Australasian Transport Research Forum 2018 Proceedings
30 October – 1 November, Darwin, Australia
Publication website: http://www.atrf.info

A Design Framework for Big Data Analysis in Railway Transport

Jingyu Zhang[1], Andrew Padley[1], Yuqi Wang[2], Shiping Chen[2, 3]

[1] NSW Trains, 470 Pitt St. Sydney, Australia 2000
[2]University of Sydney, Australia 2006
[3]CSIRO, Australia 2122

{Jingyu.Zhang, Andrew.Pedley}@transport.nsw.gov.au

Abstract

It is well-known that cloud computing has many potential advantages and a multitude of enterprise applications are currently migrating to public or hybrid cloud environments. This paper presents the workflow and implementation of a big data analysis framework with both descriptive and predictive analytics capability. This paper outlines the framework's ability to support not only a high level integrated KPI framework for enterprise reporting but also enabling a structured deployment of machine learning techniques to predict key drivers of railway performance. The two use cases highlighted are prediction of anti-social behavior incidence and train punctuality. Based on our evaluation, these two models are able to assist railway staff to estimate anti-social behavior cases and train punctuality at NSW TrainLink (NSWTL), Australia.

1.Introduction

With the increase of both data storage capability and the prevalence of distributed computing, big data is expected to change the railway transport domain completely. These technologies, like numerous systems before them are another step towards more informed decision making at both an operational and strategic level. Currently, the most popular areas for advanced techniques in rail network analytics are safety, operations, and maintenance (Ghofrani, He, Goverde, & Liu, 2018). It is vital for a railway operating in the 21st century to be able to not only adapt but capitalise on new technologies to ensure corporate decision making is timely, effective and cost-efficient. The features of big data can be described using the '5Vs' (Laney, 2001) of volume, velocity, variety, veracity, and value correspondingly.

In this paper, we address the above challenges by developing a framework for data analytics that allows multiple systems to cooperate and synchronize with each other to analyze and predict the key elements in efficient railway transport services.

4. NSW TrainLink Integrated Key Performance Indicator System

BABEL has been deployed as an enabling technology for advanced analysis and predictive modelling at NSW TrainLink (NSWTL). NSWTL is a multi-modal passenger transport service provider, providing rail and coach services across New South Wales and connecting to other states inside Australia. The main vision of the Integrated Key Performance Indicator application is to identify and analyse dominant levers of business objectives to both managers and operational staff.

Our KPI regime serves three main purposes. Firstly, it served customers as we mature the ability to enhance the elements of our service that affect them most. Second, it delivers enhanced evidence to our main stakeholders that the levers of our performance have not only been identified but are being actively managed. Third, it delivers NSWTL the enhanced capability of moving from an empirical recording of the performance to deriving general rules of operational performance. This translates to learning and proving the why behind the what.

Figures 1 and Figure are provided as illustrative examples only. Due to privacy and security issues the real data cannot be provided. To reiterate, the below data is fabricated and the plots are provided as an example of the platform outputs.

