Predicting and Analysing Indian Train Delay
Group ID - 22

Abhishek Singh
astindal@iitk.ac.in
170033

Ayush Hitesh Soneria
ayushhs@iitk.ac.in
170192

Dewansh Singh
dewansh@iitk.ac.in
170241

Jayesh Narwaria
jayn@iitk.ac.in
170325

Prajwal HG
prajwalhg@iitk.ac.in
170481

ABSTRACT

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 statistics. The purpose of this paper is to build a data-driven Train Delay Prediction System (TDPS) for large-scale Indian Railway Networks which exploits the most recent big data technologies, learning algorithms, and statistical tools.

KEYWORDS

Train Delay, Web Scraping, Data Mining, Deep Learning, Neural Networks, Gradient Boosted Trees, Prediction Systems

1 PROBLEM STATEMENT

The goal of this paper is to develop machine learning models which can accurately predict and classify Indian train delays. We also wish to analyse the vast amounts data which affect train delay.

2 INTRODUCTION & MOTIVATION

In this Paper, we focus our attention towards Indian Trains. Using historical data, we try to build a variety of Train Delay Prediction Systems.

2.1 Indian Railways

Indian Railways (IR) is India's national Train railway system operated by the Ministry of Railways. It manages the fourth-largest railway network in the world by size, with 121,407 kilometres of total track covering a 67,368-kilometre route.[19] IR runs more than 13,000 passenger trains daily, on both long-distance and suburban routes, from 7,349 stations across India. The trains have a five-digit numbering system [Train Codes]. The most common are Mail or Express Trains. In the freight segment, IR runs more than 9,100 trains daily. In the year ending March 2019, IR stated to carry 8.44 billion passengers and transport 1.23 billion tons of freight. In the fiscal year 2017-18, IR said to have earnings of Rs. 1.874 trillion (US$29 billion), consisting of Rs. 1.175 trillion (US$18 billion) in freight revenue and Rs. 501.25 billion(US$7.7 billion) in passenger revenue, with an operating ratio of 96.0 percent. [15]

CS685A: Data Mining, Oct 08–18, 2020, IIT Kanpur

IR is divided into 18 zones, which are further subdivided into 70 operating divisions [20]. In India, about 5 percent of the population (70 million passengers) use Trains daily as means to travel long-distance. So Train Delay affects a large number of people and is one of the major issues IR hasn't been able to solve. About 30 percent of all trains and about 70% of Mail and Express Trains run late. To solve this problem, we need to have a good measure first, this is where Train Delay Prediction systems have become useful to find the amount of delay of trains. If we can predict delay in advance (before a train starts its journey) and the factors affecting it in a quantitative manner, then maybe the IR can put in place delay prevention schemes which will reduce delay times significantly.

2.2 Factors Causing Train Delay

Below are some of the major factors which cause train delay. We will try to incorporate these into our models and analyze quantitatively how these factors affect train delay.

2.2.1 ***Weather:*** [8]
Fog: Visibility dips considerably due to the dense fog. Train Drivers are suggested to not risk driving at normal speeds and usually slow down or even completely stop the trains and wait out the fog. This causes severe delays, on that train and the trains following it on the same track.
Temperature: In extreme heat, there is increased track maintenance due to increased risk of Rail Track Buckling
Rainfall: If an area containing Rail Tracks is flooded with large amounts of water, there is increased risk of derailment and track circuit failure. Such accidents can lead to vast amounts of delay.
Wind: During times of strong winds or storms, more train tests and stability checks take place at intermediate stops due to increased likelihood of de-wirement (the pantograph losing contact with overhead wire) and increased possibility of train overturning leading to increased halt times and hence more delay.

2.2.2 ***Propagation Delay:*** In railway operations, there is a necessary time interval between two trains passing through the same station. If one train is delayed at a station, the following train stopping at the same station may also be delayed. This means the delay of a train can propagate to other trains, resulting in a series of train delays.

2.2.3 ***Festivals & Holidays:*** In order to avoid crowded trains during festive seasons, the Indian Railways have started festive trains or special trains to facilitate holiday travelling. Some of

these trains are scheduled on a last-minute basis, according to the how much crowd is travelling, their demand, etc. This last-minute scheduling creates a host of problems for other trains such as propagation delays and increased halt times, as some of these new trains will affect the scheduled routes.

