})$ **do**
2: $L \leftarrow$ `get-latlng`$(data)$
3: $D_{base} \leftarrow$ `get-base-data`
4: $D_{nearby\_poi} \leftarrow$ `get-nearby-poi`(D_{poi}, L)
5: $D_{nearby\_category} \leftarrow$ `get-nearby-category` (D_{poi}, L)
6: $D_{nearby\_ppt} \leftarrow$ `get-nearby-property`(D_{ppt}, L)
7: $D_{nearby\_pop} \leftarrow$ `get-nearby-population` $(D_{pop}, D_{nearby\_ppt})$
8: $D_{nearby\_cpt} \leftarrow$ `get-nearby-population` (D_{poi}, L)
9: $D_{analytical} \leftarrow D_{base} \bowtie D_{nearby\_poi} \bowtie D_{nearby\_category} \bowtie D_{nearby\_ppt} \bowtie D_{nearby\_pop} \bowtie D_{nearby\_cpt}$
10: **end for**

Algorithm 2 get-nearby-poi

Require: D_{poi}, L
Ensure: D_{nearby_poi}
1: **for** lat and lng in L **do**
2: $D_{poi\_within\_2km} \leftarrow (\texttt{haversine}(lat, lng, lat_{poi}, lng_{poi}) * 1000) \leq 2000$
3: **end for**
4: $D_{nearby\_poi} \leftarrow D_{poi} \cap D_{poi\_within\_2km}$
5:
6: **return** D_{nearby_poi}

Algorithm 2 shows the steps to generate nearby POI data within 2.0 km of base data. The haversine formula has been used in this algorithm to calculate the distance between two sets of coordinates. After calculating the distance within 2.0 km, the $D_{poi\_within\_2\ km}$ will be subsetted from D_{poi} and the counts of nearby points of interest may be extracted by counting the frequencies. Then, the nearby POI within 2.0 km will be used to create a new column, with counts of frequencies in each column. The top 5 POI with the most counts of frequencies will be selected.

Algorithm 3 get-nearby-category

Require: D_{poi}, L
Ensure: $D_{nearby\_category}$
1: **for** lat and lng in L **do**
2: $D_{category\_within\_2km} \leftarrow (\texttt{haversine}(lat, lng, lat_{poi}, lng_{poi}) * 1000) \leq 2000$
3: **end for**
4: $D_{nearby\_category} \leftarrow D_{category} \cap D_{category\_within\_2km}$
5:
6: **return** $D_{nearby\_category}$

Algorithm 3 shows the steps to generate the nearby POI category data within 2.0 km of base data. After calculating the distance within 2.0 km of two sets of coordinates by using haversine formula, the $D_{category\_within\_2\ km}$ will be subsetted from $D_{category}$ and the counts of nearby points of interest category may be extracted by counting the frequencies. After that, the nearby POI category within 2.0 km will be used to create a new column, with counts of frequencies in each column. The top 5 POI category with the most counts of frequencies will be selected.

Algorithm 4 get-nearby-property

Require: D_{ppt}, L
Ensure: D_{nearby_ppt}
1: **for** lat and lng in L **do**
2: $D_{ppt\_within\_2km} \leftarrow (\texttt{haversine}(lat, lng, lat_{ppt}, lng_{ppt}) * 1000) \leq 2000$
3: **end for**
4: $D_{nearby\_ppt} \leftarrow D_{ppt\_within\_2km}[:3]$
5:
6: **return** D_{nearby_ppt}

Algorithm 4 shows the algorithm to generate the nearby neighbourhoods within 2.0 km of base data. The L and D_{ppt} are the input of this algorithm while the output will be D_{nearby_ppt}. By utilising the haversine formula, the nearby neighbourhoods which have the short distance with base dataset will be generated. In this paper, the top 3 nearby neighbourhood will be selected and all the detail of each neighbourhood will be shown.

