Train Delay Prediction using Machine Learning

Janumpally Sravanthi1, Thalluri Prasanna2, Saddanapu Saikrupa3
Mrs.I.Anusha4

Assistant professor,Email:inagantianusha23@gmail.com

1, 2, 3, 4 Sridevi Women's Engineering College, V.N.PALLY , NEAR WIPRO GOPANANPALLY,
HYDERABAD, Ranga Reddy, 500075 ; Email : admin@swec.ac.inWebsite, www.swec.ac.in ;

Abstract:- Delay prediction is a process of estimating delay probability based on known data at a given checkpoint and is typically measured via arrival (departure) delay. The key to making delay predictions based on actual operational data involves establishing the relationship between train delays and various characteristics of a railway system.This provides a basis for the operator's scheduling decision Train delay is a significant problem that negatively impacts the railway industry and costs billions of dollars each year. In this project we have used Train delay dataset from IRTC to predict Train delays. We have used Faster RCNN algorithm to predict flight departure delay and our model can identify which features were more important when predicting Train delays. Accurate Train delay prediction is fundamental to establish the more efficient Railway business. Recent studies have been focused on applying machine learning methods to predict the Train delay. Most of the previous prediction methods are conducted in a single route or Railway Station. This project explores a broader scope of factors which may potentially influence the Train delay, and compares several machine learning-based models in designed generalized Train delay prediction tasks. To build a dataset for the proposed scheme, automatic dependent surveillance broadcast (ADS-B) messages are received, pre-processed, and integrated with other information such as weather condition, Train schedule, and Railway information. The designed prediction tasks contain different classification tasks and a regression task. Experimental results show that long short-term memory (LSTM) is capable of handling the obtained aviation sequence data, but overfitting problem occurs in our limited dataset. Compared with the previous schemes, the proposed Scheme model can obtain higher prediction accuracy (95.2% for the binary classification) and can overcome the overfitting problem.

Keywords: Train Delay Prediction, Machine Learning, Predictive Analytics, Transportation, Time Series Analysis, Rail Operations, Data Mining, Feature Engineering, Real-Time Monitoring, Predictive Modeling.

I INTRODUCTION

Railway transportation plays a vital role in the modern urban landscape, providing a convenient and sustainable mode of commuting for millions of people. However, train delays due to various factors like maintenance issues, weather conditions, or unexpected events can lead to inconvenience and frustration for passengers. The "Train Delay Prediction using Machine Learning" project addresses this challenge by employing advanced machine learning techniques to forecast delays in real-time. It has happened so many times that you have been waiting on railway station for someone to arrive and you don't have any exact information about train timing and other stuff. So here we present to you a project on Railway Tracking and Arrival Time Prediction. Using this system user's can get the information about train timing, and is it on time or not, and other information. In this, system will track the train timing at what time train departed from a particular station and pass these timing details to other station's system where it will display the timing according to train departed from previous station. If system will find any delay in train due to signal it will automatically update the train timing in next station and will be displayed to viewers.

In this system there is an admin module, who enters the detail about trains and its timing and these details will be passed through internet server and is fetched by the system on other stations, and there is other system that shows train information to the viewers on platform. Second system will get all the information of all trains but will automatically select the data that refers to particular station and shows that information on screen. Station masters on every station have a login wherein they may update train arrival time at their station when it arrives. This second System is installed on various locations on station for viewers to view the information. Admin will add information like train departed from station, expected arrival at destination, delay in the train schedule, etc. This project publishes real-time train schedule events to subscribing multiple client applications. In the present world, the major components of any transportation system include passenger Railway, cargo Railway, and air traffic control system. With the passage of time, nations around the world have tried to evolve numerous techniques of improving the Railway transportation system. This has brought drastic change in the Railway operations. Train delays occasionally cause inconvenience to the modern passengers [1]. Every year approximately 20% of Railway Trains are canceled or delayed, costing passengers more than 20 billion dollars in money and their time.

II LITERATURE SURVEY

In the area of Intelligence Transportation System (ITS) many pieces of research have been developed in the past. Following Literature review show a few researches on train delays prediction system that have been performed.

A. Hansen et al., [8] proposed a model in which two things are taken into consideration, dispatching decisions & conflicts of train path [8]. The model is built for predicting running times as well as arrival times for delayed and ontime, both kind of trains and to evaluate the effectiveness of dispatching decisions. Masoud Yaghini et al., [9] presented an ANN model with high accuracy to predict the train delays. He proposed a model that uses different strategies to define the input. To evaluate the quality of the results, they took advantage of Multinomial logistic regression models & decision tree. The model accuracy and training time may be improved over met heuristic methods such as genetic algorithms.

S. Pongnumkul et al., [10] proposed two algorithms using the average of historical travel times and average travel time of the k-nearest neighbours of the last known arrival time respectively. In this paper, a review found that both the algorithms bring in similar percent improvement in the prediction errors. There is some work to do on Thai railways for predicting train delay. They also discovered that one of the major factors for train delay is the number of stops between stations. In this report, they used 9 parameters & six months of train data to predict the inline delay.