2.3 Literature Review

Vast amounts of research have been done in the past, using Data Mining and Machine Learning Techniques in the area of Transportation Systems. Following papers are few of the prominent international Papers on Train Delay Prediction System:

2.3.1 Masoud Yaghini et al. [21]
Paper based on Iranian Railways. Paper presented an Artificial Neural Network with high accuracy for the prediction of passenger trains delay. Three diff methods were implemented to define inputs: a) Normalised Real number, b) binary coding, c) binary set encoding points. Other methodologies used were Multi-modal Logistic Regression and Decision Tree. Their models and training times could be improved through meta-heuristic methods such as genetic algorithms or particle swarm optimization.

2.3.2 S. Pongnumkul et al. [13]
Paper based on Thailand Local Passenger Trains. Authors made use of the K-NN approach and used 9 parameters with 6 months of train data. Discovered that one of the major delay factor was the no of stations in between.

2.3.3 L. Oneto et al. [12]
Paper based on Italian Railways. Paper's Major focus was on building a data-driven train delay prediction system that exploits the most recent analytical tools. Authors made use of Extreme Machine Learning, Kernel Methods, Ensemble Methods and Feed-forward neural networks. Additionally they incorporated additional database from exogenous sources,particularly weather data. They performed Simulations which suggested an improved accuracy of up-to 10%.Further accuracy may be achieved by incorporating railway assets conditions.

3 DATASETS USED

3.1 Data Scraped

We collected the following datasets from multiple sources and combined them to determine and evaluate/analyse Indian train delay patterns.

3.1.1 Train Metadata [11] [1]. We collected all possible Indian railways train numbers list and their corresponding train name, train type(Exp, Superfast, Passenger, etc.) and zone. It consists of 8526 trains running in India. The train name, zone is fetched from Open Govt Data, and the train type data is extracted by scraping NTES. The routes of train in our data is depicted in Figure 1. Fig 1. depicts reachability of Indina railway network via 11,056 trains.

3.1.2 Station Metadata [14] [7]. For every stations the following features are taken into consideration which were collected from the official Open Government Data Platform, $StationCode$, $Latitude$, $Longitude$, $StationTraffic$, $Stationdegree$, $Division$, $Zone$, $State$.

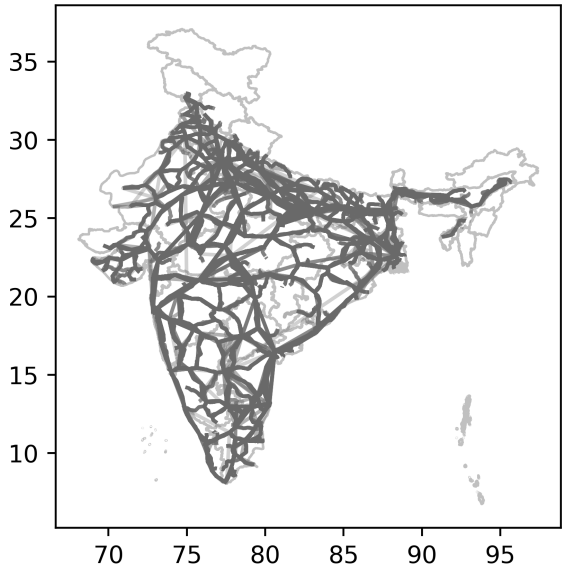

Figure 1: Train Routes in India depicting reach-ability

Latitude & Longitude data is obtained from Google Geocoding API. There are 8149 stations all over India in our data scraped.

3.1.3 Train Schedule and Delay data [16]. We scraped the train running status for every train in the Train Metadata spanned over a period of 3 years [2017-2020] from [16]. This provided us with Scheduled and Actual train arrival/departure time, the source and destination station name along with the intermediate stations. There is a lot of missing data in the data we scraped. Missing data means there is no data at that station for a train for a particular journey.Hence, we used the averages over all train runs for statistical analysis part. Due to rate limiting of requests by the website it took us average of 2 minutes to scrape schedule data of each train.