Algorithm 5 get-nearby-population

Require: D_{pop}, D_{nearby_ppt}
Ensure: D_{nearby_pop}
1: **for** lat and lng in D_{nearby_ppt} **do**
2: $D_{pop\_neighbourhood} \leftarrow (\texttt{haversine}(lat, lng, lat_{pop}, lng_{pop}) * 1000) \leq 2000$
3: $D_{nearby\_pop} \leftarrow \texttt{sum}(D_{pop\_neighbourhood})$
4: **end for**
5:
6: **return** D_{nearby_pop}

Algorithm 5 shows the function to calculate the population density of each neighbourhood. In the function, the population density within 2.0 km of nearby neighbourhood will be summed up and all details of population density will be shown.

Algorithm 6 get-nearby-competitor

Require: D_{poi}, L
Ensure: $D_{nearby\_competitor}$
1: **for** lat and lng in L **do**
2: $D_{poi\_within\_2km} \leftarrow (\texttt{haversine}(lat, lng, lat_{poi}, lng_{poi}) * 1000) \leq 2000$
3: **end for**
4: $D_{nearby\_competitor} \leftarrow D_{poi\_within\_2km}[selected\_competitor]$
5:
6: **return** $D_{nearby\_competitor}$

Algorithm 6 shows the function to generate the nearby competitor within 2.0 km of base data. The D_{poi} and L will be the input of the function and the output will be $D_{nearby\_competitor}$. The nearby POI within 2.0 km will be generated by utilising the haversine formula. From $D_{poi\_within\_2\ km,}$ users are able to selected the specific competitors (Table 5).

4 Methodology

4.1 Feature Selection

There are five types of feature selection models will be performed in this study, which included Boruta, Recursive Feature Elimination, Lasso, Fisher's Score and Information Gain. All the score of feature selection models with be ranked with descending order. The top 5 to 20 feature scores with be selected to train the models. Before the feature selection, the SMOTE process will be performed. SMOTE (Synthetic Minority Over-Sampling Technique) is a technique that used to deal with the unbalanced classification problems.

Table 5. Features and Description of Analytical Dataset

Feature	Description
address	The address of the retail store
lat	Latitude of the retail store
lng	Longitude of the retail store
point_name	Name of the retail store
city	City of the retail store
state	State of the retail store
Top 5 Point of Interest	Top 5 point of interest for each base data
Top 3 Nearby Neighbourhoods	Top 3 nearby neighbourhoods' details of the retail store, which includes name of neighbourhood, latitude, longitude, tenure of neighbourhood, type of neighbourhood and median price of neighbourhood
Population of Each Neighbourhood	The population density of each neighbourhood, which includes total density population, men population, women population, women of reproductive age population (ages 15–49), elderly population (ages 60 +), children population (ages 0–5), and youth population (ages 15–24)
Nearby Competitor	Nearby competitor for each base data

4.2 Model Construction and Evaluation

There are several classification models will be trained in this project. All the models were trained by using the important feature that selected by feature selection models. The classification models that chosen in this project are Random Forest Classifier, XGBoost Classifier, Logistic Regression, Naive Bayes Classifier, Decision Tree Classifier and MLP Classifier. By using different numbers of top features that selected by feature selection models, the machine learning models will be trained repeatedly. Parameters will be tuned by iteratively training the models with different parameters to identify the best setting for parameters to ensure the models perform at their best. Besides, the label encoder will be used to train the classification models, the non-numeric features will be converted to numeric features through label encoder.

After the model construction, the model evaluation will be performed. For classification in this study, accuracy score, precision, recall and F1 Score will be used to evaluate the models.

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Table 6. Combination of Analytical Dataset (F&B)

Dataset	Point Name
1	KFC, Nando's, Starbucks
2	KFC, Starbucks, Pizza Hut
3	Mcdonald's, KFC, Secret Recipe
4	Mcdonald's, Pizza Hut, Starbucks
5	Pizza Hut, Nando's, Secret Recipe
6	KFC, Starbucks, Secret Recipe
7	Mcdonald's, Nando's, Starbucks
8	KFC, Mcdonald's, Starbucks
9	KFC, Pizza Hut, Secret Recipe
10	Mcdonald's, Secret Recipe, Nando's

$$F1\ Score = 2\left(\frac{Precision \cdot Recall}{Precision + Recall}\right)$$

4.3 Experiment Setting

In this study, five machine learning models and one deep learning model will be developed based on the analytical dataset. There is total 6 main *point_name* to predict through the classification models, which is KFC, Nando's, Starbucks, Mcdonald's, Pizza Hut and Secret Recipe. These 6 mains *point_name* will be generated into 10 sets of analytical datasets. Table 6 shows that the combination of 6 main *point_name*.

