Comparison of Prediction Models for Delays of Freight Trains by Using Data Mining and Machine Learning Methods

Dennis Leser[1*], Matthias Wastian[2**], Matthias Rößler[2], Michael Landsiedl[2], Edmond Hajrizi[3]

[1]Inst. of Analysis and Scientific Computing, TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria; *dennis.leser@tuwien.ac.at

[2]dwh Simulation Services, Neustiftgasse 57-59, 1070 Vienna, Austria; **Matthias.wastian@dwh.at

[3]UBT Univ. of Business and Technology, Larga Kalabrija, 1000 Pristina, Kosovo

SNE 29(1), 2019, 45 - 47, DOI:10.11128/sne.29.sn.10467
Received: Nov. 20, 2018 (Selected KAS-SIM UBT 2018 Conf. Publ.); Revised: Jan. 10, 2019; Accepted; February 3, 2019
SNE - Simulation Notes Europe, ARGESIM Publisher Vienna
ISSN Print 2305-9974, Online 2306-0271, www.sne-journal.org

Abstract. On the one hand, having a tight schedule is desirable and very cost-efficient for freight transport companies. On the other hand, a tight schedule increases the impact of delays and cancellations. Furthermore, the prediction of delays is extremely complex, because they depend on many factors of influence. To address these issues, this work will show an approach to forecast delays of freight trains by using data mining and machine learning methods. For this purpose, an international freight transport company in rail traffic provided us with a huge amount of historical data of freight train runs. In order to get a suitable prediction model, we apply a knowledge discovery in databases (KDD) process, which contains the steps data selection, data preprocessing, data transformation, data mining and interpretation/evaluation. After the data selection and data preprocessing step we transform categorical features via one-hot encoding as well as via embedding with various embedding sizes. Furthermore, we apply a data transformation method for cyclical features like weekday. In the actual data mining process, we use the preprocessed historical data to perform a regression analysis, which forecasts the delays of freight trains, and compare several regression models like decision tree, random forest, extra trees and gradient boosting regression. An adequate prediction model will be integrated into an agent-based model, which tests the robustness of optimized locomotive schedules for freight trains.

Introduction

The planning of train schedules is an extremely complex task, because of the many possibilities to schedule routes and locomotives.

However, it is the daily work of freight transport companies in rail traffic. In order to reduce costs, the locomotive schedule should be as tight as possible. But a tight schedule increases the impact of delays and cancellations, especially if there are no available backup resources like traction units. Therefore, a well-balanced ratio between a tight and robust schedule is desirable. To address these issues, this work will show an approach to forecast delays of freight trains by using data mining and machine learning methods. For this purpose, an international freight transport company in rail traffic provided us with a huge amount of historical data of freight and passenger train runs. Furthermore, we apply a knowledge discovery in databases (KDD) process and compare several regression models as well as data transformation methods, in order to receive a suitable prediction model for delays of freight trains. Finally, an adequate prediction model will be integrated into an agent-based model, which tests the robustness of optimized locomotive schedules for freight trains [1].

1 The KDD Process

A knowledge discovery in databases (KDD) process is a nontrivial procedure to identify valid, novel and potentially useful patterns in data [2]. This process contains the five steps data selection, data preprocessing, data transformation, data mining and interpretion/evaluation [3]. To create a prediction model for delays of freight trains, we use the huge amount of historical data of freight train runs and follow the sequence of the KDD process.

1.1 Data selection

The data selection step includes understanding of the application domain and the relevant prior knowledge, selecting appropriate data as well as the identification of the application goal [2].

For this purpose, we cooperate closely with the international freight transport company in rail traffic. The identified application goal is the prediction of the target value "delay_ank", which is the delay at the arrival station. Furthermore, we used SQL queries to extract appropriate features from the database of the freight transport company.