Jia Hu et al, [11] proposed a prediction model that is built on the basis of Artificial Neural Network (ANN). To overcome from the endogenous drawback of ANN, further Genetic Algorithm is implemented to boost the performance of ANN. A thorough survey of the following papers was done to understand thoroughly the concept of Train delay prediction systems, datasets, its attributes and methodology used: and for related research work for flight delay prediction system. Delay in a train means that the train has not arrived at its prescheduled time. The train delay does not include unexpected stopping time near to the station or in between the station due to poor signal or unavailability of the platform. There are some important causes that leads to train delays like delay at the origin, engine breakdown, other train’s engine breakdown, waiting time at overtaking point, Climate/weather condition (temperature, wind speed, heavy rain, snowfall) and other factors (railways assets condition, festivals, strikes, national level exams, etc.). In this paper, we included most of these factors in order to predict better result. Accuracy of the model can be improved through meta-heuristic methods like genetic algorithm, hybrid methods or ensemble learning.

Tools Used In Machine Learning

By doing a comprehensive survey on various machine learning tools, we can conclude that some of the tools are user-friendly, easy to install and require less programming knowledge; while some of them might be difficult to use. Due to the user-friendly GUI, WEKA can be the easiest tool for beginners. For users with beginner or intermediate knowledge of Python language, Scikit-learn is one of the best tools to perform machine learning tasks because it contains numerous amount of machine learning libraries and can be combined with Python libraries like numpy, scipy, etc. For users with a good programming background, R and Matlab can also be one of the options but Matlab & R requires more memory space & heavy processor. In this paper, we used Scikit-learn to build a model in order to perform train delay prediction system.

III SYSTEM ANALYSIS

1 EXISTING SYSTEM:

- Supervised machine learning classifies data inputs accordingly labeled output and unsupervised learning classifies the inputs without having any labeled data. Several researchers had used machine learning algorithms to solve the classification problems in the educational domain.

- Keeping in view, the identification of Train demographic, Climate, social, personal, and others Features some latest literature proved that machine learning played a much significant role in predictive modeling.

Disadvantages of existing system:

An innovative regression algorithm used for grade prediction of a Train with an accuracy of 85%. The SVM algorithm achieved a better prediction accuracy of 96% as compared to the K- nearest neighbor to predict the attitude of Hungarian and Indian Trains towards technology.

2 PROPOSED SYSTEM:

In this project, we have used the latest optimization techniques on the MLP and compared respective optimizer with dynamic testing, activation functions, regularization parameters, etc. Afterward, we compared this optimistic MLP model with a robust RCNN algorithm. All experiments performed in the popular machine learning software Orange 3.24.1, which is opensource, hands-on, machine learning software enriched with a massive library of algorithms. Using hybrid languages. it was developed at the Train Database. For better visualization of results, we used The statistical analyses is performed with support of popular STAC web platform.

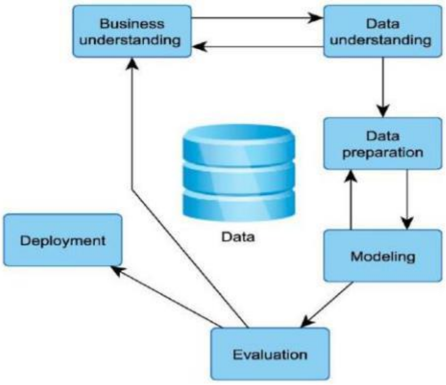
Advantages of proposed system:

- This research is accomplished to support the Railway Station's realtime Train demographic system.

- Before deploying online in a real-time environment, we need to propose an optimistic native place predictive model in the international Train environment.

- The present promising native place identification models gained the highest prediction accuracy on the prime data depicted in preliminary work with state-of-theart research inclusive mathematical equations and feature engineering.

3 SYSTEM ARCHITECTURE



Proposed Architecture

IV METHODOLOGY

1 Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

2 Data Preparation:

we will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

3 Model Selection:

While creating a machine learning model, we need two dataset, one for training and other for testing. But now we have only one. So lets split this in two with a ratio of 80:20. We will also divide the dataframe into feature column and label column.

Here we imported train_test_split function of sklearn. Then use it to split the dataset. Also, *test_size = 0.2*, it makes the split with 80% as train dataset and 20% as test dataset.

The *random_state* parameter seeds random number generator that helps to split the dataset.

The function returns four datasets. Labelled them as *train_x*, *train_y*, *test_x*, *test_y*. If we see shape of this datasets we can see the split of dataset.

We will use Random Forest Classifier, which fits multiple decision tree to the data. Finally I train the model by passing *train_x*, *train_y* to the *fit* method.

