Prediction Model to measure crowdedness within Amsterdam.
Thesis Design Proposal

Author:
Don de Lange

University of Amsterdam, Science Park 904, The Netherlands,
don.delange@student.uva.nl

Abstract. The aim of this research is to construct a model that predicts the crowdedness at selected spots within the city of Amsterdam, in the form of a time series. Historical data will be gathered from localized crowdedness measure points and will be combined with GVB data. Both a multivariable regression model and a decision tree will be used as prediction model. The models will be evaluated with the actual crowdedness counts at the selected spots in the city of Amsterdam and at the public transport spots, at the specified times in the predicted time series. The time series will be predict the crowdedness at an hourly rate each day.

Keywords: Crowdedness, Open Data, Regression, Decision Tree, Gradient Boosting, Time Series

1 Introduction

The city of Amsterdam is working towards becoming a smart city, with the use of gathered open source data from different sources within the city. Crowdedness measure is one of these instances and is aimed at predicting the degree of crowdedness within the city in real-time and making these predictions open source.

Amsterdam is a tourist attraction for tourists all over the world, increasing the crowdedness of the city. As it is unclear what the level of crowdedness will be in the near future, Amsterdam can only react to the crowdedness in the city. For example, making sure there is enough space for all the travellers during rush hour or making sure there is enough security during events.

The aim of this research is to built a time series prediction model for the crowdedness in public spaces in the city of Amsterdam, which is now unevenly measured within the city with no clear method of predicting the level of crowdedness within the city. For this project, data will be gathered from precise localized data sources and city-wide data sources and built a prediction model that can reliably predict the crowdedness within the city of Amsterdam, combined with the number of passengers at each station at an hourly rate. The main focus will lie on correctly predicting the crowdedness at the same spots as the localized

data sources and public transport stations, at an hourly rate in the form of a time series.

The main challenge of this project lies in the following points; first, the data preperation of the datasets that will be used in this research. As the data is not saved in a structure that is fully understandable for people that don't work daily with the data. In addition, there is not guarantee that the data has been saved in a consistent manner. Furthermore, the data also has to be transformed to a format that will be usable for the prediction models. And second, determining what factors have influence on the crowdedness number in the city. As this research does not use all the factors that could influence this number, it's possible that incorrect conclusions could be drawn, based on assumptions on the external environment.

The structure of this thesis design is as follows; first, the research question and its sub questions will be stated. Second, the scientific background of this research will be discussed. Third, the methadology for this research will be clearly stated. This will be done by describing the data that will be used, stating and justifying the methods that will be used to form the time series of predictions, and by describing the evaluation method used for the times series. Fourth, a risk assessment of possible problems and possible solutions for these will be given. And fifth, an initial project plan will be stated.

1.1 Research Questions

The main research question is stated as follows:

How can an accurate and fast prediction time series of the level of crowdedness within the city of Amsterdam be given, based on input from city-wide available data sources?

The main research question will be answered with the following two sub questions.

- *How can the data from the city-wide available data sources be prepared in a time efficient manner?*
- *What model could be used to provide an time series of fast and accurate predictions of crowdedness at local crowdedness measure points within the city, based on historical data, at an hourly rate?*

2 Related Work

Within the specific domain of crowdedness within the city of Amsterdam, there is no research available. So this research will serve as a first step in this area.

2.1 Crowdedness

Niu et al (2017) investigated how to estimate the crowdedness level for buses. The crowdedness is measured in real-time using sensors. This research will expand upon this by predicting the crowdedness. Ding et al (2016) seek to predict short-term ridership with the use of Gradient Boosting Decision Trees. With the algorithm the research aimed to capture the association of ridership with the other independent variables. The model is able to handle different types of predictor variables and distangle the relationship between them. This research seeks to apply a similar method in the context of the city of Amsterdam with more data.

Barth et al. (2016) presented a low cost framework to mine data obtained from passengers check in and check out data, bus stops geolocations, and buses GPS. The analysis gives greater insight into the volume and flow of passengers and the real existing demand for bus services. By mapping all the passenger data, efforts could be made to improve the flow of passengers.

2.2 Machine Learning Approaches

As there are no clear labels available for the levels of crowdedness in the city, this research uses regression and decision trees as models for prediction, rather than classification models. The prediction models will produce approximate counts for crowdedness at an hourly rate.

The Gradient boosting decision tree is a widely-used machine learning algorithm, due to its efficiency, accuracy, and interpretability (Guolin et al., 2017). The paper proposes two techniques to effectively deal with the efficiency and scalability problem when dealing with large datasets, which involve excluding parts of the dataset and reducing the number of features. The resulting algorithm reaches similar a similar accuracy score, whilst being faster. Xia et al. (2016) propose a sequential ensemble scoring model with the use of extreme gradient boosting (XGBoost). In the mdoel, the data is preprocessed and scaled, feature selection is applied on the variables, and hyperparamaters are tuned with XGBoost. The resulting optimization method outperformed random search, grid search and manual search. This research seeks to combine these different approaches into one Gradient boosted decision tree.

