Report Project II

November 2022

Pelle Kuppens, Annika Waltmann, Marta Nosowicz, Sam Groen

Introduction

This report is a follow-up to the Exloratory Data Analysis (2022) report. The previous report contained an analysis of available data about Amsterdam's health and public space conditions. It aimed to inform the governing bodies of the city about the current situation regarding these topics and support policy development, innovation, and implementation.

The analysis was carried out by analyzing three main datasets and two secondary datasets obtained through the municipality of Amsterdam using the programme RStudio. The used datasets contained information about:

- self-perceived health conditions (including, overall health, psychological health, loneliness, and obesity),
- perceived quality of public spaces and the relative amount of available hectares of greenery,
- the population's amount, age, gender and employment status

Additional datasets were analyzed to understand the discovered trends in relations between health and public spaces. Those datasets included information about: income, housing, social facilities.

Some results of the previous analysis of the data point to a relationship between the social status and health quality among inhabitants. In richer districts of Amsterdam people are more likely to perceive their health as good. Additionally, the data also showed that individuals' perception of the quality of living spaces is also related to the perception of their own health. However, some other factors such as age or the amount of facilities are also thought to have an impact on health. Furthermore, the data showed no significant relation between the amount of greenery and health, as is commonly believed. Therefore, it is important to conduct further research into this correlation.

The above-mentioned conclusions can be facilitated by many factors and might seem difficult to understand for the general public. They can create misunderstandings and lead to false implications. This in turn can cause irrelevant policy development. To clarify those results and gain a better understanding of the situation further analysis is needed.

The aim of this report is to deliver that clarification. First, a regression analysis is done to examine the relationship between health and public space, wealth and age. Secondly, a data classification is performed. A classification can help to manipulate, track and analyze individual pieces of data for forecasting and in turn draw a clearer image of what policies should be in order to support not only current trends but also continue being relevant in the future.

Regression Analysis

The results of the exploratory analysis indicated some direction for new policy development. There is a common understanding that greenery in the living environment can improve health. However, the results of the previous report showed that there might be other factors with higher significance for good health among inhabitants of Amsterdam. While it was concluded that health can to an extent be facilitated by the quality of the perceived living environment, it also became clear that there might be some other facilitators having a bigger impact on individuals' self-estimated good health. This part of the report will

bring the focus to understanding which parameters are most important determinants of health and what is the correlation between them. It will help to describe the strength and character of correlations between health (the dependent variable) and the independent variables, and thus adjust the insights accordingly and answer the following research question: What are the true facilitators of self-assessed good health in Amsterdam?

Results

The previous report indicated some variables that could be significant for the correct understanding of the health situation amid districts in Amsterdam. The forward selection procedure performed for this report made it clear that some variables showed more significance. Those variables are described as independent variables and are as follows: disposable income, perceived aesthetics of the living environment, and total amount of facilities. On the other hand, some variables that were proven to be less important were the amount of greenery in the inhabited districts and age. The following section of this report will aim at describing the accuracy of the significant variables and strength and character of correlations with health in Amsterdam's districts.

```
Call:
lm(formula = hea_good ~ (pub_setup) + (poly(inc_disposable, 2)) +
    (poly(hou_value, 2)) + poly(tot_facilities, 2), data = all_data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1328  -2.2929   0.0279   2.2087  11.4166

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 68.234      1.518  44.941  < 2e-16 ***
pub_setup2.ok                3.906      1.518   2.573 0.010205 *
pub_setup3.beautiful         4.578      1.536   2.980 0.002945 **
poly(inc_disposable, 2)1   234.577      7.394  31.727  < 2e-16 ***
poly(inc_disposable, 2)2   -57.324      5.679 -10.094  < 2e-16 ***
poly(hou_value, 2)1        -97.737      8.057 -12.131  < 2e-16 ***
poly(hou_value, 2)2         20.147      5.659   3.560 0.000386 ***
poly(tot_facilities, 2)1    96.878      4.267  22.704  < 2e-16 ***
poly(tot_facilities, 2)2   -18.533      3.634  -5.099    4e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.37 on 1131 degrees of freedom
Multiple R-squared:  0.7705,    Adjusted R-squared:  0.7689
F-statistic: 474.7 on 8 and 1131 DF,  p-value: < 2.2e-16
```

Only perceived aesthetics of the living environment, housing value, disposable income and number of social facilities, were found to have high significance. Percentage of greenery and age of the residents were found to have low significance.