3.1.4 Holidays or Major Events data [5]. We also wanted to map the possibility of Festival/Holiday rush to our Train Delay Analysis, for this we used this Kaggle Dataset to fetch the last three years of Indian Festival/Holiday data with the features being date, day, holiday, holiday type. We are considering 16 days Gazetted Holiday and 30 days of Restricted holidays per year.

3.1.5 Weather Data [2] [3] [4]. Weather being one of the most prominent reason in the delay of any transportation service, We collected the Temperature, Humidity, Visibility and 12 other features were collected from the NOAA data-set. We use data of nearest weather station on that particular day using the latitude and longitude to find the nearest neighbor.We take 203 points of weather over India and per station we assume the weather of that station on any particular day to be the weather of the nearest weather station as explained above. There were over 8000 train stations but we were able to get data for years 2017-2020 from only 203 weather stations located across India, with majority weather stations being airports. To handle the missing weather data for train stations, for each day

for each train station we employ a nearest neighbor search across all the weather station and assign its weather to the train station.

3.2 Data Collection & Preparation

For each train number in Train Metadata file we use a web scraper to get train type from NTES.
For each station we calculate the total number of trains passing through a Station and total number of direct connections of Station to other stations and store them as $StationTraffic$ & $StationDegree$ respectively. To get Latitude & Longitudinal data, we will use the Google Geo-coding API. The division and zone information is combined from Indian Railways website, for the stations not in the list, we use the latitude and longitude, and assign zones, division same as the nearest neighbor.
We have developed a web crawler to collect train schedule and delay information from site [16]. The website contains historical data of scheduled and actual train arrival and departure time for all stopping stations in its route. The data is collected for three years from 2018-2020.
The holiday data-set is fetched from Kaggle.
The data of 203 Weather stations in India for year 2018-2020 are collected from here. Latitude and Longitude data for each weather station is fetched from Google Geo-coding API, Weather for a train station not having weather station will be assigned weather of nearest weather station.

3.3 Dataset Summary

- **8526** train's delay data scraped of years 2017, 2018, 2019
- **7431** trains used after preprocessing and handling null values
- **203** weather stations considered across the India
- **16** Gazetted holiday and **30** restricted holidays considered per year
- **1974331** total number of train runs.
- **265.688** average runs considered per train
- **18** Weather Features for each location

3.4 Relational Database

As we are going to implement/review various models the input for each model will be different and also try to find patterns in propagation of delay, so instead of combining all the datasets, we will build a relational database. The database then can be easily used for further analysis or combining the tables as required by the ML models also will provide easy way to make structured query. Example to get average delay at a station on holidays we can join the Delay table, Holiday table and take average. The ER diagram for complete dataset is shown in Figure 2

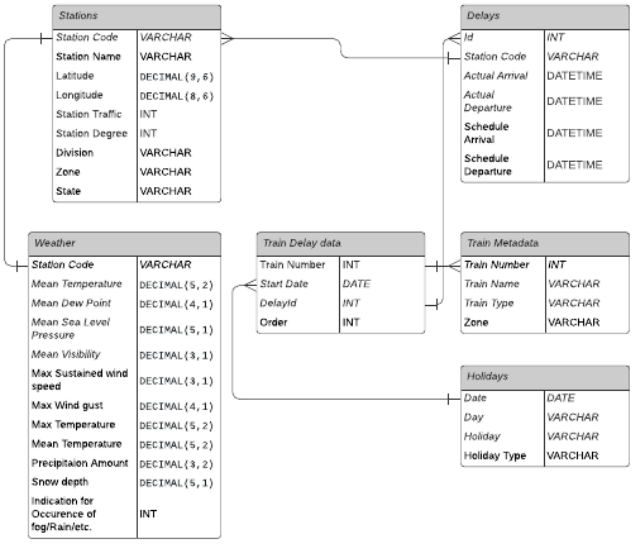

Figure 2: Database ER Diagram for the datasets

Class	Criteria (Delay in min)
0	Delay == Missing
1	0 <= Delay <= 5
2	5 <= Delay >= 30
3	Delay >= 30

Table 1: Classes according to Delay

4 METHODOLOGY

Based on the observation that a station's delay depends only on the delay of the previous stations and not on the latter stations in a train's journey. Hence, for each train's model will contain sub-models for each of its stations.

