Train Delay Prediction using Machine Learning

Lokaiah Pullagura, Jeevaa Katiravan

Abstract: Indian Railways operates both long distance and suburban passenger trains and freight services daily in the country. Trains get delayed frequently due to several reasons such as, severe weather conditions such as fog, traffic, signal failure, derailing of trains, accidents, etc, and this delay is propagated from station to station. If we can predict this in advance - it would be of great help for the commuters to plan their journey either for an earlier departure or postpone, and also lets railways to take measures to avoid delays further. In this paper, we used decision tree, a machine learning method used for predicting train delays, and Recurrent Neural Networks distinguished with various fixatures. For predicting train delays, Recurrent Neural networks with 2 layers and 22 neurons per each layer gave best results with an average error of 122 seconds.

Keywords: Train Delay(TD), Machine Learning, prediction, decision tree, RNN.

I. INTRODUCTION

Railways functions by administering trains running in their entire network. It is a typical task of managing certain aspects such as timetable, equipment, etc. To avoid collision of trains, a timetable showing schedule of arrival and departure is designed. The Indian railways uses different gauges depending on the rail traffic. The Broad Gauge is used in areas with high traffic whereas Meter gauge is used in hill regions. Train delay is a major issue in IRs. The Railway Network is an interconnected group of stations connected with links. The traffic in the Indian Railway Network(IRN) is accumulated from station to station. Giving importance to network and traffic, main stations in the route are recognized. This helped to pinpoint some of the most likely nodes of blockage in the grid. For trains functioning, a span between two trains passing through the same station is to be maintained. If a train delays at a particular station, the subsequent trains stopping at same station may also be delayed. The first station is the origin for delay of train. The delay at starting point normally leads future delays at successive stations. Train Delays (TDs) occur due to several reasons thereby leading to deviation in the stipulated schedule, thereby overwhelm railway functioning and they cause service disorder and sometimes catapult in extreme cases.

II. LITERATURE SURVEY


The remaining part of this work is organized as follows. In section 3, Data collection process is explained. In section 4 Train delay prediction model is described in detail. In section 5, experimental results are presented with and without boosting techniques as well as results of Recurring Neural Networks when tested by changing the number of hidden layers and nodes per layers.

III. DATA

Our primary focus was to predict train delays, therefore various types of data needed to be collected to make the predictions possible. The primary sources of data are websites like www.runningstatus.in and www.etrain.info/in have information about trains between stations, real-time arrival/departure information. Data collected from different origins may be in different formats. Previous weather conditions data was also collected. Data is collected on departure times for 09 major stations from Vijayawada to Chennai Central for the three-month period i.e. from Jan 2018 to March 2018. Each Train’s departure was mixed with the weather data. The combined data was split into two parts: 80% was used as training data and the rest as testing data. Data is collected for the following parameters: Station, Train_number, Date, Past_Delay, Designated_Arrival, Real_Arrival, Time of delay.

There are around 72 trains travel in the route of which 12 trains are regular. Data collected for three-month is sorted according to the train arrival. Indian Railways Website releases the delay of trains at each station. For collecting the delay information from railways websites, a Web Crawler is developed. It is observed several trains running in the route are delayed out of which 23.7% involved super fast trains.
IV. TRAIN DELAY PREDICTION MODEL

To build effective train delay predictor, it is required to model attributes like type of train, the period of its journey, the duration of its journey, area of its journey Delay Times, Train Maximum Speed, Route Maximum Speed, Speed Homogeneity, Route Departures, Vertex Departures/Arrivals, Edge Length, Percentage Double Track, etc.

Railroad is divided up into a number of paths, which meet at some locations. These paths having traffic cover a successive railroad and goes through several railway nodes. Railway network is considered to be a graph which is a collection of edges and vertices as shown in fig:

![Route Map of Vijayawada to Chennai Central stations](image)

Fig.1 Route Map of Vijayawada to Chennai Central stations

A train passes between stations. Each station contains data about its respective routes or intersections. Assuming the premise that delay is a result of infrastructure, All information about Train Departures which is downloaded was stored in a MongoDB database. Python programming language is used to develop a model as it has rich set of libraries.

