UTRECHT UNIVERSITY
Department of Information and Computing Science

Applied Data Science master thesis

Exploring the Impact of Energy Labels on Residential Properties Prices: A Data-Driven Analysis

Candidate: Danya Mawed
First examiner: Bilgecag Aydogdu
Second examiner: Prof. Dr. Albert Ali Salah
In cooperation with: STATER N.V

April 24, 2023

Abstract

In recent years, sustainable buildings that promote energy efficiency have become a major priority in the development and construction industry, due to their potential environmental, economic, and social benefits. Energy labels, which range from A (highly efficient) to G (highly inefficient), are used to assess the energy efficiency of buildings. This study focuses on residential properties in the Netherlands and aims to investigate the impact of energy labels on the sale price of buildings. By exploring this impact we seek to understand the implications of energy label upgrades on the market value of the houses. Public data provided by the Central Bureau of Statistics in the Netherlands, along with dwellings data from an online real estate listings platform, were utilized for this study. Our findings demonstrate that a house's energy label influences its sell price, indicating that energy label upgrades have the potential to increase property value. The research provides evidence of how changes in the energy label rating impact house prices. Understanding this influence offers valuable insights for homeowners, real estate professionals, investors, and policymakers. This information enables stakeholders to make better decisions in the real estate and investing markets.

This research was conducted in collaboration with the Data Lab department at STATER N.V, a Mortgages Service Company (https://stater.nl/).

Table of Contents

Introduction

The EU aims to enhance the buildings' energy efficiency and seeks to achieve highly energy-efficient and decarbonized building stock by 2050 (European Commission, 2020). The Directive on Energy Performance of Buildings (EPBD), introduced in 2003, serves as a crucial tool to promote energy efficiency in buildings. Under EPBD, Energy Performance Certificate (EPC) was established to monitor energy consumption in buildings. The EPC assigns energy labels ranging from A to G. Energy label A was recently divided into A, A+, A++, A+++, and A++++ according to rvo.nl (2023). These labels consider various factors such as insulation, heating and cooling systems, and renewable energy sources.

With a focus on the residential properties in the Netherlands, this research explores the relationship between energy efficiency of houses, which is represented by energy labels, and their sell price and examines the impact of energy-label upgrades on house value.

Economically speaking, the motivation behind investigating the connection between the energy labels and their financial impacts lies in the potential reduction in energy costs and rise in property values associated with having an efficient label. Brounen and Kok (2010) mentioned in their research that the energy label may be beneficial to both real estate investors and tenants, as energy savings from more efficient buildings may result in cheaper running expenses and better property values.

In terms of operational costs, Park et al. (2015) discussed in their research the impact of Korea's energy efficiency system on the energy consumption of a residential building. Their findings highlighted the system's impact, demonstrating that the energy-saving ratio increased as gas and district heat consumption decreased, this reduction in energy consumption subsequently resulted in lower operational costs. Similarly, Dielessen (2023) conducted a study in the Netherlands and observed a similar relationship, demonstrating that houses with higher energy efficiency have lower energy bills.

Many studies have explored the relationship between energy label, and real estate market; Aydin et al. (2019) proved the positive relationship between energy labels and the selling process, the study shows that Houses with an energy label before sale faced a decrease in selling time. Furthermore, having a high energy label speeds up the sale in comparison with the lower label, they found that houses labeled A experience an increase in sale speed by 28 percent. The results of the study conducted by Kok and Jennen (2012) concluded that less efficient property achieves lower rent compared to similar properties with a high energy label.

With regards to the energy performance impact on property value, , Aroul and Hansz (2010) research reveals a statistically significant effect on transaction prices of green residential* properties in two similar Texas cities: Frisco and McKinney. Moreover, Fuerst et al. (2015) found evidence of a positive association between price per square meter and energy efficiency labels. Likewise, Brounen and Kok (2010) reported that the property price varies with the label class, and houses with higher energy labels are associated with a higher price.