We chose to both classification and regression in this paper, because they both serve us different useful results. Classification helps us know in what range the delay in going to be in. Regression helps us predict the exact value of delay.

We convert our delay values to classes for our classification model.

While building models we only considered features of previous stations, w.r.to current station in a train's route.

We have used the following methodologies to analyse and predict the delay of the train, weather is our major parameter. We use previous delay data and use predicted weather data of the following day(station wise) to predict the delay on the following day for a particular train. Here is description of models we are going to implement with a brief introduction to the model.

4.1 Gradient Boosted Classification Trees

Gradient boosting is a machine learning technique for regression, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differential loss function. Since the delay for a train on a station does not depend on the next coming stations, To construct the training data for modelling station delay, we used the delay and weather

features for the previous stations and each station has in total of 22 features. Training and testing data was split in the ratio 80:20. We will develop a single GBRT model for each train and predict the long-term delay (3-7 days in advance) using forecasted weather. This paper [18] shows GBRT can effectively model long-term delay predictions.

Seeing figure 3, 4, we observed that delay decreases in an exponential manner with time. There is a major class imbalance for delay > 30 minutes, because of significant lack of data for each train, hence we have to oversample our data for minority class.

Through our classification tasks we also take care of the class imbalance. We do this by using SMOTE from the library 'imblearn'. SMOTE (Synthetic Minority Oversampling Technique): It works by utilizing a k-nearest neighbor algorithm to create synthetic data. SMOTE first starts by choosing random data from the minority class, then k-nearest neighbors from the dataset. Synthetic data is made between the random data and the randomly selected k-nearest neighbors.

For the purposes of this discussion, we focus our attention to four different trains.

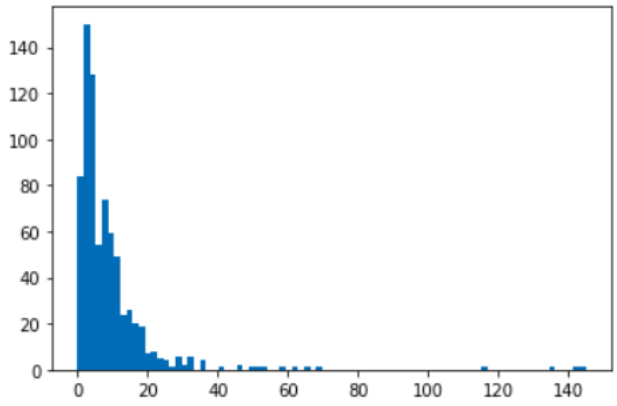

Figure 3: Train delay histogram for Train no. 12009

4.2 Feed Forward Neural Network(FFNN)

The aim is to develop a long-term prediction model for each train. A feedforward neural network (see figure 4) is composed of a large number of processors called neurons that are interconnected and work in unison to solve problems. The connections are modified in a way that minimizes the sum of the squared errors of the outputs. It doesn't have a feed back loop. There are two kinds of neurons: input neurons and weighted neurons. The former is activated through inputs and the latter is activated through the weighted connections from the previously activated neurons.
The calculation of the output can be expressed as in the equation below.

$$o_{pi} = f_i(\sum_j w_{ij} o_{pj})$$

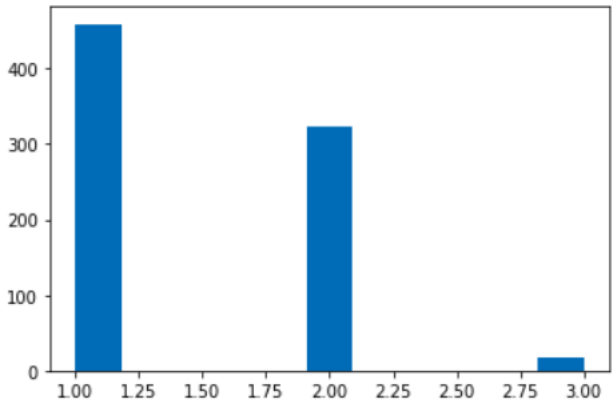

Figure 4: Delay Classes histogram for Train no. 12009

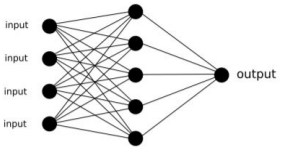

Figure 5: Example of simple neural network

where f_i is the function that activates the neuron eg - (logistic, tanh, relu) etc. We plan to use feed forward neural network which utilizes the following features, holiday,weather(station wise),day,any starting delay,other factors(which may be added). To predict the delay expected for each train. The sum calculated for each pattern p which is to be minimised is.

