Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol.6, No.2, December 2018, pp.389-393
Available online at:http://pen.ius.edu.ba

Using visualization and predictive analysis to predict train delays

Lbazri Sara[1], Oubrahim Younes[1], Rachik Amine[1],Azouazi Mohamed[1]
[1]Laboratory of Information Technology and Modeling
Faculty of science Ben Msik Hassan II University Casablanca, Morroco

Article Info

Received Jun 10, 2018

Keyword:

Railway
Visual prediction
Predictive analysis
Interactive visualization
SNCF
Open data

ABSTRACT

France has the second largest European railway network, with a total of 29,901 kilometers of railway. However, the travel experience of passengers is frequently marked by delays, late arrival of trains at stations, causing inconvenience. The purpose of this paper is to present a new approach for visual prediction of train delays. Our approach is driven by predictive analysis and interactive visualization. The study has benefitted from access to open data SNCF including information about train delays , train number , station , departure and arrival time .Based on this data we develop a new workflow for predictive analysis including visualization in all steps from data pre-processing to deployment .

Corresponding Author:

LBAZRI SARA,
Laboratory of Information Technology and Modeling
Faculty of science Ben Msik Hassan II University
Casablanca, Morroco
Email: Saralbazri@gmail.com

1. Introduction

Predictive analytics is an important part of data analysis, dealing with the problem of the quantitative or qualitative evaluation of the results of a given process [1]. Although predictive analytics has been successfully applied in many applications areas as expert tools, they usually require a good choice of models, parameters and good quality of learning data. Therefore, the user and his expertise are still not applied in this kind of analysis. In addition, the evaluation of prediction uncertainties remains difficult, especially if prediction models are applied in the black box manner. Our idea is to use visualization in predictive analytics, in order to improve the forecasting process and enable the human to interact throughout the process. In this work, we will propose a new approach to predict train delays using visualization and predictive analysis. France, and the SNCF have put in the hands of decision makers a public datasets containing information on trains, delays, schedules, information on stations, to motivate datascients to improve the field of French rail transport [2]. The contributions of this article is to propose a new workflow for predictive analysis including visualization techniques and the human in the predictions of train delays

2. Literature survey

SNCF operates 30,000 km of lines in service throughout all of France[3]. On the other hand, SNCF is still experiencing problems with train delays. Over the period from March to November 2017, 18% of TGVs and

17% of Intercités did not arrive at the scheduled times. 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 leaves the city to Marseille to reach a delay of at least 15 minutes[4].

The main causes of delay: external causes, such as weather conditions, obstacles on the tracks, suspicious packages, malevolence but also social movements (24%) and finally "rolling stock" (21%). To solve the different problems of the railway network in France, SNCF launched in 2016 a proactive transparency policy by displaying all the maintenance and modernization work carried out on the infrastructure. Today, SNCF opens 95 datasets[5], published on the platform data.sncf.com. The aim is to stimulate innovation in the railway sector. These data will be updated and enriched regularly. They can be used by professionals for the realization of data visualizations, by developers or creators of applications or simply consulted by the general public.

So our study on the prediction of train delays in France will be based on these published data, accessible from the API open data SNCF, the purpose of this work is to propose a new method for the prediction of train delays by integrating the visualisation into the prediction process.

Author names and affiliations

Where the family name may be ambiguous (e.g., a double name), please indicate this clearly. Present the authors' affiliation addresses (where the actual work was done) below the names. Indicate all affiliations with a lower-case superscript letter immediately after the author's name and in front of the appropriate address. Provide the full postal address of each affiliation, including the country name and, if available, the e-mail address of each author.

