,,

Universiteit Leiden
The Netherlands

Opleiding Informatica & Economie

Modelling Gentrification in Dutch Neighbourhoods: Investigating the Role of Neighbourhood Amenities

Floris Hessels

Supervisors:
Dr. Y. Fan & Dr. A.J. Knobbe

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

14/08/2023

Abstract

In a world characterised by rapid urbanisation and growing prosperity, gentrification has emerged as a well-known urban phenomenon affecting cities worldwide. Effectively handling gentrification relies heavily on the capacity of policymakers, planners and researchers to grasp its causes and consequences. This study aims to model gentrification in the Netherlands by using neighbourhood amenities. While gentrification has been extensively studied, limited attention has been given to the influence of amenities on gentrification, particularly through the implementation of machine learning and Explainable Artificial Intelligence (XAI). Investigating the significance of amenities on gentrification fills this research gap and could provide valuable insights for policymakers, planners and researchers. It is important to emphasise that this study operates within a correlation setting rather than a prediction setting, given the absence of a clear direction of causality between gentrification and amenities.

Data from Statistics Netherlands covering three time intervals, namely 2013-2018, 2014-2019, and 2015-2020, was used to conduct research. The amenities (features) and the gentrification score (target variable) were derived from the data. Several regression machine learning models were trained, tested and evaluated. The top-performing model achieved an R-squared of 0.4825, an MAE of 0.0479 and an RMSE of 0.0660. Additionally, these results of a black-box model were extensively interpreted using an additive explanation tool. This study potentially opens up the doors to include a broader range of amenities considered and visualising gentrification patterns on the map of the Netherlands.

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Research Question . . . 1
- 1.3 Thesis Overview . . . 2

2 Literature Review **3**
- 2.1 Gentrification in the Netherlands . . . 3
- 2.2 Amenities and Gentrification . . . 4
- 2.3 Neighbourhoods Eligible for Gentrification . . . 4
- 2.4 Consequences of Gentrification . . . 5
- 2.5 Machine Learning Approaches . . . 6
- 2.6 Research Gap . . . 7

3 Theoretical Framework **8**
- 3.1 Regression Models . . . 8
  - 3.1.1 Multiple Linear Regression . . . 8
  - 3.1.2 Random Forest Regression . . . 8
  - 3.1.3 Gradient Boosting Regression . . . 8
  - 3.1.4 XGBoost Regression . . . 8
- 3.2 Feature Selection . . . 9
- 3.3 Nested Cross-validation . . . 9
- 3.4 Metrics for Evaluation . . . 9
  - 3.4.1 R-squared . . . 10
  - 3.4.2 Mean Absolute Error . . . 10
  - 3.4.3 Root Mean Square Error . . . 11
- 3.5 Interpretability . . . 11

4 Data **12**
- 4.1 Datasets . . . 12
  - 4.1.1 Area codes . . . 12
- 4.2 Gentrification Variable . . . 13
- 4.3 Amenities . . . 14

5 Methods **15**
- 5.1 Data Preprocessing and Restructuring . . . 15
  - 5.1.1 Cleaning Data . . . 15
  - 5.1.2 Filtering Data . . . 16
  - 5.1.3 Reorganising Data . . . 18
- 5.2 Gentrification Score . . . 19
  - 5.2.1 Validity of Gentrification Score . . . 20
- 5.3 Machine Learning Model . . . 21

6 Results **22**
6.1 Model Comparison 22
6.2 Interpreting the Model 23

7 Discussion **26**
7.1 Discussion 26
7.2 Limitations 27

8 Conclusion **28**
8.1 Conclusion 28
8.2 Further Research 28

References **35**

Appendices **36**

A - Sources Information Amenities per District/Neighbourhoods Dataset **36**

B - Comprehensive Description of the Included Amenities **37**

C - Changed Neighbourhood Layout **38**

1 Introduction

1.1 Context

Gentrification is a process whereby middle-class or wealthy people move into poor neighbourhoods, resulting in changes to the physical, economic, and social landscape. The term 'gentrification' was initially documented by Ruth Glass in 1964. Despite being a subject of academic study for at least fifty years, gentrification research has seen a significant upturn recently [LSW13] [ZBC+18]. This can be attributed to researchers' bifurcated views and ideas, given that gentrification is a complex socioeconomic process with significant implications [BS17].

There is an ongoing debate about whether gentrification helps or harms neighbourhoods and their residents. Policymakers, researchers, and analysts generally agree that the effects are multi-faceted, encompassing both positive and negative impacts. On the one hand, gentrification can lead to the renewal of neighbourhood infrastructure and reduced crime rates [PSSF11]. On the other hand, negative impacts like displacement and loss of affordable housing may occur due to gentrification [Atk04].

Gentrification is influenced by various factors, including the presence of amenities in a neighbourhood. Neighbourhood amenities, such as shops, restaurants, schools, and cultural institutions, can improve the desirability of a neighbourhood to high-status individuals willing to pay more to live in a desirable location [HL16]. While it is recognised that amenities play a role in gentrification dynamics, it is essential to note that the relationship between amenities and gentrification is complex and characterised by mutual reinforcement. This makes it challenging to establish a clear direction of causality [Hyr16]. Therefore, this study aims to model gentrification and investigate the correlations between gentrification and amenities.

1.2 Research Question

The housing shortage is a highly relevant and urgent problem in the Netherlands. In 2022 the Netherlands faced a shortage of 315 thousand houses, especially affordable houses for low-income families. Therefore, the Netherlands aims to build 900 thousand houses before 2030, of which two-thirds will fall in the affordable segment [Min22].

Identifying the correlation between amenities and urban gentrification offers a policy instrument that can be used by governing authorities. Policymakers can use this knowledge to make informed decisions about the development of amenities in their communities. For example, when considering whether to grant a permit for a new amenity, policymakers can consider the potential effects on gentrification dynamics [HARH22]. Building upon this context, the research question arises:

What is the role of neighbourhood amenities on gentrification in the Netherlands?

1.3 Thesis Overview

This thesis, completed for the bachelor 'Computer Science & Economics', was supervised by Yingjie Fan and Arno Knobbe and is written for Leiden Institute of Advanced Computer Science (LIACS). This thesis consists of seven sections, each serving a different purpose. Section 1 serves as the introduction, introducing the subject, emphasising the motivation behind this research, and presenting the research question. Moving on to Section 2, which provides background information and identifies the research gap that this thesis aims to address. Section 3 provides a foundation for the subsequent sections, mainly focusing on the data science aspects. Section 4 is dedicated to describing the data used for this research and its source. The methodology of this research is illustrated in Section 5. Section 6 presents the results without interpreting these, which happens in Section 7. Furthermore, Section 7 reports the limitations of this study. Lastly, Section 8 answers the research questions and discusses possibilities for future research.

2 Literature Review

2.1 Gentrification in the Netherlands

Gentrification is a global phenomenon occurring in cities and neighbourhoods worldwide. It cannot be mindlessly assumed that the dynamics and manifestations of gentrification are the same in every country. The context, motives, and, thus, implications of gentrification can differ per country [CL95] [SMJS+20].

In the context of this study, the focus lies specifically on gentrification in the Netherlands. The concept of gentrification is not widely familiar in the Netherlands. Previously, the wealthy country had a broad social rental sector, rent protection, and controlled rent increase. These factors contributed to the country's relatively mild nature of gentrification [SD17]. However, gentrification has become state-led and is used as an urban policy. The 'state' in this context refers to the national and local governing authorities and housing associations [Tee15]. Measures undertaken are the redevelopment of housing units and urban renewal programs. Governing authorities believe state-led gentrification is a fruitful approach, stimulating positive outcomes for higher- and lower-income groups. Higher-income groups obtain suitable housing while simultaneously enabling lower-income groups to benefit from the economic and social resources brought by the influx of higher-income residents [VdGV+09]. Furthermore, according to governing authorities, this strategy tries to promote social mixing. As [HM21] argue, early gentrification will likely lead to more social mixing. However, at a later stadium, it will cause segregation and, thus, displacement. Segregation can occur within and between neighbourhoods [SBSKJ14]. Neighbourhoods with high levels of segregation may cause an absence of inspiring individuals as role models, limited opportunities to access beneficial local networks, and the stigmatisation of neighbourhoods [Wat88].

As mentioned above, gentrification takes place in the Netherlands. Subsequently, the question arises where does it take place? Although gentrification has been studied extensively in an urban context, it can occur in rural areas. Rural gentrification, however, differs significantly from urban gentrification [Phi93]. Therefore, the causes, characteristics, and implications of gentrification may vary in these different contexts [Sto10]. The Netherlands is a highly urbanised and densely populated country, resulting in few rural areas [HHG03].

Also, the boundaries of a neighbourhood can have a significant impact on how gentrification is perceived and experienced. For example, suppose the boundaries of a historically low-income and predominantly minority neighbourhood are redrawn to include more affluent areas. In that case, this can give the appearance of gentrification even if the neighbourhood has not undergone significant changes. Overall, considering changes in a neighbourhood's borders is essential in understanding gentrification and its impacts. It is necessary to carefully define and measure the neighbourhoods under analysis to ensure a fair and accurate assessment [Bar16].

Gentrification studies conducted in the Netherlands frequently used data from Statistics Netherlands [BBVSK20] [HM21] [Tee15]. This is an autonomous administrative authority which performs public service tasks but operates independently and not under the direct authority of a Dutch ministry [Sta23a].

2.2 Amenities and Gentrification

In multiple studies, the relationship between amenities and gentrification has been studied. The availability and quality of amenities have been identified as one of the critical factors associated with gentrification in neighbourhoods [HL16]. While amenities are a vital aspect correlated with gentrification, it is essential to consider the bidirectional relationship [Hyr16]. Amenities, such as shops, cultural institutions, schools, public transportation, and recreational facilities, contribute to the desirability of a neighbourhood. They are subsequently attracting higher-income individuals and families seeking a better living environment. As a result, the presence of amenities can significantly influence the decision-making process of potential gentrifiers, who are willing to pay a premium for the enhanced amenities. This deliberate investment in amenities further intensifies the process of gentrification by creating a cycle of rising demand, increased property values, and subsequent displacement of lower-income residents [CH17] [EHR19].

In addition to amenities attracting higher-income individuals seeking a better living environment, it is worth noting that this process can also happen in reverse. As higher-income individuals move into a neighbourhood, their demand for enhanced amenities can prompt local businesses and cultural institutions, among others, to fulfil their wishes regarding amenities. This can result in improved facilities. The correlative relationship between higher-income residents and amenity improvements contributes to a reinforcing cycle where increased demand and investment in amenities attract more affluent residents, driving gentrification [Hyr16].

