Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 175 (2020) 700–705

Procedia Computer Science

www.elsevier.com/locate/procedia

The International Workshop on Artificial Intelligence and Smart City Applications (IWAISCA)
August 9-12, 2020, Leuven, Belgium

Predict France trains delays using visualization and machine learning techniques

Lbazri Sara*, Ounacer soumaya ,Jihal Houda, Azouazi Mohamed

Faculty of science BenM'sik , Casablanca 20000,Morocco

Abstract

The main problem plaguing rail transport is the punctuality of trains. By complying with the French Transport Service Quality Authority (AQST) in France, 2018 is the railway group's "worst" year in terms of punctuality [1]. And according to SNCF over the period from March to November 2017, 18% of TGVs and 17% of Intercities did not arrive on schedule. The Lyon Part Dieu, Paris-Lyon and Marseille Saint Charles stations are particularly prone to these punctuality problems. For example, 32.3% of TGVs that leave Marseille to reach Lille are delayed by at least 15 minutes [2]. In France, a train is considered "late" after five minutes for a journey of less than 1.5 hours, 10 minutes for a journey from 1:30 to 3 hours and 15 minutes for a journey of more than 3 hours. The aim of our research is to use machine learning methods and advanced visualization techniques in order to predict train delays in advance and help rail users to plan their journey and know their train arrival time in advance.

Keywords: machine learning,visualization, SNCF;

1. Introduction

The French rail network covers a total length of almost 28,000 kilometers, making it the second rail network in Europe. It was also recognized as the 11th best rail network in the world in 2018 [3]. The regulatory authority for rail and road activities (Arafer) has stated that the use of the rail network has increased by 7% compared to last years with 92.4 billion passengers / km, also the number of passengers has reached nearly 6% between 2015 and 2017, i.e. 1.4 billion people who took the train. Yet 3.8 million a day [4]. And despite being named the second rail network in Europe 89% of trains in France arrive on time (less than 6 minutes late). This means that on average, 11% of trains arrive late.

* Corresponding author. Tel.: +212 6-50-77-99-92 .
E-mail address: saralbazri@gmail.com

10.1016/j.procs.2020.07.103

These statistics vary according to the type of train: the Intercities show a delay rate of 22%, and even exceed 30% between nine hours to ten hours. Conversely, TERs show a delay rate twice as low (10%) [5]. This difference is partly explained by the fact that TER trains run shorter than Intercities trains, with less risk of wasting time. To solve this problem, predicting train delays is becoming a real need for the SNCF in order to satisfy their customers and allow them to reschedule their trip. The aim of our article is to create a model that allows SNCF to predict train delays based on machine learning algorithms and advanced visualization techniques.

2. Literature survey

Before designing our train delay prediction model, we did a study on train delay prediction research that was published in order to compare the methods and dataset used. In the table below we will present each model:

Table 1: Train delay prediction related work

Author	Publication year	dataset	model	conclusion
[6]Masoud Yaghini and al.	2019	Iranian railways data	neural network	90% accuracy obtained
[7]L. Oneto, E. Fumeo, and al.	2016	Italian railways data	ELM algorithm on spark	Accuracy has been increased by about 10% using weather data
[8]Mohd Arshad, Muqeem Ahmed	2019	Indian railways data	Multivariate Regression Neural Network (NN) Random Forest	improve the accuracy of train delay prediction systems and achieve less error
[9]Emanuele Fumeo And al.	2018	Italian railways data		
[10]okaiah Pullagura, Jeevaa Katiravan	2019	Indian railways data	Decision tree with AdaBoost Recurrent Neural networks(RNN)	Compare average errors of decision tree with and without AdaBoost.
[11]Pu Wang1 and Qing-peng Zhang2	2019	China railways data	gradient-boosted regression trees	combines weather records, historical train delay records and train schedule data to determine the most important factors influencing train delays
[12]Suporn Pongnum Kul at al.	2014	Thailand railways data	KNN algorithm	Enhance the prediction error by 23%

3. DataSet

In order to predict train delays in France, several data should be collected. The main data source and the SNCF open data platform which contains 95 new datasets which relate to the description of the infrastructure (characteristics of lines, civil engineering structures, level crossings, pedestrian crossings, courses of goods, etc.) as well as maintenance and modernization work carried out every week. The remaining 19 datasets are the responsibility of SNCF Mobilities and relate to TER, Intercities and TGV fares, as well as the evolution of train journey times or the acronyms used in the company. The data collected mainly concerns the regularity of the TGV from January 2015 to December 2019, this data will be combined with meteorological data[13].