The above studies provide evidence of the positive impact of energy efficiency on the financial aspects of buildings, including property values, rental rates, operational costs, and selling time. Thus, improving buildings energy performance by upgrading their labels through enhancing insulation, heating and cooling system, or adding renewable energy systems can not only minimize the carbon emission but also will increase its market value.

We address in this study the following research question*: Does the energy label of a house significantly affect its sell price, and what is the extent of the impact of energy label improvements on house prices across different energy labels?*

To facilitate the process of addressing our research question, we translated it into four data science questions:

How does the energy label of a house correlate with its sell price?
What is the magnitude of the impact of energy labels on house prices?
What is the effect of energy label upgrades on house prices?
And what is the optimal energy label category that maximizes house prices?

To answer these questions, we employed a series of analytical methods. Initially, we assessed the influence of energy labels on house value using four Machine Learning models. These models provided us with valuable insights regarding the association between energy labels and house prices. Additionally, to understand the effect of changes/upgrades in the energy label, the best-performing machine learning model was used to apply Sensitivity Analysis, Also known as What-if Analysis. Finally, to assess the statistical significance of the sensitivity analysis results, the ANOVA test was initially attempted to be used. However, because one of the conditions of the ANOVA test was violated, a similar but non-parametric test called Kruskal-Wallis was employed.

By conducting the above approach, we aim to provide insights into the potential risks and benefits coupled with energy-efficient upgrades.

*green residential Identifiers in this research were 18 including energy features, such as solar and wind power, Energy Star certification, LEED energy rating system.

Data

Description of the data

The Data that was utilized in this study was collected from two different public sources: Funda.nl and CBS.nl. We applied data preprocessing techniques to the raw data before it was used for modeling and analysis*.

Data Collection

- Funda.nl:

Funda is a leading online platform for real estate listings in the Netherlands.It provides extensive information on properties, including their characteristics, prices, and transaction details. Using a python library called Funda Scraper, version 0.0.3, we were able to scrape data of the houses that were sold between the date of scraping and approximately one and a half years prior (between the end of 2021 and the second quarter of 2023). The scraping process was done city by city, the separated cities datasets were concatenated into a single dataset, which consists of 21,064 data points. The city, house type, price, price m2, living area, energy label, and house age are the most important variables among the many various variables that Funda data has.

- CBS.nl:

The official website of Statistics Netherlands provides reliable data on various aspects of the country. The House Price Index (HPI) data obtained from the open data portal of CBS.nl serves as a reliable indicator; HPI captures changes in the prices of residential properties over time. It provides valuable insights into the fluctuations in property prices within a specific region or country. The base year for the index is typically set at 100, and subsequent values reflect changes in prices relative to that base year. In the dataset available, the base year is 2015, and the subsequent years reflect the changes in house prices relative to 2015. The HPI data includes the price index for each year separately, as well as the price index for each quarter of the year. The data set has the region, HPI, the quarter and year variables.

For both of the datasets mentioned above, we focused solely on houses located in the cities of Amsterdam, Rotterdam, Utrecht, and Den Haag. This decision was made due to the availability of House Price Index (HPI) data specifically for these four cities. The HPI data covers the entire province of the Netherlands, but in terms of cities, HPI is only available for these specific cities. In case we wanted to include an additional city in the analysis, we would have to use the HPI of the province to which it belongs. Therefore, we preferred to keep the analysis as much accurate as we can and just include the cities stated above. Hence, HPI data set has 35 data points for the selected cities and period (the quarters of the years 2021, 2022 and 2023), this was chosen based on the period covered by the scraped Houses dataset.