In the Netherlands, starting a business often requires obtaining one or more permits or licenses. Typically, these permits must be applied for at the municipality or local government authority [Net23]. Given the correlation between amenities and gentrification, policymakers can use this knowledge to make informed decisions about the development of amenities in their communities. For example, when considering whether to grant a permit for a new amenity, policymakers can consider the potential effects on gentrification and displacement in the area, allowing them to mitigate the adverse effects of gentrification on vulnerable populations while still promoting economic development and neighbourhood improvement [HARH22].

2.3 Neighbourhoods Eligible for Gentrification

The definition of gentrification, as described in Section 1.1, implies that only some neighbourhoods are eligible for gentrification. To conduct thorough research on gentrification, it is crucial to avoid comparing neighbourhoods still eligible for gentrification with those that have already undergone gentrification or are too affluent to be gentrified [Bec20].

A foundational method to identify neighbourhoods eligible for gentrification is formulated by [GP86]. At the base year of their analysis, a neighbourhood had to meet the following criteria:

- the median value of single-family homes is less than the corresponding city-wide median;
- the median income is less than 80% of the corresponding city-wide median;
- the percentage of college-educated is less than the corresponding city-wide median; and
- the percentage of white is less than 90% of the tract population.

Researchers utilise several other methods to identify whether neighbourhoods are eligible for gentrification, some of which are inspired by [GP86]. For example, [GB16] argue that neighbourhoods with a median household income lower than the corresponding city-wide median household income are eligible for gentrification. Moreover, according to [ZB21], neighbourhoods were eligible for gentrification if the median household income was below the city-wide median and the percentage of buildings constructed in the last 20 years was below the city-wide median, which signalled disinvestment in that neighbourhood. Although there is no universal guideline to identify whether a neighbourhood is eligible for gentrification, researchers agree that not all neighbourhoods are eligible for gentrification.

2.4 Consequences of Gentrification

Previously, gentrification was described as a process whereby middle-class or wealthy people move into poor neighbourhoods. The associated consequences are changes to these neighbourhoods' physical, economic, and social landscape. However, what does this entail?

The impact of gentrification on neighbourhoods and their residents is a subject of ongoing debate. Different perspectives exist on whether it should be cheered on or discouraged [Atk04]. When viewed through a positive lens, gentrification brings about various beneficial outcomes for neighbourhoods and their residents. One of the main drivers of gentrification is urban renewal. It refers to the process of revitalising and transforming urban areas through various interventions [Hyr12] [ZSW14]. Urban renewal is linked to improved infrastructure, home renovation programs, and remodelled landscapes [EKK+13] [MMMG18]. Furthermore, crime rates in gentrifying neighbourhoods are likely to reduce. The presence of wealthier newcomers in the neighbourhood may make the neighbourhood a more attractive target for crime. In spite of that, there is a decline in property crime against pre-existing lower-income residents. In the long run, the crime rates against newcomers may reduce because of improved housing security and increased policing [Byr02]. As described in Section 2.2, improved amenities are one of the drivers of gentrification. Better amenities can enhance the quality of life for existing residents. Moreover, a major consequence of gentrification is an increase in property values, which could be seen as a benefit. These are two examples of the positive effects of gentrification on the community's residents. [Dua01] states: 'The process improves the quality of life for all of a community's residents. It is the rising tide that lifts all boats.'

However, the negative effects of gentrification on communities and their inhabitants cannot be overlooked. Gentrification and displacement are two interconnected phenomena, as gentrification often fuels the displacement of low-income residents. The displacement of residents can occur due to various factors, such as housing demolition, rising housing costs, as well as landlord harassment and eviction [NW06]. Moreover, it restricts the ability of low-income residents to move into these gentrifying neighbourhoods. The displacement of low-income residents and the loss of affordable housing may result in community conflicts, homelessness, and segregation [Atk04].

As mentioned above, gentrification has positive impacts, but these are not distributed equally. If residents are displaced from a gentrifying neighbourhood or cannot afford housing in such areas, they will not be able to benefit from the opportunities they offer. Therefore, privileged individuals are more likely to benefit from gentrification than the less fortunate [MB20].

2.5 Machine Learning Approaches

Artificial Intelligence, in short AI, refers to the development of computer systems and machines that can perform tasks that typically require human intelligence [SS18]. Machine learning is a form of AI. It focuses on the development of algorithms and models that allow computers to learn and improve from data without being directly programmed. These algorithms enable computer systems and machines to recognise meaningful relationships and patterns or make predictions. They can dynamically adjust their behaviour based on input and feedback [BGKL19] [JZH21]. Though not extensive, gentrification studies have been conducted with the support of machine learning, as it is still a relatively new field in urban studies research. These studies mainly used Python for predictive modelling, as it has extensive libraries, is easy to use, and has strong community support [Sud18].

To use a machine learning-based approach for a gentrification study, gentrification must be quantified. Quantifying gentrification and using machine learning helps identify the driving factors behind the process, thereby enabling predictions of future gentrification patterns in neighbourhoods [C+09] [RDSH19]. Calculating a gentrification score is a common practice to quantify gentrification [Eck11] [Gol16] [JPQQ21] [Tri17]. However, one may wonder how such a score can be composed.

[Fin22] reviewed quantitative methods used to define gentrification. He stated that there are many different indicators of gentrification. These can be classified according to gentrification theory, which explains that the indicators can be divided into supply and demand. On the demand side, demographic indicators describe the population moving into neighbourhoods. These demographic indicators include age, ethnicity, educational attainment, and income per capita. On the supply side, the composition and characteristics of the built environment are assessed. Typically, these indicators focus on the housing stock. Examples are property age, rent prices, or housing value [Fin22]. About 77% of the papers reviewed analyse changes in individual variables, sometimes multiple variables, over time to detect gentrification. These studies have utilised a combination of demand-side variables, supply-side variables, or both to identify gentrification. Such quantification can significantly benefit policies aimed at mitigating gentrification, as described in Section 2.2.

[RDSH19] were the pioneers in applying machine learning techniques for predicting gentrification. In order to measure gentrification, multiple indicators of gentrification were combined into a single measure of socioeconomic status using Principal Component Analysis. They used a random forest model to predict gentrification in London. Furthermore, feature importance showed the contribution of features to the model. Their predictions aimed to identify ways to improve or regenerate London without causing displacement or disconcerting social change. Since this study, numerous studies have employed machine learning in an urban studies context. [TNLP23] built a predictive machine-learning model of gentrification in Sydney. Gentrification was measured using the Socioeconomic Index for Advantage and Disadvantage, created by the Australian Bureau of Statistics. Instead of just utilising one machine learning model, this study compared several models, and ultimately, the gradient boosting machine outperformed the random forest models. The SEIFA score R2 was 0.938, and the model's accuracy was 0.747. Besides predicting gentrification, this study also focused on feature importance. The SHAPley package was used, which allows for a directional analysis of variable effects.

Studies in the field of urban studies that have utilised machine learning have demonstrated the vast potential of this technology. Researchers have noted the ability of machine learning to provide valuable insights into urban dynamics and the potential for it to contribute to urban planning and policy [AP19] [JPQQ21] [LDZ+21] [RDSH19] [TNLP23].

2.6 Research Gap

Overall, gentrification has been extensively researched, with in-depth analysis conducted on its causes, patterns, and impacts. In previous studies, the prediction of gentrification has been explored using various variables correlated to gentrification. However, the correlation between neighbourhood amenities and gentrification has received limited attention, particularly in the context of implementing machine learning. Therefore, a notable research gap exists in modelling gentrification using neighbourhood amenities and in the interpretation of this model by using Explainable Artificial Intelligence (XAI). It is important to emphasise that this study operates within a correlation setting rather than a prediction setting, given the absence of a clear direction of causality between gentrification and amenities. The geographical scope of this study is neighbourhoods in the Netherlands, as the results could benefit governing authorities in the decision-making regarding the housing shortage.

3 Theoretical Framework

The theoretical framework of this study serves as an anchor point, providing a foundation for the subsequent sections, particularly focusing on the data science aspects. By drawing upon established theories and concepts, it will inform about the approaches taken in this study and the rationale behind these choices.

3.1 Regression Models

Regression models are designed to handle continuous or real-valued target variables. In regression models, there are two main components: the features (or independent variables) and the target variable (or dependent variable). Models analyse how changes in the values of the features are associated with changes in the target variable [FKLM22]. Several different regression models were applied for this study.

3.1.1 Multiple Linear Regression

This technique uses a linear equation to model the relationship between the target variable and multiple features. It extends the simple linear regression, which can only deal with one feature, to handle cases with multiple features and a target variable [Ebe07].

3.1.2 Random Forest Regression

This algorithm combines multiple decision trees to model regression problems. It creates an ensemble of decision trees; each trained on a subset of the data. During training, each decision tree models the target variable based on a random subset of features. The final outcome is obtained by averaging the modelling outcomes of all individual trees in the forest. Combining the trees provides a more robust and accurate estimation of the target variable [Seg04].

3.1.3 Gradient Boosting Regression

This algorithm combines an ensemble of weak models to create a robust model. The model is trained in an iterative manner, where each new tree is built to correct the errors made by the previous trees. During training, the algorithm assigns higher weights to previously poorly modelled samples. The final outcome is obtained by summing the modelling outcomes of all individual trees, weighted by their contribution to the overall model. The gradient boosting regression model is known for its ability to handle complex relationships and effectively handle outliers in the data [CXZ+20] [YWX+20].

3.1.4 XGBoost Regression

The XGBoost regression model is an advanced gradient boosting algorithm. It stands for 'eXtreme Gradient Boosting' and is known for its speed and accuracy. The model builds an ensemble of weak models in a sequential manner [PNGA19]. To optimise a loss function, it uses a technique called gradient boosting (see 3.1.3), which iteratively improves the model's performance. Furthermore, XGBoost includes parallel processing and tree pruning techniques to enhance its efficiency [LLZ+18].

3.2 Feature Selection

Feature selection is the process of identifying and selecting a subset of relevant input variables from a more extensive set of available features. Feature selection aims to improve the performance of a model, reduce overfitting, and reduce computational costs [LCW+17].

