Prediction of Time Delay in Train Scheduling System using Data Analytics

Prof. Anup Bhange[1] Ms. Apurva Waghmare[2] Ms. Wanshika Giramkar[3]
Mr. Vinay Singh[4] Mr. Shubham Bhondekar[5]
[1]Assistant Prof. Department of Computer Technology
[2345]U.G.Student, Dept. of Computer Technology, and Engineering.
[12345]K.D.K College of Engineering, Nagpur, Maharashtra, India

***Abstract*—In a nation like India, where a larger part of populace relies upon the Railway foundation (approx. 25 million traveler's day by day), thinking of an application that guides to their voyaging knowledge would positively be of extraordinary help. Prepares in India get deferred regularly, and in the event that we can anticipate this is ahead of time - it would be of incredible help for the travelers to design their adventure. The objective of this project is to predict the delay in arrival of train(s) given certain features like destination, day of week etc.**

***Keywords*- linear regression, k-means clustering, train delay prediction, python**

I. INTRODUCTION

Deferral is one of the real issues in railroad frameworks everywhere throughout the world. As indicated by the British National Audit Office (NAO) [11] episodes, for example, foundation issues, armada issues, fatalities trespass still reason critical postponements to the voyaging open and incredible cost to the railroad. For case, in 2006-07, 0.8 million occurrences prompted 14 million minutes of deferral
to diversified traveler rail benefits in Great Britain, costing at least £1 billion (averaging around £73 for every moment of postponement) in the time lost to travelers in delays. Of these episodes 1376, each prompted more than 1000 minutes of deferral. Dealing with the results of occurrences and getting trains running ordinarily again is indispensable to diminishing postponements; so foreseeing traveler prepare delays is an extremely troublesome undertaking [11]. In view of the significance of this issue, Iranian Railroads dependably enlist and dissect the information of deferral regarding its date, causes, and time of postponements. In this examination, the enrolled information of traveler prepares delays in Iranian Railways from 2005 to the finish of 2009 is utilized. As indicated by the accomplished information from this database, the normal postponement from 2005 to the finish of 2009 was 18,174 hours for every year and 30 minutes for each
traveler prepare.

II. LITERATURE REVIEW

Writing survey uncovered that a couple of research on traveler prepare delays guaging has been finished. Carey and Kwiecinski [2] build up a basic stochastic technique to thump on prepare delays. Thump on defer alludes to that part of a prepare's deferral, which is caused by different prepares before it. Huisman and Boucherie [8] build up a stochastic model to anticipate the prepare delays with diverse paces. Their model can catch both booked and unscheduled prepare developments. A contextual analysis of a railroad area in the Dutch rail line arrange shows the down to earth estimation of the model, both for long and shortterm railroad arranging [8]. Subsides et al. [12] build up an insightful postponement indicator show for constant postpone observing, and timetable enhancement in the scope of prepare systems. This framework is in charge of preparing existing delays in the system to create defer expectations for depending trains in the
not so distant future. This govern based framework was utilized as a correlation with the uniquely created neural system so as to assess the precision and the staff of deliberation of such a misleadingly wise segment. Yuan [16] creates an enhanced stochastic model for prepare postponements and defer spread in stations. The most imperative logical commitment of this exploration is an imaginative diagnostic likelihood show that precisely predicts the thump on deferrals of trains, including the effect on prepare promptness at stations in view of an expansion of blocking time hypothesis of railroad tasks to stochastic wonders [16]. Yuan [17] builds up a model that arrangements with stochastic reliance in the displaying of prepare deferrals and postpone spread. The proposed model can be utilized as a part of evaluating timetable strength and foreseeing prepare reliability given essential deferrals. Display approval uncovers that the postponement gauges coordinate with genuine information exceptionally well. Briggs and Beck [1] exhibit that the dispersion of prepare delays on the British railroad arrange is precisely portrayed by q-exponential capacities. In this examination, they utilize information on flight times for 23 noteworthy stations for the period September 2005–October 2006. Daamenan et al. [5] propose a strategy to foresee thump on delays in a precise and non-discriminative way. In this examination, two principle classes of thump on delays are recognized: impediment at clashing track segments and sitting tight for planned associations in stations.

.

III. METHODOLOGY

A)Principle of multiple linear regression

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models

Linear regression has many practical uses. Most applications fall into one of the following two broad categories: (1) if the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of Y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of Y, the fitted model can be used to make a prediction of the value of Y. (2) given a variable Y and a number of variables X1 , ..., Xp that may be related to Y, linear regression analysis can be applied to quantify the strength of the relationship between Y and the Xj , to assess which Xj may have no relationship with Y at all, and to identify which subsets of the Xj contain redundant information about Y

B) K-means Clustering Algorithm

During data analysis many a times we want to group similar looking or behaving data points together. For example, it can be important for a marketing campaign organizer to identify different groups of customers and their characteristics so that he can roll out different marketing campaigns customized to those groups or it can be important for an educational institute to identify the groups of students so that they can launch the teaching plans accordingly. Classification and clustering are two fundamental tasks which are there in data mining for long, Classification is used in supervised learning (Where we have a dependent variable) while clustering is used in un-supervised learning where we don't have any knowledge about dependent variable.