Data Exploration

- **Descriptive Statistics**

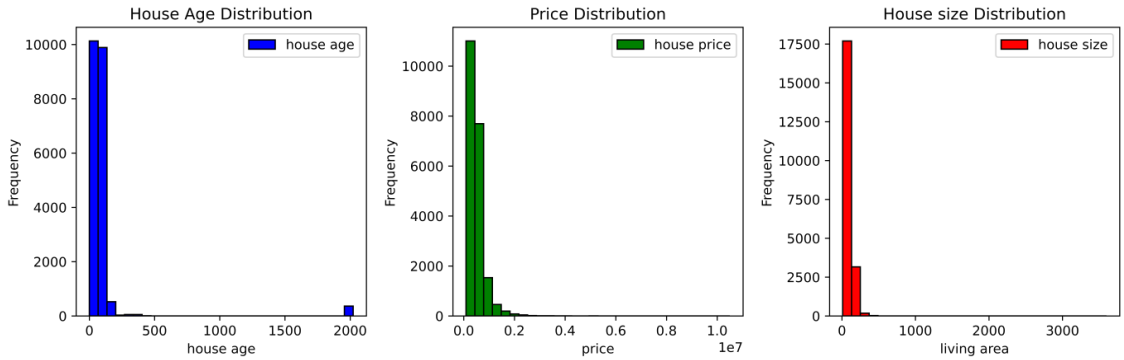
To gain a better understanding of the data, the descriptive statistics for the variables in the Funda dataset was analyzed. By examining these statistics, we can get more insights within the dataset. It is observed that the variables 'price', ‘price m2’,'living area' and 'house age' have outliers, as evidenced by a substantial difference between the mean and the maximum values.

For 'house age' variable, we observed a mean of 103.22 years, the 75th percentile was 98 with a maximum value of 2023 years. Similarly, the mean 'living area' was found to be 95.39 square units, with a 75th percentile of 115 and a maximum value of 3600 square units. Furthermore, the mean 'price' was determined to be € 506,076.1, with a maximum value of € 10,500,000, while its 75th percentile was € 589,000. The mean of 'price per m2' was €5468.23, its 75th percentile € 6820, the maximum value was € 34,426.2

However, it was also noticed that the price per m2 can be calculated by dividing the house price by the house size, which means that it is essentially redundant information. So it doesn't add extra information to the data set. For that reason we won't include it in our outlier analysis and further analysis.

These statistics indicate that there are houses with significantly higher values compared to the overall distribution of the data. The presence of outliers can impact the analysis and interpretation of the data, and it may be necessary to consider appropriate strategies to handle them.

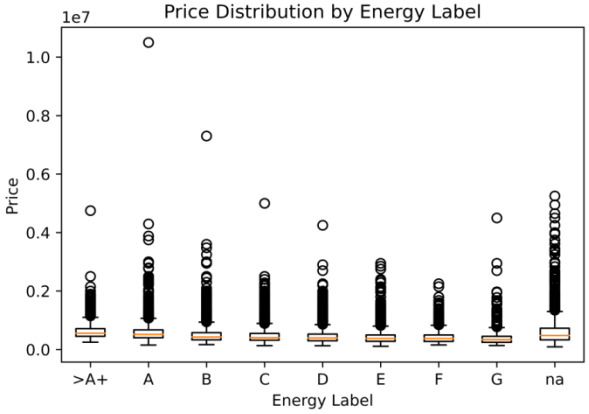
The following Histogram distributions support the above conclusion:



Histogram Distributions before performing Cleaning (Figure 1)

- **Boxplot:**

In addition to analyzing descriptive statistics, boxplot was also used to explore the relationship between prices and energy labels in Funda Dataset. The boxplot showed the distribution of prices for each energy label category in the Funda dataset.



Explore the relation between Price and Energy label (Figure 2)

We clearly can see that Energy label data has many null values in addition to an extra energy label category: >A+. These findings were taken into account in the data cleaning step.

However, due to the large amount of data and the presence of numerous outliers for each energy label, the above figure couldn't provide significant visual insights about the relationship between the price and energy label.

- **Correlation Matrix:**

The correlation matrix provides insights into the relationships between the variables. The most important observation is the strong positive correlation of 0.64 between price and living area. This indicates that the size of the property and its price have a significant relationship. As the living area increases, the price tends to increase as well.

There is a weak positive correlation of 0.10 between the price and the house age. This implies that there is a slight tendency for older houses to have higher prices, although the relationship is not strong.

