Predicting city bike usage in districts of Helsinki outside of the Helsinki city bike network using demographic data

Ville Marttila, Giulia Varvarà, Keijo Korhonen

Introduction

We have trained a machine learning model based on a random forest to predict city bike usage in different districts of Helsinki based on their demographic properties and previous bike usage patterns. The idea of the model is to be able to predict patterns of city bike usage in districts that are not yet in the Helsinki city bike network. This would make it possible to anticipate changes in usage patterns and increases in bike traffic in different districts when a new district is added or to evaluate whether adding a district is worth doing in the first place.

While the audience for such a product is not very broad, we believe that especially HSL and Helsinki city planning authorities could benefit from such predictions when considering expanding the network further and making other decisions relating to Helsinki's cycling infrastructure.

Our initial plan for the project did not include the predictions for city bike usage, but rather a simple statistical analysis of current city bike usage trends. The main goal for us was to find a link between demographics and city bike usage, such as a larger number of restaurants or low income indicating high usage. However, by analysing the data we realised that such an analysis could easily be expanded to a machine learning model and be used to describe city bike usage by district.

Data collection and preprocessing

We used the openly available city bike data which can be found on the website of HSL (https://www.hsl.fi/avoindata). This data set contains information on all city bike rides taken in 2019, with their departure and arrival times and locations as well as their distance and time taken. Unfortunately data from earlier years – although available – could not be used since the locations of the bike stations had changed and their locations for previous years were no longer available.

We also used open GIS data available at the Helsinki region infoshare (https://hri.fi/en_gb/) for the city districts to link the bike rides to specific starting and ending districts. We initially tried to use the more detailed division into 148 sub-districts, which would have given us more data points for training the machine learning model (which was based on pairs of districts and predicting the amount of bike rides between them), but the demographic data (discussed below) was not available for 2019 at this level, so we had to work on the level of the 34 base districts, of

which 23 belonged to the current city bike system and could thus be used for training data, the remaining 11 being the target of our predictions.

For demographic data we used the information available on the Helsinki region infoshare website (https://hri.fi/data/en_GB/dataset/helsinki-alueittain) , which contains information about the inhabitants and the jobs and commercial activities in the 34 major districts of Helsinki, including data about number of people living there, age distribution, number of bars, schools etc. We collected a set of 63 different potentially significant parameters for each district (the parameters are described on the project web page), and since out prediction model was based on pairs of districts, this gave us 126 demographic parameters for each pair, along with 28 parameters related to earlier city bike use and some miscellaneous geographic and other parameters, such as the distance between the districts (which turned out to be a major predictor – somewhat unsurprisingly).

Based on the GIS location data of the city bike stations, we grouped them into the different districts of Helsinki (initially into the 148 sub-districts and later into the 34 base districts) by checking their inclusion against the city district GIS data, and saved them into a PostGIS database so that we could get a starting and finishing district for each ride. Since the city bike system is shared with Espoo and many of the rides crossed the city border, but we only had the demographic data for Helsinki (and had to limit the scope somehow) we tried several approaches to dealing with the rides crossing the city border and kept them in the data (and even ran some test predictions using some of the proposed approaches), but in the end we could not find a satisfactory and statistically valid way of incorporating them in the analyses, so they were left out from the predictions entirely.

We had also intended to include the information about the intermediary districts – i.e. the districts through which the rides traveled without starting or ending in them – in the analyses and used the open Digitransit route planner API offered by HSL (https://digitransit.fi/en/developers/apis/1-routing-api/) to calculate the optimal route between each pair of bike stations and save them to a PostGIS database as Polylines to serve as district-level approximations of the route taken by each ride. Unfortunately, despite the considerable effort spent on acquiring this data using a custom Python data scraper[1], time constraints and the downgrade to the base district level meant that it was not used in the end.

For storing the different types of data, we used both csv-files (imported into (Geo)Pandas for analysis) and a local PostGIS database (which was used for some queries and data joins which would have taken too long using Pandas), the contents of which were shared as SQL-dumps through Google Drive since it was much too large to be shared over GitHub.

---

[1] https://github.com/keijokorhonen/helsinki-city-bike-data-analysis/blob/master/scripts/calculate_optimal_routes_otp.py

Data analysis

For the data analysis stage, we initially started with doing calculations on the amounts of incoming and outgoing rides for the different existing districts and calculating statistics on the distribution of rides by the time of day and day of week and visualising them using plots generated using the Bokeh visualization library (https://docs.bokeh.org/en/latest/index.html). On hindsight we could also have included a division by month, since the usage patterns for the summer months would likely have been different from those of the spring and autumn months, but since we did not include it from the beginning, time constraints meant that it was not really feasible to redo all the data processing and analysis with this additional dimension.

After initial exploratory analysis of the data, at which point we were still mainly intending to use it to describe the relationship of demographic features to city bike usage patterns in the existing city bike districts, we realised that the data would also lend itself to a more machine learning oriented approach for predicting the usage patterns of new districts based on their demographic features and moved to planning the machine learning stage.

Other interesting questions that this data could have been used to answer – and were at one point or another considered for study – perhaps with a more statistical and less machine learning oriented approach, include:

- Which districts have the most outgoing/incoming rides and what demographic features do they share? Are there individual demographic features that have a strong correlation with the frequency of city bike usage?
- What relationships do the districts have with one another? Are there clearly distinctive profiles of city bike usage that differentiate between or connect different districts?
- What are the demographics of the districts we are predicting for? What already city bike -enabled districts do they resemble in terms of their demographic profiles? (Studying this would actually have been an extremely good preliminary step for our prediction work.)

