IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

An Approach to Predict Train Delay Using Big Data Analytic Approaches

Ajay Patel[1], Manish Jaiswal[2], Rahul Kumar Chawda[3]

MCA Student, CS Department, Kalinga University, New Raipur, India[1,2]

HOD, CS Department, Kalinga University, New Raipur, India[3]

Abstract: Driven by specialized analytics systems and software, big data analytics can point the way to various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals. The public transportation industry has been at the forefront in utilizing and implementing Analytics and Big Data, from ridership forecasting to transit operations Rail transit systems have been especially involved with these IT concepts, and tend to be especially amenable to the advantages of Analytics and Big Data because they are generally closed systems that involve sophisticated processing of large volumes of data. The more that public transportation professionals and decision makers understand the role of Analytics and Big Data in their industry in perspective, the more effectively they will be able to utilize its promise. The current work aims to develop a system to predict train delay using Big Data analytic approaches.

Keywords: Big Data Analytics, Predictive modelling, Big-Data, Train Delay prediction, SVM.

I. INTRODUCTION

Big data analytics applications enable data scientists, predictive modellers, statisticians and other analytics professionals to analyse growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. That encompasses a mix of semi-structured and unstructured data -- for example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile-phone call-detail records and machine data captured by sensors connected to the internet of things. On a broad scale, data analytics technologies and techniques provide a means of analysing data sets and drawing conclusions about them to help organizations make informed business decisions. BI queries answer basic questions about business operations and performance. Big data analytics is a form of advanced analytics, which involves complex applications with elements such as predictive models, statistical algorithms and what-if analyses powered by high-performance analytics systems.

Predictive modelling is a process of creating a statistical model to predict the future behaviour. It is more of the area in data mining forecasting probabilities and trends. A predictive model is made up of predictors which are factors that influence future results. For example a retail shop's model should consider the customer's gender, age and purchase history which might be used to predict the future sale. The predictive model is described using three key features.

- The predicted outcome
- Predictors
- Creating outcome using predictors

Predicting models have become a vital part in CRM, change management, disaster recovery, security management and meteorology. Predictive analytics supports decision making by diagnosing the business. It helps the firm to eliminate processes which are time consuming. It can be used in almost all the fields with marketing and pricing using it predominantly.

- **Developing a model**: The first step in developing a predictive model is selecting relevant candidate predictor variables for possible inclusion in the model. A limited number of variables are selected from a vast list to bring bias to the selection process. Inappropriate selection of variables is an important and common cause of poor model performance. The major issue in developing a predictive model is to deal with the missing data.
- **Validating the model**: Validation can be performed using internal or external validation. A common approach to internal validation is to split the data set into two portions—a "training set" and "validation set". The objective of the external validation is to apply a previously developed model to new individuals whose data were not used in the model development, and quantify the model's predictive performance.When a validation study shows disappointing results,

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

researchers are often tempted to reject the initial model and to develop a new predictive model using the validation cohort data.

- **Assessing the performance of the model**: When assessing model performance, it is important to remember that explanatory models are judged based on strength of associations, whereas predictive models are judged solely based on their ability to make accurate predictions. The performance of a predictive model is assessed using several complementary tests, which assess overall performance, calibration, discrimination, and reclassification.

II. LITREATURE REV IEW

According to paper [1] delays can be due to various causes: disruptions in the operations flow, accidents, malfunctioning or damaged equipment, construction work, repair work, and severe weather conditions like snow and ice, floods, and landslides, to name just a few. Although trains should respect a fixed schedule called Nominal Timetable (NT), Train Delays (TDs) occur daily and can negatively affect railway operations, causing service disruptions and losses in the worst cases. According to paper [2] Rail transit systems have been especially involved with these IT concepts, and tend to be especially amenable to the advantages of Analytics and Big Data because they are generally closed systems that involve sophisticated processing of large volumes of data. The more that public transportation professionals and decision makers understand the role of Analytics and Big Data in their industry in perspective, the more effectively they will be able to utilize its promise. In paper [3] The Belgian railway network has a high traffic density with Brussels as its gravity centre. The star-shape of the network implies heavily loaded bifurcations in which knock-on delays are likely to occur. Knock-on delays should be minimized to improve the total punctuality in the network. Based on experience, the most critical junctions in the traffic flow are known, but others might be hidden. To reveal the hidden patterns of trains passing delays to each other, state-of-the-art data mining techniques are being studied and adapted to this specific problem. According to paper [4] the data and information can be outsourced to an untrusted cloud while maintaining its velocity and veracity of big data. Big data helps in decision making. It also laid foundation for advancement in science, medicine and Business. Map reduce and Hadoop frameworks are used to improve the processing speed of data. Big data analytics can improve efficiency and effectiveness across the broad range of government responsibilities, by improving existing processes and operations and enabling completely new ones.

