Rating Prediction of Tourist Destinations Based on Supervised Machine Learning Algorithms

Anupam Jamatia, Udayan Baidya, Sonalika Paul, Swaraj DebBarma and Subham Dey

Abstract The paper highlights the process of predicting how popular a particular tourist destination would be for a given set of features in an English Wikipedia corpus based on different places around the world. Intelligent predictions about the possible popularity of a tourist location will be very helpful for personal and commercial purposes. To predict the demand for the site, rating score on a range of 1–5 is a proper measure of the popularity of a particular location which is quantifiable and can use in mathematical algorithms for appropriate prediction. We compare the performance of different machine learning algorithms such as Decision Tree Regression, Linear Regression, Random Forest and Support Vector Machine and maximum accuracy (74.58%) obtained in both the case of Random Forest and Support Vector Machine.

Keywords Prediction · Machine learning · Decision tree regression · Linear regression · Random forest · Support vector machine

1 Introduction

Tourism has become one of the most popular industry in the world today. To get a share of this increasing tourism market, government and private agencies alike are always keen on investing in the right places that have the possibility of becoming

A. Jamatia (✉) · U. Baidya · S. Paul · S. DebBarma · S. Dey
Department of Computer Science and Engineering,
National Institute of Technology Agartala, Jirania 799046, Tripura, India
e-mail: anupamjamatia@gmail.com

U. Baidya
e-mail: udayanbaidya@gmail.com

S. Paul
e-mail: sonalikapaul@gmail.com

S. DebBarma
e-mail: swarajdebbarma175@gmail.com

S. Dey
e-mail: subham6666@gmail.com

H. S. Behera et al. (eds.), *Computational Intelligence in Data Mining*, Advances in Intelligent Systems and Computing 990,
https://doi.org/10.1007/978-981-13-8676-3_11

tourist attractions. A system developed to predict the rating of a site based on its attributes can be used by government agencies and private agencies alike while planning to invest in a tourist destination. This project attempts to solve this crucial problem by using machine learning models and data from websites and already present user ratings to predict new scores which can be used for proper planning and execution on the part of the agencies. Machine learning systems are being used in all fields in our modern world. The use of this technology in the field of tourism is needed for proper prediction of decisions.

The prime motivation for this research comes from the need to help the government and private agencies to plan and execute a tourism attraction project properly. For a proper tourist attraction, the location chosen is of prime importance. Such a system that can aid agencies in selecting the correct area by giving a predicted rating of a place would be of utmost importance and a boon for the planners.

The primary goal of our research is to create a system that can accurately predict the rating of a potential tourist location on a scale from 1–5. Such a system should be able to quickly give an accurate prediction about the possible score of a tourist location based on given attributes of a place. High accuracy is the primary target and so proper data collection, and adequate model training is of utmost importance. Decision Tree Regression, Linear Regression, Random Forest and Support Vector Machine are the machine learning algorithms that are used.

Previously, hand coding rules were involved in many language-processing tasks which did not give any powerful solution to process many natural language processing operations. The statistical machine-learning paradigm automatically learns such rules through the analysis of large corpora available in typical real-world examples. Different types of supervised and un-supervised machine learning algorithms have been applied to Natural Language Processing tasks. These algorithms work by taking a broad set of features as input. In our research, statistical approach using Decision Tree Regression, Linear Regression, Random Forest and Support Vector Machine are being used to solve the problem of rating prediction.

2 Related Works

Prediction of tourist locations and its rating is a very new field, and some very interesting algorithms have been applied to predict their popularity. One such algorithm is the One Slope Algorithm, which is a recommender, as used by Hu and Zhou [1], which describes how using the one slope algorithm, recommendations can be produced with very high efficiency. A good point about one slope algorithm is that it is effortless to implement. One slope algorithm is a unique form of item-based collaborative filtering (item-based CF) proposed by Lemire and Maclachlan [2]. But the way attributes are selected, and their values assigned in Hu and Zhou's paper is quite similar to the one we used in this project.