Figure 1. Customer Injury Analysis

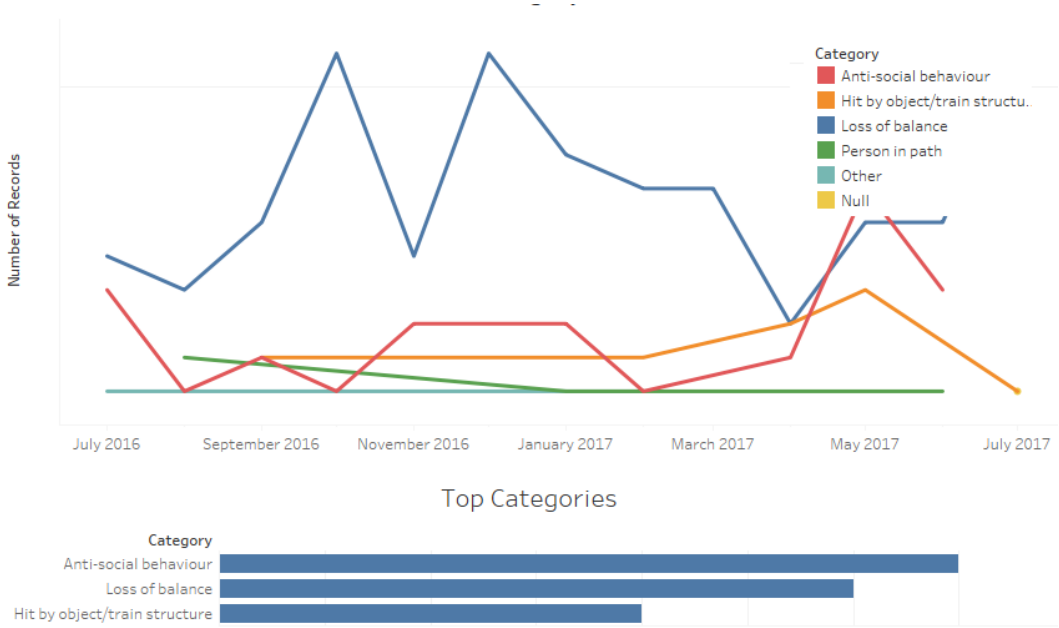
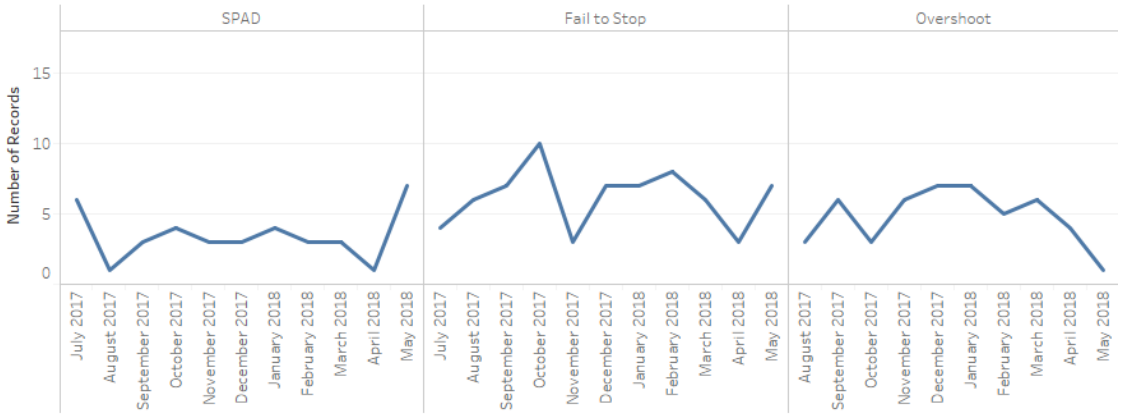


Figure 2. Technical Incidents Analysis



The overall benefit of the KPI framework was to focus enterprise efforts on the levers of business performance as all employees are provided an overview of the major drivers of NSWTL performance. The senior leadership was better able to view the business in a strategic

context and rectification efforts were able to be entered into, tracked and resolved more efficiently. A derivative benefit was the illumination of data quality holes in the business and the foundation for robust discussions around prioritization of fixes.

5. Predictive modelling with BABEL

In BABEL, one of the key functions is predictive modelling. We present two predictive modelling applications that are currently running on BABEL in this section.

5.1. Anti-Social Behaviour (ASB) prediction

Combating anti-social behavior (ASB) that allows for proper public security resource allocation is pivotal to ensure public safety on the New South Wales railway network. We are, however, recognize that the Police Transport Command has limited resources to ensure the safety of our customers and staff at all times. Rail customers who travel short or long distances by train rate safety as their highest priority (Moore, 2011). In order to enhance public safety, transport operators must find a way to allocate limited resources effectively. The introduction of predictive modelling for ASB at NSWTL is a progressive step forward to anticipate 'hot-spots' on the network and help provide a safe environment for staff, customers and the community on both stations and trains.

Machine learning is a methodology to discover latent information or patterns in datasets (Nirkhi, Dharaskar, & Thakre, 2012). It is an interdisciplinary study that involves real world situations and mining algorithms in a pure mathematical description that gives insightful analytics into various situations.Within the supervised learning area of machine learning, decision trees methods are frequently utilized. Tree-based methods for regression and classification involve stratifying and segmenting the predictor space into a number of smaller regions (James, Witten, Hastie & Tibshirani 2013).