The classification models that will be trained in this project are Random Forest Classifier, XGBoost Classifier, Logistic Regression, Naive Bayes Classifier, Decision Tree Classifier and MLP Classifier. The models will be trained repeatedly by using different numbers of top features that selected by five different feature selection. Once the models have been trained, the results will be compared to each other's to find out the best result. Table 7 shows the experiment setting of this study.

5 Findings

This section will discuss the results and findings of this study and aims to highlight the average performance as well as the best performing model.

5.1 Performance of Retail Site Selection

5.1.1 Machine Learning

Through classification, the better performance can be achieved at around 94%. The figures below show that the accuracy trend of classification models of the highest accuracy

Table 7. Experiment Setting.

Experiment	Technique Used	Dataset	Feature Selection Models	Number of Features
1	Random Forest Classifier	1 - 10	Boruta, RFE, Lasso, Fisher's Score, Information Gain	5 - 20
2	XGBoost Classifier	1 - 10	Boruta, RFE, Lasso, Fisher's Score, Information Gain	5 - 20
3	Logistic Regression	1 - 10	Boruta, RFE, Lasso, Fisher's Score, Information Gain	5 - 20
4	Naive Bayes Classifier	1 - 10	Boruta, RFE, Lasso, Fisher's Score, Information Gain	5 - 20
5	Decision Tree Classifier	1 - 10	Boruta, RFE, Lasso, Fisher's Score, Information Gain	5 - 20
6	MLP Classifier	1 - 10	Boruta	15 - 20

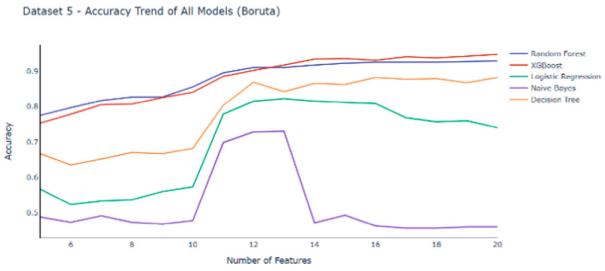

Fig. 1. Accuracy Trend of All Machine Learning Models (Boruta)

dataset by using different top features that selected by five type of feature selection (Figs. 1, 2, 3, 4 and 5).

To conclude the overall performance on classification analysis, the highest accuracy score of each classification model is shown in the Tables 8, 9, 10, 11 and 12.

By using the selected features by Boruta, XGBoost Classifier has achieve the highest accuracy score which is 0.9467 with 20 features. The runner-up model, Random Forest Classifier has the accuracy score of 0.9283 with 20 features.

By using the selected features by RFE, XGBoost Classifier has achieve the highest accuracy score among all models constructed, which is 0.9417 with 14 features. The

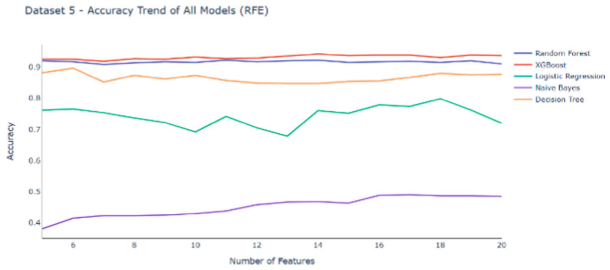

Fig. 2. Accuracy Trend of All Machine Learning Models (RFE)

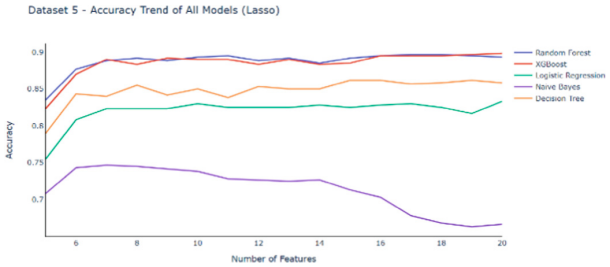

Fig. 3. Accuracy Trend of All Machine Learning Models (Lasso)