Moreover, there is a very weak negative correlation of -0.01 between the living area and the house age. This implies that there is almost no relationship between the size of the property and its age.

Preparation of the data

Data Preprocessing

Data preprocessing is an important stage because it addresses errors, missing values, inconsistencies, and noise in raw data, all of which can affect the quality and reliability of data analysis results.

During the initial exploration of the data, several oddities and interesting patterns were observed. These outliers could be attributed to factors such as data entry errors, measurement inconsistencies, or unique observations. For instance, 2023 was determined to be the maximum value for the house age variable, which appears unusual and may indicate a data entry error or confusion with the year built variable. Similarly, the maximum value of 3600 square units for the living area variable seems unusual for a residential property, suggesting the possibility of incorrect data entry, measurement errors, or a unique value. Furthermore, the price variable showed noticeably higher values than the data's overall distribution, which could indicate the presence of outliers or data collection issues.

To address these issues, data cleaning techniques can be employed to identify and correct errors or inconsistencies in the raw data. In addition to data cleaning, data transformation is another important step in data pre-processing. Data transformation involves converting data from one format or representation to another, often with the goal of making it more suitable for analysis or modeling purposes in order to maximize model performance. And since we are working with two different datasets it is crucial to apply data integration technique which involves merging multiple datasets or combining different sources of data to produce a single, combined dataset.

By performing these preprocessing steps, we can ensure that the data we have is a solid foundation for accurate and meaningful analysis in the subsequent stages of the research.

1- Data Cleaning

The Funda houses data was cleaned thoroughly because it contained the most variables and data points. However, small cleaning steps were also performed on the HPI dataset .

Firstly, the dataset was checked for duplicated houses, these duplicates were removed. The energy label category '>A+' was replaced with 'A' for the sake of uniformity, as it was unclear which specific category ('A+', 'A++', 'A+++', or 'A++++') the label represented. Since the energy labels cannot be imputed based on other variables, any entry with missing or 'na' values in the energy label field were eliminated. Additionally, entries with a house age of 2023, which could potentially be erroneous, were detected and removed from the dataset.

Secondly, Outlier Analysis: The interquartile range (IQR) method was employed instead of relying on the Z-score (which measures the number of standard deviations away a value is from the mean) to detect the outliers because the data does not adhere to a normal distribution.
In the IQR approach, The data points that fall below Q1 - 1.5 IQR or above Q3 + 1.5 IQR are considered as potential outliers, where Q1 represents the 25th percentile, Q3 represents the 75th percentile and the IQR is the difference between Q3 and Q1. Along with IQR, observation judgment that was based on domain knowledge was taken into account in the outliers' detection process.

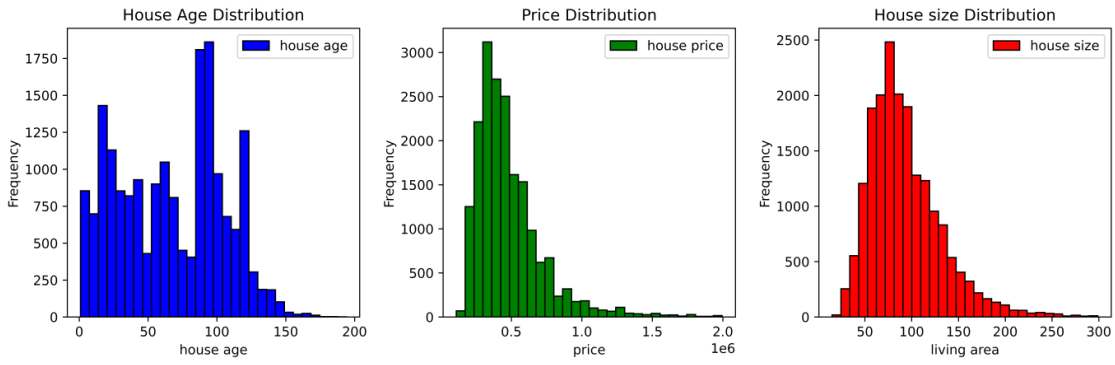
According to the IQR calculations, the price variable contained a significant number of outliers, namely 986 data points with a minimum outlier value of 955,000. While it is realistic to expect houses with values exceeding one million dollars due to location and size factors. In this event, it is crucial to consider the observational judgment in identifying and handling outliers along with statistical methods, rather than relying only on statistical calculations such as the IQR, which may lead to deleting a large number of data points that may have an important impact on the analysis.