$$\mathbf{E} = \sum_j 1/2 \sum_{i=1}^{k} (t_{pi} - o_{pi})^2$$

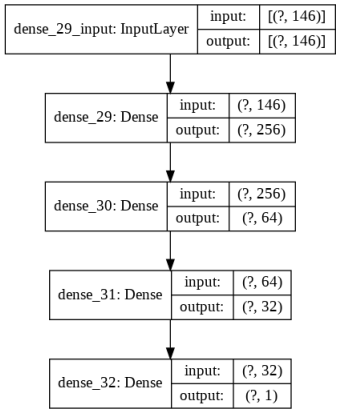

Figure 6: FFN Architecture

where t is the target and o is the output of the network.

4.3 N-Order Markov Process Regression model

In Markov Model the outcome of the current state depends only on the outcome of the previous state, in n-Markov model the outcome depends on outcome of previous n-states.
Previous done research on train delay prediction[10] used n-markov model as one of its approaches. It learnt an n-OMPR model for each of the known stations then used that model later to predict the delay. They used regression models(Random forest and Ridge) and trained it on the n-previous station data at any point of time to predict the delay. We integrated the weather data set at the previous n-stations and used Random Forest regression trees to predict further delay at any station. We found by cross validation n=3 work the best for n-OMPR model. For each train with n stations we developed n-1 models leaving the starting station.

To construct the training data for modelling station delay, we use the delay and weather features of previous n(=3) stations, totalling $n*19$ features of previous stations and 3 features of month, weekday, holiday.
Training and Test data was divided into 80:20 part, for eg. for a train with 1000 runs first 800 runs were used to train the regression model, and latter 200 were used to cross validate with mean absolute error metric used to judge the models.

5 RESULTS

5.1 Data Analysis

5.1.1 Heat Map of delay at station. To visualize the delay at each station, we calculated the average amount of delay over all the runs of trains over the last three years: 2017, 2018, 2019 and plotted the heat map. Fig 4. shows the heat map, with color mapping as follows, Blue represents low amount of average delay, while Green shows more amount of delay.
We observe that mostly for stations in North and north-west districts the train delay is very high and in south eastern coastal region the trains are mostly on time as we can observe in Figure7

5.1.2 Effect of Holidays on Train Delay. We computed the mean delays of all the trains excluding the holidays and compared it against those when we only considered holidays. See Fig4, length of red bar represents the mean delay in minutes for days which were a holiday and length of blue bar represents the mean delay for days which were not a holiday. Its clearly evident that for most of the cases the the average delay caused on a holiday is more than that of a normal day as depicted in Figure8

5.1.3 Effect of Season on Train Delay. We calculated the average delay per season by taking 600 trains complete schedule for 3 years. We observed the average delay of all the trains in winter was 15.4 minutes more than summer and the average delay order is as in the Figure9.

5.1.4 Some Insights into Data. :
Top 10 trains with highest mean delays: See **Table 1**.
Top 10 stations with highest mean delays: See **Table 2**.
More Insights.

- Mean Delay over all trains is **17.35mins**.
- StDev over all trains is **35.36 mins**.
- No of Trains with mean delay in range **0-5 mins** is **3461**.