This algorithm is based on information theory that involves the splitting of nodes to allow for optimal exploration of all available features (Rokach & Maimon, 2005). The split criteria are based on information gain as trees are expected to be simple. Given defined entropy for J classes as in Equation 1,

$$H(T) = I_E(p_1, p_2, \dots, p_i) = -\sum_{i=1}^{J} p_i \log_2 p_i \tag{1}$$

where p_i represents the i-th class' chance of appearance, the information gain can be calculated shown as in Equation 2,

$$IG(T, a) = H(T) - H(T|a) \tag{2}$$

where $IG(T, a)$ is Information Gain, $H(T)$ represents parent entropy $-\sum_{i=1}^{J} p_i \log_2 p_i$ and $H(T|a)$ states Weighted Sum of Entropy (Children) given the parent T : $\sum_a p(a) \sum_{i=1}^{J}(-\Pr(i|a) \log_2 \Pr(i|a))$.

The method of combination of several decision trees to produce better predictive performance is called an ensemble method. Gradient Boosting is an ensemble technique to a prediction that applies gradient descent to optimize boosting of the tree Equation 3 (Friedman, 2001).

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (3)$$

There are a few effective techniques to perform gradient tree boosting such as XGBoost (Chen & Guestrin, 2016), pGBRT (Tyree, Weinberger, Agrawal, & Paykin, 2011), and LightGBM (Ke, Meng, Wang, Chen, Ma, & Liu, 2017). The XGBoost algorithm grows trees depth-wise and controls model complexity by the maximum depth of each subtree. In contrast, the LightGBM algorithm uses a leaf-wise algorithm and it outperforms XGBoost in terms of speed and memory usage (Ke, Meng, Wang, Chen, Ma, Liu, et al., 2017). In this study, we used LightGBM to predict the number of ASB incidents on every railway station on a weekly basis.

5.1.1. Model evaluation

This step of the analysis is the implementation and calibration of our prediction model. LightGBM has gained a good reputation in terms of performance in accuracy and computation memory usage (Ke, Meng, Wang, Chen, Ma, Liu, et al., 2017) and was implemented in this study. K-fold cross validation is a method that evaluates predictive models by partitioning the original training sample into k equal size of subsamples. The first group is set aside for testing and the remaining k-1 subsamples are used for model building. This process repeated k times, in which each subsample uses once for validation and k-1 time for model building. In our model, a six-fold cross-validation method has been used to validate the accuracy of the classification model.

Table 1. LightGBM Core Parameters

Name of Parameter	Value of Parameter
objective	Regression method
max_depth (one tree)	3
num_leaves (one tree)	25
learning_rate	0.007
n_estimators	30000
min_child_samples	80
subsample	0.8
reg_alpha	0.02
reg_lambda	0.02

After 20 rounds, Table 1 is the best configuration that fit for our ASB forecasting. The prediction error was calculated using the Root Mean Squared Logarithmic Error (RMSLE) as follow:

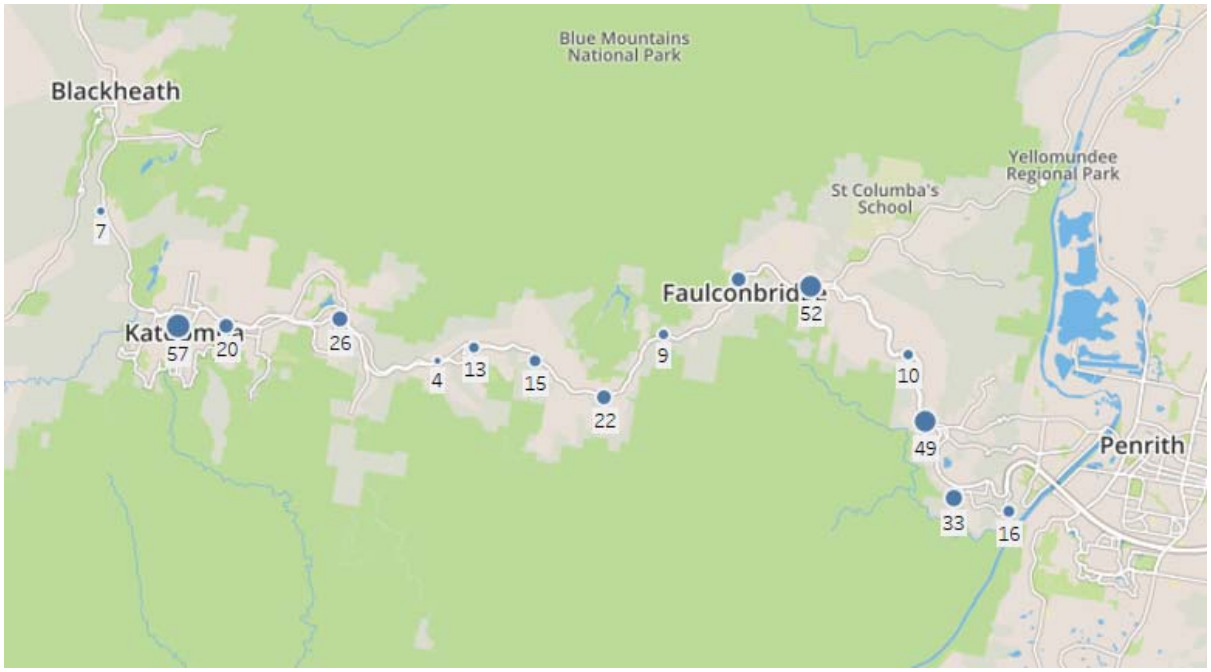
$$Prediction\ error = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i) - \log(y_i))^2} \quad (4)$$