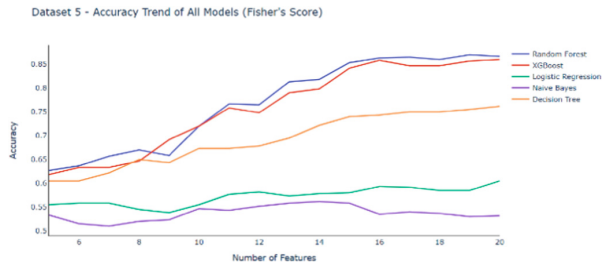

Fig. 4. Accuracy Trend of All Machine Learning Models (Fisher's Score)

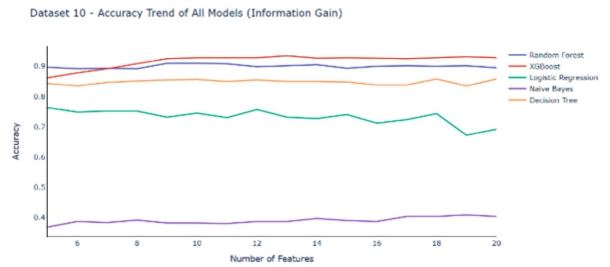

Fig. 5. Accuracy Trend of All Machine Learning Models (Information Gain)

Table 8. Performance of Machine Learning Models (Boruta)

Dataset	Classification Model	Number of Features	Accuracy Score	Precision	Recall	F1 Score
5	Random Forest Classifier	20	0.9283	1.0	0.9295	0.9279
5	XGBoost Classifier	20	**0.9467**	**0.9465**	**0.9481**	**0.9465**
5	Logistic Regression	13	0.8217	0.8209	0.82	0.8185
5	Naive Bayes Classifier	13	0.7300	0.7386	0.7147	0.6975
5	Decision Tree Classifier	16	0.8817	0.8809	0.8820	0.8805

Table 9. Performance of Machine Learning Models (RFE)

Dataset	Classification Model	Number of Features	Accuracy Score	Precision	Recall	F1 Score
5	Random Forest Classifier	14	0.9217	1.0	0.9206	0.9209
5	XGBoost Classifier	14	**0.9417**	**0.9412**	**0.9414**	**0.9411**
5	Logistic Regression	18	0.9300	0.9301	0.9303	0.9296
5	Naive Bayes Classifier	17	0.4900	0.5069	0.4734	0.4434
5	Decision Tree Classifier	6	0.8967	0.8956	0.8971	0.8959

runner-up model, Random Forest Classifier has the accuracy score of 0.9217 with 14 features.

By using the selected features by Lasso, XGBoost Classifier has achieve the highest accuracy among all models constructed, which is 0.8983 with 20 features. The runner-up model, Random Forest Classifier has the accuracy score of 0.8967 with 17 features.

Random Forest Classifier has achieved the highest accuracy among all models constructed by using the selected features by Fisher's Score, which is 0.8700 with 19 features. The runner-up model, XGBoost Classifier has the accuracy score of 0.8600 with 20 features.

XGBoost Classifier has achieve the highest accuracy among all models constructed by using the selected features by Information Gain, which is 0.9350 with 13 features. The runner-up model, Random Forest Classifier has the accuracy score of 0.9100 with 10 features.

Table 10. Performance of Machine Learning Models (Lasso)

Dataset	Classification Model	Number of Features	Accuracy Score	Precision	Recall	F1 Score
5	Random Forest Classifier	17	0.8967	1.0	0.8970	0.8960
5	XGBoost Classifier	20	**0.8983**	**0.8977**	**0.8997**	**0.8978**
5	Logistic Regression	20	0.8333	0.8323	0.8310	0.8300
5	Naive Bayes Classifier	7	0.7467	0.7462	0.7339	0.7224
5	Decision Tree Classifier	16	0.8617	0.8631	0.8641	0.8609

Table 11. Performance of Machine Learning Models (Fisher's Score)

Dataset	Classification Model	Number of Features	Accuracy Score	Precision	Recall	F1 Score
5	Random Forest Classifier	19	**0.8700**	**1.0**	**0.8707**	**0.8694**
5	XGBoost Classifier	20	0.8600	0.8604	0.8603	0.8597
1	Logistic Regression	14	0.6517	0.6517	0.6524	0.6467
2	Naive Bayes Classifier	15	0.5650	0.5850	0.5672	0.5591
5	Decision Tree Classifier	20	0.7617	0.7743	0.7637	0.7616

Through the machine learning models that trained using five different feature selection models, XGBoost consistently achieve the highest accuracy score. Moreover, Random Forest Classifier as the runner-up model also achieve higher accuracy score than others machine learning models.

