International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Smart Site Selection using Machine Learning

Jigisha Bhole[1], Sushma Nandiyawar[2], Sneha Pawar[3], Palak Vora[4]

[1-4]BE Student, Dept. of Computer Engineering, Sinhgad Institute of Technology and Science, Pune, India

---

Abstract - *Being a largely populated country, India struggles to evenly distribute the amenities. With this hurdle, it is hard to invest at a location without a proper survey. This study proposes a model that will aid the businessman in site selection. The suitability analysis is done through the Analytic Hierarchy Process (AHP) model using various factors. The unit of study is a commercial business and taking a fashion point, a food joint and a grocery store as an example, we have come up with a Machine Learning (ML) model to predict the rating of the site based on profitability.*

Key Words: Site Selection, AHP, Machine Learning

1. INTRODUCTION

Site selection is an essential component of a business expansion. Suitability analysis is the process used to establish the suitability of a system according to the stakeholder. The rapid increase of urban population has caused a high-level impact on the urban environment and is creating many problems such as unplanned sprawl, inadequate housing facilities, traffic congestion, insufficient drainage, sewerage problem, and lack of other amenities. In this context, finding a suitable area for further development is difficult. To find a suitable site for the construction of an amenity, it is required to use sophisticated analysis with consideration of large numbers of critical issues such as technical, environmental, physical, social, and many others. Site suitability analysis is the process of determining the fitness of a given tract of land for a defined use. Remote Sensing, Geographic Information System (GIS) and Analytic Hierarchy Process (AHP) method is a vital tool for identification, comparison, and Multi-Criterion Decision - Making (MCDM) analysis of urban development. The Machine Learning (ML) model will calculate the weights of different factors and compute the probability y of profit for the site. The importance of location in business success cannot be understated. An organization's place of business is where its customers evaluate and ultimately receive their product or service. Establishing business at the right location is critical for restaurants, retailers, and many service businesses. The importance of the location strategy is especially important for these small ventures because it impacts the business. Finding the right location for your business can make a huge impact on its performance, and it is more than just choosing a building.

Selecting the right location of business requires a lot of fieldwork. The survey takes up a lot of time and money. The proposed system produces accurate information in the form of rating which gives the entrepreneur a brief idea about the profitability of his/her business. In the real world, this system will allow the user to gain access to a real-time review of his/her chosen location for business without going through an extensive process of manual survey. Thus, decreasing the investment put into the selection of site without compromising the quality.

2. LITERATURE SURVEY

T. Sahin et al.[1]suggested decision supports a model for site selection of hospital-based on AHP. AHP was used for selecting the best site for hospital among various alternatives. AHP is a theory of measurement through pairwise comparisons and relies on the judgments of experts to derive priority scales. AHP is a decision-making model that consists of three parts: identifying and organizing decision objectives, criteria, and alternatives into a hierarchy; evaluating pairwise comparison; and synthesizing using solution algorithm of the result of the pairwise comparison. The authors discussed a case study of finding the best site for hospitals in the Mugla province of Turkey and consider 6 main criteria and 19 sub-criteria for evaluating the result. The 6 steps are provided by which we can solve this problem. Statistical analysis was used to determine the best site of the available locations, and also the sensitivity analysis of alternatives ranking was performed. However, a limited number of criteria and sub-criteria were taken into consideration, all criteria were not considered. Therefore, the results cannot be generalized. They discussed finding the criteria and sub-criteria, weightage to give to the criteria using the Saaty scale. Implemented AHP for deciding the criteria weights.

C. Kamps et al.[2]proposed a method to support decision-making when all the factors are not known at a time. It explains the MCDM method used in AHP with the gradient descent process. Gradient descent helps in adjusting the weightage of the factors with every calculation. C. Kamps et.al compares different existing methods and proposes an idea to overcome limitations using machine learning. The suggested method is aimed to improve the ability of the decision-making tool to predict outcome given any subset of the complete input set. Inability to improve situations when the number of available factors is low due to limited flexibility in weight distribution. As we have a large dataset, this method is efficient only to some extent. Weight adjustment technique through MCDM is used. Gradient descent is used to vary weights according to the training. Learning and filling gaps where there is an incomplete set of data.