Once the model is trained, we need to Test the model. For that we will pass *test_x* to the predict method.

Random Forest is one of the most powerful methods that is used in machine learning for regression problems. The random forest comes in the category of the supervised regressor algorithm. This algorithm is carried out in two different stages the first one deals with the creation of the forest of the given dataset, and the other one deals with the prediction from the regressor.

4 Accuracy on test set:

We got a accuracy of 95.1%,97.1%, 98.1%, 96.5%, on test set.

5 Saving the Trained Model:

Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or . pkl file using a library like pickle .

Make sure you have pickle installed in your environment.

Next, let's import the module and dump the model into . pkl file

V CONCLUSION

In this project, we use Train data, weather, and demand data to predict Train departure delay. Our result shows that the RCNN method yields the best performance compared to the SVM model. Somehow the SVM model is very time consuming and does not necessarily produce better results. In the end, our model correctly predicts 95% of the non-delayed Trains. However, the delayed Trains are only correctly predicted 41% of time. As a result, there can be additional features related to the causes of Train delay that are not yet discovered using our existing data sources. In the second part of the project, we can see that it is possible to predict Train delay patterns from just the volume of concurrently published tweets, and their sentiment and objectivity. This is not unreasonable; people tend to post about Railway Station delays on Twitter; it stands to reason that these posts would become more frequent, and more profoundly emotional, as the delays get worse. Without more data, we cannot make a robust model and find out the role of related factors and chance on these results. However, as a proof of concept, there is potential for these results. It may be possible to routinely use tweets to ascertain an understanding of concurrent Railway delays and traffic patterns, which could be useful in a variety of circumstances.

VI REFERENCES

[1] A. B. Guy, "Train delays cost $32.9 billion, passengers foot half the bill". [Online] Available : https://news.berkeley.edu/2010/10/18/Train_delays/3/. [Accessed on June 2017].

[2] M. Abdel-Aty, C. Lee, Y. Bai, X. Li and M. Michalak, "Detecting periodic patterns of arrival delay", Journal of Air Transport Management,, Volume 13(6), pp. 355– 361, November, 2007.

[3] S. AhmadBeygi, A. Cohn and M. Lapp, "Decreasing Railway Delay Propagation By Re-Allocating Scheduled Slack", Annual Conference, Boston, 2008.

[4] A. A. Simmons, "Train Delay Forecast due to Weather Using Data Mining", M.S. Disseration, University of the Basque Country, Department of Computer Science, 2015.

[5] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced Railway delays based on machine learning algorithms", Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th, Sacramento, CA, USA, 2016.

[6] L. Schaefer and D. Millner, "Train Delay Propagation Analysis With The Detailed Policy Assessment Tool", Man and Cybernetics Conference, Tucson, AZ, 2001.

[7] B. Liu "Sentiment Analysis and Opinion Mining Synthesis", Morgan & Claypool Publishers, p. 167, 2012.

[8] Statistical Computing Statistical Graphics. [Online]. Available: http://stat-computing.org/dataexpo/2009/the-data.html. [Accessed on April 2017].

[9] FAA Operations & Performance Data. [Online].Available: https://aspm.faa.gov/.[Accessed on April 2017].

[10] B. Bailey, "Data Cleaning 101". [Online]. Available: https://towardsdatascience.com/data-cleaning-101-948d22a92e4. [Accessed on March 2018].

[11] P. Panov, L. Soldatova and S. Džeroski, " OntoDM-KDD: Ontology for Representing the Knowledge Discovery Process", Discovery Science 2013, Volume 8140, pp. 126-140, 2013.

[12] Bureau of Transportation Statistics. [Online]. Available: https://www.transtats.bts.gov/carriers.asp. [Accessed on 2 April 2017].

[13] How to Predict Yes/No Outcomes Using Logistic Regression. [Online]. Available: https://blog.cleaarbrain.com/posts/how-to-predict-yesno-outcomes-using-logistic-regression [Accessed on 3 Feubrary 2018].

[14] S. Polamuri, "How The Random Forest Algorithm Works In Machine Learning". [Online]. Available: https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674. [Accessed January 2018].

[15] S. Ray, "Understanding Support Vector Machine algorithm". [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector- machine-example-code/.[Accessed November 2017].

[16] OneHotEncoder. [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. [Accessed on March 2018].

[17] R. Vasudev, "Why and When do you have to use OneHotEncoder?".[Online].Available: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do- you-have-to-use-it-e3c6186d008f. [Accessed on March 2018].

[18] Twitter API Twitter. [Online]. Available: https://developer.twitter.com/en/docs.

[19] S. Loria , "TextBlob: Simplified Text Processing", 2016. [Online]. Available: http://textblob.readthedocs.io/en/dev/ [Accessed on December 12, 2017].

[20] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data," Columbia University, New York, December, 2011.

[21] V. A. Kharde and S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (0975 – 8887), Volume 139, no.11, p.11, April 2016.