3. Methodology

3.1. Dataset

In this study we will focus only on trains of type TGV because the different types of trains do not have the same specificities, routes, and causes of the delay. This analysis is based on two datasets. The first dataset represents the regularity of the TGV since 2015. On TGV, the regularity is calculated at the arrival of the train at the last station of its course (terminus). This method of calculation, which does not include the intermediate stations, proposes the accumulated delay on the whole of a route. A train is considered late if it arrives five minutes after its scheduled time for a journey of less than 1:30, ten minutes for a trip from 1:30 to 3am, and fifteen minutes for a trip of more than 3 hours[6]. This dataset also presents the Arrivals and departures stations as well as the expected arrival time and departure time. It also represents the causes of delays (traffic, infrastructure, equipment, etc.), it can be downloaded in JSOn, CSV or excel format. The second dataset offered data about the weather in many areas in France, such as temperature, humidity, precipitation, wind speed and more, it can also be downloaded in Json or CSV format. Data from both datasets will be combined to use as input for our model. For the first dataset only information on delays due to the weather will be used because weather data was the first factor causing delays in trains in France. The information from both datasets will be combined and placed in a single tabular file that contains the necessary information about each start and the weather for that particular moment.

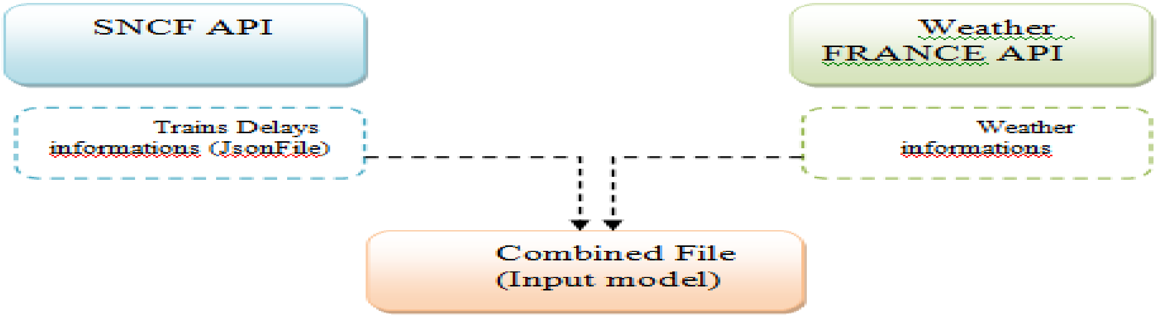
SNCF API
Weather FRANCE API
Trains Delays informations (JsonFile)
Weather informations
Combined File (Input model)

Figure 1: the paths taken by the data derived from SNCFAPI and weather API

3.2. Proposed model

There are several studies done to predict train delays. The goal of our research is not only to predict train delays, but also to improve the process of predictive analytics with visualization techniques in order to have better results for the prediction of trailing delays. The aim of our study is to involve humans in solving the problems of train delays. This study is based on the premise that humans are better than machines to solve some tasks, also to guarantee a better understanding of the data and a quick interpretation. In this section, we explain our proposed model based on predictive analysis and visualization.

