Feasibility study on Machine Learning techniques applied to predict train incidencts: a catenary occurrence log case study

Haritz Laboa and Marina Aguado*

November 23, 2018

Abstract— **The railway sector will be the most rising means of transport in the coming years. However, incidences are still frequent, and pose a problem both security and economic. In order to improve the traditional methods of prevention and maintenance, Data Science and, especially, Machine Learning algorithms, have supposed a paradigm shift.**

This paper describes the use of Data Science to implement a predictive model related to railway incidents, specifically, those incidents linked to the catenary. In this sense, in the last years multiple projects of predictive models on railway incidents have already been addressed. However, previous studies on occurrences where carried out under the trainset consideration and linked to features linked to these trainsets (speed, type, rolling stock age, etc). In contrast, this research focuses on the railway infrastructure and its related potential features that may have contributed and that have not been previously taken into consideration, such as the geographic characteristics of the railway lines.

I. INTRODUCTION

In a time where sustainability and pollution have become a pressing issue, the need to carry goods and people without CO2 emissions is turning essential. In relation to this point, nowadays the train is the most efficient means of transport. Over one third of the energy used in the railways is electric, and one quarter of worldwide lines are electrified. This differs from the outlook drawn by road vehicles, responsible for a quarter of the total greenhouse gas emissions. Therefore, the railway, which houses 8% of world transport [3], will be a fundamental element for social economic development in coming years, becoming increasingly important.

However, this growth contrasts with the fact that incidences continue already being frequent. Only in the Basque Country (region of Spain with 2 million people) occur over 1,000 incidents every year. Among them, those with the most impact are linked to the catenary. This kind of events occur little regarding the rest, and despite this, their transcendence is greater; from snags that cause long delays to accidents and derailments.

To shed light on these issues, the expert systems used by railway companies usually include traditional statistical methods. Although, these type of approaches demand too many resources and often extract no knowledge.

For this reason, in recent years data science has been breaking into the railway environment with increasing relevance. More research is being done using Machine Learning algorithms to make predictions. Most of these studies take the rolling stock as the central axis. They analyse their characteristics (speed, antiquity, etc.), and based on them try to figure out if the trains will suffer any mishap. Nevertheless, many of the incidents of catenary are independent of trains. Thus, when analysing catenary incidents, a new approach is required.

This paper describes a catenary occurrence research focused on data related to the infrastructure itself, and not on the rolling stock. The study done on it demonstrates that when adding meteorological and geographical data to catenary occurrence logs, over 15% of the total occurrences could be predicted.

*This work is supported by the I2T Research Group, University of Basque Country, Bilbao, Spain. http://i2t.ehu.eus

II. RELATED WORK

This section describes related work in using Data Science in rail transportation, after a brew view of the relevance of data mining nowadays.

A. *Machine Learning nowadays*

Machine Leaning is already ingrained in our daily lives. Recommendation systems are held by Machine Learning algorithms. Anti-spam or anti-virus systems are also based on Machine Learning, as well as face detection or speech recognition systems.

In recent years are proliferating applications based on this advanced data methods. Due to this, in the last decade public statements and big corporate around the world have begun to offer their information through Open Data platforms. This big availability of data has led to all type of analyzes and predictive studies [2] [1] and leaving contributions in many areas.

Thanks to their predictive capacity, these machines are generating a real revolution in many organizations decision making. Companies from different sectors have already approached data analytics. Thus, studies to predict under what conditions will clients leave, or preventive maintenance plans based on conclusions given by intelligent systems [4], are increasingly common.

B. *Machine Learning in the railway environment*

The railway environment has not been kept apart from the potential of using these advanced techniques. Parallel to the growth of Data Science itself, in the last years have emerged many projects based on Machine Learning.

An example is a project called *Train Tracker* [5], where incidents reports sent by train users themselves are used to predict delays. Another interesting case could be the study by Donald E. Brown [6], who applies text mining to incident registers for discovering new factors related to American rail accidents.

Other researchers have developed predictive systems using Machine Learning models. There is, for example, the study carried out by Christoph Bergmei [7], who makes use of statistical regression methods to develop maintenance plans for high-speed trains. Or the Hongfei Li study [8], which combines the speed information of the trains and the climatology to create a classifier that allows elaborating a plan of incident prevention.

The aforementioned investigations elaborate predictive models based on Machine Learning, as in this project. In the case of Hongfei Li, for example, even adds meteorology features in the same way done in this research. However, the work presented in this paper differs from those studies in the focus used to build instances of the dataset. They take the rolling stock as the central axis (speed, class, rolling stock age, etc.), while here the focus is placed on the railway infrastructure itself.