![Illustration of a railway graph data model](image)

Fig:2 Illustration of a railway graph data model

V. EXPERIMENTAL RESULTS

Decision tree was tested with variation in maximum depth setting as shown in table 4.1. The maximum depth 4 gave the accurate result with average errors 166. The tests were run once for each setting.

<table>
<thead>
<tr>
<th>Maximum depth</th>
<th>Average error(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>178</td>
</tr>
<tr>
<td>2</td>
<td>181</td>
</tr>
<tr>
<td>3</td>
<td>201</td>
</tr>
<tr>
<td>4</td>
<td>166</td>
</tr>
<tr>
<td>5</td>
<td>168</td>
</tr>
</tbody>
</table>

Table I: Average errors for various tree depths without using boosting.

A. Decision tree with AdaBoost

Decision tree using AdaBoost was added with 5 estimators. The result is shown in table 4.2. The Maximum depth that gave the most accurate result was 5. The average errors of decision tree with and without AdaBoost are compared in figure 4.1. For each testing the tests were performed once.

<table>
<thead>
<tr>
<th>Maximum depth</th>
<th>Average error (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>346</td>
</tr>
<tr>
<td>2</td>
<td>314</td>
</tr>
<tr>
<td>3</td>
<td>271</td>
</tr>
<tr>
<td>4</td>
<td>181</td>
</tr>
<tr>
<td>5</td>
<td>145</td>
</tr>
</tbody>
</table>

Table II: Average error for different tree depths when using AdaBoost with 5 estimators.

![Two graphs showing the average prediction errors in seconds while using decision tree with and without AdaBoost](image)

Fig:3 Two graphs showing the average prediction errors in seconds while using decision tree with and without AdaBoost.

B. Recurrent Neural networks (RNN)

RNNs are typical neural networks where the output of previous layer are fed as input to the current layer. The main and most important feature of RNN is Hidden state, which remembers some information about a sequence.
RNN was tested by changing the number of hidden layers and nodes per layer. This is shown in table 4.3 and figure 4.2. The most accurate configuration was 2 hidden layers with 22 nodes per layer. For each testing the tests were performed once.

**Table III: Average error for different layer configurations in neural network.**

<table>
<thead>
<tr>
<th>Hidden layers</th>
<th>Nodes per layer</th>
<th>Average error (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>680</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1150</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>186</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>322</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>346</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>227</td>
</tr>
</tbody>
</table>

**VI. CONCLUSION**

In this paper, one of the important goals is to predict delays by choosing the accurate method in machine learning. From results it is understood that Recurrent Neural Network is the most accurate method. However, the study of data considered was very short period. Another goal of this paper is to predict train delays using machine learning. Results obtained are vague. This is because duration of time was very short. A Delay prediction model developed displays statistics about train delays.

In a further study, the work done in this paper can be extended to collect data for a one or more years thereby giving machine to find the determinants and remove the noise data. It can be extended to RNN with more hidden layers. In this paper we have used two machine learning methods and a boosting technique. More investigation can be done for other machine learning methods and boosting technique.

**ACKNOWLEDGMENT**

I would especially like to thank Dr. Jeevaa Katiravan., for guiding me throughout my work. As my teacher and mentor, he has taught me more than I could ever give him credit for here. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models. Most importantly, I wish to thank my loving and supportive wife, Phebe, and my sons, Sanjay Bissy and Joy Bhushan.

**REFERENCES**

4. Pu Wang and Qingsheng Zhang: “Train delay analysis and prediction based on big data fusion” Transportation Safety and Environment, 2019, Vol. 00, No. 00 1–10.
17. Luca Oneto, Emanuele FumeoGiorgio ClericoRenzo, CanepaFederico, PapaCarlo, DambraNadia. Mazzino Davide Anguita “Delay Prediction System for Large-Scale Railway Networks Based on Big Data Analytics".
Train Delay Prediction using Machine Learning

AUTHORS PROFILE

Lokaiah Pullagura, Research Scholar, Department of CSE, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India. He has published several papers in International journals. His research area of interest is in Machine Learning.

Dr. Jeeva Katiravan, He is currently Associate Professor in IT Department. He has rich experience in Research/Teaching and published several papers in International and National journals. He has guided several research scholars in the area of Information Security and Machine Learning. He is a Doctoral Committee Member at Anna University, Sathyabama University and VIT University and acted as External member For Ph.D Viva at Anna University and St. Peter’s University.