To further understand those relationships and their significance a multiple linear regression model was built. It can be said that the model is reliable because of its high R2 value and symmetrical distribution of the scatterplot and the histogram. The residuals seem to be symmetric. It appeared that the interaction effect between health, wealth, and quality of the living environment is statistically significant. This model concludes that disposable income, housing value, and social facilities have the strongest impact on residents' health. People with higher incomes can afford better healthcare and personal care. This can positively add to their self-estimated health. Furthermore, more access to social facilities creates more opportunities for physical and social activities which are known to be good for one's health. Surprisingly, the model shows a lower

significance of the quality of the living environment on the residents' self-assessed good health in comparison to the other factors.

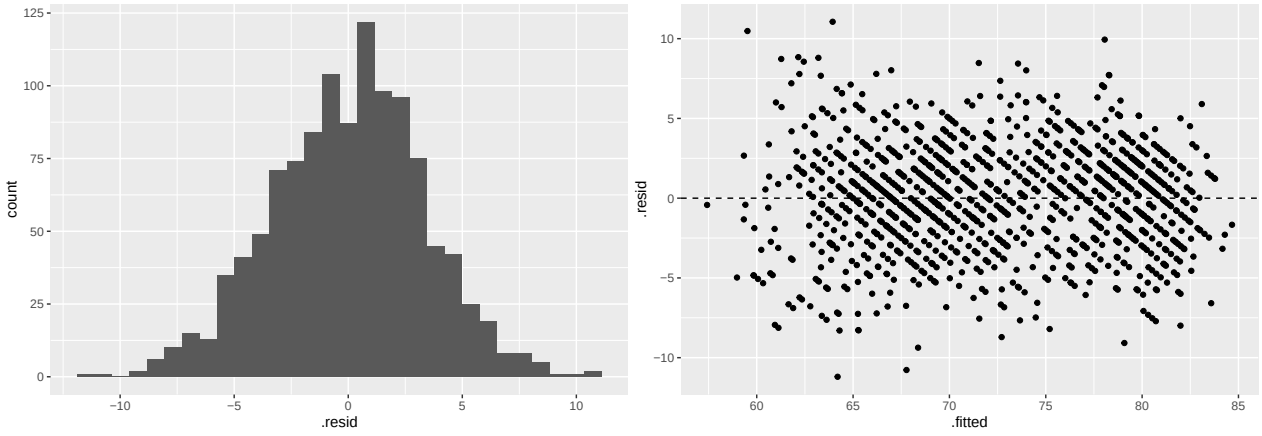
ANOVA Analysis

```
Anova Table (Type II tests)

Response: hea_good
                                       Sum Sq   Df  F value    Pr(>F)
pub_setup                               158.5    2   7.0851 0.0008754 ***
poly(inc_disposable, 2)               11926.7    2 533.1169 < 2.2e-16 ***
poly(hou_value, 2)                     1361.4    2  60.8548 < 2.2e-16 ***
tot_facilities                         5665.9    1 506.5299 < 2.2e-16 ***
pub_setup:poly(inc_disposable, 2)       119.4    4   2.6687 0.0310036 *
poly(inc_disposable, 2):tot_facilities  389.1    2  17.3940 3.633e-08 ***
Residuals                             12595.2 1126
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table was created to test hypotheses about group means while controlling for other factors and assess the importance of the factors next to the regression model. The analysis of the table shows once again that income, housing value, and social facilities appeared to be the important predictors. Quality of the living environment was proven to be less significant. Below you can see that the model that was created as a result is of high quality due to the normal distribution of the histogram and the symmetrical distribution of the scatter plot.

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Conclusion