J. A. Parry et al.[3]suggested an AHP model using a set of geophysical and socioeconomic variables. These variables include slope, altitude, land use/land cover, and existing

amenity status. The use case studied is a municipal ward. For better urban planning and suitable decision-making, the study provides information not only on the existing urban land use pattern and existing amenity status, but also on the suitability of land for the establishment of urban amenities in the future. Different factors, and the method for data collection was recognized from this. The study was very effective for its cause but urban planning is a future scope of the proposed system. As it is only based on a hilly region, the study is not applicable all over India.

G. Zhao et al.[4]discussed a system that recommended places for meteorological observation stations. The study involved a prediction model along with the recommendation model. The prediction model concept was taken up by our system. The authors also made it possible to add different factors for consideration of the recommendation which allowed the system to be more configurable.

A. Tiroshi et al.[5]theorized a system which uses a graph-based approach to predict ratings of a business. This study is useful for recognizing business-location and business-category relation. Although the system was not validated and the results were not checked with real-life applications.

L. Hu et al.[6]considered both extrinsic and intrinsic variables for rating of a venture. The study mainly focused on the geographical neighbors. The proposed system gives a rating for a new location whereas, the study discussed computes rating of an existing business. The factors taken into consideration were used for the proposed system.

M. H. Satman et al.[7]proposed a system that uses Google places Application Program Interface (API) for geographical data collection. The data collected merged with data of the retail stores already present gives the rating of a location concerning how profitable the site could be. Artificial Neural Network (ANN) is used to optimize the system into a more sophisticated one with the feature of establishing a relationship among the factors taken into consideration. This process resembles the proposed system concerning the concept but using Google Places API on a large scale requires buying the privileges. The free version can only be used for a certain amount of requests and with a limitation to the access. Artificial Neural Network (ANN) cannot work in all cases and as their survey also yielded that it didn't work for all the retail stores.

A. Rikalovic et al.[8]proposed GIS are used in conjunction with other systems and methods such as Decision Support System (DSS) for decision-making and the method for MCDM for Industrial site selection. The focus was whether the MCDM methods can be efficiently used as a decision support tool for industrial site selection concerning the proposed model. Multi-Criteria Decision Analysis (MCDA) which includes a complex array of factors involving multiple factors. The MCDA in GIS should be viewed as a process of the conversion of data to information that adds extra value to the original data. There are many multi-criteria decision methods that are nowadays in use in the GIS environment. The most commonly used analysis is: AHP, Weighted Linear Combination (WLC) and Ordered Weighted Averaging (OWA). The final results MCDA in GIS are a recommendation for future action for decision-maker presented in the form of a suitability map using the AHP method. However, the visualization of results was not that clear to take the decision efficiently. As future research involves exploring new methods and developing software that gives better clarity for taking decisions.

R. B. Bhagat [9]summarizes a system that help in restructuring of the local governance. The proposed system uses the urban and rural classification method to recognize the type of place the venture is being established in. As the author only focuses on any of the rating factor, it is a misfit with the proposed system. The study also focuses on planning and structuring of a city which will help the government. Much of the data has been collected through a census that occurs in India every 10-11 years which was a major data source for the proposed system.

M. Vlachopoulou et al.[10]proposed the method that can be used for the warehouse site selection decisions. The authors used the geographic decision support system that is the combination of the GIS and the decision support system. GIS is a system designed to capture, store, manipulate, and analyze data. As they research Warehouse site location analysis, particularly with the increased availability of computer-based techniques, can provide invaluable information to assist warehouse and marketing management with their decision-making process. Site selection can be done by considering some factors. This paper aims to develop a geographic decision support system for the warehouse site selection process. They suggested using the multi-criteria evaluation model which involves selection of factor from the database, evaluation, and the scoring of the factor values and then pairwise comparison of weights, and the factors, the weights starting from minimum value to maximum value, calculation of Consistency Ratio (CR) is done to get the location of the warehouse with the highest score tested. One important advantage of the model is that, once the relevant factors, and their weightings have been set up, the model can be used for site assessment by the personnel who have little knowledge of site location theory. However, the implementation of image mapping is expensive. There is no clear dataset for India. Multiple criteria evaluation model can be used in the project. As it contains pairwise comparison of different factors with their weights. This model consists of the weights starting from minimum value to maximum value. Calculation of CR is done to get the location of the warehouse with the highest score tested.