5.1.2 Deep Learning

Through MLP Classifier, the better performance can be achieved at around 56%. The figures below show that the accuracy trend of classification models of different dataset by using different top features that selected by Boruta (Fig. 6).

Dataset 10 has achieved the highest accuracy score among all dataset, the highest accuracy score of dataset 10 is 0.5600 with 18 features. Besides, Dataset 6 also has a high accuracy score of 0.5150 with 17 features.

Table 12. Performance of Machine Learning Models (Information Gain)

Dataset	Classification Model	Number of Features	Accuracy Score	Precision	Recall	F1 Score
5	Random Forest Classifier	10	0.9100	1.0	0.9068	0.9079
10	XGBoost Classifier	13	**0.9350**	**0.9366**	**0.9334**	**0.9343**
5	Logistic Regression	16	0.7717	0.7738	0.7701	0.7697
6	Naive Bayes Classifier	11	0.6817	0.6848	0.6748	0.6607
5	Decision Tree Classifier	20	0.8633	0.8623	0.8626	0.8623

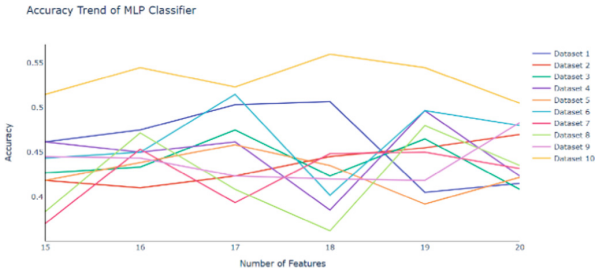

Fig. 6. Accuracy Trend of MLP Classifier (Boruta)

5.2 Deployed Dashboard

In this study, a web-based dashboard will be deployed to show the output of predictive model by using Streamlit. The highest accuracy model among all the predictive models will be chosen as the model of the dashboard.

The model that selected as the predictive model of dashboard is XGBoost Classifier with top 20 features by using Boruta Feature Selection and the dataset that chosen is dataset 5 which is the combination of Pizza Hut, Nando's and Secret Recipe as the base data. The figure below shows that the overview of dashboard. The dashboard enables users to view all of the POI and details of nearby properties by clicking on the map. The map is interactive which allows users to drag and zoom the map for a better view of the map. Moreover, once the map is clicked or selected, the coordinate of the pinpoint will be prompt with its longitude and latitude for its exact location (Fig. 7).

6 Conclusion

This paper had aimed to construct the analytical datasets to construct the models for retail site selection. The raw datasets were obtained from different sources and transformed into analytical datasets. Feature Selection was done in this paper to select the

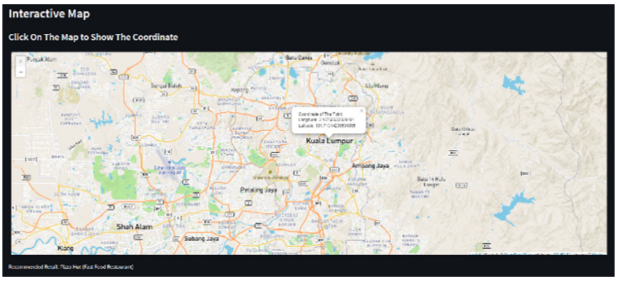

Fig. 7. Interactive Map of Dashboard

important features, using Boruta, Recursive Feature Elimination, Lasso, Fisher's Score and Information Gain. Besides, the models were evaluated by using accuracy score, precision, recall and F1 Score. Lastly, the dashboard has been deployed by using the highest accuracy model.