An example of a feature selection method is SelectKBest from the scikit-learn framework. This feature selection technique ranks the features based on their statistical significance with respect to the target variable. The number of selected top features can be determined by the k. The scoring function is 'f_regression' as the model is a regression model. According to [DSY20] and [ZKB+21], the SelectKBest feature selection technique has been demonstrated to be effective in improving a model's performance.

3.3 Nested Cross-validation

Cross-validation is a technique used in machine learning to evaluate the performance of a model. Nested cross-validation is a commonly used cross-validation technique. This approach is used for model selection and hyperparameter tuning that addresses the issue of overfitting the training dataset. It integrates model hyperparameter tuning within the broader K-fold Cross Validation process [Bro20].

The nested cross-validation process works optimally as follows. At first, the model is split into a training and a testing set. The testing set is held out and, therefore, not seen by the model until the final model is made. Before the model is trained, the training dataset is shuffled and split into k subsets. Each subset, or fold, is used as the validation set once, while the remaining subsets are the training data. This process is repeated k times, with each subset serving as the validation data exactly once [MK09]. As a hyperparameter optimising procedure, grid search is chosen. Each training dataset is provided with a grid search, which finds a set of parameter values from the parameter grid that is the 'best'. However, the parameter grid is arbitrarily chosen and does not necessarily lead to the optimal hyperparameters. Hence, whenever a minimum or maximum value in the parameter grid was found to be the best, the parameter grid was refined iteratively.

3.4 Metrics for Evaluation

To evaluate the quality and performance of the regression models, multiple regression metrics were applied, resulting in a more comprehensive understanding of the model's performance on different aspects [SSSB+15].

3.4.1 R-squared

This measure represents the proportion of variance in the target variable explained by the features in a regression model [Kas19]. Equation 1 presents the formula of R-squared.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{SSR}{SST} \tag{1}$$

$$(\text{worst value} = -\infty; \text{ best value} = +1)$$

where:

- SSR: Sum of squares of the residual errors
- SST: Total sum of the errors

A higher R-squared value indicates a better fit of the model to the data. For example, an R-squared of 0.75 indicates that 75% of the variation in the output can be explained by the input variable.

3.4.2 Mean Absolute Error

The mean absolute error (MAE) is a metric that calculates the average absolute difference between the modelled values by the regression model and the actual values [WM05]. The formula of the MAE is presented in Equation 2.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$(\text{best value} = 0; \text{ worst value} = +\infty)$$

where:

- n: Number of observations
- y_i: Actual output value
- $\hat{y}_i$: Modelled output value

It is important to note that the MAE is measured on the same scale as the target variable being modelled. Consequently, there is no universal guideline for interpreting the ranges of MAE values. Overall, the closer to 0, the higher the accuracy of the model.

3.4.3 Root Mean Square Error

The Root Mean Square Error (RMSE) is a metric that measures the average difference between the modelled values and the actual values. The differences are squared to emphasise larger errors [CD14]. The formula of the RMSE is presented in Equation 3:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\text{(best value} = 0\text{; worst value} = +\infty\text{)}$$

where:

- n: Number of observations
- y_i: Actual output value
- $\hat{y}_i$: Modelled output value

A model with a lower RMSE means that the modelled outcomes are closer to the actual values. This implies higher accuracy, as the model's errors are minimised.

3.5 Interpretability

There is an increasing conflict between machine learning models' accuracy and interpretability. Therefore, there are methods to explain the outcome of models. In most studies, understanding and interpreting the best-working model and its outcome is essential [LL17].

In Section 2.5, there was a short introduction to an interpretability approach taken by [TNLP23]. The SHAP (Shapely Additive exPlanations) method is employed to analyse the black-box model results. Applying this method explains the outcome of models and includes Explainable Artificial Intelligence (XAI) into studies. XAI has undergone increased interest in recent years due to the realisation of the ethics, trust, and bias of Artificial Intelligence [DR20].

From the SHAP method, specifically the summary plot interpretation technique is used. This technique combines the feature importance with the feature effect, with the features being ordered based on their importance to the model. Furthermore, the summary plot has the capability to analyse feature effects directionally. Additionally, the relationship between amenities and gentrification is studied using scatter plots, which can be made with Matplotlib.

4 Data

This section will provide a comprehensive overview of the datasets utilised for this research and its source. Furthermore, an explanation of the variables used to quantify gentrification will be given, followed by a description of the amenities incorporated in this study.

4.1 Datasets

Data is necessary to conduct research about gentrification in neighbourhoods in the Netherlands. The data used for this research is retrieved from Statistics Netherlands. As mentioned in Section 2.1, this is a trustworthy data source utilised in numerous studies on gentrification in the Netherlands, establishing its reliability. The following datasets are used for this research:

- key figures districts and neighbourhoods; and
- proximity amenities per district/neighbourhood.

These datasets are published every year. For this research, data for three time intervals is used: 2013-2018, 2014-2019, and 2015-2020. There are several reasons why the first year taken into account is 2013 and the last year taken into account 2020. First of all, the key figures datasets were restructured in 2013, resulting in a new layout with information per year. Secondly, although the key figures datasets of 2021 and 2022 are available, there needs to be more information available to call these datasets informative. In a majority of cases, column values are empty. Lastly, the majority (65%) of gentrification studies have been conducted for approximately a decade [Fin22]. So, there are two datasets gathered per year, resulting in a total of twelve separate datasets.

4.1.1 Area codes

Since 1968, the Ministry of the Interior and Kingdom Relations has given, in cooperation with Statistics Netherlands, every municipality, district, and neighbourhood a unique name and a code. Both datasets used in this study contain such a unique code for every level of aggregation. This enables the datasets to be merged based on this area code, facilitating the integration process.

The area codes have a structured correlation: the first four digits of the municipality code match the corresponding district codes, and the first six digits of the district code align with the neighbourhood codes. Furthermore, in front of the digits there are abbreviations present. 'GM' is short for 'gemeente', which translates to 'municipality' in English. 'WK' stands for 'wijk', which means 'district' in English. 'BU' represents 'buurt', which corresponds to 'neighbourhood' in English. The standard format is presented in Table 1.

Table 1: Examples of standardised area codes.

Type of area	**Code format**	**Code example**
Municipality	GM-XXXX	GM1234
District	WK-XXXXXX	WK123456
Neighbourhood	BU-XXXXXXXX	BU12345678

As mentioned above, three different types of areas exist: municipalities, districts, and neighbourhoods. Municipalities are the highest regional level of these three. They are made up of at least one district. Similarly, districts are made up of at least one neighbourhood, which is the lowest regional level [Sta]. The corresponding municipality defines the layout of the districts and neighbourhoods. Every used dataset for this research has a column that contains codes of municipalities, districts, and neighbourhoods.

4.2 Gentrification Variable

The key figures datasets are created to gain insight into the demographic and socioeconomic characteristics of municipalities, districts, and neighbourhoods. Furthermore, it grants the possibility to compare these areas with each other based on area codes, as described in Section 4.1.1.

In order to create the key figures dataset, information is needed. Statistics Netherlands uses several different techniques for the collection of data. One part of the data is derived from other surveys or databases. Examples are the Personal Records Database, Social Statistical Database, Integral Income and Wealth Statistics, and the Registration of Addresses and Buildings. Not all the data is derived from other surveys. Whenever this is not the case, Statistics Netherlands either uses comprehensive observations or samples with a large sample size. The variables used from the key figures datasets are presented and described in Table 2.

Table 2: Description of the variables used from the key figures datasets.

Variable name	**Description**
Area code	This is the unique code of all municipalities, districts, and neighbourhoods
Neighbourhood layout	This number indicates changes in area code, or area layout from the previous year.
Housing value (x €1000)	The average housing value of the area based on the Real Estate Assessment (WOZ value).
Income per capita (x €1000)	The average income per person based on the total population in private households.
Degree of urbanity	Each area is assigned an urbanity class based on address density (addresses per square kilometre).

The variable *neighbourhood layout* can have three different values, each has a different meaning:

- a '1' means that the layout of the neighbourhood has not changed since the previous year;
- a '2' means that the layout of the neighbourhood has not changed since the previous year. Only the code of the neighbourhood changed; and
- a '3' means that the layout of the neighbourhood has changed since the previous year.

When the layout of the neighbourhood has changed since the previous year, it was at least a five-meter shift of the boundaries. However, the size of the change is not traced by Statistics Netherlands.

4.3 Amenities

In this study, information about amenities is used in order to model gentrification. Consequently, the amenities data derived from the amenities dataset are the features. The goal of the dataset is to gain insight into the travelling distance of residents to a range of amenities. Furthermore, the number of amenities within a specific neighbourhood radius is calculated. By considering both the travelling distance to amenities and the number of amenities within a certain radius, the accessibility and availability of amenities are captured. This research has been conducted annually in the Netherlands since 2006. Consequently, it is possible to compare municipalities, districts, and neighbourhoods.

To generate the amenities dataset, the acquisition of information is necessary. Information on residents, amenities and infrastructure should be gathered. A summary of the sources of information is given and can be found in Appendix A Table 8. Some amenities were not used in this study because little information was known in the datasets. Therefore, the sources of these amenities are left out.

The distance to the nearest amenity is measured by taking the distance from every address to that amenity. The calculation of this distance is based on paved roads per address. The calculated distance gets assigned to every resident of that specific address. Finally, the average distance per type of area to a particular amenity is calculated by taking the average of the calculated distance for all the residents in that area. This process is done for every type of amenity (see Appendix B Table 9). The statistical unit used is kilometres.

The average amount of certain amenities within a fixed distance is calculated by taking the average of the measured amounts of amenities per resident. This is done for every resident in that type of area. It can be interpreted as follows: the more amenities present of that specific amenity within the radius, the more choice residents of that area have for these amenities. In this study, the smallest possible radius is chosen for every amenity, as the chances of that amenity being in the area itself are the highest. According to [Sta23b] a valid address was determined for 99.7% of the population in the Netherlands. Additionally, the publication of results occurred only when the address of 90% or more of the residents in a specific area was determined. For every neighbourhood researched in this study, the travelling distance to amenities and the number of amenities within a certain radius were used. Appendix B Table 9 provides a comprehensive description of each of the amenities used.

5 Methods

This chapter outlines the research methodology utilised in this study to answer the research question based on the gathered data. It provides a clear framework for reproducing and validating the research. As stated in Section 2.5, Python is a strong language for modelling. Therefore, the programming language Python was used in this study, and libraries such as Pandas and NumPy were employed for data processing and analysis.

First of all, the process of engineering the dataset will be discussed. Secondly, how the gentrification score is composed is covered. Finally, we will elaborate on the modelling steps used in this study.