The initial hypothesis and conclusions of the previous report are clarified. Amid various districts of Amsterdam wealth and access to social facilities correlate with higher percentages of inhabitants who perceive their health as good. Quality of the living environment is less significant. In comparison to the results of the previous report it is now clear that good health amid citizens of Amsterdam is not as much facilitated by the quality of the public space that they live in, but more so by their financial situation.

Classification Models

In order to make effective investments, ACME bases their investments on geographical data. However, this geographical data is only provided until the next calendar year. As a result, ACME only has data, without corresponding geographic aeras. However, by using historical data a prediction can be made to link the most

recent data to a geographical area. By using the datasets used in earlier parts of this report, several classifiers will be built to assess whether it is possible to do this effectively. The question is, which classifier should be used to predict geographical areas of the most recent data?

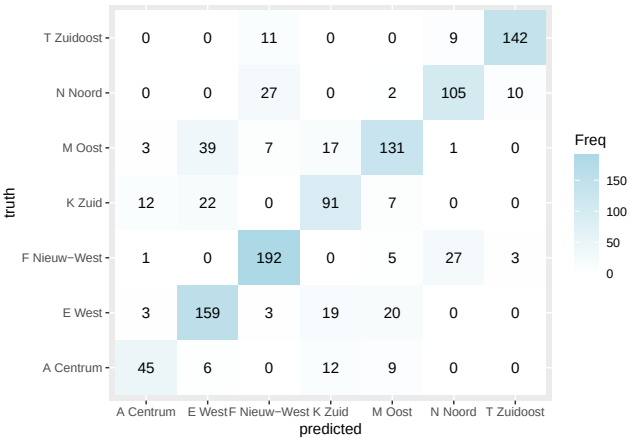
To assess and compare the different classifiers, a look is taken into two performance metrics. First, overall accuracy was taken into account as it is important to have an accurate classifier when matching geographical areas with the most recent data. Second, a look is taken into the f-score range, which takes into account the imbalance of the classifier as it is possible that the classifier is very good in predicting a certain district, but bad at predicting another district. For all classifiers, a confusion matrix is shown to give an overview of the classifier.

For every classifier, it was decided to use the following inputs, based on the models built earlier. First, a look was taken into the interaction effects of all health components. Since this is only possible for a multinomial regression classifier, these interaction effects were dropped for the other two classifiers. Then, disposable income, housing value and total social facilities were added. In addition, the aesthetics of the public setup and the average amount of greenery were added. At last, the percentage of population with an age above 60 was added to complete the model.

Multinomial Regresson

The first classifier that was built is a so-called multinomial regression classifier. This type of classifier can be used to predict categorial variables based on the input of multiple dependent variables. This is done by making complex calculations.

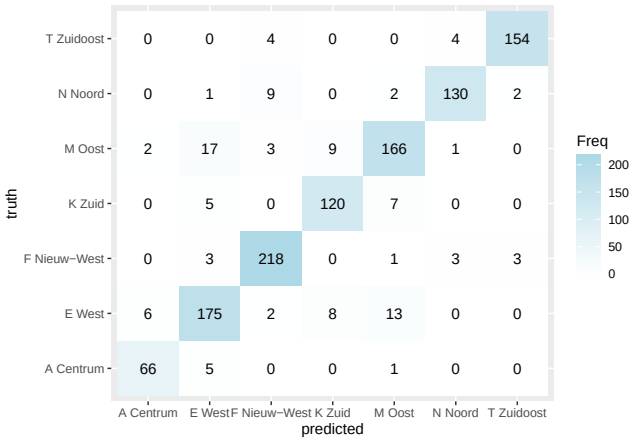