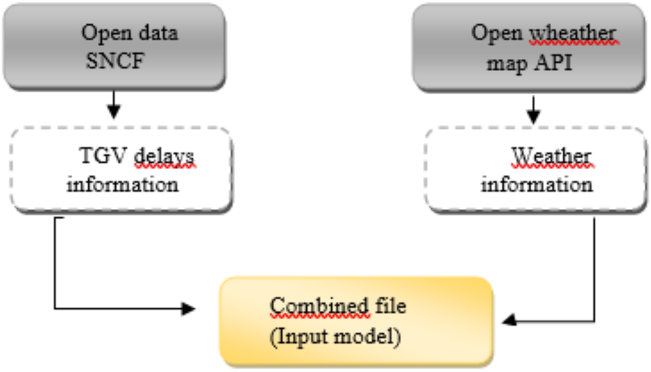

Fig. 1. DataSet Collected from API

The data collected from both dataSet was split into two parts: 80% was used as training data and the rest as testing data and will be combined to provide us with the following features:

- Train's information: It contains the train numbers, their types, their specifics.
- Road's information: It contains the departure and arrival stations, as well as the time of departure and arrival
- Weather information: Contains the weather in different regions in France, snow, humidity, rain ..

4. Proposed model

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it – they are set to work by themselves.

There are 5 steps in the machine learning prediction process :

- Data collection
- Data cleaning
- Training of model
- Evaluation of result
- Deployment

In order to refine and improve this traditional process of predictive analysis we will add visualization or descriptive analysis in order to better understand the problem.

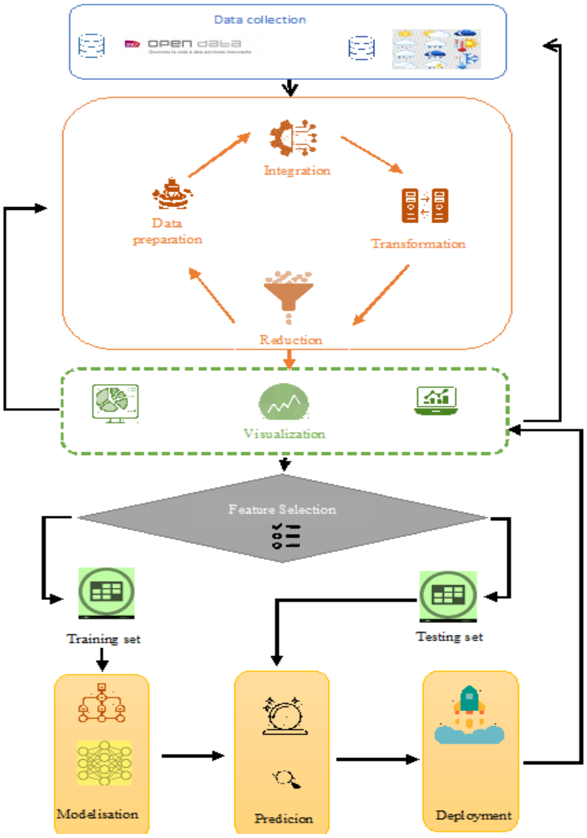

Fig. 2. Proposed model

Our model uses the data collected from the SNCF open data API, this data is cleaned and prepared. Then to better select the features we use visualization techniques using the python library to better understand the data. These data will be separated into two parts: training set and testing test. The first part was modeled as classes and we used the two methods Random Forest and SVM.

4.1. Random forest

It is a classification algorithm that reduces the variance of forecasts from a single decision tree, thereby improving their performance[14]. To do this, it combines many decision trees in a bagging-type approach.For our analysis, we chose 60 and we randomly took 3 entities to divide the tree.The average accuracy of the drive assembly was 71.03% and the error of the test assembly was 81.46%.

4.2. SVM

SVMs are a family of machine learning algorithms that solve classification, regression, and anomaly detection problems. They are known for their solid theoretical guarantees, their great flexibility as well as their ease of use even without great knowledge of data mining[15].

We separated the data into two classes not delayed (delay from 0 to 15 minutes) and delayed (> 15 minutes of delay), in order to know the factors corresponding to the delay. We were using a Gaussian kernel. This algorithm was better for predicting delays, but it took a long time to practice.