where the x_i is the forecasts and y_i is the observed values. From our model the six-fold values are 0.26820, 0.26167, 0.26978, 0.26467, 0.27509, and 0.26762 respectively. Production of the machine learning model is the next step after a model been trained.

5.1.2. Model publication

The best model was selected from all the trained models. Deploying a machine learning model into a production environment is a very important step for a workflow. We used a batch off-line prediction method to publish our prediction result because our ASB model is run on a weekly basis.

Figure 3. Visualisation of the Anti-Social Behaviour Predictions



Our forecasting results are shown in

Figure . This map presented ASB cases alone with the location of cities, towns, and regional boundaries. This is a fully interactive map which allows for user interactivity and ensures consistency and efficiency in visualisation. It should be noted at this stage that we have, omitted Central station from this process as the high volume of incidents not only drowns the algorithm (i.e. all resources would tend to concentrate at that location) but also because a large number of incidents occur that aren't related to rail operations.

5.2. Predicting Punctuality

On complex railway networks, passengers are frequently confronted with train delays, especially during peak hours of demand. Delays can be caused by a number of factors, including bad weather, freight movements, signal problems, customer or staff injuries, or over-crowding leading to longer dwell times. A train delay or cancellation will lead to deterioration

in the quality of service which impacts not only the customers' satisfaction but also the reputation of the train operator. Using the collected data, machine learning models can be trained to predict train punctuality. Automated data collection systems collect daily data on how passengers use the transit system. Thus, this train network can be abstracted as a time series, which is an important form of structured data for a train delay. The punctuality time series is of fixed frequency, which means the data points occur at regular intervals according to a given operation pattern.

A powerful type of algorithm that is designed to handle sequential data is called Long Short Term Memory (LSTM). LSTM is a prominent variant of the recurrent neural network algorithm used in deep learning to take care of input space with latent connections in sequence (Hochreiter & Schmidhuber, 1997). A typical neuron of LSTM consists of a cell, an input gate, an output gate and a forget gate, that allow for flexible manipulation of input samples in sequence. Given an input vector x_t at the timestamp t and hidden unit h_t with corresponding W, U and b for weights and bias to be learned in training, the forward pass of an LSTM unit equipped with a forget gate can be calculated in the following way.

The forget gate activation vector $f_t \in R^h$:

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (5)$$

The input gate activation vector $f_t \in R^h$:

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (6)$$

The output gate activation vector $f_t \in R^h$:

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (7)$$

The cell state vector $f_t \in R^h$:

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (8)$$

where $\circ$ denotes Hadamard product. The hidden unit is a Hadamard product result of o_t and $\sigma_h(c_t)$. Cell states store historical information about units on the previous timestamp, whereby LSTM is able to have a comprehensive consideration over input spaces with latent sequential connections.

Figure 4. The Map View Forecasted Punctuality at Key Network Points.

From the above (Figure 4), the regression analysis is used in order to estimate the punctuality at for interchange stations where passengers change trains from the regional network to the city network. According to our experimental results, the produced train delay model, which is trained on limited features from our raw data, is good at predicting train punctuality.

6. Conclusion

In this paper, we have proposed a novel framework to enable big data analysis within the operations of a railway. Our framework blends heterogeneous data sources to provide descriptive and predictive analytics. In the descriptive space, we designed and implemented an integrated Key Performance Indicator regime over the top of the KPI framework. In the predictive analytics space we showcased both an anti-social behavior prediction engine as well as a train punctuality prediction capability. These capabilities were built to ultimately deliver an evidence base that would flow through to both strategic and tactical level decision making within railway operations. By using leading edge technology within a very traditional business we have been able to drive better decisions, enhanced engagement with key stakeholders and focus efforts and resources in the pursuit of increased community value through a more successful railway.

