Data Mining in Indian Railways: A Survey to Analyze Applications of Data Mining

Jyoti Verma
Assistant Professor, ROFEL, Shri G.M. Bilakhia College of Applied Sciences (BCA), Vapi
Research student, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA)
Faculty of Computer Science and Applications (FCA)
Charotar University of Science & Technology (Charusat)

Jaimin N. Undavia, PhD
Associate Professor, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications (CMPICA)
Faculty of Computer Science and Applications (FCA)
Charotar University of Science & Technology (Charusat)

ABSTRACT

There are many means of transportation in the world but the most affordable and cheap mode of transportation is the railways in any part of the world. People focus on travelling by the trains the most during the vacation period or for any pre-planned meet. Here we have tried to study various papers which are concerned with the study related to railways in various parts of the world but I would like to focus on the Indian railways. Unusual patterns are unveiled in the papers through different predictions. The ticket booking patterns can be studied to find some unusual relation like level of satisfaction is based on the family type & marital status, revenue generated after refund of cancelled tickets, TATKAL booking ticket auction, delay in the train based on previous delay information, fraud detection in the ticket.

Keywords

RAC (Reservation against Cancellation), KNN (K-Nearest Neighbor), ARP-Advanced reservation period, clustering, Vickery Clarke Groves (VCG) mechanism, TTE, TATKAL

1. INTRODUCTION

Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data or pattern analysis. Data mining cannot be thought of as a single subject field it is a combination of various interdisciplinary subjects like statistics, image processing, pattern detection, medicine and many more. To discover the hidden pattern in the data the data is trained based on specific data mining algorithms. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

In India as per the information published in Wikipedia the India railways is the fourth largest railway network in the world by size, with a route length of 95,981-kilometre (59,640 mi) as of March 2019. People can book the tickets through the ticket window or online through the https://www.irctc.co.in/. the online ticket booking has proved to be a boon for the passengers travelling by train but it has also given the rise to overbooking of seats. Even the scheme of the government of booking the tickets beforehand means booking tickets before 3 months has also led to the cancellation of the tickets. In the computerized Passenger Reservation System (PRS), confirmed berths/seats are allotted on first come first served basis till the availability and thereafter Reservation Against Cancellation (RAC)/Waiting List tickets are issued. The status of RAC/Waiting List tickets get automatically updated against the cancellation of confirmed berths/seats and also against release of unutilized reservation quotas.

To facilitate those passengers who have to undertake journey at short notice and to save such passengers from the clutches of unscrupulous elements/touts, Tatkal scheme of reservation is available where the accommodation becomes available for booking on the previous day of journey from train originating station. Further, with a view to providing confirmed accommodation to waiting list passengers and to ensure optimal utilization of available accommodation, a scheme known as Alternate Train Accommodation (ATAS) known as “VIKALP” was introduced as a pilot project in November, 2015. This scheme has been expanded to cover all type of train on all sectors from 01.04.2017. For this facility, waiting list passenger has to give an option at the time of booking of ticket & passengers with waiting list status at the time of preparation of first reservation charts are shifted to other trains, subject to availability of vacant accommodation.

2. LITERATURE REVIEW

Here in the paper the authors (Budhkar & Das, 2017)have studied the reservation pattern of the passengers. The authors have studied whether the advanced booked tickets do get confirmed or not and what percent of the reservations are confirmed. The authors themselves have kept the track of the seat availability by logging into the IRTC website. The authors have also tried to predict the chances that whether the RAC tickets were confirmed or not. Only 3 trains are considered for the study on fixed dates and the route is also fixed. They have considered that the passengers had no other mode of transportation apart from the trains. [2]

The paper (Satyakrishna et al., 2018) shallow and deep machine learning technique used to predict rain delay. The factors responsible for train delay our climate conditions, train movement but additional factors like riots, strikes, mmalfunctioning of the trains are not considered for predicting the train delay and the prediction is region centric means only one station is considered this restriction can be removed and one or two more parameters may be added in further research. The prediction used to maintain the delay time but are made but these predictions can be made useful for maintaining the infrastructure. [8]

In this paper (Premsanthi & Sivakami, 2016) authors have tried to find from the data whether there is some direct correlation between the personal profile and relation of ticket satisfaction also daughters studied the problem faced by the passengers The mode of collecting the percentage or data is the questioner. The authors have used Chi square method for the purpose of research and they have considered only one station for the study.[7]

Authors(Zhi-Xin Liang et al., 2013) have studied the ticket overbooking trend during the spring festival in China. Because of this excess booking of the tickets the train seats remain vacant on the date of journey. If such overbooking can be prevented and the vacant seats can be allotted to a person who wants to travel on the given date than the railways would generate a good revenue.[9]

The authors (Mukhopadhyay et al., 2010) have proposed an auction based scheme for the TATKAL bookings keeping in mind the idea dead person will be ready to pay any amount to get the reservation but the current at card reservation works on 1st come 1st services which medic pride the person who desperately wants the ticket the auction based game is developed using VCG days algorithm Is the scheme is applied by any of the villages elation it will lead to profit of billions of these barrier the authors have considered the multiunit auction scheme. The passenger has to bid for the ticket and the one whose price is high gets the ticket.[6]