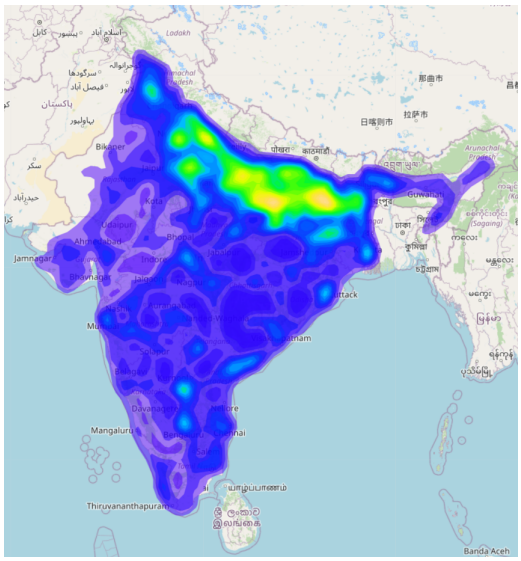

Figure 7: Heat Map of average station delay over 3 years in India

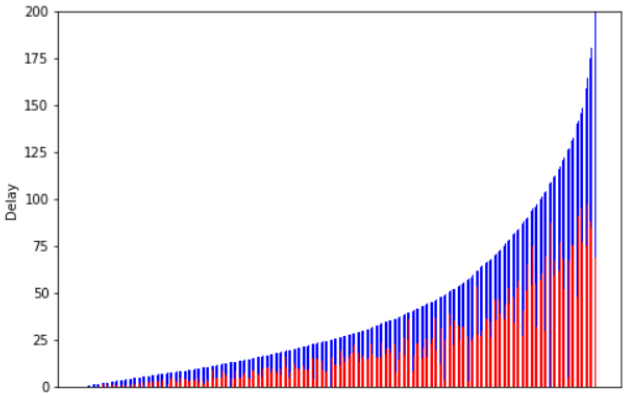

Figure 8: Mean Delays for holidays and normal days(Red:Holidays,Blue:Normal Days)

- No of Trains with mean delay in range **5-30 mins** is **2854**
- No of Trains with mean delay greater than **30 mins** is **1116**
- **Mean Delay** over all Stations is **35.04 mins**
- **Standard deviation** of delay over all Stations **29.47 mins**

5.2 Prediction Models

For each train, the number of models that we build is equal to the number of total stations that train traverses. So each model will be for each station.
We build two classification and one regression model.
We convert our delay values to classes for our classification model. The classes are given in Table 3

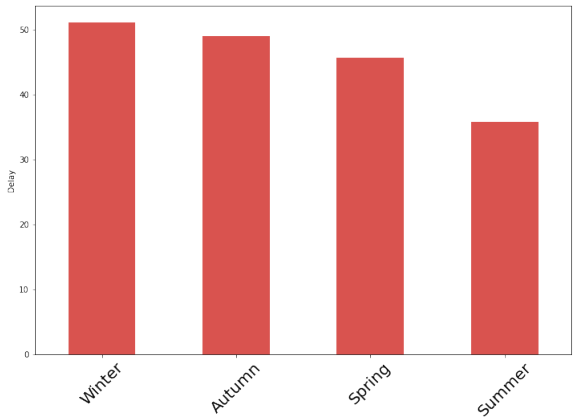

Figure 9: Delay Season Wise

TrainNo	TrainName	Delay(in min)
12598	Antyodaya Express	495
12592	Gorakhpur Express	413
58428	GNPR-PSA PASS	409
11108	GWALIOR Bundelkhand Express	401
15652	Lohit Express	379
15530	Jan Sadharan Express	378
15654	Jammu Ghy Express	376
11123	GWALIOR Mail Express	354
12175	Chambal Express	352
19666	Udz Kurj Express	339
22405	Bgp Garib Rath	326

Table 2: Top 10 trains with highest mean delays

StationCode	StationName	Delay(in min)
SCC	SITAPUR CANT	463
RBGJ	ROBERTS GANJ	457
KRTR	KAROTA PATRI	412
SKTN	SHAKTINAGAR	282
JBN	JOGBANI	279
NKMG	NEW KARIMGANJ	270
DAM	DHAMORA	265
DUN	DUGANPUR	265
DAN	DHANETA	264
PKRA	PARSA KHERA	263
SAR	Shahzad Nagar Railway	262

Table 3: Top 10 stations with highest mean delays

5.2.1 Feed Forward Neural Network Classification. We use Tensorflow > 2.0's [6] classical neural net models. We use the 'to_categorical' to convert the classes in Label-Encoded form and feed this to our model. After training the models, we get the following test accuracy's, given in table 4.