With this model, an overall accuracy of 75.8% was achieved. What is interesting is that the range of the f-scores of the model is quite big as it ranges from 66.1% to 89.5%. The model does not predict well for the Centrum district and Zuid district, but does a very good job at predicting Nieuw-West and Zuidoost.



Decision Tree

Secondly, a random forest model was built. This model is built upon the principle of a decision tree. As the term explains, a 'forest' is made by making different decision trees randomly. Then, a prediction is made based on the 'votes' of the decision trees, with the majority being the answer. For this classifier, the same parameters were used. However, decision tree classifiers do not work with interaction effects, so these are dropped.

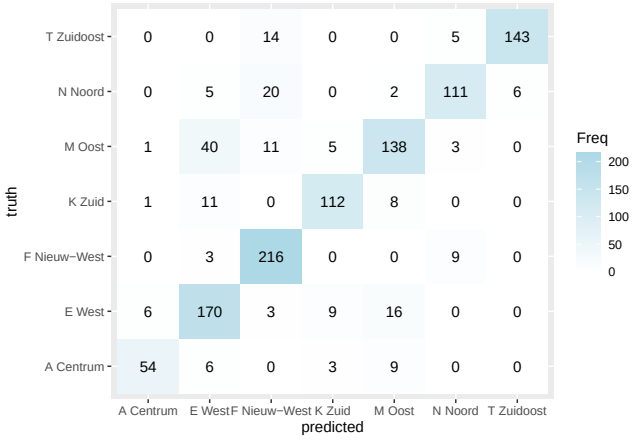
Since this classifier is based on randomness, an approximate accuracy of 90% was achieved. This classifier performs significantly better than the multinomial regression classifier. In addition, the f-score range is smaller as well. With this classifier, an f-score range of 85.1% to 95.0% was achieved. The classifier predicted the worst for the West district, while it did the best with the Zuidoost district.



Support Vector Machine

At last, a support vector machine was built. This model is built around separating data points by drawing so-called hyperplanes between different categories. When the data is not easy enough to split in a straight line, several complex calculations in a multidimensional space are performed and reversed to make a non-linear border between the data points.

With this model, an accuracy of 82.8% was achieved. An f-score range of 74.4% to 91.9%. The classifier predicts the worst for West, and the best for Zuidoost. This classifier thus outperforms the multinomial regression model, but is underperforming compared to the random forest classifier.



Cross Validation

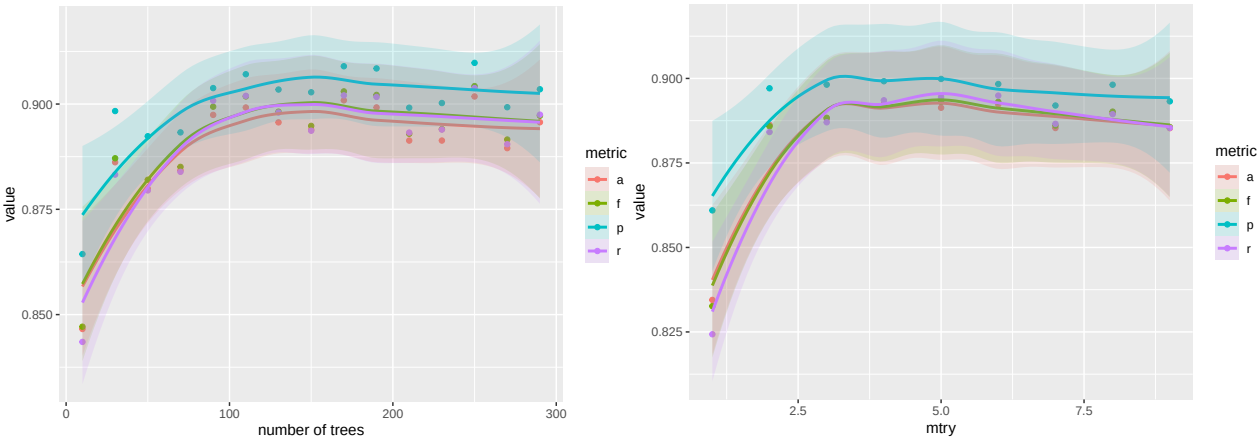
To optimize the performance of the classifiers mentioned above, several cross validations are performed. Cross validations help to optimize the classifiers by tuning so-called hyperparameters of the classifier. Optimizing can be done in terms of accuracy, but also for compute time.

It was decided to drop the multinomial regression classifier as this classifier did not match the performance of the other two classifiers.