Machine Learning

Although it was only discovered at a relatively late stage in the project – which unfortunately meant that we did not have as much time to develop and especially make use of it as we would have wanted and also wasted time on some data collection and wrangling that did not get used in the end – the machine learning tool ended up being central to the project. After having considered the parameters of our situation – need for a regression algorithm for predicting continuous values (instead of classification), a relatively large number of features and a relatively small data set, and a predictive rather than descriptive modeling task – we ended up choosing Random Forest as the most suitable candidate for the machine learning predictor. Furthermore, its benefits included reduced risk for overfitting and excessive variance due to its bootstrap

aggregation approach, and its ability to provide information about the significance of different features (which in itself is information that we were interested in).

The method of prediction we used was to use pairs of districts as data points, one of them being defined as the "base" and the other as the "comparison". The features thus consisted of the demographic features of both districts as well as some aggregate bike usage statistics of the comparison district (the unknown districts to be predicted were always defined as the base). This gave us 529 data points for our training set, 20 % of which were used for a test set. For each of these data points, we had 56 separate labels – the numbers of inbound and outbound rides for 28 combinations of days of the week and times of day – which not only provided more fine-grained predictions, but also allowed for the possibility that the amounts of rides at different times were conditioned by different features.

Our initial idea was to improve the accuracy of the Random Forest algorithm – for which we used the scikit-learn RandomForestRegressor – and to get information about the importance of the various features at the same time by using importance pruning – a technique where the explanatory power of each feature is evaluated after each iteration of the regression algorithm and the features with the lowest importance are pruned out and the process repeated until only the most significant predictors are left. For calculating feature importance we tried several different methods, including the native feature importance of the Random Forest algorithm, Permutation feature importance, and Drop Column feature importance and ended up choosing the last of these for the final trials, due to its theoretical accuracy and despite its high time cost.[2] However, upon extensive testing, we found that it did not actually seem to improve the accuracy of the model with our data while multiplying the training time hundredfold. Thus we ended up using the Random Forest algorithm without feature optimisation, as this seemed to give better results in a fraction of the time, which unfortunately meant that the considerable time spent on building and testing the importance pruning system was more or less wasted.[3]

After experimentation, we ended up using a simple Random Forest algorithm with 100 predictors, a maximum tree depth of 5 and a fixed random factor, since this seemed to give consistently better results than either more or fewer predictors, the average test scores (calculated over the 56 labels we were predicting) varying between 0.6 and 0.8 with an average training score of about 0.90-0.95. The sometimes quite noticeable variation and fluctuation in the accuracy score and results is most likely a result of the relatively small data set; using a larger data set – for example using the dub-districts would have given us over 20 000 data points instead of the 500 or so – would most likely have helped here. However, despite the limitations,

---

[2] The features of the different feature importance methods are outlined for example in this article: https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e.

[3] The feature importance pruning and simple versions of the machine learning model are contained in https://github.com/keijokorhonen/helsinki-city-bike-data-analysis/blob/master/scripts/random_forest_prediction_final.py (with feature pruning) and https://github.com/keijokorhonen/helsinki-city-bike-data-analysis/blob/master/scripts/random_forest_prediction_noimp.py (no feature pruning).

the accuracy scores of 0.80 that we quite frequently achieved as well as the predicted usage patterns themselves seem to be at least somewhat reasonable, indicating that this system could in fact be a viable means of prediction.

Communication of results

As the audience for our project is mainly HSL and our results are basically a type of raw data, we wanted to present our data as some kind of data analysis instead of an application. Without going into much detail, we aimed to describe the most important district features we used as well as the general methodology to show why our results should be somewhat accurate.

For visualising, we would have wanted to do a full interactive map of all districts of Helsinki, with various kinds of data indicated by colouring and tooltip info-boxes, but due to time constraints we had to settle for illustrating some example cases using these kinds of interactive plots and tables givind aggregates of the result data. The reason why we wanted the map to be interactive, is that including the district names or IDs in a readable way proved to be quite difficult with the amount of districts. Moreover, interactive maps are more readily understandable even for the target audience and can deliver the findings in a more immediate way.

Conclusion

To improve this model we could find a way to collect more detailed data, so to be able to divide also the demographic data in subdistricts, as we already did with the bike data. Dividing all the data in sub-districts would allow the machine learning algorithm to have more data to train on and also the predictions could be made for smaller areas, making them much more accurate. However, as a proof-of-concept the data setup and the prediction system seems both solid enough and practically useful, and thus well worth further development.

In addition to being used for Helsinki, this model could just as well be used in other cities that are looking into expanding their bike sharing network, since the model could be retrained for with the demographic and bike sharing data available in any city. Moreover a city similar in composition to the city of Helsinki could use this model to also create a bike sharing network from scratch. If a city is assumed to be similar to the city of Helsinki, it could be assumed that the usage of city bikes should behave in a similar way with regard to the parameters that we used in our model. This means that our model could be used also for this kind of purpose by plugging in the demographic data of the city, to estimate the probable bike traffic and choose where bike stations will be most needed.

Perhaps the biggest learning experience of this project was that data science projects – especially exploratory and experimental ones – can take a lot of time and require some space for exploration and "wasted" effort, which is why a tight schedule is even a bigger problem than in other types of projects.