5.1 Data Preprocessing and Restructuring

5.1.1 Cleaning Data

Initially, for each year ranging from 2013 to 2015 and from 2018 and 2020, two separate datasets were collected: the key figures dataset and the amenities dataset. The datasets included data that was not relevant to this study. As a result, the datasets were cleaned.

First of all, unnecessary rows and columns were removed from the datasets. As mentioned in section 4.1, three different levels of aggregation are present in the datasets, namely municipalities, districts, and neighbourhoods. Data about districts and neighbourhoods are necessary to analyse gentrification in neighbourhoods in the Netherlands [Wil15]. However, municipality data is irrelevant for this study, so these rows were filtered out of the datasets. Furthermore, only some of the columns in key figures datasets were relevant to this study. Table 2 presents the columns kept for this research; the others were dropped. Section 5.2 will elaborate on this choice.

Secondly, the readability of the datasets was enhanced. Each column name was translated into English and renamed adequately. Additionally, to improve the legibility, a suffix of every year was added to all the columns in every dataset. For example, in the amenities dataset of 2015, the column *distance_supermarket* was transformed into *distance_supermarket_2015*.

Thirdly, data cleaning was performed on the data points. Whitespaces were present in both the key figures and amenities datasets. These whitespaces could occur before or after a value, and multiple whitespaces could exist within a single value. This could pose a potential issue at a later stage. A comparison failure could occur when datasets are merged on the area code (a string) [Dek20]. Therefore, all the whitespaces were stripped from the datasets. Whenever a data point was unavailable, Statistics Netherlands placed a '.' as the corresponding value. These values were replaced by a NaN (Not a Number) value using NumPy.

At last, all the datasets per year were merged using the area codes (see 4.1.1), generating a total of six datasets.

To provide a comprehensive overview of the cleaned datasets, an illustrative example dataset of 2015 is presented in Table 3. The corresponding suffix is not added in the column names in this example. This table serves to depict the current structure and contents of the datasets following the aforementioned data cleaning procedures. Note that dummy data is present in the table and that *amount amenity* and *distance amenity* can represent any amenity from all the amenities in the dataset (see Table 9).

Table 3: Illustrative and concise example of the cleaned 2015 dataset.

neighbourhood	**layout**	**urbanity**	**housing value**	**income capita**	**amount amenity**	**distance amenity**
WK000500	1	3	275.8	22.5	2.7	0.5
BU00050000	1	2	280.4	21.4	0.4	3.6
BU00050001	1	1	295.5	23.6	1.3	1.1
BU00050002	1	4	266.5	24.7	1.2	1.9
BU00050003	1	5	243.3	20.3	0.9	2.4

5.1.2 Filtering Data

In short, three main filters were implemented: eligibility criteria, neighbourhood boundaries, and urbanity. Neighbourhoods not meeting the eligibility criteria were filtered out, neighbourhoods with altered boundaries were filtered out, and non-urban areas were excluded for urban gentrification focus. This section elaborates on the implementation of these filters and the reasons behind their usage.

First of all, the eligibility criteria were implemented. As mentioned in Section 2.3, not all neighbourhoods are eligible for gentrification. A foundational method to identify neighbourhoods eligible for gentrification was proposed by [GP86]. This study uses these conditions as a reference point but not simply copied without adjustment. First, although the percentage of college-educated could be a helpful indicator, it is not available in the dataset provided by Statistics Netherlands [BFR16]. Furthermore, only the mean of the housing value and income per capita is available. The last condition about race is not considered because racial or ethnic population shifts are not a necessary component of gentrification [SKK22]. Moreover, the study of [GP86] is from 1986, and the stereotype that white people are gentrifying black neighbourhoods is not thoroughly researched in today's society. According to [Lee16], there is a need for a more extensive and comprehensive exploration concerning the interplay between race and class within gentrification processes.

The conditions that Dutch neighbourhoods will have to meet in 2013 (first year of study) to be eligible for gentrification are:

- the mean housing value is less than the corresponding district-wide mean; and
- the mean income per capita is less than the corresponding district-wide mean.

It should be noted that neither the *mean housing value* column nor the *mean income per capita* column can contain NaN values. The presence of NaN values in these columns would introduce uncertainty regarding the eligibility of a neighbourhood. Therefore, any rows that contain a NaN

value in either of these columns for a particular year are excluded from the analysis.

Secondly, the boundaries of neighbourhood during the analysed period should be considered. As mentioned in Section 2.1, the boundaries of a neighbourhood can have a significant impact on how gentrification is perceived and experienced. According to Statistics Netherlands, the corresponding municipality defines the neighbourhood layout. The boundaries of a neighbourhood can be altered each year. For a detailed example of changed neighbourhood layouts and possible reasons for this, see Appendix C Figure 6. Section 4.2, explains the variable *neighbourhood layout* and its possible values. Neighbourhoods having a *2* or *3* in the column *neighbourhood layout* are discarded from the dataset. Whenever the layout of a neighbourhood was altered, it is no longer possible to accurately analyse this neighbourhood.

Thirdly, the filter to ensure only urban neighbourhoods were analysed in this study was implemented. Section 2.1 states that rural gentrification differs significantly from urban gentrification. The causes, characteristics, and implications of gentrification may vary in these different contexts. In this study, urban gentrification is analysed. Therefore, rural neighbourhoods are not taken into account. The degree of urbanity variable mentioned in Table 2 represents the number of addresses per square kilometre and is measured on an ordinal scale. According to Statistics Netherlands, a non-urban neighbourhood has less than 500 addresses per square kilometre. To analyse gentrification in an urban context, the neighbourhoods labelled non-urban (a *5*) were excluded from the dataset.

After these filtering steps, the datasets of every year are merged into one dataset. This dataset containes the cleaned and filtered key figures datasets and the amenities datasets of every year. Furthermore, a gentrification score was composed and added to the dataset. The way in which this was done is described in Section 5.2. At last, the mean housing value columns and the mean income per capita columns were discarded from the dataset.

To provide a comprehensive overview of the cleaned and filtered dataset, we present an illustrative example dataset with dummy data in Table 4. This table serves to depict the current structure and contents of the datasets following the aforementioned data filtering procedures. Note that *amount amenity 'year'* and *distance amenity 'year'* can represent any amenity present in the dataset (see Table 9) from 2013 to 2015 and from 2018 to 2020. Furthermore, the *GS 'year'* refers to the gentrification score calculated for the years 2013 to 2015 and for the years 2018 to 2020.

Table 4: Illustrative and concise example of the cleaned and filtered dataset.

neighbourhood	**amount amenity 'year'**	**distance amenity 'year'**	**GS 'year'**
BU00700004	1.2	3.1	1.1
BU00802103	2.3	0.7	1.3
BU02440001	0.4	6.5	0.99
BU07721614	0.8	2.2	1.05

5.1.3 Reorganising Data

To answer the research question, a machine learning model must be built. This should model the correlation between the gentrification score and the amenities. In order to do so, the dataset has to be restructured. Each row in the restructured dataset corresponds to a specific neighbourhood in a particular year, capturing the gentrification score and the amenities. Moreover, the amenity values for the years 2018, 2019, and 2020 represent the differences between these years' values and the corresponding base years' values (2013, 2014, and 2015). This new structure is suitable for training machine learning models to model the correlation between neighbourhood amenities and the gentrification in these urban neighbourhoods.

After reorganising the dataset, there was one last cleaning step to be made. As mentioned in Section 5.1.1, the dataset contains NaN values. Whenever gentrification scores contained NaN values, these rows were dropped. For NaN values present as amenity values, a different approach was taken. Initially, it was found that 50 rows in the entire dataset consisted of just NaN values, representing entire neighbourhoods with missing information for the period of 2014 until 2020. Consequently, these 50 rows were removed from the dataset. After this process, some NaN values remained in the feature values. To address these missing values, linear interpolation and extrapolation methods were applied. This technique helped fill in the missing data points by estimating values based on the neighbouring observations [Hua21] [NYRAB14]. In some cases, entire amenities data was missing for particular neighbourhoods throughout the specified time period. Neither interpolation, nor extrapolation was an option. Therefore, the missing values were filled with zeros to indicate no change since 2013, 2014 or 2015.

Restructuring the dataset and appropriately handling the NaN values ensured a completer and more reliable dataset. Consequently, the dataset is suitable for modelling effectively. To provide a comprehensive overview of the reorganised dataset, an illustrative example dataset with dummy data is presented in Table 5. This table serves to depict the current structure and contents of the datasets following the aforementioned data reorganising procedures. Note that *amount amenity* and *distance amenity* can represent any amenity present in the dataset (see Table 9).

Table 5: Illustrative and concise example of the cleaned, filtered and reorganised dataset.

neighbourhood	**year**	**GS**	**amount amenity**	**distance amenity**
BU00740101	2018	1.1	1.2	4.0
BU00740101	2019	1.3	1.5	2.5
BU00740101	2020	1.25	2.0	3
BU01060401	2018	0.9	0.2	1.5
BU01060401	2019	1.01	0.8	1.0
BU01060401	2020	1.15	1.1	1.0

5.2 Gentrification Score

There is no clear way to measure gentrification, as it is a complex and multifaceted process that can manifest differently depending on the context [MSI+19]. As mentioned in Section 2.5, about 77% of the papers analyse changes in individual variables, sometimes multiple variables, over time to detect gentrification. These studies have utilised a combination of demand-side variables, supply-side variables, or both to identify gentrification. Therefore, in this study, both demand- and supply-side variables were chosen to measure gentrification in neighbourhoods. The chosen indicators are:

- Housing value; and
- Income per capita.

As pointed out in Section 2.5, calculating a gentrification score is a method to quantify gentrification. In this study, to measure the degree of gentrification, the indicators mean housing value and mean income per capita are combined into a gentrification score [JPQQ21]. To determine the extent of gentrification, a focus is placed on fixed time intervals of 5 years. Specifically, gentrification scores are computed for the periods 2013-2018, 2014-2019, and 2015-2020. To calculate the gentrification score a base year for each 5-year interval is established. For the first interval (2013-2018), the base year is 2013, and assigned a gentrification score of 1.0. Similarly, for the second interval (2014-2019), the base year is 2014, and for the third interval (2015-2020), the base year is 2015. For each interval, the mean housing value of the given year is divided by the mean housing value of the corresponding base year. Likewise, the income per capita of the given year is divided by the income per capita of the corresponding base year. The resulting values from both indicators are then summed together and divided by two, as two indicators are being used [Tri17].