Further observation revealed that a small number of data points had prices above 2 million dollars (76 data points) which is far less than amount of data that could be erased if the IQR method alone was used. Thus, considering the house values above 2,000,000 as extreme outliers is more appropriate.

Moreover, a common characteristic among these outliers is their larger sizes and their older ages. In terms of house size, we determined points with area above 300 m2 as outliers, which was confirmed through observational judgment for the same reason of the price variable. IQR method identified 654 houses as outliers, with minimum outlier value of 185 m2, while when taking into consideration domain knowledge 53 data point were treated as outliers.

Where in the context of house age, houses that surpassing 198 years, as determined by the IQR method where identified as extreme outliers. This selection obtained via the IQR method aligned closely with the observations made.

The following Histogram distributions of the variables after deleting the values we considered as outliers:



Histogram Distributions after Cleaning steps (Figure 3)

Thirdly, the HPI dataset was cleaned, the values of region column that have the cities names, were cleaned to be aligned with cities names in the houses data set. The HPI value column name was changed to make it shorter. As the values of the HPI represent the change in House prices relative to a specific base year, in the analysis it is more meaningful to use the HPI as a percentage relative which will provide more interpretable information about the change in the price.

2- Data Integration

The merging process was based on the selling period and the city that each house is located in. To accomplish merging, we extracted the quarter and the year of sale into separate column and match it with a similar column in the HPI data.

Then we assigned the HPI value to each house based on the quarter and the year in which it was sold along with the city name. These steps allowed us to merge the datasets effectively, and get a single data set to apply our further work on.

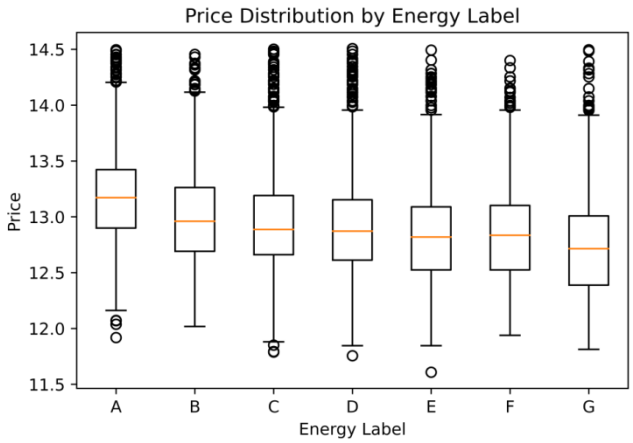
3- Data Transformation

It can be inferred out of Figure 1 that the three aforementioned variables exhibit right-skewness. When we have variables with different scales and skewed distributions, it is beneficial to apply scaling. Scaling helps to ensure that all the variables are on a similar scale and no single variable dominates the learning process. It also improves the interpretability of the model coefficients. Based on some modeling tests we have done on the house age and house area variables (as independent variables in the model) it was observed that scaling enhanced model performance

Scaling the dependent variable (in our case the price variable) can also be helpful, because without scaling the model will be trained on a larger amount of modestly priced houses, which means it may have less accuracy in forecasting the prices of the most expensive houses (Jermain, 2019). By scaling, we can potentially improve the performance of my model.

Considering that the variables have skewed distributions, using a log transformation is the most popular approach. The log transformation helps to reduce the skeweness of the variables and reduces the impact of extreme values. This, in turn, leads to more accurate and reliable results.

