Detecting Fraudulent Companies Using Online Data

Author: Don de Lange
don.delange@student.uva.nl
11156414
Github Repository
Universiteit van Amserdam
Amsterdam

Supervisor: Noud de Kroon
a.a.w.m.dekroon@uva.nl
Universiteit van Amserdam
Amsterdam

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. The predictions with the given data will be made with both a multivariate regression model and a gradient boosted decision tree. The predicted crowdedness of both models will be visualized in separate time series over a given period of time. The two models will be compared in the evaluation based on the loss.

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.

The aim of this research is to built time series prediction models 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, the count of persons at given localized spots will be gathered. These counts will be used by the predicition models to predict the number of people at these localized spots, 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 point. The data preparation of the datasets that will be used in this research. There is no guarantee that the full dataset will be preprocessed into a usable format for this research, as the full dataset still needs to be prepared and gathered within the city of Amsterdam. So it is possible there will be some inconsistencies or missing values within the data.

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 methodology 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 Question

The main research question is stated as follows:

How can a prediction 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.

- *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?*
- *How can the data be structurally visualized in a time series?*

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 this 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. From this research I could use the data preparation method and the SVM model and apply it in this context.

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

Given that the crowdedness is measured as a natural number and not stated as a label, a regression model would be the best fit.

Ing and Ing (2018) constructed a multivariate predictive model with ten different variables for the nomogram of GCA. This research aims to use the nomograms mentioned in this research as explainability for the given prediction from the regression model.

As an alternative classification method the gradient boosting decision tree could be used. 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 a similar accuracy score, whilst being faster. This would be useful for this research as there is a lot of data that needs to be run through the model. Increasing the time efficiency would make the model more feasible for the given time frame.

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 use XGBoost to optimize the previously mentioned gradient boosting 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 use their feature selection method to test which features combination leads to the most significant increase in performance compared to the performance of a random guessing model.

3 METHODOLOGY

3.1 Data

The table below (see table 1) gives an overview of the entire dataset, what features it consists of and what the input features will be for the models.

Firstly, the dataset *Camera Count* gives a count of objects that passed a localized points in the city. These objects are counted with the use of camera recognition and saved as a JSON dump file. For this research the data will be used to predict the number of people at the same localized points the data is gathered. This research will use the aggregated count of people that passed the localized points in the time span of an hour.

Secondly, the dataset *IP Router* gives a count of devices that connect to IP routers at localized points. These IP Routers detect what devices are connected with them and log these. These logs will be used as data source. Similarly to the above given dataset, the number of smarthphones that connected at the localized points will be predicted. This research will use the aggregated count of connected smartphone devices in a time span of an hour.

Thirdly, the dataset *GVB Passenger Data* contains the number of passsengers at each station at an hourly rate, within the city of Amsterdam. This research will use this data to predict the number of passengers at these given stations, at an hourly rate. .

And thirdly, 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.

Table 1: Overview Data Dataset

Data Source	Features	Input Data Models
Camera Count	• Object • Time • Camera ID • Time • Count	Aggregated count people per hour per camera
IP Router	• Object • Time • Router ID • Time • Count	Aggregated count people per hour per Router
GVB Passenger Data	• Station • Hour slot • Number of Passengers	Number of passengers per station per hour
Amsterdam Event Data	• Location Event • Data event • Name Event • Coordinates Event	Data Events and their location

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

Firstly, Telecom data of certain points within the city. Everytime someone 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.

Secondly, weather 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.

And thirdly, twitter data of tweets send from within the city. The number of tweets send from within amsterdam could be counted, using the previously mentioned Telecom data. On these tweets, a topic and sentiment analysis could be performed to detect whether the tweets about crowdedness are positive, negative or neutral.

3.2 Methods

For this research, the following two methods will be used and compared to each other. Feature selection will be used to select the optimal subset of features from the data (Glanz et al., 2018).

The first prediction method is a *Multivariate Regression Model*, where the model seeks to predict the numerical value of crowdedness at a given time on a given day. The categorical features will be hot encoded into binary values of 0 or 1. The performance will be explained with the use of Nomograms (Ing and Ing, 2018)

And the second method is a *Decision tree with Gradient Boosting*, where the tree uses regression to estimate the crowdedness (Guolin et al, 2017). The categorical features are one hot encoded into binary values of either 0 or 1. The performance will be optimized with the use of XGBoost (Xia et al., 2016).

In case there is enough time, *SVM Classifier* could be added as possible third method to this research. As described in related work, Niu et al. (2017) used this model to estimate the crowdedness level of buses, given sensory data. The model takes the number of people from a single localized point as input. The model is then trained to determine the optimal crowdedness level, given the input data. Once the model is trained, the 'best' crowdedness level is chosen, given the number of people at a given time.

The prediction of the above given model will be visualized in a time series over a given period of time. The prediction will be made at an hourly rate.

3.3 Evaluation

As there is no similar research nor model available to compare the results with, this research has no baseline. The crowdedness count of the localized data count points will be used as the ground truth for the models. 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 ground truth.

Second, the *Decision Tree* will be evaluated with the loss function of mean squared error between the estimation and the ground truth.

And third, optionally, the *SVM Classifier* will be evaluated with the loss function hinge loss.

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.

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.

And the fourth risk is that the data is too dirty and will not be usable for the model. In this case, I will use a subset of the data and perform the data preparation on this subset. This subset will then be used for the prediction and evaluation part of this research. Alternatively, the dataset could also be excluded from this research and be replaced by another dataset.

5 PROJECT PLAN

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

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.

[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.