Two cross validations were done for the random forest classifier. First, cross validation was done to assess what the optimal number of trees. This helps to improve computational efficiency of the classifier. As can be seen on the left graph below, the performance of the classifier drops after 150 trees. As a result, it is more efficient to use 150 trees than the 500 trees used in the initial random forest classifier. Calculating the extra 350 trees is a waste of time.

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

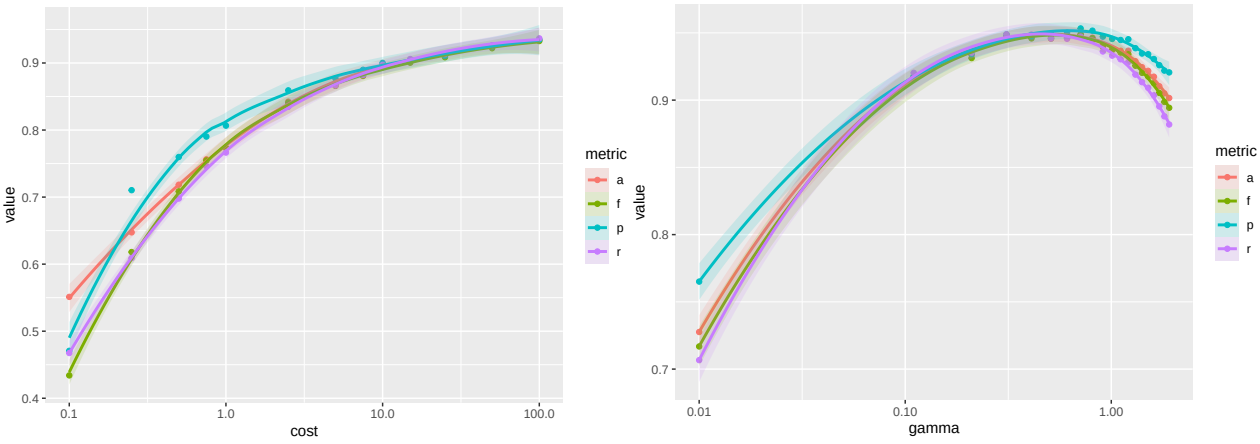
```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Second, a cross validation was performed to find what the optimal number of features for new branches was for the random forest. As can be seen on the right graph above, making branches based on 3 features seems to result in the best Random Forest classifier with an accuracy of approximately 90%. As a result, a random forest with 150 trees and 3 features for new branches is optimal.

For the support vector machine two cross validation were performed as well. First, one in which the classifier is punished when it falsely predicts a district. As can be seen in the left graph below, the performance of the model keeps rising as this 'cost' is increased. However, a cost that is too high will result in a less generalizable model. Therefore, it was chosen to keep the cost at 10, as increasing it any higher also results in diminishing results.

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Second, cross validation was performed on another hyperparameter, gamma, which affects the curvature of the hyperplane. As can be seen on the right graph above, a gamma of 0.95 seems to result in an optimal model, with an accuracy of 95%. As a result, the optimal hyperparameters are a cost of 10, and a gamma value of 0.95.

Recommendation

Three different classifiers were developed in order to match data points to a certain geographical location. It was found that the multinomial regression classifier was unable to match the performance of the other two classifiers. After performing several cross validations to optimise the model, it was found that the support vector machine classifier was the best performing classifier, with an accuracy of 95%. Therefore, it is

recommended to use this classifier to predict the geographical area of the most recent data.

Discussion

It must be noted that all classifiers are trained on an imbalanced dataset. As a result, the classifiers are biased towards the districts that have the most data points. Also, due to randomness, the performance of the Random Forest classifier is subjected to change with every run. As a result, the performance will slightly change compared to the performance in this report.