Inflation is the increase in the general price level of goods and services over time, which means that the purchasing power of money decreases. To make the gentrification index robust to inflation, the indicators should be adjusted for inflation [JCLK22] [Wil15].

The gentrification score, as described above, results in Equation 4:

$$Gentrification\ Score\ (y) = \frac{\left(\frac{HV_y}{HV_b} + \frac{IC_y}{IC_b}\right)}{2 \cdot CPI_g} \tag{4}$$

where:

- HV_y: Mean housing value in the given year
- HV_b: Mean housing value in the base year
- CPI_y: Consumer Price Index, calculated for each year relative to the base year
- IC_y: Mean income per capita in the given year
- IC_b: Mean income per capita in the base year

The gentrification score can be interpreted as follows. If the index value is above 1.0, the neighbourhood is gentrifying. Conversely, if the index value is below 1.0, the neighbourhood is not gentrifying [JPQQ21] [Tri17].

5.2.1 Validity of Gentrification Score

The gentrification score is the target variable of this study and calculated as shown in Equation 4. To determine the validity and correctness of the formulated gentrification score, neighbourhoods with the highest scores were visually examined using Google Street View. According to [ISZ19], gentrification results in visible neighbourhood changes. In Figure 1, three neighbourhoods with a high gentrification score were captured in Google Street View.

Figure 1: From the top down: BU03440243 (Queeckhovenplein in Utrecht), BU03070401 (Eemplein – Nieuwe Stad in Amersfoort), and BU05032805 (TU-Campus in Delft).

The neighbourhood called 'Queeckhovenplein' in Utrecht has undergone urban renewal since 2009. In this neighbourhood, old buildings and amenities were renovated. The play area was relocated and renewed, and several flats and houses have been completely renovated. Nowadays, it is a trendy neighbourhood, and it can be assumed that gentrification has occurred. In 2011, the neighbourhood 'Eemplein – Nieuwe stand in Amersfoort was mostly in construction. Over the following years, amenities and houses were built in this neighbourhood, spurring gentrification. According to the municipality of Amersfoort, this neighbourhood can be described as a creative hotspot where collaboration, sustainability, and innovation take centre stage.

The last reviewed neighbourhood is 'TU-Campus' in Delft. This neighbourhood is located on the campus of the Delft University of Technology. Subsequently, this is a favourable place for students to live. In 2013, there was unused land in the middle of the neighbourhood. After a couple of years, student housing flats were built, surrounded by amenities such as cafes, a botanic garden, a bar, and a parking garage.

5.3 Machine Learning Model

Section 3 provided a foundation of the approaches taken to build a machine learning model, which is able to model gentrification in Dutch neighbourhoods using neighbourhood amenities.

Firstly, the feature selection method SelectKBest from the scikit-learn framework was implemented, where the top 20 amenities (features) were kept, resulting in $k = 20$. Secondly, the regression models were initiated by using the scikit-learn framework. Subsequently, these models (as described in Section 3.3) were trained, and the hyperparameters were tuned. To accomplish this, nested cross-validation was employed, integrating model hyperparameter tuning within a broader K-fold Cross Validation process. Prior to training the models, the training dataset was shuffled and split into ten subsets: therefore, $k = 10$. Once the models were trained on the training data, correlations were modelled on the held-out test data. The aim was to analyse the relationships between the gentrification score and the amenities. The quality and performance of the models were evaluated by the regression metrics outlined in Section 3.4. Lastly, the top-performing model was interpreted by a SHAP summary plot.

6 Results

This section presents the results of this study that aim to address the research question: *'What is the role of neighbourhood amenities on gentrification in the Netherlands?'*. First of all, the results of the models are showcased. The results of these models are categorised into two groups: default models and models that underwent hyperparameter tuning. Secondly, the model is interpreted by using a SHAPley summary plot. Lastly, the relationship between several amenities with a high feature importance and gentrification is analysed.

6.1 Model Comparison

In order to assess the performance of the different models, a comprehensive comparison was conducted using the evaluation metrics described in Section 3.4. Table 6 shows the performance of the four different models before the hyperparameters were tuned.

Table 6: Performance of the untuned regression models.

Model	**R-squared**	**MAE**	**RMSE**
Linear regression	0.1168	0.0601	0.0862
Random forest	0.4440	0.0499	0.0684
Gradient boosted machine	0.3827	0.0536	0.0721
XGBoost	0.4347	0.0498	0.0690

Section 3.3 explains the process of tuning the hyperparameters of the different machine learning models. The performance of the models after hyperparameter tuning is presented in Table 7. Note that the linear regression model is not shown in this table, as it is impossible to tune this model's hyperparameters.

Table 7: Performance of the tuned regression models.

Model	**R-squared**	**MAE**	**RMSE**
Random forest	0.4684	0.0497	0.0675
Gradient boosted machine	0.4599	0.0489	0.0674
XGBoost	0.4825	0.0479	0.0660

When comparing the untuned models, the linear regression model performed the worst on every evaluation metric. The random forest, gradient-boosted machine, and XGBoost models performed relatively better, with R-squared values ranging from 0.3827 to 0.4440. The MAE and RMSE values did not vary widely. However, after tuning the hyperparameters of the models, improvements were observed across all metrics. The R-squared values now range from 0.4599 to 0.4825. Moreover, the MAE and RMSE showcased progress. Notability, the XGBoost regression model achieved the

highest R-squared value of 0.4825. This indicates mediocre modelling accuracy. The results suggest that the XGBoost regression model is the top-performing model in our tests for modelling the target variable.

6.2 Interpreting the Model

As stated in Section 6.1, the XGBoost regression model is the best-performing model in our tests. Therefore, the decision has been made to interpret this model thoroughly. Figure 2 presents a SHAPley summary plot of this model, containing the 20 used features. It visually represents the relationship between these features and the gentrification score. Each dot in the summary plot is a SHAP value for a feature and for an instance. The feature importance determines the position on the y-axis, while the position on the x-axis is determined by the SHAP value (impact on model output). Additionally, the colour of each point corresponds to the value of the feature, ranging from low to high.

Furthermore, the summary plot has the capability to analyse feature effects directionally. For example, an increase in the number of cafes in a neighbourhood is associated with a lower gentrification score. This suggests that the presence of cafes may hinder the gentrification process. On the other hand, a decrease in the distance to a train station is linked to a higher gentrification score, indicating that improved access to this amenity may contribute to the desirability of a neighbourhood.

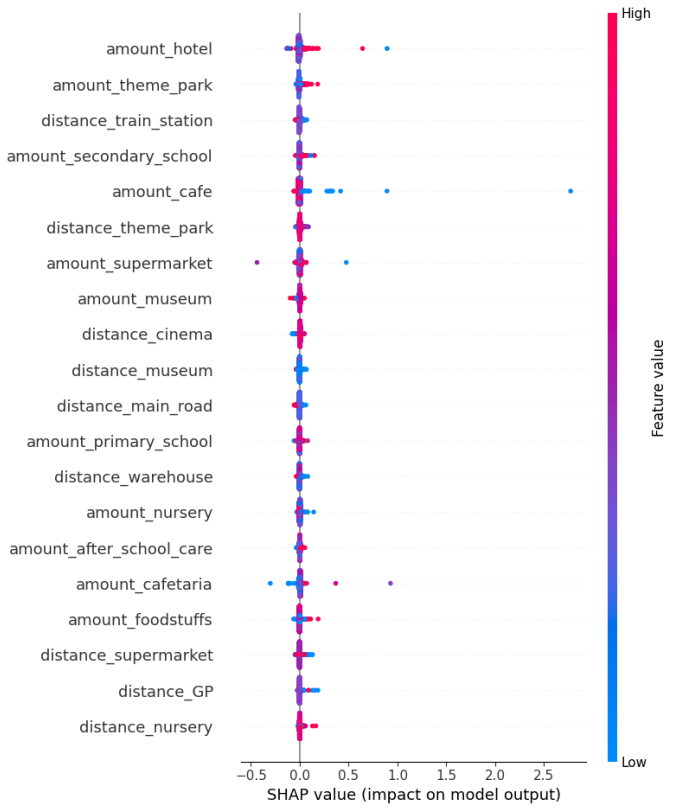

Figure 2: SHAPley summary plot containing all the features

Section 6 demonstrated that the linear regression model did not perform well on the data. This suggests the absence of individual variables showing a trend where higher or lower amounts of this amenity, or proximity to this amenity, correspond to higher or lower gentrification scores. As shown in Figure 2, the three most important features are *amount_hotel*, *amount_theme_park*, and *distance_train_station*. According to feature importance, these amenities correlate the strongest with the gentrification score. To further investigate the relationship and better understand the trend between these amenities and the gentrification score, scatter plots were made.

The amenity with the highest feature importance is the number of hotels in a neighbourhood. Figure 3 displays a scatter plot of *amount_hotel* against the gentrification score, indicating that as the number of hotels increases, there is a tendency for gentrification also to increase. However, the relationship appears to be somewhat scattered, and the trend is not convincingly clear.

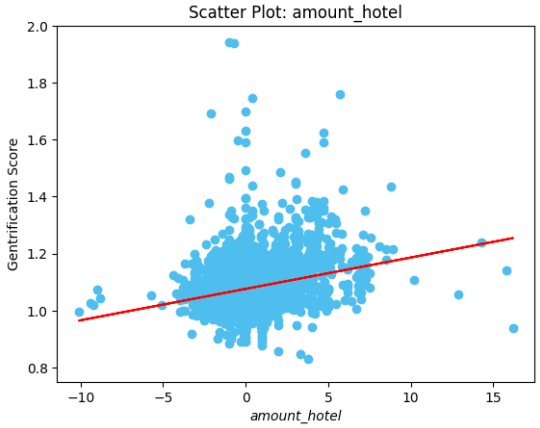

Figure 3: Scatter plot of the amount of hotels against the gentrification score.

Figure 4 presents a scatter plot of *amount_theme_park* against the gentrification score. This amenity has the second highest feature importance. The plot reveals a less scattered distribution compared to Figure 3, with a noticeable linear trend indicating that as the number of theme parks increases, gentrification tends to rise accordingly.