III. DATASET STRUCTURE

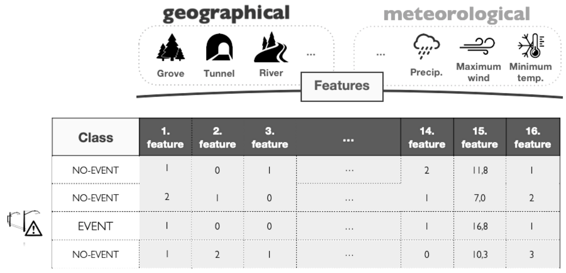
This project has focused on constructing a classifier capable of predict catenary incidents. For this purpose, we required a dataset with labelled instances. But since this dataset did not exist, it has been necessary to create it.

We built the skeleton of the dataset creating a set of instances. After that, we labelled each of these instances, specifying in each case, whether the instance corresponds to normal behaviour, or an incident. Hence, the label in this project is binary: ”event, or ”no-event. This label it is the variable we want to predicted. To figure out this incognita, the classifier requires features so it can find patterns and guess the correct value of the label in each case.

Nevertheless, a railway infrastructure manager provided a set of registers of real incidents, but we had no data that reflects normal behaviour. Having both class of data was essential since the classifier learn to predict with the differences between them. Therefore, a question arises: how to create data that correspond to ”non-events, that can be used to build a dataset?

We thought first about creating instances depending on the trains. Each instance would correspond to a rolling stock. The value of its label would be ”event if the rolling stock has suffered an incident during its journey, and ”no-event if it had reached its destination with no anomaly. Features would be factors related to the rolling stock (speed, type, antiquity, etc.).

However, although some catenary incidents are connected with rolling stocks - such as the fusion of contact wires with the pantograph - many other are independent of trains. Thus, catenary incidents must be studied from another perspective: putting the central axis on the infrastructure itself. As a result, we discarded incorporating train features, and we created the dataset including 2 major types of features: meteorological features, and geographical features.



Class	1. feature	2. feature	3. feature	...	14. feature	15. feature	16. feature
NO-EVENT	1	0	1	...	2	11,8	1
NO-EVENT	2	1	0	...	1	7,0	2
EVENT	1	0	0	...	1	16,8	1
NO-EVENT	1	2	1	...	0	10,3	3

Fig. 1. Dataset structure

IV. DATASET CONSTRUCTION

A. Instance construction

The construction of the dataset has been started creating the instances. This process makes up the skeleton of the dataset and has been carried out from a temporal segmentation and geographical segmentation.

Thus, each instance represents the snapshot of each railway track section, on a specific day. With the division of the railway lines under study, have been created 94 sections. The research interval, in turn, corresponds to 7 years. Therefore, 240,358 instances have been built, as a result of multiplying 94 sections with 2,557 days.

Segmenting railway lines into section allow specifying where an incident has occurred. In addition, track sections permit establishing the climatic conditions with accuracy, since make possible to choose the nearest meteorological station. Also, dividing railway tracks enable the inclusion of geographical features which describe the location (if it is an urban environment, if there are tunnels, etc.).

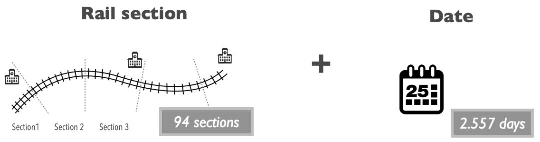


Fig. 2. Dataset skeleton

B. Geographical feature construction

The geographical features have been generated synthetically. This construction has been done using Google Earth - except for those features related to train traffic. Thus, all the railway track sections have been plotted on a map. After that, sections have been flown, pointing in each case the geographical characteristics. With this information have been created the following 10 features:

- GROVE: Expresses if the section crosses a grove. (No - A little - Yes)
- URBAN: Expresses if the section crosses an urban area. (No- A little - Yes)
- TUNNEL: Expresses if the section has a tunnel. (No - Short - Long)
- RIVER: Expresses if the section crosses or adjoins a river. (No - Yes)
- OVERPASS: Expresses the number of overpasses included in the section. (None - A little - Many)
- STATION: Expresses if the section corresponds to a station, or to the route that links two stations (No - Yes)

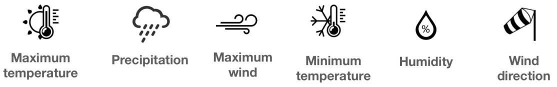
- DISTANCE: Expresses the distance of the section in km
- COMMUTER TRAIN: Expresses the number of commuter trains that cross that section, in the course of a week
- LONG-DISTANCE TRAIN: Expresses the number of long-distance trains that cross the section, in the course of a week
- FREIGHT TRAIN: Expresses the number of freight trains that cross the section, in the course of a week

C. *Meteorological feature construction*

Climatological information has been obtained through Open Data. As a result, have been gathered over 50 million readings, related to meteorological stations near the railway lines.