Feature name	Description
abfbsstelle_id	Station ID of the departure station
ankbsstelle_id	Station ID of the arrival station
plananfahrt	Planned departure time
planankunft	Planned arrival time
lon_abf	Longitude of the departure station
lat_abf	Latitude of the departure station
lon_ank	Longitude of the arrival station
lat_ank	Latitude of the arrival station
border_abf	Indicates if dep. stat. lies on the border
border_ank	Indicates if arrival stat. lies on the border
region_abf	Region of the departure station
region_ank	Region of the arrival station
meter	Distance between two stations
tz	Indicates the position of the traction unit
reihe	Series of the traction unit
orndr	ID inside the series of the traction unit
produktname	Indicates the type of the train operation
delay_abf	Delay at the departure time
altitude_diff	Difference betw. altitudes of two stations

Table 1: Chosen features for the further KDD process and their descriptions.

For the feature "altitude_diff", we used the open elevation public API to receive the altitudes of the stations and we calculated their altitude differential.
The chosen features for the further KDD process are shown in Table 1. To investigate the correlations between the chosen numerical features and the target value "delay_ank", we use a correlation matrix (Figure 1).

As expected, there exists a strong correlation between delay at the departure station and delay at the arrival station. In order to get more information about the importance of the other chosen numerical features, we apply the random forest method without the feature "delay_abf" (Figure 2).

1.2 Data preprocessing

Data preprocessing, also called data cleaning includes strategies for handling missing data fields and if appropriate removing noise to obtain consistent data [4].
The huge amount of historical data of freight train runs are real data including missing and wrong entries.

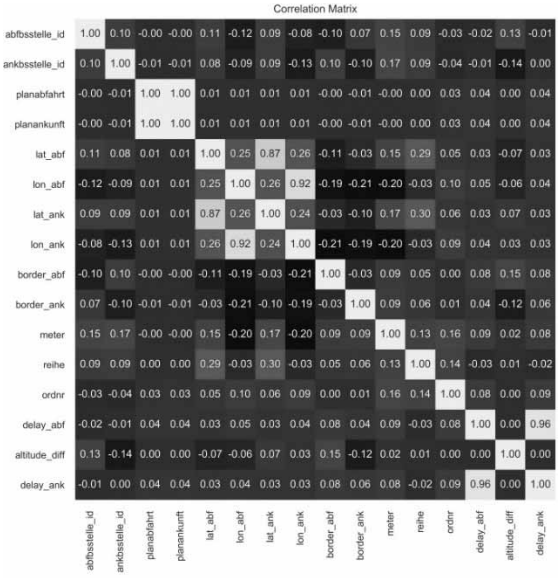

Figure 1: Correlation matrix, which shows the correlations between the chosen numerical features and the target value "delay_ank".

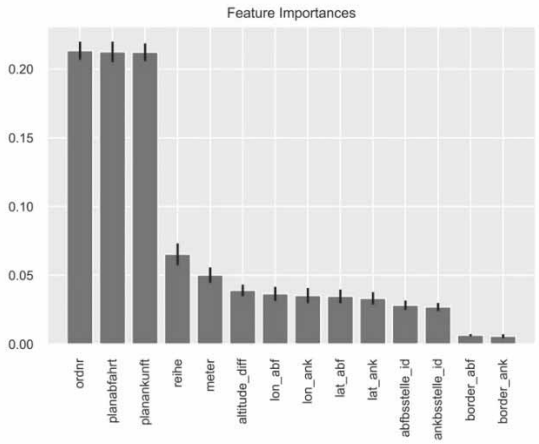

Figure 2: Feature importances of the chosen numerical features without the feature "delay_abf" in descending order of importance.

Some wrong entries were easy to adjust, like swapped geographic coordinates. But for example, samples including planned arrival time before planned departure time or no distance between two different stations couldn't be adjusted that easily. For the further steps of the KDD process we don't consider samples with missing or wrong entries, which we couldn't replace or adjust.