There are a lot of methods to predict ratings based on attributes and community provided scores, as discussed by Marović et al. [3], where they discussed various

ways. One such method is through decision trees proposed by Li and Yamada [4] for movie rating prediction. Another technique uses neural networks to provide web page rating [5]. Algorithms such as k-Nearest neighbors heuristic methods can be used to cluster the ratings based on attributes and then give predictions based on these clusters. Resnick et al.[6] used such a system for news filtering based on predicted ratings. Accurate prediction of tourism data is utmost essential for the success of tourism industry. Time Series prediction has been used to significant effect in tourism industry predictions. Koutras et al.[7] suggested various linear and non-linear models using systems like multi-layer perceptron models, support vector regressor with polynomial kernel models, Support Vector Regressor with Radial Basis Functions Kernels (SVR-RBF). Item-based collaborative filtering technique has been used in the past for recommending tourist locations as seen in this paper by Chen et al. [8]. Collaborative filtering works by collecting preferences from many users, it is based on the fact that users with similar tastes in a particular paper might have the same feeling on another topic and that can be used for giving the recommendation. A few websites that use collaborative filtering system are DieToRecs, TripAdvisor, Heracles, TripSay.

Automatic feature extraction methods are increasingly being used for extracting relevant features from text for many classification and regression tasks. Automatic feature extraction has shown to produce better results than using manual features for a variety of tasks. Metrics such as word frequency, tf-idf [9] and even modified metrics using tf-idf like delta tf-idf have been used to score features in text as show by Martineau et al. [10].

3 Methodology

The objective of this paper is to present a methodology to predict the success of a city as a tourist destination. The machine learning approaches used in this paper are Decision tree regression, Linear Regression, Random Forest and Support Vector Machine.

Decision Tree: Decision Tree Learning [11] uses a decision tree data structure where decisions are made on each node of the tree to arrive at conclusions about the item's target value which are represented as the leaf nodes. It is one of the predictive modeling techniques being used in many fields of mathematics and computer science like statistics, data mining and machine learning. Tree models can be of two types—classification trees and regression trees. A tree model where the target variable can take a discrete set of values are called classification trees, in this case leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable is of continuous values are called regression trees. A decision tree, in other words, is an acyclic graph with a fixed root. A node describes an attribute in the data, and the edges define a decision based on this attribute. Even though most examples use binary decisions, it is important to note that a node can have as many edges as they want. In operations research, these

trees are understood as decisions and consequences. In general, the tree is traversed from the root, using the attribute in the node to choose a new node. Depending on the task, the leaf has a different meaning. Decision trees used in data mining are two main types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g., the price of a house, or a length of day at a hotel).

Decision tree regression is being used for our research. Regression differs from classification algorithms as in regression we predict a continuous output on a given range and not from a discrete set of values. Regression with decision trees is very similar to classification since the trees are taking data and classifying it to a regression model or even previously computed value.

Linear Regression is used in statistics as a linear approach for modelling the relationship between a scalar dependent variable y and one or more independent variables denoted X. Regression which consists of only one independent variable is called Simple Linear Regression. For more than one independent variable, the process is called multiple Linear Regression. Multiple Linear Regression is distinct from multivariate Linear Regression, where multiple correlated dependent variables are predicted, rather than a single dependent variable [12]. In linear models, linear predictor functions are used to model relationships in linear regression whose unknown model parameters are estimated from the data. Linear Regression focuses on the conditional probability distribution of the output given a set of input, rather than on the joint probability distribution of the all the variables inputs and outputs, which is the domain of multivariate analysis. The practical applications of Linear Regression may fall into one of the following two categories:

- Linear Regression may be used to fit a predictive model to an observed data set of y and X values if the goal is prediction or forecasting or error reduction. After developing such a model, if an additional value of X is then provided without its accompanying value of y then the fitted model can be used to make a prediction of the value of y.
- If we are provided a variable y and a number of variables X_1, ..., X_p that may be related to y, Linear Regression analysis can be applied to estimate the strength of the relationship between y and the X_j, to evaluate which X_j may have no relationship with y at all, and to identify which subsets of the X_j contain inessential information about y.