Clustering helps to group similar data points together while these groups are significantly different from each other.

1) K-Means Clustering

There are multiple ways to cluster the data but K-Means algorithm is the most used algorithm. Which tries to improve the inter group similarity while keeping the groups as far as possible from each other.

Basically K-Means runs on distance calculations, which again uses "Euclidean Distance" for this purpose. Euclidean distance calculates the distance between two given points using the following formula:

Euclidean Distance = $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$

Above formula captures the distance in 2-Dimensional space but the same is applicable in multi-dimensional space as well with increase in number of terms getting added. "K" in K-Means represents the number of clusters in which we want our data to divide into. The basic restriction for K-Means algorithm is that your data should be continuous in nature. It won't work if data is categorical in nature.

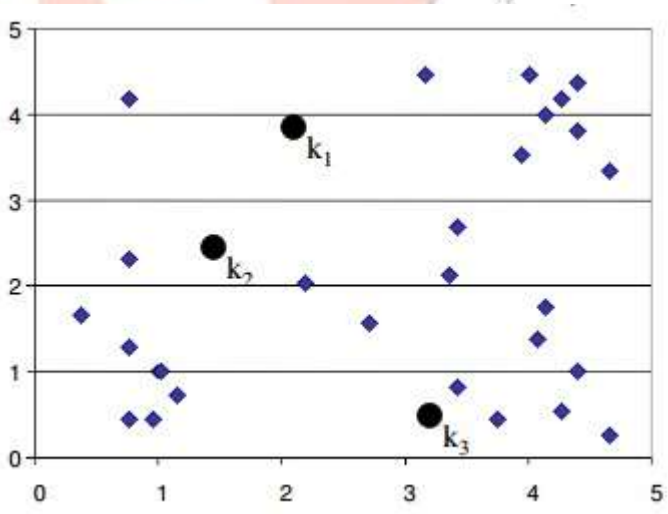
2) Data Preparation

As discussed, K-Means and most of the other clustering techniques work on the concept of distances. They calculate distance from a specific given points and try to reduce it. The problem occurs when different variables have different units, e.g., we want to segment population of India but weight is given in KGs but height is given in CMs. One can understand that the distance matric discussed above is highly susceptible to the units of variables. Hence, it is advisable to standardize your data before moving towards clustering exercise.

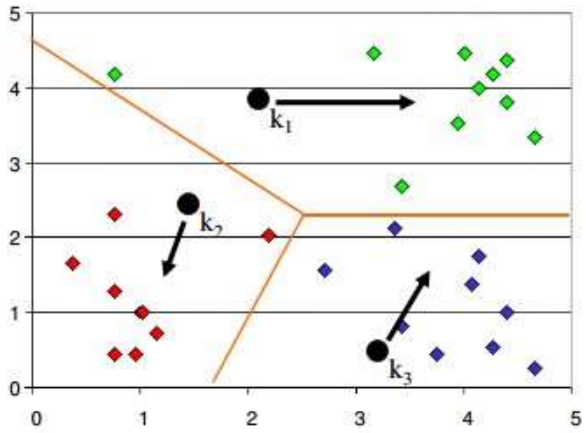
3) Algorithm

K-Means is an iterative process of clustering; which keeps iterating until it reaches the best solution or clusters in our problem space. Following pseudo example talks about the basic steps in K-Means clustering which is generally used to cluster our data

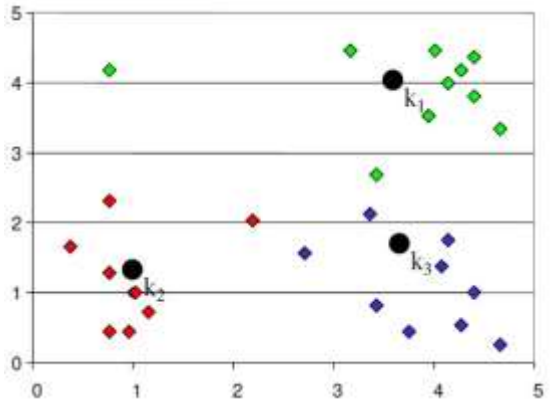
1. Start with number of clusters we want e.g., 3 in this case. K-Means algorithm start the process with random centers in data, and then tries to attach the nearest points to these centers



2. Algorithm then moves the randomly allocated centers to the means of created groups



3. In the next step, data points are again reassigned to these newly created centers



4. Steps 2 & 3 are repeated until no member changes their association/ groups

IV. EXPERIMENTAL RESULTS

From this system, the results achieved are feasible and accurate enough to predict delay. At the beginning the dataset is preprocessed to identify the outliers. The preprocessed dataset is then given to the model. Models for predicting train delay are developed using the data from the Bureau of Transportation Statistics (BTS).