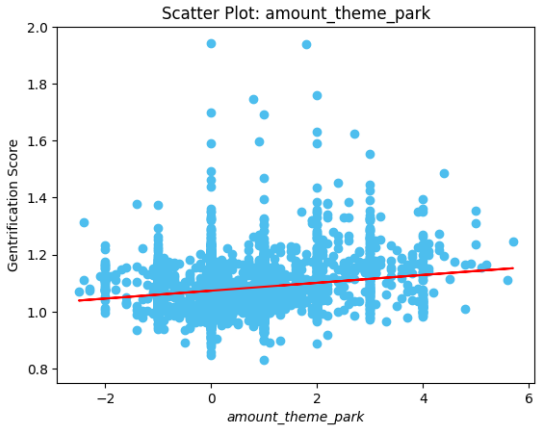

Figure 4: Scatter plot of the amount of theme parks against the gentrification score.

The amenity ranked third highest in terms of feature importance is the distance to a train station. Figure 5 displays a scatter plot of *distance_train_station* against the gentrification score, revealing two distinct clusters. One cluster shows a slight decrease or increase in distance to a train station, while the other centres around a 10 kilometre increase. Although a linear trend is not apparent, drawing a trend line shows that a neighbourhood closer to a train station correlates with a higher gentrification score.

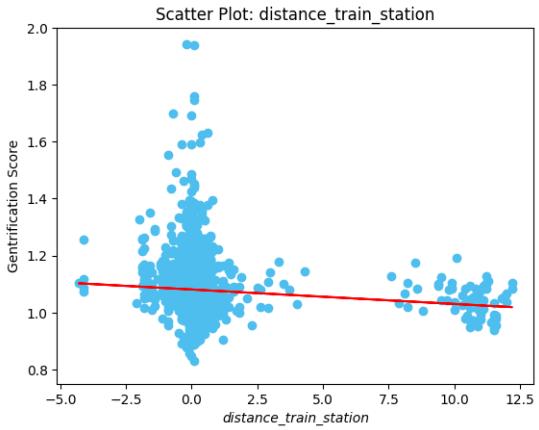

Figure 5: Scatter plot of the distance to a train station against the gentrification score.

In summary, the scatter plots of the amenities *amount_hotels*, *amount_theme_parks*, and *distance_train_station* highlighted varying relations to the gentrification score. While some relationships were more pronounced and linear, others displayed scattered patterns, indicating a more complex interplay between amenities and gentrification. Notably, the trend lines in the scatter plots align with the correlation direction depicted in Figure 2.

7 Discussion

7.1 Discussion

Gentrification is a complex socioeconomic process with significant implications. Urban planners, researchers, and policymakers are continuously researching gentrification and its implications. These studies can be divided into qualitative research, quantitative research, spatial mapping, community engagement, and historical analysis [LDZ+21]. Recently, machine learning techniques have been adapted to analyse the aspects of gentrification and provide new insight into its causes, patterns, and impacts.

This study explores the promising potential of applying machine learning in gentrification studies. Modelling based on neighbourhood amenities can provide valuable insights for decision-making processes related to urban development and housing policies. Additionally, using Explainable Artificial Intelligence (XAI) techniques enhances the understanding of the effects of amenities, which are seen as one of the driving factors of gentrification [HL16].

The study results indicate that the XGBoost regression model, after hyperparameter tuning, achieved the highest performance among the tested regression models. This is reflected in an R-squared value of 0.4825, which suggests that approximately 48.25% of the variance in the gentrification score is explained by the features used in the model. Furthermore, the mean absolute error (MAE) of 0.0479 indicates that, on average, the modelled output value made by the model differs by 0.0479 units from the actual values of the gentrification score. Lastly, the root mean square error (RMSE) of 0.0660 indicates the average magnitude of errors in the original scale of the gentrification score.

According to [CWJ21], an MAE and an RMSE with a value of 0 implies a perfect fit. However, there is a caveat: the target variable. Considering that the gentrification score has a relatively small range, with values mostly centred around 1 and ranging from 0.75 to 1.75, the MAE and RMSE values should be interpreted cautiously. Drawing definitive conclusions solely based on these metrics is not appropriate. Nonetheless, the results of the best-performing predictive model can still be considered promising within the context of the specific scale and domain of the target variable.

The summary plot generated using the SHAP method allows for further interpretation of the results. Notably, the presence of amenities has an impact on gentrification. On the one hand, certain amenities like cafes and nurseries may hinder the gentrification process, while on the other hand, amenities such as hotels and secondary schools may drive the gentrification process. Additionally, there are other amenities, such as theme parks, cinemas and after-school care, that surprisingly hold significant importance in the model. These amenities were interestingly enough barely discussed in the literature regarding the contribution of amenities to gentrification.

In this study, a machine learning model was built to investigate the correlations between amenities and gentrification. Although the model may not have achieved the same level of performance as models in other studies that included many diverse features, the tuned XGBoost regression model can still serve as a tool for modelling gentrification in the Netherlands based on neighbourhood

amenities. Furthermore, building upon the findings of XAI, conducting further analysis could shed light on the interaction between the presence of neighbourhood amenities and gentrification. Overall, we believe, together with other researchers, that machine learning can provide valuable insights into urban dynamics and its potential to contribute to urban planning and policy.

7.2 Limitations

This study aimed to model gentrification in Dutch neighbourhoods using amenities. To comprehensively communicate the results of this study, it is essential to acknowledge the study's limitations. Moreover, acknowledging the limitations may provide insights for further research.

In order to conduct this study, open-source data from Statistics Netherlands was used. Although this is a trustable data source, there were certain constraints regarding the utilised data. As mentioned in Section 5.1, some data was missing, accounting for approximately 10% of the entire dataset. This was tackled by using linear interpolation and extrapolation. However, this method is not entirely reliable. Besides missing some data, it could be argued that taking more years into account could lead to trustworthy results.

In Section 2.2, it is stated that amenities are recognised as one of the critical factors driving gentrification in neighbourhoods. In this study, amenities were used to model gentrification in Dutch neighbourhoods. Appendix B Table 9 presents the different sorts of amenities used in this study. Although there are 19 unique amenities used, there is always potential for further expansion. Including additional amenities, such as green spaces and sports facilities, could enhance the model's performance and its interpretation.

The target variable of this study was the gentrification score. This was composed of the mean housing value and the mean income per capita, see Equation 4. Instead of using the means of these indicators, the median could be a better approach. The median housing value and the median income per capita would be a more robust measure for centrality, as they are less affected by outliers. Additionally, the results of the gentrification score were analysed using Google Street View. Even though the neighbourhoods shown in Section 5.2.1 showed significant signs of gentrification, only twenty were reviewed. Therefore, it cannot be concluded that the gentrification score is a reliable indicator of gentrification.

8 Conclusion

8.1 Conclusion

This thesis addressed the research question of what the role is of neighbourhood amenities on gentrification in the Netherlands. The study used data from Statistics Netherlands, a trustworthy autonomous administrative authority. The research focused on three time intervals, namely 2013-2018, 2014-2019, and 2015-2020.

In order to get insights into the role of neighbourhood amenities on gentrification, the aim was to model gentrification and investigate the correlations between gentrification and amenities. Several regression models were tested to answer the research question, including Multiple Linear Regression, Random Forest Regression, Gradient Boosting Regression, and XGBoost Regression. Performance evaluation metrics such as R-squared, mean absolute error, and root mean square error were implemented. The best-performing model was further evaluated using the SHAP method. Additionally, the study conducted a detailed investigation of the three amenities (*amount_hotels*, *amount_theme_park*, and *distance_train_station*) with the highest feature importance.

Based on the findings from this study, we cautiously answer the research question: *What is the role of neighbourhood amenities on gentrification in the Netherlands?* Our analysis revealed that the XGBoost model, after hyperparameter tuning, was the top-performing model in terms of accuracy. It achieved mediocre results with an R-squared of 0.4825, indicating a moderate modelling ability. Additionally, the model demonstrated an mean absolute error of 0.0479, and an root mean squared error of 0.0660, further supporting its precision in modelling gentrification. Thus, our findings suggest that neighbourhood amenities correlate with gentrification and can be indicators for modelling gentrification in the Netherlands. Figure 2 provides a comprehensive summary of the amenities' importance to the model, their relationship with the gentrification score, and the direction of the amenities' effects studied in this research. However, the role of the studied amenities on gentrification was not entirely linear, indicating a complex interplay between amenities and gentrification. Therefore, the relationships between amenities and gentrification are nuanced and may vary based on other factors.

In conclusion, this thesis contributes to understanding gentrification in the Netherlands by examining the modelling power of neighbourhood amenities and assessing the relationship between neighbourhood amenities and gentrification. The results highlight the potential impact of specific amenities on the gentrification process, and the insights provide valuable information for understanding the drivers of the model's outcomes. Furthermore, it can inform decision-making processes related to urban development and planning.

8.2 Further Research

For future work, there are several directions and ways to explore and build upon the results of this research. This study aimed to model gentrification in Dutch neighbourhoods using two key factors: the number of amenities present and the distance to those amenities. Despite the new insights, unexplored topics remain that offer promising opportunities for future research.

In future research, an alternative approach to studying gentrification could involve predictive modelling instead of traditional modelling based on amenities. By using predictive models, research can try to forecast and estimate gentrification trends and patterns in neighbourhoods. This can provide valuable insights for urban planning and policy-making.

In Section 7.2, the limitations of this study have been discussed. Data from Statistics Netherlands is used for the model's features and the target variable. In the future, indicating gentrification by using a gentrification score could be further enhanced. Including more indicators of gentrification when composing a gentrification score could improve its reliability. Moreover, expanding the range of amenities in the research offers potential for further investigation. This can be achieved in two different ways:

- Adding different amenities, such as green spaces, sports facilities, financial institutions, et cetera.
- Further specifying amenities, for example, splitting the 'restaurant' category into more specific types, such as vegan restaurants, Michelin-starred restaurants, fast-food restaurants, et cetera.

After conducting such a study, which builds upon this study, urban planners and governing authorities gain valuable insights into the different types of amenities and their impacts on gentrification. Armed with this knowledge, they can make informed decisions and formulate effective strategies to improve the quality of life in neighbourhoods.

Another potential future research is the visualisation of gentrification patterns on the map of the Netherlands. By geographically representing the neighbourhoods experiencing gentrification over time, researchers can identify hotspots, observe gentrification patterns, and comprehend gentrification's spatial impacts. Creating visualisations on a map for an extended period would be convenient for capturing the dynamic of gentrification.

References

[AP19] Yesenia Alejandro and Leon Palafox. Gentrification prediction using machine learning. In *Advances in Soft Computing: 18th Mexican International Conference on Artificial Intelligence, MICAI 2019, Xalapa, Mexico, October 27–November 2, 2019, Proceedings 18*, pages 187–199. Springer, 2019.