Random Forests, also defined as random decision forests are a type of ensemble learning method which can be used for classification, regression, etc; they operate by constructing a plenitude of decision trees during the training time and outputting the class that is the mode of the classes (as used in random forests for classification) or mean prediction (as used in random forests for regression) of the individual trees. Random decision forests are used to overcome decision trees' habit of over fitting to their training set. Random Forests use the popular method of bagging to train the

model. It works by taking a majority vote on each of the decision tree values and outputting the majority voted prediction as the prediction of the model [13].

In machine learning, Support Vector Machines (SVMs, also Support Vector Networks) are supervised learning models used for classification and regression analysis. SVM uses the training data and optimizes support vector points to create a boundary separating various classes in a 2D space. Each new point is mapped to an existing class in this 2D space and that is how classification works in a Support Vector Machine [14]. SVMs can be used to resolve various real world problems such as:

- SVMs can be used for text and hypertext categorization as their usage can significantly reduce the need for labeled training specimens in both the standard inductive and transductive settings.
- Image classification can also be done using SVMs. Experimental results show that SVMs are capable of achieving significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevant feedback.
- SVM can be used to recognize hand-written characters.
- SVM have been used in Biology and other sciences where they've been used to classify proteins with an higher accuracy of the compounds.

4 Datasets, Experiments and Results

This section provides data description, experimental implementation and results. The experiments were conducted on a dataset whose data points were collected from www.mouthshut.com.

4.1 Data sets

For the experiment, we used Beautiful Soup which is python package for parsing to collect the details of a total of 590 cities from www.wikipedia.org. The percentage of cities chosen from various regions are shown in Table 1.

Table 1 Percentage of cities from each continent

North America (%)	South America (%)	Africa (%)	Europe (%)	Asia (excluding India) (%)	Australia (%)	Antarctica (%)	India (%)
2.71	0.68	3.39	10	9.83	2.20	0	71.19

Table 2 Distribution of ratings of the tourist destinations when floored to integer values

Classification	1	2	3	4
% of Distribution	1.36	3.79	25.08	69.83

Table 3 Sample of the final dataset

City	Attributes					
	Mountain	Desert	Waterfall	...	Concert	Rating
Cape Town	1	0	0	...	1	4.83
Cairo	0	1	0	...	1	4.71
Lanzarote	1	1	0	...	0	4
Mauritius	1	0	0	...	1	3.84
...	...	...	...	...	...	...
Manchester	1	0	0	...	1	4.33
London	1	0	0	...	1	4.4

Ratings from 1 to 5 for the corresponding cities were collected from www.mouthshut.com. We have chosen a total of 20 attributes such as Mountains, River, Waterfall etc, that can be used to form our dataset which would be fed into the machine learning algorithm. Table 2 shows the distribution of ratings of the tourist destinations when floored to integer values.

To prepare the dataset for the experiment, we checked that if the cities have those pre-selected attributes present. A true and false value in the dataset indicates the presence or absence of the attributes for a particular city. The ratings are also inserted in the dataset which looks like the table shown in Table 3:

In total 20 attributes were chosen for the experiment. These attributes are mountain, desert, waterfall, beach, river, worship place, climate, zoo, park, travel, archaeological site, festival, pollution, tourist, cuisine, safety, museum, stadium, market, concert. These particular attributes have been chosen among the various other features of a location keeping in mind the factors which mostly affect tourism patterns. Natural Beauty such as mountain, beach, waterfall, desert etc are favourable. Hill stations generally have scenic beauty and cool climate which attract many tourists. Hence cool climate is added as one feature. A place with good connectivity will also have more tourists visiting that place. Amenities like Zoo, Park, Museum, Stadium will boost tourism. Further pollution and safety of people also determine number of tourist visiting. Concerts and grand festivals also contribute to tourism growth. In addition, tourists are quite likely to visit archaeological sites. A place with superb local cuisine is more likely to attract tourists. Keeping these factors in mind, we have chosen the aforementioned attributes.

The data set is used to train the machine learning model. The data set is split into two parts of training and test data set in an 80:20 ratio. Splitting of the data set is done on a random basis. Initially, the training set is selected as 80% of the corpus,

Table 4 Test set prediction instances for manual features

City	Golden value	Predicted value
Cape Town	4.83	4.01
Mauritius	3.84	4.01
Johannesburg	3.6	4.09
Nairobi	4.57	4.1
...	...	...
Oxford	4.0	4.21
London	4.4	4.21

and then the remaining 20% part is used as the test set. The training data set is used to train the various machine learning model. For each of the test set instances, the rating is predicted based on how the machine learning model is trained. A sample of the test set prediction instances for manual features using Decision Tree Regression is shown in Table 4.