References

Chen, T. and Guestrin, C. (2016) 'XGBoost : Reliable Large-scale Tree Boosting System', *arXiv*. doi: 10.1145/2939672.2939785.

Cozens, P. *et al.* (2002) 'Managing crime and the fear of crime at railway stations-A case study in South Wales (UK)', *International Journal of Transport Management*. doi: 10.1016/j.ijtm.2003.10.001.

Friedman, J. H. (2001) 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics*. doi: DOI 10.1214/aos/1013203451.

Ghofrani, F. *et al.* (2018) 'Recent applications of big data analytics in railway transportation systems: A survey', *Transportation Research Part C: Emerging Technologies*, 90, pp. 226–246. doi: https://doi.org/10.1016/j.trc.2018.03.010.

Goverde, R. M. P. and Hansen, I. A. (2000) 'TNV-Prepare: Analysis of Dutch railway opeartions based on train detection data', *Computers in Railways VII*. doi: 10.2495/CR000751.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*. doi: 10.1162/neco.1997.9.8.1735.

Ke, G., Meng, Q., Wang, T., Chen, W., Ma, W. and Liu, T.-Y. (2017) 'A Highly Efficient Gradient Boosting Decision Tree', *Advances in Neural Information Processing Systems 30*.

Ke, G., Meng, Q., Wang, T., Chen, W., Ma, W., Liu, T.-Y., *et al.* (2017) 'LightGBM: A highly efficient gradient boosting decision tree', *Advances in Neural Information Processing Systems*.

Kecman, P. and Goverde, R. M. P. (2012) 'Process mining of train describer event data and automatic conflict identification', in *WIT Transactions on the Built Environment*. doi: 10.2495/CR120201.

Kianmehr, K. and Alhajj, R. (2008) 'Effectiveness of support vector machine for crime hot-spots prediction', *Applied Artificial Intelligence*. doi: 10.1080/08839510802028405.

Kouziokas, G. N. (2017) 'The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment', in *Transportation Research Procedia*. doi: 10.1016/j.trpro.2017.05.083.

Laney, D. (2001) '3-D Data Management: Controlling Data Volume, Velocity, and Variety', *META Group Res Note 6*, 6.

Moore, S. (2011) 'Understanding and managing anti-social behaviour on public transport through value change: The considerate travel campaign', *Transport Policy*. doi: 10.1016/j.tranpol.2010.05.008.d

Neale, R. *et al.* (2004) 'Tackling crime and fear of crime while waiting at Britain's railway stations', *Journal of Public Transportation*. doi: 10.5038/2375-0901.7.3.2.

Nirkhi, S. M. S. M. S. M., Dharaskar, R. V. and Thakre, V. M. (2012) 'Data Mining: A Prospective Approach for Digital Forensics', *International Journal of Data Mining & Knowledge Management Process*.

Oneto, L. *et al.* (2017) 'Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout', *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. doi: 10.1109/TSMC.2017.2693209.

Pongnumkul, S. *et al.* (2014) 'Improving arrival time prediction of Thailand's passenger trains using historical travel times', in *2014 11th Int. Joint Conf. on Computer Science and Software Engineering: 'Human Factors in Computer Science and Software Engineering' - e-Science and High Performance Computing: eHPC, JCSSE 2014*. doi: 10.1109/JCSSE.2014.6841886.

Rokach, L. and Maimon, O. (2005) 'Top-Down Induction of Decision Trees Classifiers—A Survey', *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*. doi: 10.1109/TSMCC.2004.843247.

Tyree, S. *et al.* (2011) 'Parallel boosted regression trees for web search ranking', in *Proceedings of the 20th international conference on World wide web - WWW '11*. doi: 10.1145/1963405.1963461.

Wang, R. and Work, D. B. (2015) 'Data Driven Approaches for Passenger Train Delay Estimation', in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. doi: 10.1109/ITSC.2015.94.

Zheng, X., Cao, Y. and Ma, Z. (2011) 'A mathematical modeling approach for geographical profiling and crime prediction', in *ICSESS 2011 - Proceedings: 2011 IEEE 2nd International Conference on Software Engineering and Service Science*. doi: 10.1109/ICSESS.2011.5982362.