III.PROBLEM IDENTIFICATION

Major challenges faced in predictive modelling are:

- One of the considerations in any analytics solution is the quality of data. Useless data provides poor results, while critical information can offer key findings that will shape major insurer decision making. However, too much focus on the quality of data is the challenge that businesses have to face, according to the news source.
- Predictive analytics best practices dictate that providers should allow these solutions to monitor themselves. This will minimize the complexity of the reports generated without stunting the power of the useful intelligence tools.
- Predictive analytics solutions are best known for taking complex models and generating easy to understand results from them. However, firms can get carried away when developing their models and expect too much of the tools used simply because they know the solutions are capable of handling the workload.
- According to the news source, models have to be able to be put to use for practice purposes, not just to generate reports. This focus will assist in reducing complexity and optimize the actionable intelligence garnered from predictive analytics.

Current train delay prediction systems do not take advantage of state-of-the-art tools and techniques for handling and extracting useful and actionable information from the large amount of historical train movements data collected by the railway information systems. Instead, they rely on static rules built by experts of the railway infrastructure based on classical univariate statistic.

IV.METHODOLOGY

A.Train Delay Dataset

The dataset used consists of following features:

- Train Route features: These include the times of departure and arrival, the destination station, and the distance covered by the train.
- Train features: These include the train no and type of train.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

- Weather features: The weather data includes several categorical features indicating the presence of snow, hail, thunder, rain and tornado warnings. It also contains a few numeric features such as wind speed, temperature and humidity.

B. Preliminary analysis using Linear Regression: By using this model first, we want to use the regression parameters to identify prominent factors affecting delays. Second, we would like to get a baseline estimate of how accurately delays can be predicted. Third, we want to analyse how the prediction error changes with the size of the training set.

C. Features affecting delays the coefficients of the linear regression give us a good idea of the relative importance of various features. It is found that train delay is normally due to bad weather condition as well as due to technical fault.

D. Predicting delay with classifiers: In order to predict whether a train will be delayed or not, we model the problem as a classification with two classes: delayed for trains with delays above 10 minutes, and non-delayed otherwise.

E.The classifier used is SVM classifier. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. In two dimensional spaces this hyper plane is a line dividing a plane in two parts where in each class lay in either side.

F. Regression SVM**:** *I*n a regression SVM, you have to estimate the functional dependence of the dependent variable y on a set of independent variables x. It assumes, like other regression problems, that the relationship between the independent and dependent variables is given by a deterministic function f plus the addition of some additive noise:

$$y = f(x) + \text{noise} \quad \ldots\ldots (1)$$

The task is then to find a functional form for f that can correctly predict new cases that the SVM has not been presented with before. This can be achieved by training the SVM model on a sample set, i.e., training set, a process that involves, like classification, and the sequential optimization of an error function. For this type of SVM the error function is:

$$\frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i + C\sum_{i=1}^{N} \xi^*_i \quad \ldots\ldots\ldots\ldots\ldots (2)$$

Which we minimize subject to:

$$w^T \phi(x_i) + b - y_i \le \varepsilon + \xi_i^*$$
$$y_i - w^T \phi(x_i) - b_i \le \varepsilon + \xi_i$$
$$\xi_i, \xi_i^* \ge 0, i = 1,\ldots, N \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

G. Weka Toolkit: Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Five features of Weka that I like to promote are:

- **Open Source**: It is released as open source software under the GNU GPL.
- **Graphical Interface**: It has a Graphical User Interface (GUI). This allows you to complete your machine learning projects without programming.
- **Command Line Interface**: all features of the software can use from the command line. This can be very useful for scripting large jobs.
- **Java API**: It is written in Java and provides a API that is well documented and promotes integration into your own applications. Note that the GNU GPL means that in turn your software would also have to be released as GPL.
- **Documentation**: There books, manuals, wikis and MOOC courses that can train you how to use the platform effectively.