4.2 Experiments

Using automatic feature extraction would make our algorithm compatible with a variety of machine learning tasks and variety of different corpus. Thus our algorithm would be a general fit for many tasks and in the process would usually yield better accuracy than manual selection of features. Keeping that in mind we decided to go for automatic feature extraction using tf-idf.

TFIDF or tf-idf stands for term frequency-inverse document frequency. In information retrieval, it indicates how important a word is to a document in a collection or corpus. It is most often used as a weighting factor in searches of information retrieval, text mining. The value of tf-idf is proportional to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, this helps to decrease the score for the words that appear in many documents in the corpus.

The tf-idf is the product of two statistics, term frequency and inverse document frequency.

– Term frequency The term frequency indicates raw count of a term in a document, i.e. the number of times that term t occurs in document d.
Therefore, Term Frequency

$$tf_{t,d} = \frac{\text{Number of times the given token } t \text{ appears in the document } d}{\text{Total number of tokens in the document } d}$$

Table 5 20 attributes selected by automatic feature extraction

Serial No.	Feature	Score
1	Area	0.131
2	Population	0.131
3	State	0.121
4	District	0.114
5	World	0.112
6	Century	0.097
7	Town	0.090
8	Government	0.089
9	South	0.085
10	Temple	0.079
11	Capital	0.073
12	North	0.073
13	University	0.067
14	Region	0.065
15	River	0.064
16	Year	0.063
17	East	0.062
18	Island	0.614
19	School	0.059
20	Airport	0.058

– Inverse document frequency The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It can be calculated as the logarithm of the fraction of the number of documents in the corpus to the number of documents containing the given term.
Therefore, inverse domain frequency

$$idf_t = \log_{10} \frac{\text{Total number of documents in the corpus}}{\text{Number of documents in which the term } t \text{ appears}}$$

The 20 attributes that were automatically selected by the tf-idf feature extractor from our corpus are listed in Table 5.

The next step was the dataset creation which was done the same way as the manual features were used for creating the dataset. After the dataset was created the next step was the machine learning model creation using the new dataset.

4.3 Results

The algorithm is run through different metrics and the accuracy score, mean squared error, F-measure score, precision score and recall score is found out.

In real life regression problems, one significant challenge always occurs in calculating the accuracy score. A regressor model working on floating point data can give varied results which have a very low probability of exactly matching with the golden value. The predicted values were rounded off to the nearest integer and then compared them with rounded off golden values that we had acquired from www.mouthshut.com. The different metrics scores using manual features and automatically generated features are shown in Table 6.

Finally, we see that the decision tree algorithm produces an accuracy of 54.24% using manual features and 70.34% using automatic features. The accuracy is calculated by seeing the number of instances that our algorithm could predict correctly to the total number of instances given to the algorithm for testing. This, in turn, was represented as a percentage. This accuracy implies that the algorithm can be used in real life scenarios where it is meant to be used to guide and give a brief idea to decision-makers about the probability for a tourist destination being successful. The predictions give a good idea about the mood of a reviewer about a particular place based on the attributes being provided. We did achieve a mean squared error score of 38.98% which should be reduced further. We converted our ratings into floored discrete integer value for use by algorithms such as SVM and random forests and we get an accuracy of 74.58%. Such accuracy is totally valid in cases where floating points are not a concern and the user just wants a brief idea of the possible popularity of a tourist destination. The F-measure value, recall, and precision are other metrics that can also be used to measure the effectiveness of the machine learning model thus created. The score achieved for F-measure was quite low, and this should

Table 6 Different metrics scores for manual features and automatically generated features

Feature extraction style	Metrics	Decision tree algorithm (%)	Linear regression algorithm (%)	Random forest algorithm (%)	SVM algorithm (%)
Manual	Accuracy	54.24	55.08	74.58	74.58
	MSE	48.31	51.69	34.75	34.75
	Precision	27.34	27.76	18.64	18.64
	Recall	28.53	29.42	25.00	25.00
	F-measure	27.90	28.51	21.36	21.36
Automatic	Accuracy	70.34	62.71	74.58	74.58
	MSE	38.98	39.83	34.75	34.75
	Precision	29.26	26.92	18.64	18.64
	Recall	29.06	27.72	25.00	25.00
	F-measure	29.12	27.29	21.36	21.36

be increased. A low precision and recall and comparatively higher accuracy show that we have less number of true positives and comparatively more number of true negatives for some classes. A proper statistical measure of the effectiveness of a model is always helpful to find out how good an algorithm is and how much more improvement can be made over it.