1.3 Data transformation

In this step, we transform the nominal features "abfbsstelle_id", "ankbsstelle_id", "region_abf", "region_ank", "tz", "reihe", "ordnr" and "produktname", to use them for the data mining process and to increase the accuracy of the prediction model. For this purpose, we

apply the data transformation method one-hot encoding, which creates for each possible value of a nominal feature a new dummy feature and returns a quite sparse array. But each of the nominal features "abfbsstelle_id", "ankbsstelle_id", "reihe" and "ordnr" contains several hundred possible values.

In order to reduce the number of dimensions to represent these nominal features, we use an embedding with various embedding sizes, which is a further data transformation method. The embedding learns to map each possible value of a nominal feature into a vector with the length of a given embedding size [5, 6].

Furthermore, we used the features "planabfahrt" and "planankunft" to extract the cyclical features day of the year "jahrestag", weekday "wochentag" and minute of the day "tagmin" for the planned departure and arrival time.

In order to further increase the accuracy of the prediction model, we transform these cyclical features by using a sine and cosine transformation [7].

1.4 Data mining

The actual data mining process includes selecting an adequate model, for example classification, clustering or regression, and choosing the data-mining algorithm(s) to find patterns of interest and finally to achieve the application goal [8]. Because our target value "delay_ank" is a continuous numerical value, we chose a regression analysis and split the prepared data into training and test data. With the training data, we train a dummy regressor model, decision trees, random forests, extra trees and gradient boosting regression models.

For the comparison of prediction models, we apply the test data and evaluate for each regression model as well as different data transformation methods the mean squared error between the prediction and the target value "delay_ank". We use the dummy regressor which always predicts the mean of the training targets as a baseline to compare the mean squared errors.

1.5 Interpretation and evaluation

This step consists of interpreting the found patterns, evaluating the prediction models and acting on the discovered knowledge. Proper interpretation of data mining results requires a high degree of domain knowledge. For this reason, we cooperate closely with experts of the freight transport company to interpret the results. For the theoretical evaluation of the prediction models, we used 2-fold cross-validation.

2 Results

Table 2 shows the comparison of the mean squared errors of the different regression models and the data transformation methods one-hot encoding as well as embedding by using three different embedding sizes. The prediction model with the highest accuracy is the gradient boosting regression model with the embedding and embedding sizes 25 or 50.

Regression model	One-hot	Emb. 10	Emb. 25	Emb. 50
DummyRegressor	8329	8329	8329	8329
Decision tree	1119	1462	1363	1504
Random forest	696	683	690	696
Extra trees	749	684	692	695
Gradient boosting	643	632	629	629

Table 2: Comparison of different regression models and data transformation methods.

3 Conclusion and Outlook

This work has shown a comparison of prediction models for delay of freight trains by using data mining and machine learning methods. For further work, we will use the gradient boosting regression model with the embedding and embedding size 25. We will integrate an adequate prediction model into an agent-based model [1] to test the robustness of optimized locomotive schedules for freight trains.

References

[1] Rößler, M., Wastian, M., Landsiedl, M., & Popper, N. (2018). An Agent-based model for robustness testing of freight train schedules. In Proc. MAS, Hungary, Budapest.

[2] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

[3] Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

[4] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[5] Teschner, F. (2018). Exploring Embeddings for Categorical Variables with Keras. Online at http://flovv.github.io/Embeddings_with_keras/ (Retrieved 1 October 2018 at 13:12 CET).

[6] Satnalika, M. (2018). On learning embeddings for categorical data using Keras. Online at https://medium.com/@satnalikamayank12/on-learning-embeddings-for-categorical-data-using-keras-165ff2773fc9 (Retrieved 2 October 2018 at 15:31 CET).

[7] London, I. (2017). Encoding cyclical continuous features – 24-hour time. Online at https://ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/ (Retrieved 2 October 2018 at 16:03 CET).

[8] Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). *Data mining: a knowledge discovery approach*. Springer Science & Business Media.