Glanz et al (2018) developed a predictive model that predicts the probability that a patient will have an opioid overdose, based on five different variables. The paper offers a method for the construction and validation of a predictive model. This research aims to do use a similar method in constructing a predictive model, within the context of crowdedness within the city of Amsterdam.

Ing and Ing (2018) constructed a multivariable predictive model with ten different variables for the nanogram of GCA. This research aims the apply a similar predictive model in the context of crowdedness in the city.

2.3 Prediction Models

Ribeiro et al. (2016) offer two approaches that help in understanding why given models make a certain prediction.

3 Methodology

3.1 Data

The dataset consists of the following parts. Ideally the resulting model will output a time series of crowdedness per hour each day.

First, a daily count of the number of people at given data gathering points. The data is gathered in two different ways. The first is a Count Camera, which uses image recognition over a certain area on the street at given points in the city. The camera identifies objects, and counts them if they move a predetermined minimal distance over the street area. For this research only objects classified as *people* will be used. And the second a Wifi Sensor, which counts and identifies the devices it connects to within a predetermined radius. For this research only objects that identify as mobile devices will be used.

Second is public transport data of the GVB subway within Amsterdam of the number of passengers. This data is anonymized and aggregated, meaning that only the number of passengers at a certain station is visible. Given that the number was higher than five.

And third, event data of Amsterdam will be used as a separate category in the dataset. As events are exceptions to the normal days in Amsterdam in terms of crowdedness. For example, King's day and Christmas day.

The input data for the models is as follows; first, the localized data sources give the precise counts of people measured per hour, the date of that measurement and the place of the measurement. Second, the GVB data gives the counts of passengers, the hour, the station, and the date. And third, the event data, gives the dates for special events within Amsterdam and their location.

The dataset could later be expanded upon with the following data sources.

First, Telecom data of certain points within the city. Everytime someonse sends a requests over the internet, this request is picked up and saved. Think a person sending an app, posting a picture on Instagram, or searching dogs on google.

And second, wheather data of Schiphol from the last few months. There is no KNMI data available for Amsterdam, but Schiphol is close enough to count as an estimation.

3.2 Methods

For this research, the following prediction one of the following methods will be used. Or two methods will be used and compared to each other in terms of performance. The crowdedness count of the localized data count points will be used as the ground truth for the model.

The first prediction method is a *Multivaraite Linear Regression Model*, where the model seeks to predict the numerical value of crowdedness at a given time on a given day. The categorical values will be transformed into binary values 0 and 1.

And the second method is a *Decision tree with Gradient Boosting*, where the tries to predict the level of crowdedness and improve the results with gradient boosting. The categorical values will again be converted to binary values of either 0 or 1.

3.3 Evaluation

As there is no similar research nor model available to compare the results with, this research has no baseline. The models will be evaluated in the following ways:

First, the *Linear Regression Model* will be evaluated by comparing the predicted crowdedness number with the actual crowdedness number.

And second, the *Decision Tree* will be evaluated by checking whether the instances in the given clusters are similar to each other in terms of given crowdedness.

4 Risk Assessment

There are three main foreseeable risks with this project:

The first risk is that the data is not delivered on time. To make sure I'll be able to start if this happens, I have acquired a dataset with older, incomplete and less relevant data. This will enable me to start with my research on time and incorporate the real dataset later.

The second risk is that my computer lacks the computational power to run the model. If this happens, I will be able to run my model on one of the computers within the company. These computers are used to run their high computational models, so it will be able to run mine.

And the third risk, is that the data preparation takes too long with the large dataset. In this case, I will only use a small subset of my data to run the model on. This will give me some results needed to be able to finish my research on time.

5 Project Plan

References

1. Niu, Xiaoguang, et al. "A hierarchical-learning-based crowdedness estimation mechanism for crowdsensing buses." 2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC). IEEE, 2017.
2. Glanz, Jason M., et al. "Prediction model for two-year risk of opioid overdose among patients prescribed chronic opioid therapy." Journal of general internal medicine 33.10 (2018): 1646-1653.

Week Number	Date	Achievement
Week 1	01/04 - 05/04	Related Literature
Week 2	08/04 - 12/04	Data Preparation
Week 3	15/04 - 19/04	Data Preparation
Week 4	22/04 - 26/04	Methodology
Week 5	29/04 - 03/05	Evaluation
Week 6	06/05 - 10/05	Initial Results
Week 7	13/05 - 17/05	Incorporate Feedback
Week 8	20/05 - 24/05	Finalize Methods
Week 9	27/05 - 31/05	Final Evaluation results
Week 10	03/06 - 07/06	Draft Version Thesis
Week 11	10/06 - 14/06	Finalize Thesis
Week 12	17/06 - 21/06	Project Finished

3. Ing, Edsel B., and Royce Ing. "The use of a nomogram to visually interpret a logistic regression prediction model for giant cell arteritis." Neuro-Ophthalmology 42.5 (2018): 284-286.
4. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.
5. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems. 2017.
6. Ding, Chuan, et al. "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees." Sustainability 8.11 (2016): 1100.

7. Barth, Raul S., and Renata Galante. "Passenger density and flow analysis and city zones and bus stops classification for public bus service management." SBBD. 2016.