train from a station on a certain day. It continuously polls the iRail API to check for new occupancy records. When such a record is found, it is automatically processed and added to our NoSQL MongoDB. Every night, two processes are run. On the one hand our predictive model is re-trained with the newly collected data. On the other hand, the hyper-parameters are tuned using Bayesian optimization. The API is accessible through the following IP: 193.190.127.247

# 6. CONCLUSION AND FUTURE WORK

In this paper, the first steps towards a system that can predict the occupancy level of a train in the nearby future based on query logs are presented. Such a system can have a significant positive impact on the quality of service while decreasing the operational costs. We discussed the different phases of constructing such a system: (i) adding a functionality to a widely used application in Belgium in order to collect data through crowd-sourcing; (ii) extracting numerical features from these raw JSON logs and (iii) creating a predictive model on this extracted data. Moreover, an API was created in order to expose the predictions of our model and a Kaggle competition was set up to enable collaborative benchmarking.

We conclude that, in this early phase, our predictive model, which is trained on a limited amount of data, is good at predicting trains with a low occupancy. This comes at no surprise, as the low occupancy of trains outside peak hours is easy to predict and as it is the largest populated class (currently, around 41% of all samples have the low occupancy label). When more samples are collected, we are convinced that the system's predictive performance will increase. The strength of the approach in this paper is that the data used can be gathered for any public transport system. At this moment, data has only been collected over a limited timespan. The current dataset thus contains only a limited amount of samples, but is growing steadily with more than 1000 query logs per month.

# 7. ACKNOWLEDGMENTS

Gilles Vandewiele is funded by a PhD SB fellow scholarship of FWO (1S31417N). Thank you to iRail, TreinTramBus and Metro Time, and crowd-funding supporters for their time and financial effort in the Spitsgids campaign. Thank you to SNCB for the support to gather first data. Thank you Serkan Yildiz, Stan Callewaert and Arne Nys for their enthusiasm implementing the features in the apps during open Summer of code. Thank you Kris Peeters, Nathan Bijnens and other Twitter users who helped discussing the data publicly.

# 8. REFERENCES

[1] M. Cantwell, B. Caulfield, and M. OâĂŹMahony. Examining the factors that impact public transport commuting satisfaction. *Journal of Public Transportation*, 12(2):1, 2009.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[4] P. Colpaert, A. Chua, R. Verborgh, E. Mannens, R. Van de Walle, and A. Vande Moere. What public transit api logs tell us about travel flows. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 873–878. International World Wide Web Conferences Steering Committee, 2016.

[5] Y. Kim and J. Kim. Gradient lasso for feature selection. In *Proceedings of the twenty-first international conference on Machine learning*, page 60. ACM, 2004.

[6] M. B. Kursa, A. Jankowski, and W. R. Rudnicki. Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285, 2010.

[7] U. Lundberg. Urban commuting: Crowdedness and catecholamine excretion. *Journal of Human Stress*, 2(3):26–32, 1976.

[8] R. Martinez-Cantin. Bayesopt: a bayesian optimization library for nonlinear optimization, experimental design and bandits. *Journal of Machine Learning Research*, 15(1):3735–3739, 2014.

[9] M. Milkovits. Modeling the factors affecting bus stop dwell time: use of automatic passenger counting, automatic fare counting, and automatic vehicle location data. *Transportation Research Record: Journal of the Transportation Research Board*, (2072):125–130, 2008.

[10] A. Nuzzolo, U. Crisalli, L. Rosati, and A. Ibeas. Stop: a short term transit occupancy prediction tool for aptis and real time transit management systems. In *Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on*, pages 1894–1899. IEEE, 2013.

[11] A. Puong. Dwell time model and analysis for the mbta red line. *Massachusetts Institute of Technology Research Memo*, 2000.

[12] R. Silva, S. M. Kang, and E. M. Airoldi. Predicting traffic volumes and estimating the effects of shocks in massive transportation systems. *Proceedings of the National Academy of Sciences*, 112(18):5643–5648, 2015.

[13] A. Tirachini, D. A. Hensher, and J. M. Rose. Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. *Transportation research part A: policy and practice*, 53:36–52, 2013.

[14] N. Zhang, H. Chen, X. Chen, and J. Chen. Forecasting public transit use by crowdsensing and semantic trajectory mining: Case studies. *ISPRS International Journal of Geo-Information*, 5(10):180, 2016.