Classifying train delays based on environmental variables
Machine Learning for Earth and Environmental sciences (Fall 2022)

Max Henking [*] [1 2 3]

Abstract

This paper tries to implement Machine Learning to classify delays by using environmental variables such as rain, sun and temperature.

1. Introduction

Train systems are central pieces of the mobility infrastructure. The size of such infrastructures is rapidly increasing in recent years due to the development high quality services. To sustain such growth in this type of mobility, it is important to maintain a high quality of service. This quality is largely influenced by the frequency of trains and also the reliability in terms of departure/arrival times. Delays can cause a decrease in the trust of the users and as such can lead to a decrease in the number of users and a shift to other modes of transport (cars, planes).

Service disruptions are often times the cause of train delays. These can be due to infrastructure failures, accidents and weather variations. To reduce the impact of such events, delay prediction is to be used as a tool to plan the (re-) scheduling of trains.

The main objective of this report is to identify if it is possible to predict train delays based on environmental variations such as snow, temperature and sunlight.

For this project, the first part will be to evaluate our variables by using a baseline model (Linear Regression). Secondly, a Random Forest algorithm will be applied so as to be able to classify delays.

2. Data

The data used in this project is from the 15th of December 2022. It contains 63137 instances. The departure delay is the dependant variable and the environmental variables are the independant variables.

Feature Name	Unit
tre200h0	°C
sre000h0	minutes
rre150h0	mm/h
train cancelled	boolean
station ID	abbreviation
train type	Regio, ICE,...
+3min departure delay	boolean

Table 1. variables

tre200h0 corresponds to the temperature at 2m above ground, sre000h0 to the duration of sunlight per hour, rre150h0 is the volume of rain per hour. train cancelled is a boolean variable stating if the train was cancelled or not. The station ID is a variable containing the station where the measure was taken (exemple : ZH for Zurich). The train type is a variable describing to which train type the measure was related (Regio for RegionalBahn, ICE for InterCity Express) The choice of the dependant and independant variables is based on Lapamonpinyo and al. (2022).

2.1. Delay data

The delay data was obtained via the Open Data portal of the swiss rail services. The data was extracted for the 15th of November and contains multiple informations regarding each train.

2.2. Meteorological data

The environmental variables for this project were obtained via the Idaweb website of the meteorological office (MeteoSuisse). The data are taken every hour for each station.

3. Methodology

The whole project is done using the programming language Python for the machine learning part. The pre-processing was done using QGIS and Python. The first part of this project was the pre-processing where a join between the meteo stations and the train stations had to be done in order to get data for each instance. In the case of this study, the data was joined based on the nearest neighboring meteo

[*]Equal contribution [1]University of Lausanne, FGSE [2] [3]. Correspondence to: Max Henking <max.henking@unil.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

station to get the most accurate data possible.

3.1. Baseline model

To analyse the link between the dependant variable and the independant variables, a Logistic Regression (LR) can be used. LR tries to predict a class based on continuous or categorical variables. The resulting line from the LR can be interpreted with the following formula : $y = ax + b$, where a is the slope and b is the intercept. In the case of this report, a multiple logistic regression is used, this variant follows this formula :

$$y_i[0-1] = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i, \text{ i = 1,2,..n}$$

Where i is the number of observations, y_i is the dependant variable, x_i are the independant variables, β_0 is the intercept, β_p are the coefficients for the dependant variables.

3.2. Classification algorithm

The support vector machine classification (SVC) was used for this project to classify delay based on the above-mentionned variables. The SVC was implemented using the scikit learn package in Python.

3.3. Code

The code for this project is avalaible at : https://github.com/MaxHenk/2022_ML_EES/blob/c1400c3dd9c5ae97af4070e0091ae154edce1d63/ML_PROJECT_MH.ipynb

4. Results

4.1. Logistic Regression

The LR gives pretty good results and does not show signs of overfitting as we can by figures 1-3. We can see that the model has more trouble correctly labelling trains having delays and is accurate when predicting on-time trains. This result is potentially due to the higher number of trains having less than 3 minutes delay. The model tends to accurately predict delays but tends to predict too many delayed instances compared to what is true (higher true 0 to predicted 1 than true 1 to predicted 0).

4.2. SVC

The SVC has pretty similar results compared to the LR. The only significant variation is related to a slight decrease in the prediction accuracy on the validation set.

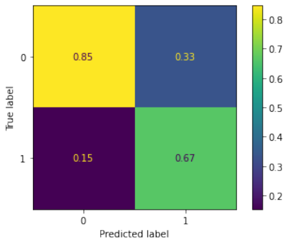

Figure 1. LR training accuracy

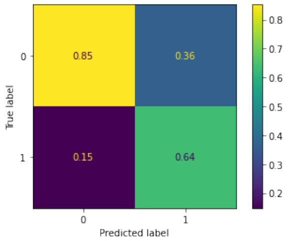

Figure 2. LR validation accuracy

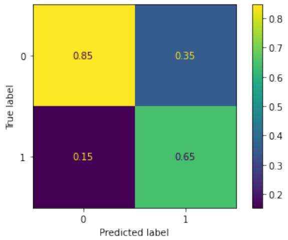

Figure 3. LR test accuracy

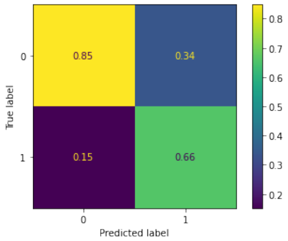

Figure 4. SVC train accuracy

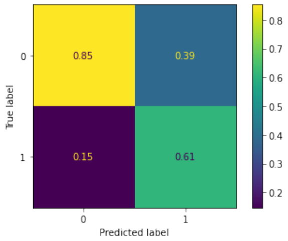

Figure 5. SVC validation accuracy

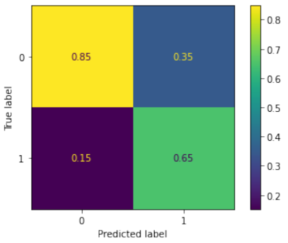

Figure 6. SVC test accuracy

5. Conclusion

Overall both models performed pretty well and did predict more than 80% off our instances correctly. Those results are not foolproof as the chosen independant variables are not the most relevant for this kind of study. The riderhsip data is one part of the information which could be very useful in ascertaining if a train would be late. As such our results may be biased and only show a small part of reality.

	Training	Validation	Test
LR	0.8406	0.8452	0.8394
SVC	0.8410	0.8445	0.8405

Table 2. Overview of accuracy scores

Further work could be to analyse the feature importance to see if a delay is more related to environmental variables or to the station caracteristics. This could prove to be useful in identifying problematics nodes in the swiss rail system.

References

- Lapamonpinyo, P., Derrible, S., Corman, F. (2022). Real-Time Passenger Train Delay Prediction Using Machine Learning: A Case Study With Amtrak Passenger Train Routes. IEEE Open Journal of Intelligent Transportation Systems, 3, 539-550. https://doi.org/10.1109/OJITS.2022.3194879

- Y. Ding, “Predicting flight delay based on multiple linear regression,” IOP Conf. Ser.: Earth Environ. Sci., vol. 81, p. 012198, Aug. 2017, doi: 10.1088/1755-1315/81/1/012198.