Predicting & Analysing Indian Train Delay

CS685A: Data Mining,
Fall 2020,
Prof. Arnab Bhattacharya,
IIT Kanpur

Group ID - 22

Abhishek Singh	170033	astindal@iitk.ac.in
Ayush Hitesh Soneria	170192	ayushhs@iitk.ac.in
Dewansh Singh	170241	dewansh@iitk.ac.in
Jayesh Narwaria	170325	jayn@iitk.ac.in
Prajwal H G	170481	prajwalhg@iitk.ac.in

Motivation

- In India, about 5 percent of the population (70 million passengers) use Trains daily as means to travel long-distance .
- About 30 percent of all trains and about 70% of Mail and Express Trains run late.
- Train Delay affects a large number of people and is one of the major issues IR (Indian Railways) hasn't been able to solve.
- 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.

Problem Statement

- ★ Develop machine learning algorithms to accurately predict and classify Indian Train Delays

- ★ Analyse the vast amounts of data which affect the train delays

Datasets: Sources

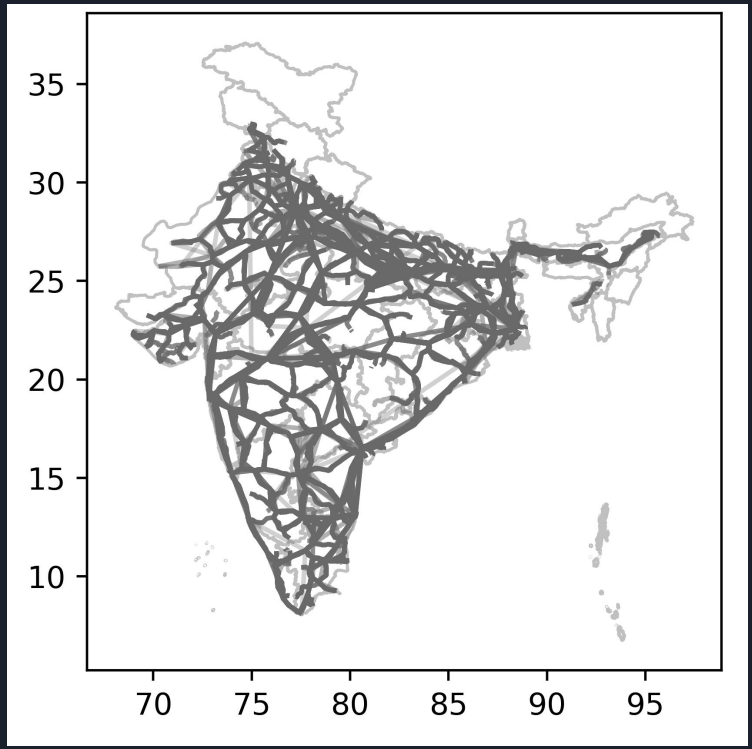
- The train name, zone, station metadata is fetched from Open Govt Data.
- Train type data is extracted by scraping NTES
- Station Latitude & Longitude data is obtained from Google Geocoding API.
- Train Metadata is scraped from running status website.
- Kaggle Dataset to fetch the last three years of Indian Festival/Holiday .
- Weather data is scraped from noaa.gov for the weather stations.

Datasets: Sources

The train name, zone is fetched from
Open Govt Data, and the train type data is extracted by scraping
NTES
Station Latitude & Longitude data is obtained from Google Geocoding API
We scraped the train running status
Kaggle Dataset to fetch the last three years of Indian Festival/Holiday
Weather data :-ftp://ftp.ncdc.noaa.gov/pub/data/gsod

Datasets Used

- Number of trains scraped = 8526
- Number of trains of after processing = 7431
- Total number of stations =8149
- Total No of weather stations = 203
- Percentage missing data = 12.48
- Holidays:- Gazetted = ~16 days
  Restricted =~30 days
- Total no. of runs=1974331
- Avg Runs = 265.688



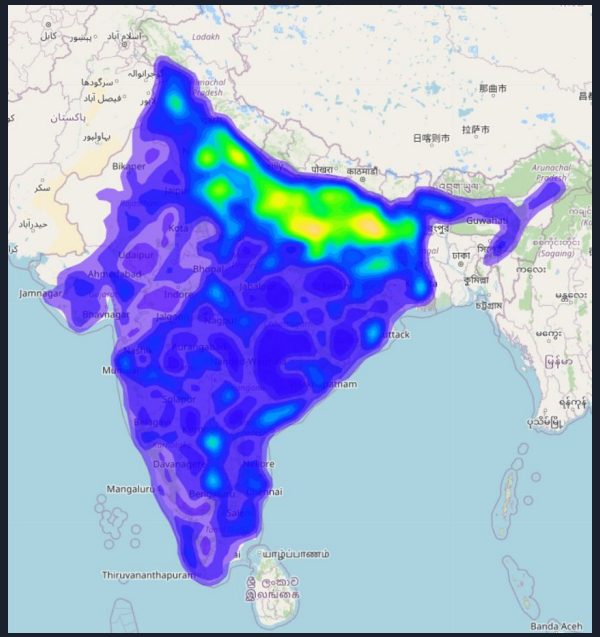
Map of train routes

Methodology

- ➔ Two Delay Classification Models:
  - ◆ Feed Forward Neural Network (FFN)
  - ◆ XGBoost Trees Classifier
- ➔ One Delay Regression Model:
  - ◆ N-Markov (RandomForestRegressors)

Results: Data Analysis

- Train runs late in winter the most(15.4 minutes average late than summer)
- Mean Delay over all trains is 17.35mins and StDev over all trains is 35.36 mins.
- Mean Delay over all Stations is 35.04 mins and Standard deviation of delay over all Stations 29.47 mins.



Results: Prediction

- Our accuracy for all trains is in the range of 70-80 % even though the number of train runs is significantly low.
- The XGBoosted trees classifier performs better than the FFN model.
- SMOTE oversampling technique only improves XGBoost Classifier.
- N-Markov Regression Model gives Mean Absolute Error in Delay in the proximity of 10 minutes for most of the trains.

Thank You

Q & A