To establish the climatic conditions with accuracy, each track section has been associated with the nearest meteorological station. Once association was made, data has been processed, grouping the readings of each day and establishing the maximum, average, etc. in each case. Before completing the construction of the meteorological features, generated data has been cleaned, resolving missing values, and checking that values were in a correct range.

In total, 6 meteorological features have been created:



- MAXIMUM TEMPERATURE: Expresses the maximum air temperature, expressed in degrees centigrade, in the interval of 1 day
- PRECIPITATION: Expresses the maximum intensity of the precipitation in the interval of 1 day, expressed by a num. code (*table I*)
- MAXIMUM WIND: Expresses the maximum horizontal wind streak expressed in kilometers per hour, in the interval of 1 day
- MINIMUM TEMPERATURE: Expresses the minimum air temperature, expressed in degrees centigrade, in the interval of 1 day
- HUMIDITY: Expresses the average of the relative humidity of the air in the interval of 1 day, expressed as a percentage
- WIND DIRECTION: Expresses the average direction of the wind in the interval of 1 day, expressed by the cardinal points

TABLE I

CONVERSION OF PRECIPITATION

0 *mm/h*	WITHOUT RAIN	(0)
> 0 *mm/h* & ≤ 2 *mm/h*	WEAK	(1)
> 2 *mm/h* & ≤ 15 *mm/h*	MODERATE	(2)
> 15 *mm/h* & ≤ 30 *mm/h*	STRONG	(3)
> 30 *mm/h* & ≤ 60 *mm/h*	VERY STRONG	(4)
> 60 *mm/h*	TORRENTIAL	(5)

D. *Dataset labeling*

For the development of this project, the railway infrastructure manager has provided 708 incidents of the catenary. However, not all incidences met the needs of the study. Therefore, the first step has been selecting only those incidents related to the infrastructure.

table

Those instances coincident in date and section with the filtered incidents have been labeled as ”event - 405 instances -, labelling the rest of instances as ”no-event - 239.908 instances.

V. MODELLING

A. *Dataset balancing*

Before training the predictive classifier, the dataset has been divided with the following proportion: 75% for training, 25% to test the model implemented. However, the dataset that has been built is unbalanced. There are extremely few incidents (events) compared to instances that represent normal behaviour (non-events). Specifically, for each event, there are about 500 events. This imbalance is a problem when implementing a predictive classifier since an inequitable distribution of the instances promotes the learning of one class to the detriment of the other.

To resolve the imbalance, it has been used a hybridization between the techniques of over-sampling and under-sampling, since it has shown better results than a simple re-sampling. Specifically, has been selected a hybridization between Smote and Tomek-Links techniques: through Tomek-Links, the “non-events” have been subtracted, and through Smote, the ”events have been replicated. As a result, it has been possible to balance the distribution of the instances of the training set before proceeding to train the model.

B. Construction of the predictive classifier

Once the training set has been balanced a Random Forest multi-classifier has been created by assembling 15 trees. To decide the number of trees, 25 iterations have been carried out, measuring the prediction error in each case, and choosing the number from which the error was stabilized. The evolution of the prediction error can be seen in the following figure:

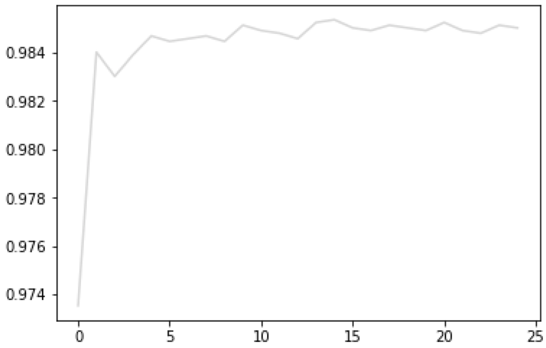


Fig. 3. Evolution of pred. error related to number of trees

Finally, one of the great features offered by Random Forest, is that the algorithm can offer an approximation of the importance of each feature when generating a prediction. This approximation is calculated based on the variation of the prediction error when swapping a feature while keeping the rest unchanged.