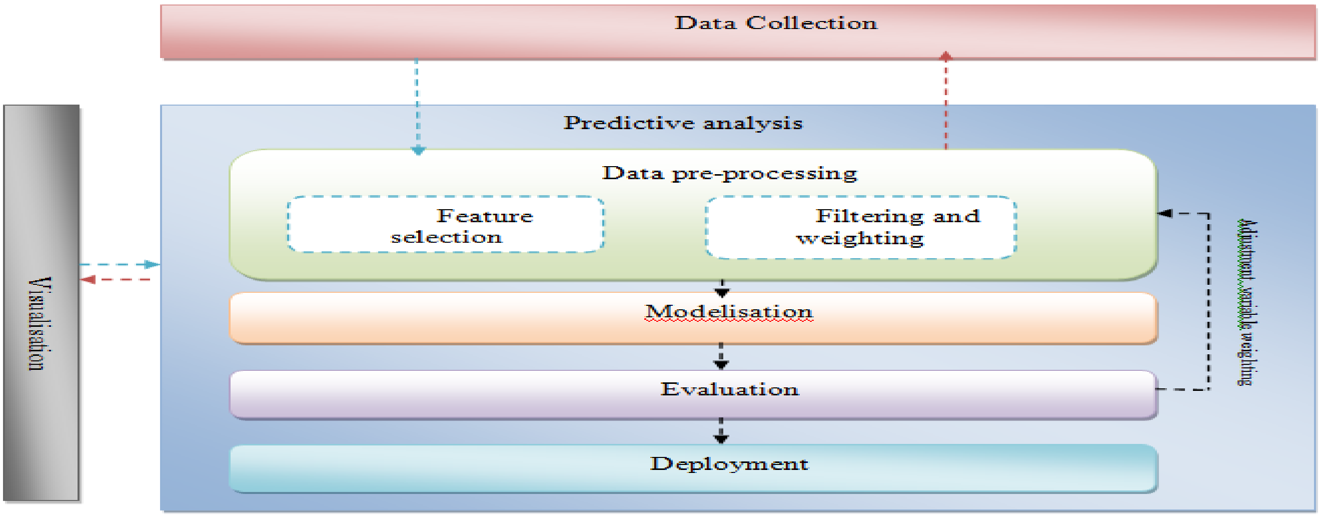

Figure 2: Proposed model using visualisation and predictive analysis

3.3. Data Preprocessing and visualization

Data preprocessing is the process of manipulating the data before the actual extraction steps. it aims to transform raw data into an understandable format. The pretreatment of the data can be divided into several categories: cleaning, integration, transformation, reduction and compression. Our goal is to use visualization techniques, in order to make a human analysis combined with a computer analysis of the data.

Visualization of data also helps to identify data that needs special attention, such as outliers, which may affect our model later. It also helps us to understand the factors that influence the results: In our case we can detect that the weather is the main factor causing train delays. Visualization can also identify different clusters in a dataset, which can be difficult to identify through simple files.

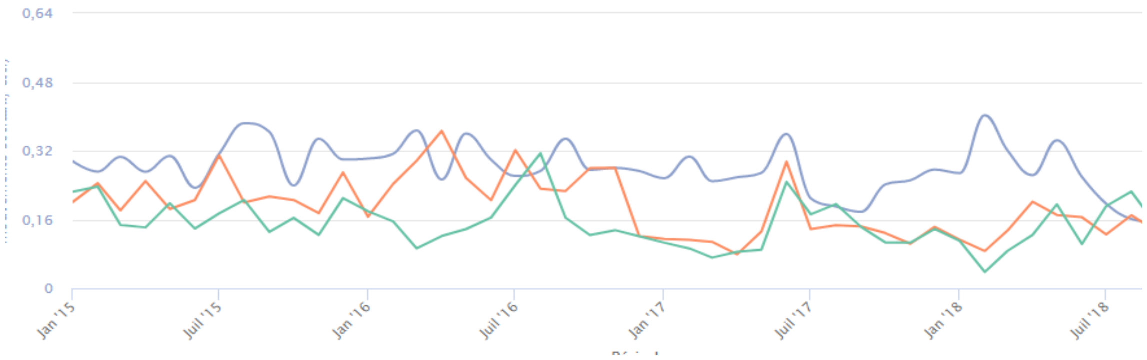

Figure 3: Visualisation average of the trains delays by causes per mounth

The blue line represents the average of the trains in meteorological delays, the orange line shows the average of the trains in delays because of the management of the stations and equipment, and the green line presents the average of the trains in delays due to travelers account. This visualization allowed us to deduce that the weather is the main factor in train delays.

3.4. Modelisation and visualization:

Once the data is processed and prepared, we can start modeling. Many methods have been developed to visualize the models; the objective is to adapt a representation on known data in order to predict unknown data. For example , in regression modeling visual analysis methods have focused on the exploration of data subspaces, the modification of training sets, the removal of outliers, the adjustment of model parameters and modeling with different targets and different optimization functions. Guo et al. presents a visual analysis system that helps analysts discover linear models and extract subsets of data the reasons. They integrate automatic discovery of linear trend and interactive exploration of the multidimensional attribute space, support the refinement of the model, and the selection of subsets of data. Visualization in this step allows the exploration of data subspaces, modification of learning sets, remove outliers (which do not fit well into clusters or predefined classes) ,setting model parameters and modeling with different targets and different discover hidden groups of related items within your data (clustering) and finally ,compare the forecast results and the quality of the model under different parameterizations.