3. METHODOLOGY

The proposed system consists of a mixture of different algorithms namely AHP and Random Forest. AHP for making dataset and Random Forest for predicting the rating. For research purposes other algorithms like Multilinear classifier and decision tree have also been used. Initially, the system needs a good amount of data from various sources.

3.1 Data Gathering

This step focuses majorly on gathering data based on different factors. The factors decided for every use case has to have large amounts of data rows for the model to learn properly. Data collection is done through a combination of

different sources and dynamic scraping. Factors considered for every factor cannot have the same amount of weight or the same importance. Deciding which factor weighs more for which business is a crucial part of this system. Manually assigning the weights be very risky and can result in the building of a wrong model. To get real-time importance of the factors, a quick survey is passed on to various businessmen. The results of this survey can determine the appropriate position of the factors. The importance was taken on a scale of 1 to 5.

Table -1: Factor Importance Scale

Importance Scale	Value
1	Most Important
2	More Important
3	Important
4	Less Important
5	Least Important

Collected data is then processed, cleaned, and verified. Only the data which has been validated can go forward as the input for the model. This will help the model in the correct computation as data is a major part of every machine learning model.

3.2 Preparing Dataset using AHP for Model Analysis

The average importance for each factor is considered from the data collected through the survey. Using the importance the AHP algorithm gives the final criteria weights of the factors. The large amount of data row collected through scrapping and different resources, which is the quantity of each factor near the site location. Each row corresponds to one site location. The data is processed, cleaned, and verified. The criteria weights obtained are used for computing the rating of each row and the dataset is prepared. Only the data which has been validated can go forward as the input for the model. This will help the model for correct computation as data is a major part of every ML model.

3.3 Model Comparison

There are many algorithms currently being used for different ML models. To decide the correct algorithm concerning the type of data used is again a very important task. This can be decided through given different algorithms the same input dataset with the same weights of the factors. The accuracy of these models can be compared to come up with the best algorithm. To determine that the accuracy is the best, various methods like increasing data size, changing the number of factors, etc. can be used

4. IMPLEMENTATION

For this particular system, three different algorithms are used. All of them have specific properties that suit the varied factors chosen the best. The algorithms used include: Random Forest, Multi-Linear Regression and Decision Tree. The library scikit-learn makes ML coding easy and provides customization as and when needed. Using the library all the algorithms were customized for the specific model.

Special function like: fit() and predict() were used in this system. While fit() is used to fit the model, predict() helps to get results after a model is trained. Random Forest scikit-learn provides a RandomForestClassifier function which can be defined by the user following the guidelines. It is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size. The first step in the implementation of the Random Forest algorithm is initializing the function which is used from the predefined library.

RandomForestRegressor(n_estimators=n_estimators, random_state=random_state) For training and testing to the system uses the predefined functions which can be changed according to the requirements.

fit(train_features,train_target) and predict(test_features) Similarly, for decision trees the DecisionTreeClassifier function provided by scikit library is used. The library has few predefined fucntions which can be used to our advantage. The functions used in this system include:

DecisionTreeRegressor(max_features=max_features max_depth= max_depth)

fit(train_features,train_target)

A Decision Tree is made for this algorithm which can be done by following steps. Get a list of rows (dataset) that are taken into consideration for making a decision tree (recursively at each node). Calculate uncertainty of our dataset or Gini impurity or how much our data is mixed up etc. Generate a list of all questions which need to be asked at that node. Partition rows into True rows and False rows based on each question asked. Calculate information gain based on Gini impurity and partition of data from the previous step. Update the highest information gain based on each question asked. Update the best question based on information gain (higher information gain). Divide the node on the best question. Repeat from step 1 until we get pure node (leaf nodes).

predict(test_features)

LinearRegrression is a function used to implement Multi-Linear regression in this system. The initialization is done with the given definition :

LinearRegression(fit_intercept = fit_intercept, normalize = normalize)

The next step is to fit the dataset into the model which is attained by defining the fit function.

fit(train_features,train_target)