In this way, a first model has been generated to measure the importance of each feature, and once the most relevant variables have been discovered, definitive model has been constructed and trained. Distribution of the approximate weight of each feature can be seen in the next graph:

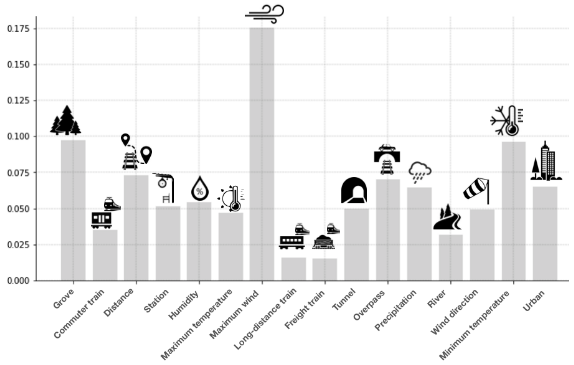


Fig. 4. Feature importance

VI. Metrics

To test the quality of a classifier, one of the most used tools in Machine Learning is the confusion matrix. This matrix serves to visualize the predictions made together with the real values. In this project presented, the obtained confusion matrix is the following:

TABLE II

Confusion matrix

		Real	
		event (+)	**no-event** (−)
Pred	**event** (+)	29	22
	no-event (−)	165	11.719

Using the confusion matrix it is possible to get metrics that measure the efficiency of a classifier. The most common metric for assessing the quality of a predictive classifier is accuracy.

$$\mathit{Accuracy} = \frac{TruePos. + TrueNeg.}{Total\ predictions} = 98,43\%$$

Accuracy is useful and significant in most cases, however, this metric is not too relevant in unbalanced scenarios, since it does not distinguish between classes, and normal behavior predictions are treated the same as incident predictions. Therefore, for this project, precision and recall metrics are more appropriate, since they do measure the ability to detect an event (incidence).

$$Precision = \frac{TruePos.}{TruePos. + FalsePos.} = 56,87\%$$

$$Recall = \frac{TruePos.}{TruePos. + FalseNeg.} = 15,51\%$$

Can be seen how the precision of the model is greater than 50%. This means that in more than half of occasions, when the model predicts an incident it is really an incident. Attending to the recall we see it is close to 15%, so of all the catenary incidents that occur, the classifier is able to predict 15%.

VII. Conclusions and Future Research

The results presented in the previous section show the feasibility of the new approach proposed in this paper. Focusing the study on the railway infrastructure and data related instead of the rolling stock, over 1 out 10 catenary occurrences can be predicted. And although this research has been carried out collecting real data from a railway infrastructure manager in a specific region, the methodology here reflected could be easily used by any railway infrastructure manager since most of the features are commonly used and available by any of them.

Thus, throughout this exposition we have seen an example of how to face the construction of a dataset with this new point of view. Starting with the construction of a dataset as the result of combining the information available on the incidents, along with geographic data and climatological data, and ending with a resolution of the great imbalance between the number of incidents and the number of cases that correspond to normal behavior.

However, there is a lot of additional work that needs to be done before we can deploy the predictive model that has been implemented in this project. In order to develop a visual tool capable of forecasting incidents in real time, it is essential that the railway infrastructure managers monitor the status of the catenary using sensors in the infrastructure. In addition, it would also be convenient to record weather information at different points on the line. Above all, the speed of the wind.

On the other hand, itis worth mentioning the wide margin for improvement of the results expressed in this article. Although 15% recall is considerable, in this study other algorithms apart from Random Forest have not been considered, with which better predictions may be obtained. In addition, the chosen temporal segmentation was 1 day, but perhaps creating instances with finer granularity could improve the results. Finally, in view of the close relationship between wind speed and many of the incidences of the catenary, a more detailed study in this aspect could also contribute to make better catenary incident predictions.

References

[1] *La guerra en Siria y refugiados, un drama humanitario reflejado en datos,* URL = http://bilbaodatalab.wikitoki.org/

[2] *Bilbao DataLab,* URL = https://data-speaks.luca-d3.com/2016/12/la-guerra-en-siria-y-refugiados-un.html

[3] URL = https://www.eldiario.es/edcreativo/viajes/sostenibilidad/futuro-movilidad-sostenible-viaja-tren_0_686981444.html

[4] Anna Corazza ; Francesco Isgr, *"A machine learning approach for predictive maintenance for mobile phones service providers"*

[5] URL = http://www.melbournetraintracker.com.au

[6] Donald E.Brown, *"Text Mining the Contributos to Rail Accidents"*

[7] Christoph Bergmeir ; Gregorio Sinz ; Carlos Martnez Bertrand ; Jos Manuel Bentez, *"A Study on the Use of Machine Learning Methods for Incidence Prediction in High-Speed Train Tracks"*, Lecture Notes in Computer Science, 2013

[8] Hongfei Li ; Dhaivat Parikh ; Qing He ; Buyue Qian ; Zhiguo Li ; Dongping Fang; Arun Hampapura, *"Improving rail network velocity: A machine learning approach to predictive maintenance"*, 2014