After scaling the variables, let's revisit the relationship between prices and energy labels, but this time with clearer visuals:



Explore the relation between Price and Energy label (Figure 4)

Visual examination of Figure 4 reveals that homes with higher energy labels tend to have slightly higher mean prices than homes with lower labels. This observation suggests that energy efficiency, as indicated by the energy label, may play a role in determining house prices.

After performing the above data preprocessing steps the dataset that we are going to use in our further analysis was reduced to 17,832 data points.

Methods

In order to facilitate the use of data science methods for investigating the research topics of this study, we can develop the following data science questions:

How does the energy label of a house correlate with its sell price?

What is the magnitude of the impact of energy labels on house prices?

What is the effect of energy label upgrades on house prices?

What is the optimal energy label category that maximizes house prices?

Data Modeling

To answer the first two questions, we started by studying the influence of energy labels on house sell prices using machine learning models. However before Modeling we needed to do two more steps in order to enhance the model learning effectiveness:

1- **Selection of the independent variables:**

To study the effect of energy labels on house sell prices, the energy labels should be treated as the independent variable, while the price of the house is the dependent variable.

To incorporate the effect of time on the house price, the price index variable was included in the model. This variable allows us to account for changes in prices over time and capture the temporal dynamics of the housing market. However, it is ineffective to predict the price of a specific house using only HPI because it is a rough indicator derived from all transactions (Truong *et al.*, 2020).

For that reason, it is logical to acknowledge that other factors may also impact sell prices. Thus, considering the conducted data exploration and the insights we obtained from the correlation matrix, the chosen independent variables were: City, House Price Index, Energy Label, House Living Area, House Age, and House Type. These variables were identified as having a logical contribution to the house price, while other variables were deemed less relevant in this context.

City, Energy labels, and House type are categorical variables. In order to handle them, dummy variables were created. This approach enables us to represent categorical variables as binary indicators, making them suitable for inclusion in the modeling process.

Note: In this study, we created dummy variables for the energy label variable. While other methods for representing the energy label variable, such as encoding it as numeric values, are possible, our decision to use dummy variables was driven by the specific requirements of the data analysis step, which is dependent on the chosen machine learning model and its input variables.

2- **Training and Test sets:**

To evaluate the performance of the models accurately, the dataset was divided into two subsets: a training dataset 80% and a testing dataset 20%. The training dataset is used to build and train the models, while the testing dataset serves as an independent set for assessing the models' predictive performance. This separation helped us evaluate how well the models generalize to unseen data.

- **Machine Learning Models:**

To help answer the above questions, four different models were trained and evaluated to identify the best-performing model for further analysis. Each of these models provided valuable insights and

contributed in different ways to address the research questions. We were better able to interpret the dataset and its characteristics by utilizing these models, which helped us gain a deeper understanding of them.

- **The initial model** that was utilized in this research is Linear Regression. LR model was chosen for its simplicity and interpretability. This choice was informed by the observed linear relationships between the dependent variable and variables such as living area, house age (as observed in the correlation matrix), and energy labels (as shown in the boxplot).
  The resulting R2 score indicates that the model explains a moderate level of the variance in the dependent variable.
- **The second model** that was employed in this study is Polynomial Regression, specifically using a second-degree polynomial; it was chosen for its ability to capture nonlinear relationships between variables. The choice to use polynomial regression was driven by the finding that the linear regression model had a moderate score of R2, which indicated that there might be additional nonlinearity or complexity in the relationship between the independent variables and the dependent variable that could be better captured by a polynomial approach.
  The resulting R2 score explains around 78% of the variance in the house prices, indicating a good level of predictive accuracy. It is worth mentioning that attempts were made to experiment with higher polynomial degrees, but the improvement in performance was minimal.
- **The third model** that was used is Random Forest. Several reasons contributed to the selection of RF: its ability to handle large datasets, explore complex relationships, reduce overfitting, and generate accurate predictions. The number of estimators was set to 100, which proved to be sufficient for obtaining optimal results from this model. The R2 score for RF model was higher than the previous model which refers to a high level of prediction accuracy.
- **The fourth and last model** that was utilized in this research is the LightGBM Regressor, developed by (Microsoft, LightGBM), which stands for Light Gradient Boosting Machine. LightGBM is a powerful machine learning model, built on the concept of gradient boosting, which combines weak learners (decision trees) to form a powerful ensemble model.
  "LightGBM contains two novel techniques: Gradient-based One-Side Sampling and Exclusive Feature Bundling to deal with large number of data instances and large number of features respectively" (Ke et al. 2017). LightGBM is suitable for large-scale datasets and real-time applications as it enhances speed by utilizing leaf-wise tree growth and gradient-based one-side sampling and It optimizes memory usage and computation efficiency through exclusive feature bundling
  For this project, the number of estimators was set to 110, which yielded the best outcome for our model.