[Atk04] Rowland Atkinson. The evidence on the impact of gentrification: new lessons for the urban renaissance? *European journal of housing policy*, 4(1):107–131, 2004.

[Bar16] Michael Barton. An exploration of the importance of the strategy used to identify gentrification. *Urban Studies*, 53(1):92–111, 2016.

[BBVSK20] M Bockarjova, WJW Botzen, MH Van Schie, and MJ Koetse. Property price effects of green interventions in cities: A meta-analysis and implications for gentrification. *Environmental science & policy*, 112:293–304, 2020.

[Bec20] Brenden Beck. Policing gentrification: Stops and low–level arrests during demographic change and real estate reinvestment. *City & Community*, 19(1):245–272, 2020.

[BFR16] Jörg Blasius, Jürgen Friedrichs, and Heiko Rühl. Pioneers and gentrifiers in the process of gentrification. *International Journal of Housing Policy*, 16(1):50–69, 2016.

[BGKL19] Qifang Bi, Katherine E Goodman, Joshua Kaminsky, and Justin Lessler. What is machine learning? a primer for the epidemiologist. *American journal of epidemiology*, 188(12):2222–2239, 2019.

[Bro20] Jason Brownlee. Nested cross-validation for machine learning with python. *Machine Learning Mastery. https://machinelearningmastery. com/nested-cross-validation-for-machine-learning-with-python*, 2020.

[BS17] Japonica Brown-Saracino. Explicating divided approaches to gentrification and growing income inequality. *Annual review of sociology*, 43:515–539, 2017.

[Byr02] J Peter Byrne. Two cheers for gentrification. *Howard LJ*, 46:405, 2002.

[C+09] Karen Chapple et al. Mapping susceptibility to gentrification: The early warning toolkit. *Berkeley, CA: Center for Community Innovation*, 43, 2009.

[CD14] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific model development discussions*, 7(1):1525–1534, 2014.

[CH17] Victor Couture and Jessie Handbury. Urban revival in america, 2000 to 2010. Technical report, National Bureau of Economic Research, 2017.

[CL95] Juliet Carpenter and Loretta Lees. Gentrification in new york, london and paris: an international comparison. *International Journal of Urban and regional research*, 19:286–286, 1995.

[CWJ21] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

[CXZ+20] Jianchao Cai, Kai Xu, Yanhui Zhu, Fang Hu, and Liuhuan Li. Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. *Applied energy*, 262:114566, 2020.

[Dek20] Vaclav Dekanovsky. Dealing with extra white spaces while reading csv in pandas, Jul 2020.

[DR20] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[DSY20] T Desyani, A Saifudin, and Y Yulianti. Feature selection based on naive bayes for caesarean section prediction. In *IOP Conference Series: Materials Science and Engineering*, volume 879, page 012091. IOP Publishing, 2020.

[Dua01] Andres Duany. Three cheers for" gentrification". *The American Enterprise*, 12(3):36–36, 2001.

[Ebe07] Lynn E Eberly. Multiple linear regression. *Topics in Biostatistics*, pages 165–187, 2007.

[Eck11] Adam Eckerd. Cleaning up without clearing out? a spatial assessment of environmental gentrification. *Urban Affairs Review*, 47(1):31–59, 2011.

[EHR19] Ingrid Gould Ellen, Keren Mertens Horn, and Davin Reed. Has falling crime invited gentrification? *Journal of Housing Economics*, 46:101636, 2019.

[EKK+13] Matt Egan, Srinivasa Vittal Katikireddi, Ade Kearns, Carol Tannahill, Martins Kalacs, and Lyndal Bond. Health effects of neighborhood demolition and housing improvement: a prospective controlled study of 2 natural experiments in urban renewal. *American Journal of Public Health*, 103(6):e47–e53, 2013.

[Fin22] Nicholas Finio. Measurement and definition of gentrification in urban studies and planning. *Journal of Planning Literature*, 37(2):249–264, 2022.

[FKLM22] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D Marx. Regression models. In *Regression: Models, methods and applications*, pages 23–84. Springer, 2022.

[GB16] Joseph Gibbons and Michael S Barton. The association of minority self-rated health with black versus white gentrification. *Journal of urban health*, 93:909–922, 2016.

[Gol16] Shaun A Golding. Gentrification and segregated wealth in rural america: Home value sorting in destination counties. *Population Research and Policy Review*, 35:127–146, 2016.

[GP86] George Galster and Stephen Peacock. Urban gentrification: Evaluating alternative indicators. *Social Indicators Research*, 18:321–337, 1986.

[HARH22] Jason Hawkins, Usman Ahmed, Matthew Roorda, and Khandker Nurul Habib. Measuring the process of urban gentrification: A composite measure of the gentrification process in toronto. *Cities*, 126:103708, 2022.

[HHG03] Tialda Haartsen, Paulus PP Huigen, and Peter Groote. Rural areas in the netherlands. *Tijdschrift voor economische en sociale geografie*, 94(1):129–136, 2003.

[HL16] Jackelyn Hwang and Jeffrey Lin. What have we learned about the causes of recent gentrification? *Cityscape*, 18(3):9–26, 2016.

[HM21] Cody Hochstenbach and Sako Musterd. A regional geography of gentrification, displacement, and the suburbanisation of poverty: Towards an extended research agenda. *Area*, 53(3):481–491, 2021.

[Hua21] Guilin Huang. Missing data filling method based on linear interpolation and lightgbm. In *Journal of Physics: Conference Series*, volume 1754, page 012187. IOP Publishing, 2021.

[Hyr12] Derek S Hyra. Conceptualizing the new urban renewal: Comparing the past to the present. *Urban Affairs Review*, 48(4):498–527, 2012.

[Hyr16] Derek Hyra. Commentary: Causes and consequences of gentrification and the future of equitable development policy. *Cityscape*, 18(3):169–178, 2016.

[ISZ19] Lazar Ilic, Michael Sawada, and Amaury Zarzelli. Deep mapping gentrification in a large canadian city using deep learning and google street view. *PloS one*, 14(3):e0212814, 2019.

[JCLK22] Glen D Johnson, Melissa Checker, Scott Larson, and Hanish Kodali. A small area index of gentrification, applied to new york city. *International Journal of Geographical Information Science*, 36(1):137–157, 2022.

[JPQQ21] Shomik Jain, Davide Proserpio, Giovanni Quattrone, and Daniele Quercia. Nowcasting gentrification using airbnb data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[JZH21] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[Kas19] Eiiti Kasuya. On the use of r and r squared in correlation and regression. Technical report, Wiley Online Library, 2019.

[LCW+17] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[LDZ+21] Li Lin, Liping Di, Chen Zhang, Liying Guo, and Yahui Di. Remote sensing of urban poverty and gentrification. *Remote Sensing*, 13(20):4022, 2021.

[Lee16] Loretta Lees. Gentrification, race, and ethnicity: towards a global research agenda? *City & Community*, 15(3):208–214, 2016.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[LLZ+18] Yahui Liu, Huan Luo, Bing Zhao, Xiaoyong Zhao, and Zongda Han. Short-term power load forecasting based on clustering and xgboost method. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 536–539. IEEE, 2018.

[LSW13] Loretta Lees, Tom Slater, and Elvin Wyly. *Gentrification*. Routledge, 2013.

[MB20] Lauren E Mullenbach and Birgitta L Baker. Environmental justice, gentrification, and leisure: A systematic review and opportunities for the future. *Leisure Sciences*, 42(5-6):430–447, 2020.

[Min22] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. Grootschalige woningbouwgebieden. https://www.volkshuisvestingnederland.nl/onderwerpen/grootschalige-woningbouwgebieden, Nov 2022.

[MK09] Zafar Mahmood and Salahuddin Khan. On the use of k-fold cross-validation to choose cutoff values and assess the performance of predictive models in stepwise regression. *The International Journal of Biostatistics*, 5(1), 2009.

[MMMG18] Roshanak Mehdipanah, Giulia Marra, Giulia Melis, and Elena Gelormino. Urban renewal, gentrification and health equity: a realist perspective. *The European Journal of Public Health*, 28(2):243–248, 2018.

[MSI+19] Mahasin S Mujahid, Elizabeth Kelley Sohn, Jacob Izenberg, Xing Gao, Melody E Tulier, Matthew M Lee, and Irene H Yen. Gentrification and displacement in the san francisco bay area: a comparison of measurement approaches. *International journal of environmental research and public health*, 16(12):2246, 2019.

[Net23] Netherlands Chamber of Commerce. Permits for your business. https://business.gov.nl/starting-your-business/first-steps-for-setting-up-your-business/permits-for-your-business/, 2023.

[NW06] Kathe Newman and Elvin K Wyly. The right to stay put, revisited: Gentrification and resistance to displacement in new york city. *Urban studies*, 43(1):23–57, 2006.

[NYRAB14] MN Noor, AS Yahaya, Nor Azam Ramli, and AM Al Bakri. *Filling missing data using interpolation methods: Study on the effect of fitting distribution*, volume 594. Trans Tech Publ, 2014.

[Phi93] Martin Phillips. Rural gentrification and the processes of class colonisation. *Journal of rural studies*, 9(2):123–140, 1993.

[PNGA19] Jessica Pesantez-Narvaez, Montserrat Guillen, and Manuela Alcañiz. Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, 7(2):70, 2019.

[PSSF11] Andrew V Papachristos, Chris M Smith, Mary L Scherer, and Melissa A Fugiero. More coffee, less crime? the relationship between gentrification and neighborhood crime rates in chicago, 1991 to 2005. *City & community*, 10(3):215–240, 2011.

[RDSH19] Jonathan Reades, Jordan De Souza, and Phil Hubbard. Understanding urban gentrification through machine learning. *Urban Studies*, 56(5):922–942, 2019.

[SBSKJ14] Patrick Sturgis, Ian Brunton-Smith, Jouni Kuha, and Jonathan Jackson. Ethnic diversity, segregation and the social cohesion of neighbourhoods in london. *Ethnic and Racial Studies*, 37(8):1286–1309, 2014.

[SD17] Mariska van der Sluis and Wenda Doff. De invloed van sterke schouders, 2017.

[Seg04] Mark R Segal. Machine learning benchmarks and random forest regression. 2004.

[SKK22] Michelle Stuhlmacher, Yushim Kim, and Ji Eun Kim. The role of green space in chicago's gentrification. *Urban Forestry & Urban Greening*, 71:127569, 2022.