In the paper presented here the authors (Khan, Volume 8, No. 5, May-June 2017) have proposed the seat allocation method. The person who wishes to travel a long-distance journey always tries to get the ticket booked in the train. but the reservation ticket can be confirmed, waiting, RAC. If the ticket is confirmed than there is no issue but is the ticket is waiting or RAC that there is no surety that the ticket gets confirmed. but a person with waiting or RAC ticket is allowed to travel because when the chart is prepared the TTE gets the perfect number of the vacant seats and then he can allot the seat to the passenger. This allotment is totally in the hands of the TTE which can be biased by bribing the TTA to prevent this the authors have proposed a method of automatic seat allotment to only those having the waiting and RAC tickets using the clustering method and regularly updating the information of the seats allotted and vacant.[12]

The authors (Mohd Arshad, Vol.-7, Issue-2, Feb 2019) have tried to predict the delay of the train by applying three machine learning algorithms. They found that major train delay in particular region was due to the climate condition of that region. The variables responsible were collected from the past data. The multivariate regression, neural network and random forest algorithm are used for the study. The study can be expanded to other regions also and other factors responsible can also be considered.[13]

The authors (Prof.Ashish Saxena, Volume 5, Issue 03, March -2018) have proposed the system of booking the tickets where even if the tickets are not confirmed the passenger gets a waiting ticket. The information of the vacant available seat gets updated periodically and if the allotted waiting ticket has the chances of getting confirmed is notified of the seat availability. All the paper work done by the staff is moved to the software to reduce the work and make fair decisions.[14]

The ticket bookings are done online and multiple users access the website for the ticket reservation, which leads to the fact that very few passengers get the confirmed tickets and others either get the waiting or RAC tickets. The cause of such situation is that people are not aware of the fact when the ticket booking is accessible. The authors have considered the past data of southern railways and have used the linear and probability equation to predict the train details where the seat availability is fetch based on user's search. The search included the route, source and destination. Based on this random value would be selected from past data and used to predict the train information (Raparthi abhinay, Volume 13, Number 12 (2018) pp) [15]

The authors(Ma et al., 2014) have used the short term forecasting for the revenue management as it helps to cut short the extra prices in managing the revenue and also prevent the excess booking of the tickets . the data on which the users have relied upon the historical data model when they had considered the advanced booking model and combined booking models. The authors have used the clustering algorithm and used the Chinese historical railway data for the study. [5]

The paper uses the data mining to detect the malpractices in the ticket booking like false identity, fake booking and so on the authors have detected the fraudulent pattern in the reservation based on the hypothetical data and have not implemented the research on real world data. The authors have just used the pattern matching for their findings (Rasika Ingle, Volume 3, Issue 3, March, 2013) [15]

The authors(Gaigowad et al., 2014) have proposed a method for automatic ticket confirmation based on the association rule. The rule detects the fraud and prompts the user so that necessary steps can be taken. The authors have used the data railway website. The association rules have been designed for the effective discovery of the pattern. [4]

3. METHODS

3.1. Naive Bayes [11]

It is a technique based on Bayes' Theorem. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This model is easy to build and particularly useful for very large datasets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods

$P\ C|X = P(X|C) * P(C)/P(X)$ (1)

$P(C|X)$ is posterior probability of class C

$P(C)$ is prior probability of class C

$P(X|C)$ is probability of predictor given the class.

$P(X)$ is prior probability of predictor

3.2. K- Nearest Neighbor [11]

K-NN is the simplest of all machine learning algorithms. The principle behind this method is to find a predefined number of training samples closest in distance to the new point and predict the label from these. The number of samples can be a user-defined constant or vary based on the local density of points. The distance can be any metric measure. Standard Euclidean distance is the most common choice for calculating the distance between two points. The Nearest Neighbours have been successful in a large number of classification and regression problems, including handwritten digits or satellite image processing and so on

3.3. Decision tree [10]

The series of questions and their possible answers can be organized in the form of a decision tree, which is a hierarchical structure consisting of nodes and directed edges. The tree has three types of nodes:

- A root node that has no incoming edges and zero or more outgoing edges.
- Internal nodes, each of which has exactly one incoming edge and two or more outgoing edges.
- Leaf or terminal nodes, each of which has exactly one incoming edge and no outgoing edges.