FFN	Accuracy (without SMOTE)	Accuracy (with SMOTE)
12009	0.67	0.42
11001	0.48	0.43
11015	0.50	0.46
12512	0.63	0.59

Table 4: FFN Classification Results

XGB	Accuracy (without SMOTE)	Accuracy (with SMOTE)
12009	0.66	0.78
11001	0.61	0.72
11015	0.62	0.78
12512	0.73	0.84

Table 5: XGBoost Tree Classification Results

N-Markov	MAE (in min)
12009	4.66
11001	14.32
11015	7.98
12512	32.48

Table 6: N-Markov Regression Results

5.2.2 XG Boost Trees Classification. We use xgboost's models. After training the models, we get the following test accuracy's, given in table 5.

5.2.3 N-Markov Regression. We use sklearn's randonforestregressor models. After training the models, we get the following Mean Absolute Error (MAE) in delay, given in table 6.

6 DISCUSSION

We see that the accuracy increases for XGBoosted trees when we apply SMOTE, but accuracy decreases for Feed Forward Neural Network when SMOTE is applied. This shows us the intrinsic difference in nature of the two different algorithms.
We also see that XGBoost Trees performs better than the FFN for all four trains. Hence our preferred model would be XGBoost trees.

7 CONCLUSION

Indian train delay prediction can be effectively modelled using weather, delay and holiday features. We used 203 weather data points to interpolate the 8451 stations. If we come across more accurate weather data then our models will perform very well.
Factors such as season and holidays/festivals affect train delay significantly.
The northern eastern part of India has most train delay.

8 FUTURE DIRECTION

8.1 Multitask Learning

Multi-task learning (MTL) is a subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks. This can

result in improved learning efficiency and prediction accuracy for the task-specific models, when compared to training the models separately.[9]
Most of the previous work on train delay develop a single delay prediction model for each train. We suspect there is a direct correlation between delay of trains running in same division. To exploit this commonality, we are going to learn a prediction model for a division giving output of delay prediction of train in each division. Here the prediction of each train in a given division can be considered as related tasks, which may benefit from selective information sharing across the tasks. We are going to explore a Multitask Deep Neural Network Model for prediction on a division.

8.1.1 Transforming inputs. Each train passing through *n* divisions will be split into *n* sub-trains, each sub-train is then treated as new unique train. All the sub-trains passing through a division will be included as features.

8.1.2 Model Architecture. A simple starter architecture is shown 10, which can/will be later modified to get optimum performance and correctly handle input format.

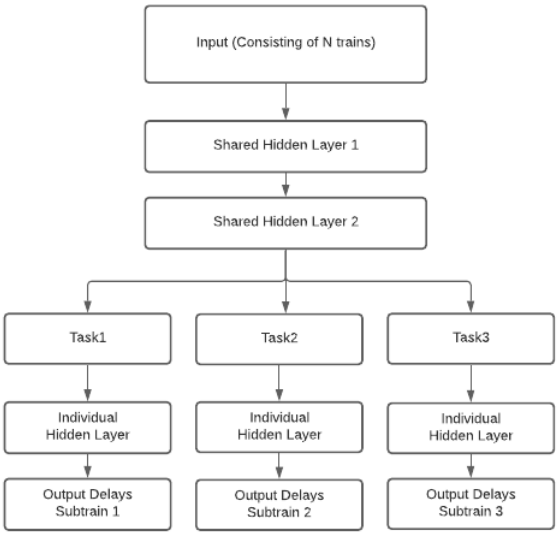

Figure 10: Architecture of simple MT-DNN

8.1.3 Output. The delay prediction for each sub-train is treated as a separate learning task. The output of the model will be predictions of delay at stations for each sub-trains in the division.

8.2 Online Learning RNN

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows RNN to memorize previous delays using its internal memory, and as delay in train is a continuously changing process we can try to model an Online learning RNN which can continuously learn from new inputs, and model the dynamic pattern of train delay.

