Forecasting bus ridership with trip planner usage data: a machine learning application

Jop van Roosmalen, Chintan Amrit, Engin Topan and Niels van Oort.

Context

Currently, public transport gives much attention to environmental impact, costs and traveler satisfaction. Good short-term demand forecasting models can help improve these performance indicators. It can help prevent denied boarding and overcrowding in buses by detecting insufficient capacity beforehand. It could be used to operate more economically by decreasing the frequency or the size of the bus if there is overcapacity. Moreover, it could help operators plan their buses during incidental occasions like big public events where little information is known. Finally, it could be used to reliably inform the travelers on the current crowdedness (Ohler et al., 2017; Van Oort et al., 2015a; Pereira et al., 2015).

This study investigates the usefulness of a new data source; the usage data of a trip planner. In the Netherlands there are multiple trip planners available for users to help find the most optimal (multimodal) journeys. These trip planners require a date, a time and an origin and destination, which they use to construct multiple alternative journeys from which the user can choose. For this study the data of 9292 was used. 9292 is one of the major trip planners in the Netherlands and includes all public transport modes for the whole country.

For the study we used data of 20 lines (urban and regional) for the first three months of 2017. A regression analysis is used to determine the forecasting potential of the trip planner usage data. This data is regressed towards smart card transaction data.

Practical challenges

A few challenges had to be overcome in order to perform the study. Firstly, the data had to be modified by the organizations to oblige by the privacy laws (Van Oort, 2015b).

Secondly, the data that is logged by 9292 is not optimized to be used for forecasting demand: It is unknown if two requests are made by the same person (viewing an alternative journey plan is logged as a separate request). There is no identifier for the bus trip stored (only line number). And it is difficult to match the trip planner trips with bus trips, since, over time, the 9292 private bus stops database evolved differently and there is no information stored on the actual delay which are used in the journey plans.

Thirdly, everyone has his own strategy (for different scenarios) in planning a journey and will use the trip planner to fulfill his needs. The user interface design and functionality of the trip planner influence this behavior and therefore directly impact the usage data. Furthermore, it is unknown if a travel plan is made for one person or for a group of people.

Finally, 9292 is not the only major trip planner used in the Netherlands. This dataset is therefore an incomplete source. Including usage data from other major trip planners could improve the usefulness of the data source.

Research gaps

This study is one of the first in assessing trip planner usage data for predicting short-term ridership of buses. There is limited academic research published where trip planner usage data is utilized for forecasting public transport demand.

The following research question is defined: "Can one forecast ridership of buses using data containing the consulted travel advices from a widely used trip planner for public transport and what accuracy can one achieve?"

Methodology

We developed a model for forecasting the number of people boarding and a model for forecasting the number of people alighting at a certain stop. These forecasts are defined at the vehicle-stop level. By counting the number of people boarding and subtracting the number of people alighting along the trip, the forecasted number of passengers after a stop can be calculated (Ohler et al., 2017).

We compare five different machine learning models: multiple linear regression, decision tree, random forests, neural networks and support vector regression with a radial basis kernel (Zhang et al., 2017; James et al., 2013). We compare these models with two simple rules: 1 predict the same number as last week, and 2 predict the historic average as number. The models are implemented in the Scikit-Learn library of Python (Pedregosa et al., 2011). The data is stored in a PostgresSQL database.

The trip planner datasets and smart card dataset are merged and preprocessed. The resulted dataset is rather sparse; a lot of stops have zero passengers boarding or alighting or requests suggesting doing so. Therefore, we investigated if subsampling is needed. From the datasets useful data is selected and

features are constructed. The features are standardized. Different number of features are tested, these features are selected based on recursive elimination using a simple random forests model. Finally, the hyperparameters of the models are tuned and the optimal configurations are stored. The scores are validated by using cross validation.

Results

We used the trips of one route during the morning peak to test our models. We used different kind of data partitions to train these models. All models are constructed with a planning horizon of 15 minutes. In most cases the best performing model used 20 features, the max number that was allowed.

The random forests model predicted the number of people boarding most accurate with a Root Mean Squared Error (RMSE) of 2.55 (R2 of 0.76). The random forests model forecasted the number of people alighting most accurate as well, with an RMSE of 2.20 (R2 of 0.76). The lower RMSE indicates that the number of people alighting is more predictable. In both cases the best version of the other models outperformed the forecasts of rule 1 and 2. It was discovered that subsampling had a slight negative effect.

When combining the boarding and alighting model, random forests outperforms the other machine learning models with an RMSE of 8.72. However, rule 2 has an RMSE of 8.603. When looking at the percentage of trips correctly forecasted within an absolute error of 5 passengers, rule 2 outperforms the random forests model with 84.08% against 58.9%. Thus, rule 2 outperforms the machine learning models when it comes to forecasting the number of passengers. Combining the best performing boarding and alighting model does not lead to the best forecast for the number of passengers. When looking at the percentage of correct max loads predictions of trips – the most important indicator for adjusting the size of the bus –, the forecasts of rule 2 and the random forests model severely underestimated (more than 10 passengers lower as the real value) the max load for more than 27% of the trips.

The two most important features are the historical average of the number of people boarding (or alighting) and the number of requests for the same line aggregated over a window of 3 hours. The first feature was included to give an adequate baseline. The disaggregated version of the second feature is probably too noisy and fluctuates too much. Aggregating this feature over time helps to reveal the underlying trend more reliably.

Implications for practice and science and recommendations for further research

The trip planner usage data is an interesting source to detect the number of additional people boarding or alighting. Especially, since this process could be fully automated. However, the different organizations should adjust their data structures in order to construct more useful features, do more valuable analysis and to streamline the whole data preprocessing process of merging the different datasets.

Researchers could help this process by further developing these forecasting models, testing more features and models, testing the models in different scenario's and by researching models that forecast the max number of passengers using a different method (since combining the boarding and alighting model leads to interference errors).

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.

Ohler F., Krempels K. and Möbus S. (2017). Forecasting Public Transportation Capacity Utilisation Considering External Factors. In Proceedings of the 3rd International Conference on Vehicle Technology and Intelligent Transport Systems ISBN 978-989-758-242-4, pages 300-311.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Pereira, F. C., Rodrigues, F., & Ben-Akiva, M. (2015). Using data from the web to predict public transport arrivals under special events scenarios. Journal of Intelligent Transportation Systems, 19(3), 273-288.

Van Oort, N., Drost, M., Brands, T., & Yap, M. (2015a, July). Data-driven public transport ridership prediction approach including comfort aspects. In Conference on Advanced Systems in Public Transport, Rotterdam, The Netherlands.

Van Oort, N., Brands, T., & de Romph, E. (2015b). Short-term prediction of ridership on public transport with smart card data. Transportation Research Record: Journal of the Transportation Research Board, (2535), 105-111.

Zhang, J., Shen, D., Tu, L., Zhang, F., Xu, C., Wang, Y., ... & Li, Z. (2017). A real-time passenger flow estimation and prediction method for urban bus transit systems. IEEE Transactions on Intelligent Transportation Systems, 18(11), 3168-3178.