In a decision tree, each leaf node is assigned a class label. The non-terminal nodes, which include the root and other internal nodes, contain attribute test conditions to separate records that have different characteristics. Classifying a test record is straightforward once a decision tree has been constructed. Starting from the root node, we apply the test condition to the record and follow the appropriate branch based on the outcome of the test. This will lead us either to another internal node, for which a new test condition is applied, or to a leaf node. The class label associated with the leaf node is then assigned to the record

3.4.Cluster Analysis [3]

Classification and prediction analyze class-labelled data objects while clustering analyses data objects without consulting a known class label. In many cases, class labelled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters. Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together

3.5.Classification [3]

Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data). In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analysing or "learning from" a training set made up of database tuples and their associated class labels. In the second step the model is used for classification. First, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the classifier's accuracy, this estimate would likely be optimistic, because the classifier tends to overfit the data (i.e., during learning it may incorporate some particular anomalies of the training data that are not present in the general data set overall). Therefore, a test set is used, made up of test tuples and their associated class labels. They are independent of the training tuples, meaning that they were not used to construct the classifier. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The associated class label of each test tuple is compared with the learned classifier's class prediction for that tuple.

3.6. Association rule[1]

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value

3.6.1. Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

Artificial neural network

Analogous to human brain structure, an ANN is composed of an interconnected assembly of nodes and directed links.

4. CONCLUSION

The paper has done the review of various papers where the data mining techniques are used to detect the fraud done by the TTE for the ticket allocation, detect the reason for the delay on the train and also predict the delay time of the trains including the delay time, the satisfaction of the customer by the services given by the railways. Thus the paper has supported the idea of implementing the data mining in the field of the railway ticket booking. Papers under the consideration have used the data which is either historical record or prepared the dummy data for the study. Research can be implemented to predict the confirmation of the wait listed or Reservation Against Cancellation tickets. This kind of the research can serve the passenger to get accurate information about the changes of the confirmed tickets. The research conducted by the authors can be implemented on the real world data for much accurate result and findings.

5. REFERENCES

[1] Bharati, M., & Ramageri, B. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering, 1.

[2] Budhkar, A., & Das, S. (2017). Finding trend of advanced ticket booking in Indian railways. Transportation Research Procedia, 25, 4822–4831. https://doi.org/10.1016/j.trpro.2017.05.492

[3] Data Mining: Concepts and Techniques 3rd Edition. (n.d.). DATA MINING, 560.

[4] Gaigowad, A., Deote, P., Badge, P., & Giradkar, R. (2014). Effective Use of Pattern Discovery for Detection of Fraudulent Patterns in Railway Reservation. 1(1), 4.

[5] Ma, M., Liu, J., & Cao, J. (2014). Short-Term Forecasting of Railway Passenger Flow Based on Clustering of Booking Curves. Mathematical Problems in Engineering, 2014, 1–8. https://doi.org/10.1155/2014/707636

[6] Mukhopadhyay, S., Mukherjee, N., Bhattacharjee, J., Ghosh, D., Saha, M., & Choudhury, P. (2010). An Efficient Auction Based TATKAL Scheme for Indian Railway. 2010 International Conference on Innovative Computing and Communication and 2010 Asia-Pacific Conference on Information Technology and Ocean

Engineering, 153–157. https://doi.org/10.1109/CICC-ITOE.2010.47

[7] Premsanthi, P., & Sivakami, M. (2016). A STUDY ON TRAIN PASSENGERS SATISFACTION AND PROBLEMS OF TICKET RESERVATION IN ERODE DISTRICT. 1(10), 7.

[8] Satyakrishna, J., Sagar, R. K., & Tech, M. (2018). Train Delay Prediction Systems Using Big Data Analytics. 6(3), 7.

[9] Zhi-Xin Liang, Shuo Wen, & Feng-Wen Yang. (2013). Research on the optimal train ticket overbooking strategy for transportation during the Spring Festival. 11th International Symposium on Operations Research and Its Applications in Engineering, Technology and Management 2013 (ISORA 2013), 177–180. https://doi.org/10.1049/cp.2013.2279

[10] PANG.NI NG TAN , MICHAEL STEINBACH, VI PI N KU MAR. Introduction to data mining

[11] Palak Baid , Apoorva Gupta, Neelam Chaplot. Sentiment Analysis of Movie Reviews using Machine Learning Techniques.

[12] Khan, J. A. (Volume 8, No. 5, May-June 2017). Waiting Ticket Optimization using Reservation Chart Cluster for Indian Railway. International Journal of Advanced Research in Computer Science, 913-916.

[13] Mohd Arshad, M. A. (Vol.-7, Issue-2, Feb 2019). Prediction of Train Delay in Indian Railways through Machine Learning. International Journal of Computer Sciences and Engineering, 405-411.

[14] Prof.Ashish Saxena, M. N. (Volume 5, Issue 03, March - 2018). Indian Railway System Monitoring Regards Passenger Allocation and Reservation. International Journal of Advance Engineering and Research , 608-611.

[15] Raparthi abhinay, T. y. (Volume 13, Number 12 (2018) pp). Passenger Flow Analysis in Metro Network Using Machine Learning Technique. International Journal of Applied Engineering Research, 10604-10606.

[16] Rasika Ingle, M. K. (Volume 3, Issue 3, March, 2013). An Approach for Effective Use of Pattern Discovery for Detection of Fraudulent Patterns In Railway Reservation Dataset. International Journal Of Computational Engineering Research, 26-29.