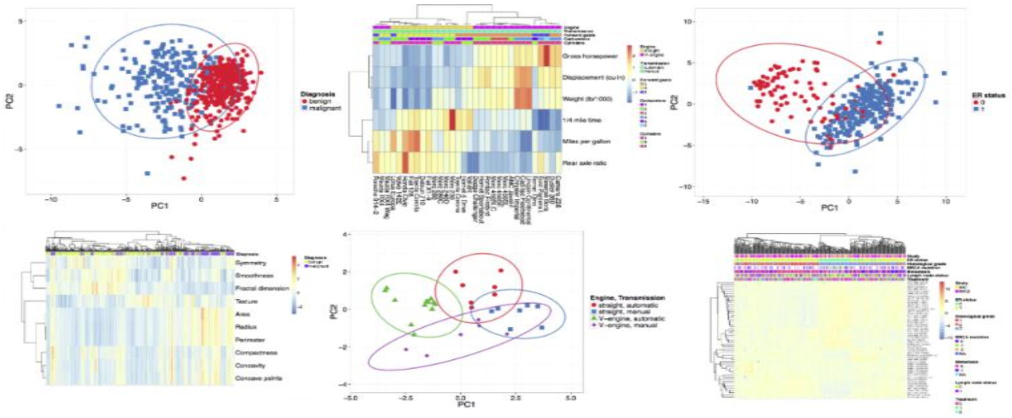

Figure 4: visualization example for cluster with ClustVis Web tool

4. CONCLUSION

We presented a new approach to predict train delays in France. This approach is focused on predictive analytics and visualization. Based on our work and the evaluation of this approach, we have also developed a new predictive analysis process based on visualization. In addition, the results of our research will be discussed during the ICCMIT day in Vienna.

5. REFERENCES

Yamamura, A., 2012. Identification of causes of delays in urban railways. 1st ed. WIT Transactions on The Built Environment, Vol 127: WIT Transactions on The Built Environment, Vol 127.

Arafer.2016. The French passenger rail transport market. [ONLINE] Available at: http://www.arafer.fr/wp-content/uploads/2018/01/bilan-ferroviaire-2015-2016-versionanglaise.pdf. [Accessed 5 February 2019].

Vromans, 2005. Reliability of railway systems. Doctoral thesis, Erasmus Research Institute of Management, Erasmus University Rotterdam, Rotterdam

Xia, Y., Van Ommeren, J. N. , Rietveld, P. & Verhagen, W. (2013). Railway infrastructure disturbances and train operator performance: the role of weather. Transportation Research Part D, vol. 18, pp. 97102.
Zakeri, G. & Olsson, N. O. E. (2017). Investigation of punctuality of local trains: The case of Oslo area. Paper presented at EURO Working Group on Transportation Meeting 2017 (EWGT 2017), September, 4-6, Budapest, Hungary
Bergström, A., 2013. Modeling Passenger Train Delay Distributions – Evidence and Implications. 1st ed. Centre for Transport Studies SE-100 44 Stockholm Sweden: CTS Working Paper.
GUO Z., WARD M. O., RUNDENSTEINER E. A.: Model Space Visualization for Multivariate Linear Trend Discovery. In IEEE Symposium on Visual Analytics Science and Technology (2009), IEEE, pp. 75–82. 7
F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
KUHN M., JOHNSON K.: Applied Predictive Modeling. Springer, 2013.
LU J., CHEN W., MA Y., KE J., LI Z., ZHANG F., MACIEJEWSKI R.: Recent Progress and Trends in Predictive Visual Analytics. Frontiers of Computer Science (2016)
KAPOOR A., LEE B., TAN D. S., HORVITZ E.: Performance and Preferences: Interactive Refinement of Machine Learning Procedures. In AAAI (2012), Citeseer
Zeiler, Matthew D. and Fergus, Rob. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013

Web sites:
Web-1: http://www.data.gouv.fr,consulted 18 april 2018.
Web-2: http://data.sncf.com/explore/
Web-3: https://www.apixu.com/