The techniques and approaches used throughout this paper have built a foundation in retail site selection. Despite the potential shown, the work of this research study is fraught with challenges. Future works of the paper should consider including other related features and variables. The retail market is seemingly more complex as it tends to be affected by factors with the exception of the features considered in this paper. Some of the geographical and demographic features should be consider as the future research in this paper, such as nearby transportation, education and economy features.

References

1. Ganesan, S., George, M., Jap, S., Palmatier, R. W., & Weitz, B. (2009). Supply chain management and retailer performance: emerging trends, issues, and implications for research and practice. Journal of retailing, 85(1), 84–94
2. Guy, C. M. (1980). Retail location and retail planning in britain. Gower Publishing Company, Limited.
3. Mui, L. Y., & Ghafar, A. (2003). Retail activity in malaysia: from shophouse to hypermarket. In Pacific rim real estate society 9th annual conference (Vol. 20, p. 22).
4. Hossin, M. S., Sentosa, I., & Miah, M. S. (2020). The impact of covid-19 outbreak on human resource operation: Empirical evidence from the perspective of malaysian retail employees in klang valley. International Journal of Organizational Leadership, 9(4), 304–320.
5. Ahmad, N. N., Hanafi, W. N. W., Abdullah, W. M. T. W., Daud, S., & Toolib, S. N. (2020). The effectiveness of additional prihatin sme economic stimulus package (prihatin sme+) in malaysia post-covid-19 outbreak: A conceptual paper. Global Business & Management Research, 12(4).
6. Maliszewska, M., Mattoo, A., & Van Der Mensbrugghe, D. (2020). The potential impact of covid-19 on gdp and trade: A preliminary assessment. World Bank Policy Research Working Paper(9211).
7. Wang, L., Fan, H., & Wang, Y. (2018). Site selection of retail shops based on spatial accessibility and hybrid bp neural network. ISPRS International Journal of Geo-Information, 7(6), 202.
8. Ting, C.-Y., Ho, C. C., Yee, H. J., & Matsah, W. R. (2018). Geospatial analytics in retail site selection and sales prediction. Big data, 6(1), 42–52.

9. Liu, J., Li, Q., Qu, M., Chen, W., Yang, J., Xiong, H., . . . Fu, Y. (2015). Station site optimization in bike sharing systems. In 2015 ieee international conference on data mining (pp. 883–888).
10. LaGro Jr, J. A. (2013). Site analysis: Informing context-sensitive and sustainable site planning and design. John Wiley & Sons.
11. Erbıyık, H., Özcan, S., & Karaboˇga, K. (2012). Retail store location selection problem with multiple analytical hierarchy process of decision making an application in turkey. Procedia-Social and Behavioral Sciences, 58, 1405–1414.
12. Xu, M., Wang, T., Wu, Z., Zhou, J., Li, J., & Wu, H. (2016). Demand driven store site selection via multiple spatial-temporal data. In Proceedings of the 24th acm sigspatial international conference on advances in geographic information systems (pp. 1–10).
13. Lin, J., Oentaryo, R., Lim, E.-P., Vu, C., Vu, A., & Kwee, A. (2016). Where is the goldmine? finding promising business locations through facebook data analytics. In Proceedings of the 27th acm conference on hypertext and social media (pp. 93–102).
14. Mazhi, K. Z., Suryana, L. E., Davi, A., & Dewi, W. R. (2020). Site selection of retail shop based on spatial analysis and machine learning. In 2020 international conference on advanced computer science and information systems (icacsis) (pp. 135–140).
15. Wang, J., Tsai, C.-H., & Lin, P.-C. (2016). Applying spatial-temporal analysis and retail location theory to public bikes site selection in taipei. Transportation Research Part A: Policy and Practice, 94, 45–61.
16. Kamali, M., Alesheikh, A. A., Khodaparast, Z., Hosseinniakani, S. M., & Borazjani, S. A. A. (2015). Application of delphi-ahp and fuzzy-gis approaches for site selection of large extractive industrial units in iran. Journal of settlements and Spatial planning, 6(1), 9–17.
17. Jelokhani-Niaraki, M., & Malczewski, J. (2015). A group multicriteria spatial decision support system for parking site selection problem: A case study. Land Use Policy, 42, 492–508.
18. Ting, C.-Y., Ho, C. C., & Yee, H.-J. (2020). Geospatial insights for retail recommendation using similarity measures. Big data, 8(6), 519–527.
19. Cabello, J. G. (2019). A decision model for bank branch site selection: Define branch success and do not deviate. Socio-Economic Planning Sciences, 68, 100599.
20. Niu, H., Liu, J., Fu, Y., Liu, Y., & Lang, B. (2016). Exploiting human mobility patterns for gas station site selection. In International conference on database systems for advanced applications (pp. 242–257).
21. Erdin, C., & Akba¸s, H. E. (2019). A comparative analysis of fuzzy topsis and geographic information systems (gis) for the location selection of shopping malls: A case study from turkey. Sustainability, 11(14), 3837.
22. Rohani, A. M. B. M., & Chua, F.-F. (2018). Location analytics for optimal business retail site selection. In International conference on computational science and its applications (pp. 392–405).
23. Hui-Jia, Y., Choo-Yee, T., & Ho, C. C. (2018). Human elicited features in retail site analytics. In Top conference series. earth and environmental science (Vol. 169).
24. Damavandi, H., Abdolvand, N., & Karimipour, F. (2019). Utilizing location-based social network data for optimal retail store placement. Earth Observation and Geomatics Engineering, 3(2), 77–91.
25. Yee, H.-J., Ting, C.-Y., & Ho, C. C. (2019). Retail site selection using machine learning algorithms.
26. Miao, J., & Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91, 919–926.
27. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70–79.
28. Nti, K. O., Adekoya, A., & Weyori, B. (2019). Random forest based feature selection of macroeconomic variables for stock market prediction. American Journal of Applied Sciences, 16(7), 200–212.

29. Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and pca. Informatics in Medicine Unlocked, 19, 100330.
30. Haq, A. U., Zhang, D., Peng, H., & Rahman, S. U. (2019). Combining multiple feature-ranking techniques and clustering of variables for feature selection. IEEE Access, 7, 151482-151492. doi: https://doi.org/10.1109/ACCESS.2019.2947701
31. Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2020). Feature selection using an improved chi-square for arabic text classification. Journal of King Saud University-Computer and Information Sciences, 32(2), 225–231.
32. Chavent, M., Genuer, R., & Saracco, J. (2021). Combining clustering of variables and feature selection using random forests. Communications in Statistics-Simulation and Computation, 50(2), 426–445.
33. Rado, O., Ali, N., Sani, H. M., Idris, A., & Neagu, D. (2019). Performance analysis of feature selection methods for classification of healthcare datasets. In Intelligent computing-proceedings of the computing conference (pp. 929–938).
34. Senvar, O., Otay, I., & Bolturk, E. (2016). Hospital site selection via hesitant fuzzy topsis. IFACPapersOnLine, 49(12), 1140–1145.
35. Bhole, J., Nandiyawar, S., Pawar, S., & Vora, P. (2020). Smart site selection using machine learning.
36. Chang, K.-T. (2016). Geographic information system. International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology, 1–9.
37. Olson, D. L. (2004). Comparison of weights in topsis models. Mathematical and Computer Modelling, 40(-8), 721–727.
38. Aviso, K. B., Janairo, J. I. B., Promentilla, M. A. B., & Tan, R. R. (2019). Prediction of co 2 storage site integrity with rough set-based machine learning. Clean Technologies and Environmental Policy, 21(8), 1655–1664.
39. Al Garni, H. Z., & Awasthi, A. (2017). Solar pv power plant site selection using a gis-ahp based approach with application in saudi arabia. Applied energy, 206, 1225–1240.
40. Sahin, T., Ocak, S., & Top, M. (2019). Analytic hierarchy process for hospital site selection. Health Policy and Technology, 8(1), 42–50.
41. Kumar, S., & Bansal, V. (2016). A gis-based methodology for safe site selection of a building in a hilly region. Frontiers of architectural research, 5(1), 39–51.
42. Messaoudi, D., Settou, N., Negrou, B., Rahmouni, S., Settou, B., & Mayou, I. (2019). Site selection methodology for the wind-powered hydrogen refueling station based on ahp-gis in adrar, algeria. Energy Procedia, 162, 67–76.