IP 3: Predicting production and attraction of zones based on POI's and population density with an xgBoost model.

# Description of the project

The aim of this project is to predict the production and the attraction per zone by using a machine learning model. The data used for this is the information from the Point Of Interest (POIs)[[[1]](#footnote-1)] and the density per zone. The POIs for the studied area were made available through Lotte Notelaers and the densities per zone have been calculated using the public available data of the number of inhabitants per ha in Flanders[[[2]](#footnote-2)]. As actual values, the production and the attraction calculated according to the department Mobiliteit & Openbare Werken (MOW) are assumed.

An XGboost model is used as the machine learning model since it is a widely used model in the literature. SHAP values are used to interpret the results of the model. To evaluate the performance of the model, it is compared with the linear model of the paper of Notelaers, L. et al. [1]

# Research questions

Four models will be made to help answer the research questions. A small model will be trained on the data of Antwerp. A full model will be trained on the zones of Antwerp and smaller cities such as Mechelen. These two models will be used to predict the production and attraction of the zones.

The main goal is to create a model that is reliable and gives decent results. The following research questions are examined using this model:

1. Does a machine learning model provide better estimates than a linear model?
2. How does the full model perform on different sizes of cities?
3. How does the small model perform on smaller cities?
4. Is it useful to add additional data from smaller cities to predict a larger city?

# Files

## Overview files:

The files for this project are as follows:

* **SHAP.ipynb**: Explanation and background information on SHAP values
* **density\_per\_zone.ipynb**: File to compute the density per zone
* **combine\_data.ipynb:** File that creates the dataset for the models. It combines the data of MOW, POIs and the density per zone to one dataset.
* 4 Models:
  + **Full\_model\_attraction.ipynb:** Model to predict the attraction based on training data of multiple cities
  + **Full\_model\_production.ipynb:** Model to predict the production based on training data of multiple cities
  + **Small\_model\_attraction.ipynb:** Model to predict the attraction based on training data of Antwerpen
  + **Small\_model\_attraction.ipynb:** Model to predict the production based on training data of Antwerpen

## Reading files

The following order is recommended to read the files:

1. **SHAP.ipynb:** If desired, read first the background information on SHAP values  
   Note: This is the information already presented in March
2. **density\_per\_zone.ipynb :** If desired, read the file to calculate the density per zone  
   Note: Not essential as this also comes back briefly in the presentation and this is not necessary for understanding the rest of the project
3. **combine\_data.ipynb:** If desired, read the file that creates the dataset for the model

Note: This file contains no insights. This file does the following:

* Calculates production and attraction per zone using the data from MOW
* Calculates totals per zone of the POIs' categories by using the POIs
* Creates one large dataset combining the information of the density, the production and attraction and the POIs of the zones of interest

1. **Models:**

It is recommended to read the full models first. These files are fully commented and contain all the descriptions.The small models have the same comments as the full models. However no elaborate descriptions are added as the small models serve as support to answer the research questionss. Moreover, it is also recommended to read the attraction model first and then the production model, as the production model contains some comparisons with the attraction model. In summary, the recommended order is:

* **Full\_model\_attraction.ipynb**
* **Full\_model\_production.ipynb**
* **Small\_model\_attraction.ipynb**
* **Small\_model\_attraction.ipynb**

# Answers research questions

Before reading the answers to the research questions, it is recommended to read the files of the models. These answers contain the information extracted from these files.

## Does a machine learning model provide better estimates than a linear model?

To evaluate the performance of the machine learning model, the full model is used. A comparison is made with how the models perform on test data. Since only results from the linear model are known for the test on Ghent, the performance is compared with this city.

### Attraction

Table 1 and Table 2 show the result of the metrics for the test on Gent for the prediction of the attraction. It is notable that the linear model has a bias of overprediction, while the full attraction model has a bias of underprediction which is about half of the magnitude. In addition, it can be seen that the difference in RMSE is a bit lower in the full model. However, the full model does have a much lower MAE. It can be concluded that the full model performs better in predicting the attraction on Gent than the linear model.

Table 1: Metrics linear attraction model: Gent

Afbeelding met tekst, schermopname, nummer, Parallel

Automatisch gegenereerde beschrijving

Table 2: Metrics full attraction model: Gent

Afbeelding met tekst, schermopname, nummer, Lettertype

Automatisch gegenereerde beschrijving

### Production model

Table 3 and Table 4 show the result of the metrics for the test on Gent for the prediction of the production. It is notable that the linear model has a bias of overprediction, while the full production model has a bias of underprediction that is about the double of the magnitude. So note that again it is the linear model that has overprediction and the full model has underprediction. In addition, it can be seen that the difference in RMSE is a bit lower in the full model. However, the full model does have a much lower MAE. It can be concluded that the full model seems to performs a bit better in predicting the production, certainly in the intervals up to a 1000 produced trips, on Gent than the linear model. However, for the production the conclusion is less strong than the for the attraction. This is because, on the one hand, the linear model already makes rather accurate predictions and, on the other hand, there is only one very dominant factor, namely 'inhabitants'.