Fig 4.1 Input Data and Shape

Fig 4.1 Shows that input data and shape which includes the train transportation data.

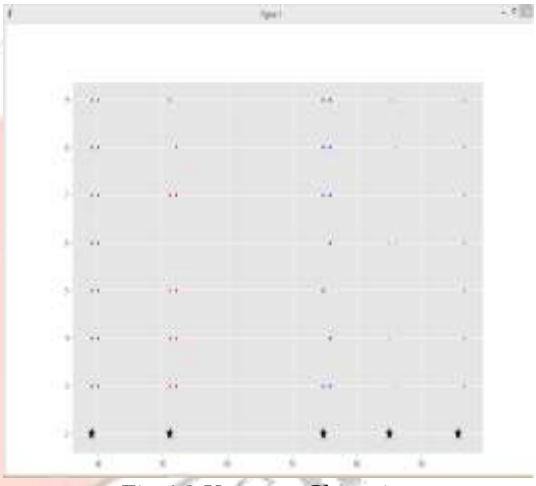


Fig 4.2 K-means Clustering

This Fig 4.2 shows that the pre-processing of input data in which k-means closeting algorithm is applied on the input data file and form 5 diiferent no.of clusters.

Fig 4.3 Test Data and Shape

This Fig 4.3 shows that Test data and shape of input data. It includes outcomes of and coeffecients of data to predict train delay.

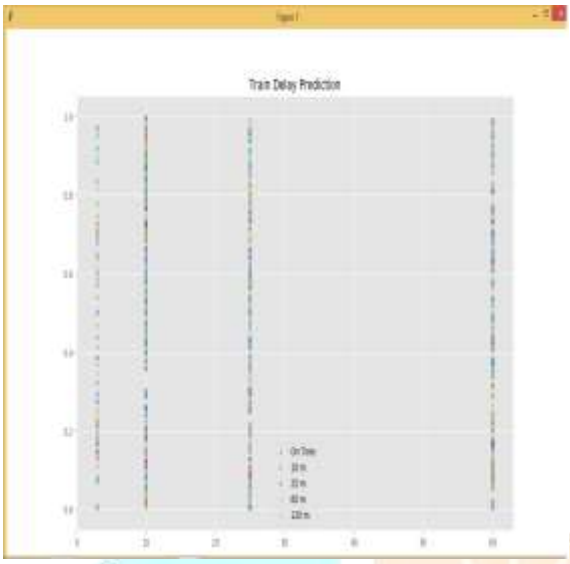


Fig 4.4 Train Delay Prediction

This Fig shows that the actual output i.e train delay prediction. In this Fig multiple linear regression is applied and formed a classification of data into some different classes like on time, 10 m, 25m, 60, and 120m.

V. CONCLUSION

In this paper for having the capacity to anticipate traveler prepare delays in Iranian Railroads, a neural system show with high exactness is exhibited. In the proposed display, we utilize three diverse ways to deal with characterize inputs including standardized genuine number, double coding, and parallel set encoding input esteems. Finding a fitting engineering for the traveler prepare expectation neural arrange show, different techniques are explored. Anticipating traveler prepare delays, the enlisted information of in Iranian Railways from 2005 to the finish of 2009 year is utilized. To assess the nature of the outcomes, we exploit choice tree and multinomial strategic relapse models. These correlations among results uncovered that the proposed demonstrate has awesome precision and low preparing time and as the outcome great arrangement quality. Postpone forecast

VI. FUTURE WORK

Future research will progress in two directions: improving training time and improving prediction accuracy. The model accuracy may be improved through metaheuristic methods such as genetic algorithms or simulated annealing or hybrid algorithms to find a better network architecture. Training time can be improved through other meta-heuristic methods such as particle swarm optimization or continuous ant colony optimization.

VII. REFERENCES

[1] Carey, M., Kwiecinski, A. (2007). Stochastic approximation to the effects of headways on knock–on delays of trains, *Transportation Research*, vol. 28, August, pp. 251–267.

[2] Chen, M., Liu, X., Xia, J., Chien, S.,(2004). A Dynamic Bus-Arrival Time Prediction Model Based on APC Data, *Computer-Aided Civil and Infrastructure Engineering*, vol. 19, no.5 , p.p. 364–376.

[3] Chen, M., Yaw, J., Chien, S., Liu, X. (2007). Using automatic passenger counter data in bus arrival time prediction, *Journal of Advanced Transportation*, vol. 41, no.3 , p.p. 267–283.

[4] Daamen, W., Goverde, R., Hansen. (2009). Non–Discriminatory Automatic Registration of Knock–On Train Delays, *Networks and Spatial Economics*, vol. 9, 23, November, pp. 47–61.

[5] Freedman, D., Pisani, R., Puves, R.(2007). *Statistics*, 4th edition, W. W. Norton & Company.

[6] Han, J., Kamber, M. (2006). *Data Mining Concepts and Techniques*, 2nd edition, Morgan Kaufmann Publishers.

[7] Huisman T., Boucherie, R. (2001). Running times on [8] railway sections