V. EXPERIMENTAL RESULTS

By using the Weka toolkit with SVM classifier the predicted data will help to get more accurate train delay prediction system. In this experiment by using lower rate of prediction we can get most accurate values that are by using the lower prediction rate we can recommend items of the customers interest by using the higher accuracy.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

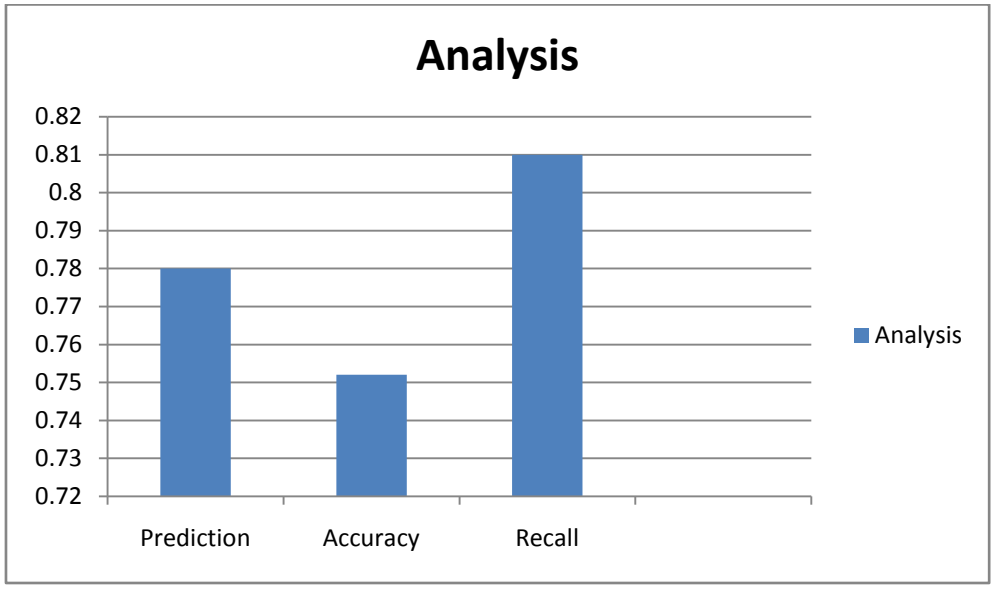

Figure: Proposed approach analysis graph

VI.. CONCLUSION

Predictive modeling is one of major areas which has numerous number of applications like weather prediction, stock prediction etc. The aim of the paper is to investigate options for predicting delays in transportation sector like railways, airlines etc. According the proposed approach the prediction can be done by utilizing no of features retrieved from train delay dataset like train features, route features , weather related features etc. On the basis of the identified features an analysis is done in order identify the most suitable ones for right predictions. The output of the analysis is forwarded to SVM classifier in order to predict whether the train will be delayed or not. The investigation results show that the feature accuracy greatly affects the prediction accuracy. In future more study is needed in order to identify more such features for improving the accuracy of predictions.

REFERENCES

[1] L.Oneto,E.Fumeo.et.al, "Train Delay Prediction Systems: a Big Data Analytics Perspective", *Big Data Research May 25, 2017*.
[2] A.Thaduri, D. Galar .et.al, " Railway assets: A potential domain for big data analytics",ELSEVEIR 2015
[3] R.Kaur,"BigData – is a Turnkey Solution",ELSEVIER 2015
[4] E. Fumeo, L. Oneto, D. Anguita, Condition based maintenance in railway transportation systems based on big data streaming analysis, in: The INNS Big Data conference, 2015.
[5] H. Li, D. Parikh, Q. He, B. Qian, Z. Li, D. Fang, A. Hampapur, Improving rail network velocity: A machine learning approach to predictive maintenance, Transportation Research Part C: Emerging Technologies
[5] C. Snijders, U. Matzat, U.-D. Reips, Big data: Big gaps of knowledge in the field of internet science, International Journal of Internet Science 7 (1) (2012) 1–5