Online learning algorithms take an initial guess model and then picks up one-one observation from the training population and re-calibrates the weights on each input parameter.Time to time due to many factors the actual time schedule or delay vary over time due to many factors such as railway department optimising the schedule, new routes ,newer engine leading to lesser duration to travel. So the data which is recent along with the weather effect we plan on trying to predict the delay.
The goal of online deep learning is to learn a function.
$\mathbf{F}$:$\mathbb{R}^d \longrightarrow \mathbb{R}^c$ based on the sequence of training samples.
$D = \{(x_1, y_1), (x_2, y_2), ..., (x_T, y_T)\}$ that arrive sequentially.
Where $x_t \in \mathbb{R}^d$ d dimensional representation of features.
The prediction $\hat{y}_t$ and the performance of the learnt functions are usually evaluated based on the cumulative prediction error:
$\epsilon = 1/T \sum_{t=1}^{T} \mathbb{I}_{\hat{y}_t \neq y_t}$
where $\mathbb{I}$ is the indicator function[17].

When the data set observed is less it uses shallow networks but as data set observed gets bigger it uses deep networks.So even for newer trains we should be able to directly use this method to predict its delay.

REFERENCES

[1] [n. d.]. Data Gov Indian Railways. https://data.gov.in/node/4223341.
[2] [n. d.]. Weather Dataset. ftp://ftp.ncdc.noaa.gov/pub/data/gsod.
[3] [n. d.]. Weather Dataset. https://www7.ncdc.noaa.gov/CDO/GSOD_DESC.txt.
[4] [n. d.]. Weather Dataset. https://rp5.ru/Weather_archive_in_Patna_(airport),_METAR.
[5] 2018-2019. Festival Data. https://www.kaggle.com/sakethramanujam/nationalholidaysindia.
[6] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://tensorflow.org/ Software available from tensorflow.org.
[7] Google Geocoding API. [n. d.]. https://developers.google.com/maps/documentation/geocoding/start.
[8] Chris baker. 2010. *Climate change and the railways.* https://www.unece.org/fileadmin/DAM/trans/doc/2010/wp5/Workshop_PPP_05_Baker.pdf
[9] Rich Caruana. 1997. Multitask Learning. *Mach. Learn.* 28, 1 (July 1997), 41–75. https://doi.org/10.1023/A:1007379606734
[10] R. Gaurav and B. Srivastava. 2018. Estimating Train Delays in a Large Rail Network Using a Zero Shot Markov Model. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC).* 1221–1226. https://doi.org/10.1109/ITSC.2018.8570014
[11] Train Metadata. [n. d.]. https://enquiry.indianrail.gov.in/ntes/NTES?action=getTrainData&trainNo=16086.
[12] L. Oneto, E. Fumeo, G. Clerico, R. Canepa, F. Papa, C. Dambra, N. Mazzino, and D. Anguita. 2016. Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA).* 458–467.
[13] S. Pongnumkul, T. Pechprasarn, N. Kunaseth, and K. Chaipah. 2014. Improving arrival time prediction of Thailand's passenger trains using historical travel times. In *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE).* 307–312.
[14] Indian Railways. 2015-2019. Station Metadata. https://indianrailways.gov.in/StationRedevelopment/AI&ACategoryStns.pdf.
[15] INDIAN RAILWAYS. 2018-2019. YEAR BOOK. https://indianrailways.gov.in/railwayboard/uploads/directorate/stat_econ/Year_Book/Year%20Book%202018-19-English.pdf.
[16] RunningStatus. 2015-2019. Train Delay Data. https://www.runningstatus.com.
[17] Doyen Sahoo, Quang Pham, Jing Lu, and Steven Hoi. 2017. Online Deep Learning: Learning Deep Neural Networks on the Fly. (11 2017).
[18] Pu Wang and Qing-peng Zhang. 2019. Train delay analysis and prediction based on big data fusion. *Transportation Safety and Environment* 1, 1 (02 2019), 79–88.

https://doi.org/10.1093/tse/tdy001

[19] Wikipedia contributors. 2020. Indian Railways — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Indian_Railways&oldid=983079386. [Online; accessed 18-October-2020].

[20] Wikipedia contributors. 2020. Zones and divisions of Indian Railways — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Zones_and_divisions_of_Indian_Railways&oldid=982285881. [Online; accessed 18-October-2020].

[21] Masoud Yaghini, Mohammad M. Khoshraftar, and Masoud Seyedabadi. 2013. Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation* 47, 3 (2013), 355–368. https://doi.org/10.1002/atr.193 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/atr.193