Table 3: Metrics linear production model: Gent

Afbeelding met tekst, schermopname, nummer, lijn

Automatisch gegenereerde beschrijving

Table 4: Metrics full production model: Gent

Afbeelding met tekst, schermopname, Lettertype, nummer

Automatisch gegenereerde beschrijving

## How does the full model perform on different sizes of cities?

To evaluate the performance of the full model on different size of cities, the results on the test data is checked.

## Attraction

Table 5 shows that attraction predictions get better as cities get smaller. At first glance, this is a somewhat surprising result if one takes into account that more than half of the observations of the training model come from Antwerp (152 zones out of 268 zones). An explanation for this is that the model mainly has difficulties in predicting high values as shown in Table 6. The smaller the city, the fewer or no high values there are, leading to more accurate overall predictions.

Table 5: Metrics full attraction model: Totals per city

Afbeelding met tekst, schermopname, Lettertype, nummer

Automatisch gegenereerde beschrijving

Table 6: Metrics full attraction model: overall

Afbeelding met tekst, schermopname, nummer, Lettertype

Automatisch gegenereerde beschrijving

## Production

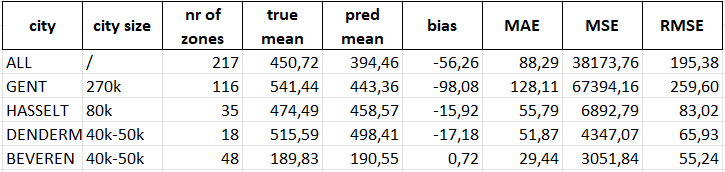
For production, it is also the case the smaller the city, the more accurate the predictions are in general. Note the accuracy of the total predictions for all cities, except Ghent. For production, the remarkable conclusion can even be taken that the lower the production value, the more accurate the prediction.