5 Conclusion

Predicting user ratings for tourist places presents an interesting and well-formed problem. The report has shown that machine learning and natural language processing technologies can be used for predicting the rating of a particular location and hence gather an idea about how popular the place would be if the place would be turned into a tourism hub. Decision tree regression has given a accuracy of 54.24% with manual features and 70.34% with automatic features extraction method when using decision trees. Using algorithms such as SVM we get an accuracy of 74.58%. The experiment result shows that this simple method is effective. Our work is just beginning, we will continue to improve the system, and improve its accuracy. This study can be of great significance for the practical application of assisting Government of India in deciding about investment in locations. This research can act as a benchmark to compare and produce a more accurate system in the future.

In the future, we will focus our work on using universal dependencies and using them to not include negated words. When the description of a location contains a sentence like- "The area is not mountainous", it would not be picked up by algorithm as a mountainous area as it is negated in the current context. Universal dependencies or Stanford dependencies would work well for this purpose. The Stanford typed dependencies representation was designed to provide an easy to understand system to extract relationships between different phrases and words for people not having linguistic expertise [15]. Also better data and attributes can be acquired for the dataset to further reduce the mean squared error, increase the accuracy, precision, and F-measure.

Acknowledgements Thanks to all the anonymous reviewer for extensive and helpful comments.

References

1. Hu, H., Zhou, X.: Recommendation of tourist attractions based on slope one algorithm. In: 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 1, pp. 418–421. IEEE (2017)
2. Lemire, D., Maclachlan, A.: Slope one predictors for online rating-based collaborative filtering. In: Proceedings of the 2005 SIAM International Conference on Data Mining, pp. 471–475. SIAM (2005)

3. Marović, M., Mihoković, M., Mikša, M., Pribil, S., Tus, A.: Automatic movie ratings prediction using machine learning. In: MIPRO, 2011 Proceedings of the 34th International Convention, pp. 1640–1645. IEEE (2011)
4. Li, P., Yamada, S.: A movie recommender system based on inductive learning. In: 2004 IEEE Conference on Cybernetics and Intelligent Systems, vol. 1, pp. 318–323. IEEE (2004)
5. Pazzani, M., Billsus, D.: Learning and revising user profiles: the identification of interesting web sites. Mach. Learn. **27**(3), 313–331 (1997)
6. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, pp. 175–186. ACM (1994)
7. Koutras, A., Panagopoulos, A., Nikas, I.A.: Forecasting tourism demand using linear and nonlinear prediction models. Acad. Tur.-Tour. Innov. J. **9**(1) (2017)
8. Chen, J.H., Chao, K.M., Shah, N.: Hybrid recommendation system for tourism. In: 2013 IEEE 10th International Conference on e-Business Engineering (ICEBE), pp. 156–161. IEEE (2013)
9. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the First Instructional Conference on Machine Learning, vol. 242, pp. 133–142 (2003)
10. Martineau, J., Finin, T., Joshi, A., Patel, S.: Improving binary classification on text problems using differential word features. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 2019–2024. ACM (2009)
11. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: icml, vol. 99, pp. 124–133 (1999)
12. Freedman, D.A.: Statistical Models: Theory and Practice. Cambridge University Press, Cambridge (2009)
13. Ho, T.K.: Random decision forests. In: Proceedings of the Third International Conference on Document analysis and Recognition, vol. 1, pp. 278–282. IEEE (1995)
14. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
15. De Marneffe, M.C., Manning, C.D.: The stanford typed dependencies representation. In: Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation, pp. 1–8. Association for Computational Linguistics (2008)