The weights are measured by MSE (Mean Squared Error) and adjusting them to get the best possible Linear line. To gain optimal result we need to minimize MSE. So, to minimize this error or MSE we use gradient descent to find the weights after MSE or error rate calculation. Lastly, the weights are updated every time to reduce the error. After the model is

trained, the next part is testing the system and see if the desired results are obtained or not. It is done by passing the test dataset to the predict function provided by the scikit-learn library.

predict(test_features)

As we can see that most of the function definition is the same, and thus it is very efficient and easy to use for many algorithms. The system is built to predict the user data value using the model which gives the best accuracy concerning the factors chosen by the client. This helps various options and personalized optimization according to the category of the business and needs of the venture. After careful comparison the above-mentioned algorithms were used as they fit the requirements the best. The trained model can be stored using pickle for it to be loaded later and used for prediction.

5. CONCLUSION

This paper puts forward a system that provides the degree of profitability of a given business location. Categories of businesses explored here are Fashion Point, Food Joint, and Groceries. Different factors for each of these use cases are taken into consideration. The best ML algorithm among compared algorithms will be used to predict the degree of profitability of the business site. Users will be able to compare which location is best for the specified category of venture.

6. FUTURE SCOPE

The proposed system will include consideration of only commercial businesses (Grocery, Fashion Point, Food Joint). In future research, we recommend to consider more business types and make the system generalized for businesses. The location will be analyzed based on various factors, and a rating of the desired location will be output. The output will contain only the main factors which affected the rating. This proposed system will store the previous rating for types of businesses and extract information from the same. Limitations for this proposed system is it cannot calculate the future rating for the business site.

ACKNOWLEDGEMENT

We would like to offer our sincerest gratitude to our mentor Mr. Hardik Gandhi (Founder, CEO Zestl Software Pvt. Ltd.) who from the beginning motivated and guided us.

REFERENCES

[1] T. Sahin, S. Ocaka, M. Topb, "Analytic hierarchy process for hospital site selection", Health Policy and Technology 8, 2019, pp. 42–50.

[2] C. Kamps, R. Jassemi-Zargani, "Weight adjustment using machine learning applied to analytical hierarchy process", International Symposium on the Analytic Hierarchy Process, Hong Kong, HK., 2018.

[3] J. A. Parry, S. A. Ganaie, M. S. Bhat, "GIS based land suitability analysis using AHP model for urban services planning in Srinagar and Jammu urban centers of J&K, India", Journal of Urban Management 7, 2018, pp. 46-56.

[4] G. Zhao, T. Liu, X. Qian, T. Hou, H. Wang, X. Hou, Z. Li, "Location recommendation for Enterprises by Multi-source Urban Big Data Analysis", IEEE Transactions on Service Computing, Manuscript, 2017.

[5] A. Tiroshi, S. Berkovsky, M. A. Kaafar, D. Vallet, T. Chen. T. Kuflik, "Improving business rating predictions using graph based features", !9th International Conference on Intelligent User Interface, 2014, pp. 17-26.

[6] L. Hu, A. Sun, Y. Liu, "Your neighbours affect your ratings: on geographical neighbourhood influence to rating prediction", 37th international ACM SIGIR Conference on Research and Development in Information Retrieval, 2014, pp. 345-354.

[7] M. H. Satman, M. Altunbey, " Selecting Location of Retail Stores Using Artificial Neural Networks and Google Places API", Int. J. of Statistics and Probability, Vol 3, 2014.

[8] A. Rikalovic, I. Cosic, D. Lazarevic, "GIS Based Multi-Criteria Analysis for Industrial Site Selection", 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013, pp. 1054 – 1063.

[9] R. B. Bhagat, "Rural – urban classification and municipal governance in India", Singapore Journal of Tropical Geography, 26(1), 2005, pp. 61-73.

[10] M. Vlachopoulou*, G. Silleos, V. Manthou, "Geographic information systems in warehouse site selection decisions", Int. J. Production Economics 71, 2001, pp. 205-212.

BIOGRAPHIES

Jigisha Bhole
BE Computer Engineering Student

Sushma Nandiyawar
BE Computer Engineering Student

Sneha Pawar
BE Computer Engineering Student

Palak Vora
BE Computer Engineering Student