Table 7: Metrics full production model: overall

Afbeelding met tekst, schermopname, Lettertype, nummer

Automatisch gegenereerde beschrijving

Table 8: Metrics full production model: Totals per city



## How does the small model perform on smaller cities?

## Attraction

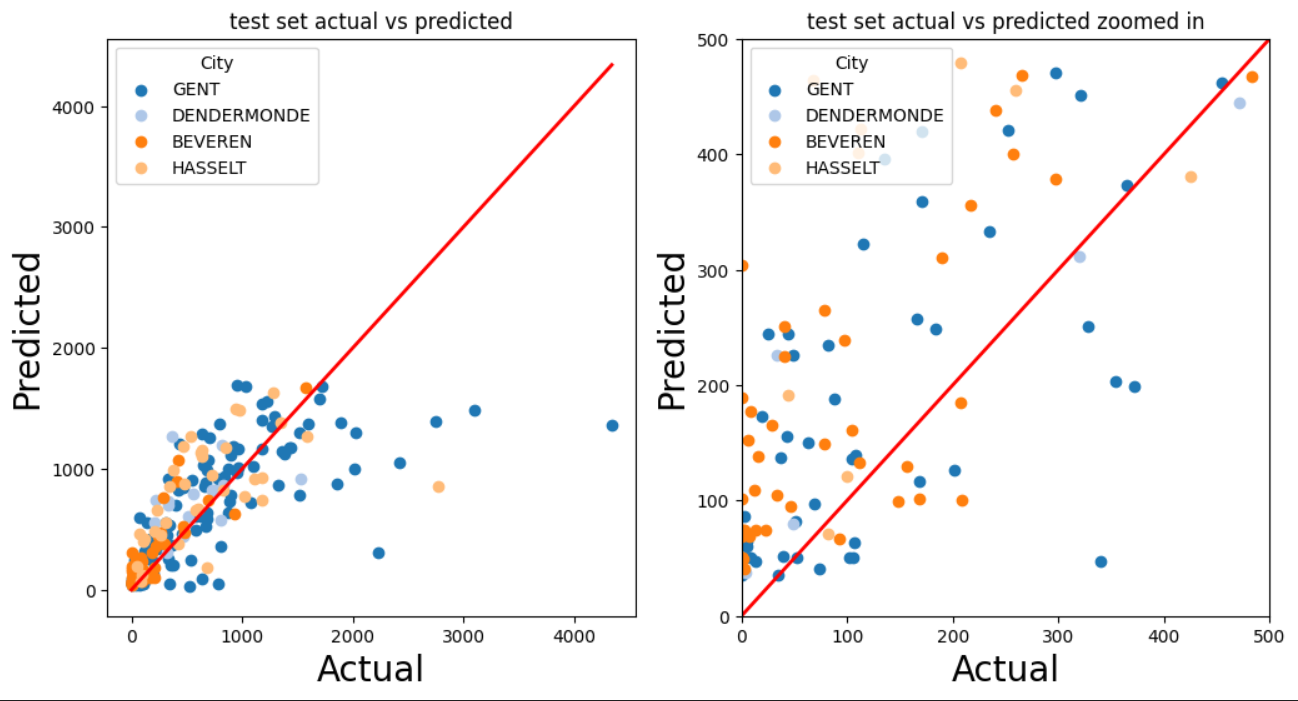


Figure 1: Predicted vs actual values of the test set of small attraction model

To answer this question, a look is taken at the plot of the predicted values of the small attraction model, see Figure 1, and the metrics. Figure 1 immediately shows that the model trained on Antwerp mostly overestimates. Certainly on the smaller cities the predictions are mostly above the red line, meaning the predicted values are overestimations.

All zones in the intervals up to a 1000 trips show a very high bias. Other intervals lack information to make conclusions.

What causes this? A possible explanation is that the small model puts more importance on the POI features compared to the full model. This causes predictions to increase rapidly when a zone contains a high amount of POIs. For example, the force plot of the prediction of zone 2304 in Dendermonde can be seen in Figure 2. This zone has a real attraction of 360. The model predicted 1267. This plot shows that Services, School and Industry have a high positive impact on the prediction. However this impact is too much, finally leading to an overestimation.

Afbeelding met tekst, schermopname, Lettertype

Automatisch gegenereerde beschrijving

Figure 2: Force plot of zone 2304

## Production

Afbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figure 3: Predicted vs actual values of the test set of small production model

For production a first look is taken at Figure 3. This figure shows decent predictions for all cities. In contrast to the attraction model, the small production model actually performs well on the smaller cities.

The metrics in Table 9 show that the absolute bias and the MAE increases slightly in Dendermonde when the model is only trained on Antwerp, compared to metrics in Table 10. However, Dendermonde has the lowest R^2 score of 0.85. This is still high. The explanation for these good predictions is because of the high feature importance of 'inhabitants', as mentioned earlier. Beveren and Hasselt show similar behaviour.

Table 9: Metrics of Dendermonde of small production model

*Afbeelding met tekst, schermopname, nummer, Parallel

Automatisch gegenereerde beschrijving*

Table 10: Metrics of Dendermonde of full production model

*Afbeelding met tekst, schermopname, nummer, Lettertype

Automatisch gegenereerde beschrijving*

## Is it useful to add additional data from smaller cities to predict a larger city?

## Attraction

The metrics, shown in Table 11 and Table 12, show that the full model has slightly better metrics, with a lower RMSE and a lower MAE. But when dividing over the zones, the behaviour between the two models is different. The small model predicts higher values over most zones. Which means there’s more overestimation in the lower intervals but less underestimation in the higher intervals. This causes the small model to have a bias closer to 0 than the full model.

So for the attraction model it is hard to conclude if it’s useful to add extra smaller cities to the training data for predicting a bigger city. So, no strong conclusion can be made.

Table 11: Metrics of Gent of small attraction model

## Afbeelding met tekst, schermopname, nummer, lijn Automatisch gegenereerde beschrijving

Table 12: Metrics of Gent of full attraction model

*Afbeelding met tekst, schermopname, nummer, Lettertype

Automatisch gegenereerde beschrijving*

## Production

The metrics, shown in Table 13 and Table 14, show that no significant changes can be seen. All metrics are very close. The predictions for Gent didn't get better when extra training data was added. In conclusion, adding extra training data of smaller cities did not give better predictions for a bigger city.

Table 13: Metrics of Gent of small production model

## Afbeelding met tekst, schermopname, nummer, Lettertype Automatisch gegenereerde beschrijving

Table 14: Metrics of Gent of full production model

Afbeelding met tekst, schermopname, nummer, lijn

Automatisch gegenereerde beschrijving

1. Notelaers, L., Verstraete, J. and Tampère, C.M.J. (2023). A Demand Modelling Framework based on OpenStreetMap [Manuscript submitted for publication]. KU Leuven Insitute for Mobility (CIB), KU Leuven. [↑](#footnote-ref-1)
2. Vlaamse overheid. (n.d.). Inwonersdichtheid per ha - vlaanderen - toestand 2019. https://www.vlaanderen.be/datavindplaats/catalogus/inwonersdichtheid-per-ha-vlaanderen-toestand-2019 [↑](#footnote-ref-2)