At the end, to select the final model we used cross validation. The figure below shows the average of training and test set accuracy

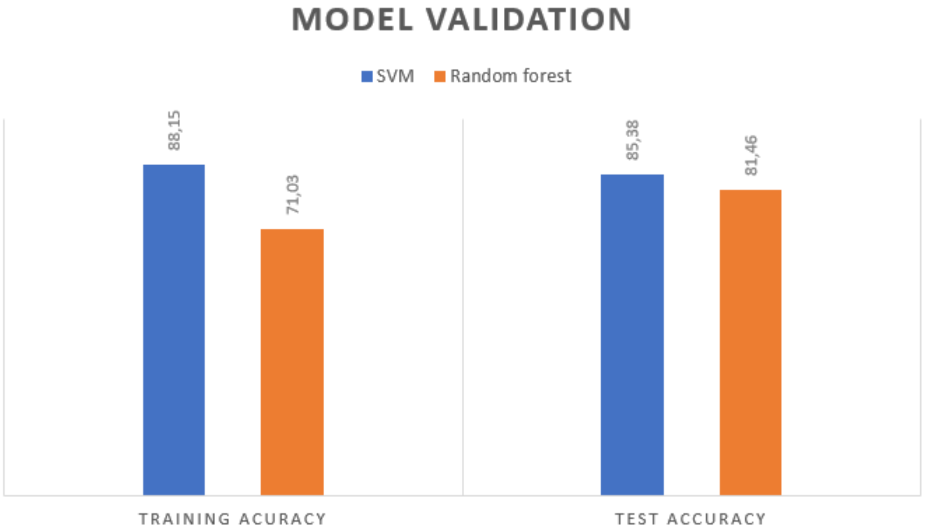

Fig. 3. Average of training and test set accuracy

5. Conclusion

In this article, we used machine learning methods to predict train delays in France. Despite the fact that the study period was very short, we were able to test two SVM and random forest methods which led to very good results. The work done in this article can be extended to use more advanced visualization techniques throughout the predictive analysis process from data preparation to deployment. In this document, we have used only two machine learning methods, other methods can be applied to get better results.

References

[1] Qualitetransports.gouv.fr. 2018. Bilan 2018 De La Qualité De Service Des Transports De Voyageurs En France. [online] Available at: <http://www.qualitetransports.gouv.fr/IMG/pdf/ppt_qst2018_v190417-2.pdf> .

[2]Statistiques.developpement-durable.gouv.fr. 2018. Chiffres Clés Du Transport. [online] Available at: <https://www.statistiques.developpement-durable.gouv.fr/sites/default/files/2018-10/datalab-31-chiffres-cles-transport-mars2018-c.pdf> .

[3] raffin, N., 2017. Retards, Subventions... Les 5 Chiffres Insolites Du Transport Ferroviaire. [online] 20minutes.fr. Available at: <https://www.20minutes.fr/economie/2170531-20171116-trains-retard-subventions-nombre-passagers-cinq-chiffres-insolites-transport-ferroviaire> [Accessed 20 April 2020].

[4] Déléaz, P. and Déléaz, T., 2018. Plus D'un Train Sur Dix Est Arrivé En Retard En 2017. [online] Le Point. Available at: <https://www.lepoint.fr/societe/plus-d-un-train-sur-dix-est-arrive-en-retard-en-2017--12-12-2018-2278808_23.php> .

[5] Sorrell, Steve (2009) "The Rebound Effect: definition and estimation", in Joanne Evans and Lester Hunt (eds) *International Handbook on the Economics of Energy*, Cheltenham, Edward Elgar

[6] Y. Masoud et al. " Railway passenger train delay prediction via neural network model" JOURNAL OF ADVANCED TRANSPORTATION, 47:355–368, (2013)

[7] L. Oneto et al. "Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data," IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, 2016, pp. 458-467. 2016

[8] Arshad, Mohd & Ahmed, Muqeem. (2019). Prediction of Train Delay in Indian Railways through Machine Learning Techniques. International Journal of Computer Sciences and Engineering. 7. 405-411. 10.26438/ijcse/v7i2.405411.

[9] Fumeo, Emanuele & Oneto, Luca & Anguita, Davide. (2015). Condition Based Maintenance in Railway Transportation Systems Based on Big Data Streaming Analysis. Procedia Computer Science. 53. 437-446. 10.1016/j.procs.2015.07.321.

[10] R. Nilsson and K. Henning, "Predictions of train delays using machine learning," Dissertation, 2018.

[11] Wang, P. and Zhang, Q., 2019. Train delay analysis and prediction based on big data fusion. Transportation Safety and Environment, 1(1), pp.79-88.

[12] S. Pongnumkul et al. "Improving arrival time prediction of Thailand's passenger trains using historical travel times," 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 307-312, 2014

[13] Observatoire-transports-hauts-de-france.fr. 2017. Open Data SNCF - Observatoire Régional Des Transports Hauts-De-France (ORT). [online] Available at: <http://www.observatoire-transports-hauts-de-france.fr/open-data-sncf-a71.html> .

[14] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., … & Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, *56*(1), 116-124.

[15] Wang, J., Chen, Q., & Chen, Y. (2004, August). RBF kernel based support vector machine with universal approximation and its application. In *International Symposium on Neural Networks* (pp. 512-517). Springer, Berlin, Heidelberg