[SMJS+20] Alina S Schnake-Mahl, Jaquelyn L Jahn, SV Subramanian, Mary C Waters, and Mariana Arcaya. Gentrification, neighborhood change, and population health: a systematic review. *Journal of urban health*, 97:1–25, 2020.

[SS18] Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE, 2018.

[SSSB+15] Martin Spüler, Andrea Sarasola-Sanz, Niels Birbaumer, Wolfgang Rosenstiel, and Ander Ramos-Murguialday. Comparing metrics to evaluate performance of regression methods for decoding of neural signals. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1083–1086. IEEE, 2015.

[Sta] Statistics Netherlands. Opbouw regionale indelingen. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/informatie-voor-gemeenten/opbouw-regionale-indelingen.

[Sta23a] Statistics Netherlands. Organisation. https://www.cbs.nl/en-gb/about-us/organisation, Jun 2023.

[Sta23b] Statistics Netherlands. Proximity statistics. https://www.cbs.nl/en-gb/our-services/methods/surveys/brief-survey-description/proximity-statistics, Feb 2023.

[Sto10] Aileen Stockdale. The diverse geographies of rural gentrification in scotland. *Journal of rural studies*, 26(1):31–40, 2010.

[Sud18] Kalyan Sudhakar. Python vs. r programming language. *Journal Homepage: http://www. ijmra. us*, 8(8), 2018.

[Tee15] Annalies Teernstra. Contextualizing state-led gentrification: Goals of governing actors in generating neighbourhood upgrading. *Environment and Planning A: Economy and Space*, 47(7):1460–1479, 2015.

[TNLP23] William Thackway, Matthew Ng, Chyi-Lin Lee, and Christopher Pettit. Building a predictive machine learning model of gentrification in sydney. *Cities*, 134:104192, 2023.

[Tri17] Kate Trigg. Quantifying urban inequality: An investigation of the wicked problems of gentrification, 2017.

[VdGV+09] Peter Van der Graaf, Lex Veldboer, et al. The effects of state-led gentrification in the netherlands. *City in sight: Dutch dealings with urban change*, pages 61–80, 2009.

[Wat88] Bernard C Watson. The truly disadvantaged: The inner city, the underclass, and public policy., 1988.

[Wil15] Kristin Williams. Toward a universal operationalization of gentrification. *Sociation Today*, 13(2):1, 2015.

[WM05] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

[YWX+20] Fangfang Yang, Dong Wang, Fan Xu, Zhelin Huang, and Kwok-Leung Tsui. Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model. *Journal of Power Sources*, 476:228654, 2020.

[ZB21] Kasey Zapatka and Brenden Beck. Does demand lead supply? gentrifiers and developers in the sequence of gentrification, new york city 2009–2016. *Urban Studies*, 58(11):2348–2368, 2021.

[ZBC+18] Miriam Zuk, Ariel H Bierbaum, Karen Chapple, Karolina Gorska, and Anastasia Loukaitou-Sideris. Gentrification, displacement, and the role of public investment. *Journal of Planning Literature*, 33(1):31–44, 2018.

[ZKB+21] Md Sabab Zulfiker, Nasrin Kabir, Al Amin Biswas, Tahmina Nazneen, and Mohammad Shorif Uddin. An in-depth analysis of machine learning approaches to predict depression. *Current research in behavioral sciences*, 2:100044, 2021.

[ZSW14] Helen Wei Zheng, Geoffrey Qiping Shen, and Hao Wang. A review of recent studies on sustainable urban renewal. *Habitat International*, 41:272–279, 2014.

Appendices

A - Sources Information Amenities per District/Neighbourhoods Dataset

Table 8: The sources of all the types of features.

Features	**Source**
Addresses of persons	Personal Records Database (BRP)
Addresses and locations	Land Registry and Mapping Agency
Roads	National Road Register
Addresses of general practices	Nivel
Addresses of hospitals and outpatient clinics	National Institute for Public Health and the Environment
Addresses of retail shops, hospitality industry, and cinemas	LOCATUS
Addresses of childcare	National Childcare Register
Addresses of schools	Service Execution Education
Locations of train stations	Rijkswaterstaat and Prorail
Addresses of museums	Museum Association
Locations of performing arts	VSCD and VNPF

B - Comprehensive Description of the Included Amenities

Table 9: Description of the amenities used from the amenities datasets.

Features	**Description**
Area code	This is the unique code all the municipalities, districts, and neighbourhoods.
Distance to general practice and number of general practices within 1 km.	A building where one or more general practitioners work .
Distance to hospital and number of hospitals within 5 km.	Hospitals including outpatient clincis.
Distance to supermarket and number of supermarkets within 1 km.	These are big supermarkets with a minimum surface area of 150 m2.
Distance to other foodstuffs and number of other foodstuffs within 1 km	Shops for everyday foods. Think of bakeries, butcheries, liquor stores, et cetera.
Distance to warehouse and number of warehouses within 5 km	Stores with at least 50 or more employees and diverse products.
Distance to café and number of cafés within 1 km	Cafés, coffeeshops, nightclubs, pubs, sex clubs and party centres.
Distance to cafeteria and number of cafeterias within 1 km	Fast food restaurants, grillrooms, lunchrooms, pancake restaurants and ice cream parlours.
Distance to restaurant and number of restaurants within 1 km	Restaurants, café-restaurants, and takeaway/delivery places.
Distance to hotel and number of hotels within 5 km	Hotels, hostels, and hotel-restaurants.
Distance to nursery and number of nurseries within 1 km	Places where children ranging from 0 to 4 years are taken care of.
Distance to after-school care and number of after-school cares within 1 km	Places where primary school children are taken care of before and/or after school.
Distance to primary school and number of primary schools within 1 km	Primary schools excluding special primary schools.
Distance to secondary school and number of secondary schools within 3 km	Secondary schools excluding special secondary schools.
Distance to main road	Distance to main roads calculated over the road.
Distance to train station	Distance to train stations calculated over the road.
Distance to museum and number of museums within 5 km	An institution serving society and its development.
Distance to performing arts and amount of performing arts within 5 km	Performing arts carried out by professional actors for an audience excluding festivals.
Distance to cinema and number of cinemas within 5 km	Theatres where films are shown for public entertainment.
Distance to theme park and amount of theme parks within 10 km	These are theme parks, zoos, and indoor playgrounds.

C - Changed Neighbourhood Layout

There can be various reasons for a municipality to change the layout of a neighbourhood. A municipality may want to make policy at a finer level and be able to assess it. Therefore, they would reorganise the neighbourhoods differently. It may also have to do with new construction, for example, if a new neighbourhood is being built, and that area fits better with neighbourhood X than with neighbourhood Y. Moreover, municipal reorganisation can also play a role. When municipalities merge, it is often the moment when the organisation is adapted to the new situation.

Figure 6 shows a map of municipalities in the Netherlands. There are three different line colours:

- red is the neighbourhood layout in 2020;
- blue is the neighbourhood layout in 2021; and
- black is the municipality layout in 2021.

In Culemborg, in 2021, one boundary line was added to divide the neighbourhood into two parts, while in Vijfheerenlanden, the boundaries underwent a complete change between 2020 and 2021. Comparing the neighbourhoods of 2020 with 2021 would lead to an inaccurate assessment of gentrification.

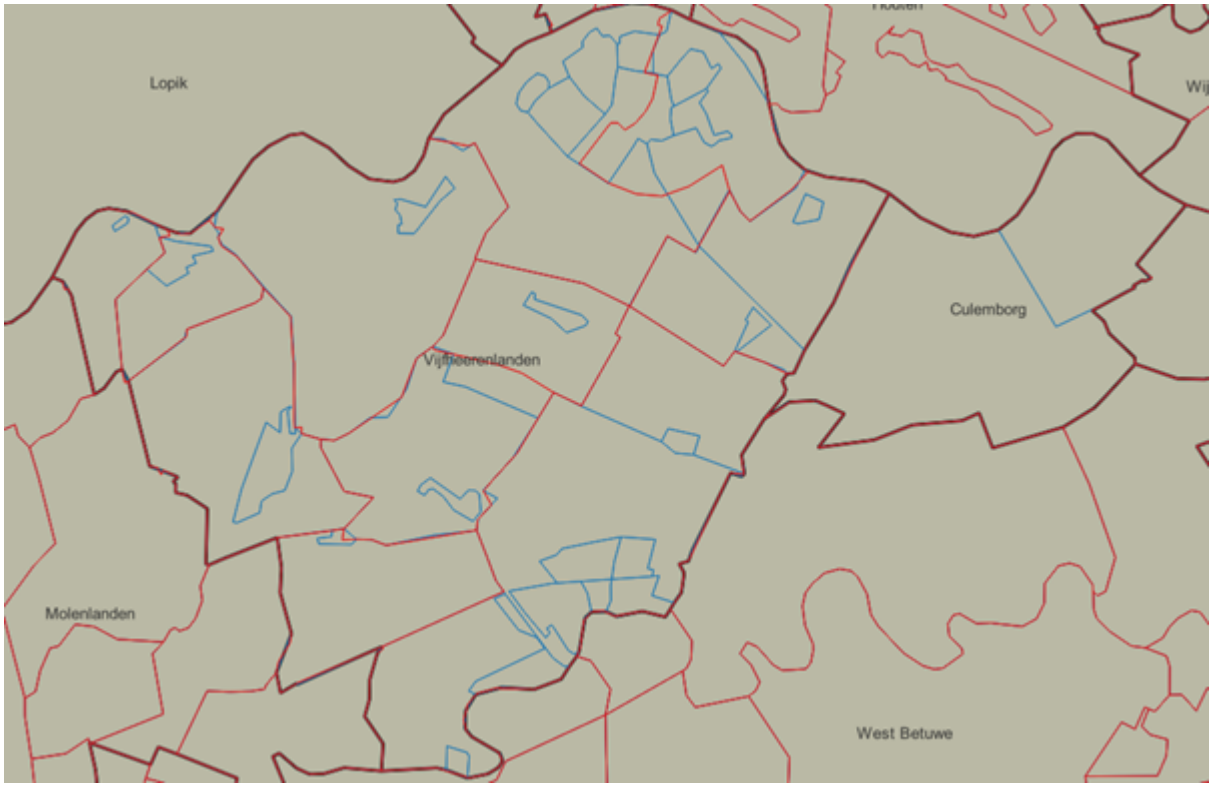

Figure 6: A map that lines out the borders of municipalities and neighbourhoods.