- **Models Evaluation:**

It is important to assess and compare the performance of different machine learning models in order to select the most appropriate one to be used as a foundation for next steps.

To evaluate the above four models performance we used the MSE (mean squared error) metric, which measures the average squared difference between the predicted and actual values, RMSE (Root Mean Squared Error) displays the average size of prediction errors in the target variable's

original units, and the R2 score indicates the proportion of the variance in the dependent variable, it provides a measure of how well the regression model fits the data

Among the 4 ML models, the LGBM model had the highest R2, hence explaining a larger proportion of the variance in the dependent variable compared to the other models. Furthermore, it scored the lowest MSE, indicating that it performed better in terms of predicting the dependent variable with less error. Therefore, Light Gradient Boosting Machine Model was selected to serves as the foundation for the subsequent analysis step, known as Sensitivity Analysis.

Data Analysis

To continue answering the remaining 3rd and 4th questions, A technique called Sensitivity Analysis was employed. This technique is used to study and understand how the changes in input variable can affect the model output (Loucks and van Beek, 2005). Sensitivity Analysis was also defined as a series of investigations of potential changes and assess whether these changes leads to different conclusions (Pannell,1997; Thabane et al., 2013). Based on Pannell (1997), SA is used for a wide range of applications such as Assessing the 'riskiness' of a strategy or scenario ,allowing decision makers to select assumptions, estimating relationships between input and output variables, understanding relationships between input and output variables, and developing hypotheses for testing. In the context of our research we are using SA to assess the influence of certain input variable on the outcome variable.

One-at-a-time Sensitivity Analysis was used, which is the simplest way to analyze a model. In general, it consists of (Glen, S.):

1- All variables are held constant except one, which is altered while new readings are taken.
2- Iterating the first step multiple times until all of the desired changes are applied.

The above steps were applied as the following:
Baseline Prediction: we started by training the LGBM model using the train dataset and obtaining baseline predictions of House prices for the test dataset that has Houses with energy label D. These baseline predictions serve as our reference point for comparison.

Changes: Next, we modified the energy label variable in the test dataset. Specifically, six modifications were employed by changing the energy label from D to A, B, C, E, F, and G, respectively. Each change represents a different energy label assigned to the houses in the test dataset.

Prediction Comparison: we generated new predictions for each change, where the energy label has been modified. We compared these new predictions with the baseline predictions to observe the effect of upgrading or downgrading the energy label on house prices.

In the end, we applied statistical tests to evaluate and assess the meaningfulness of the results we observed in the sensitivity analysis and to statistically demonstrate that there is a difference in the house prices across different energy label.

Initially, ANOVA test was intended to be used, but before applying it was important to consider the assumptions associated with this test (independence, normality, and homogeneity of variances).

Although the independence and homogeneity assumptions were met, the normality assumption was violated despite various attempts at data transformations. Therefore, the ANOVA test results could not be relied upon. As an alternative, the Kruskal-Wallis test, a non-parametric test that compares the medians of three or more independent groups, was used. This test is suitable when the

normality assumption is violated and is less sensitive to outliers. Since our data met the assumptions of the Kruskal-Wallis test (independence, similar shape, equal variability, continuous dependent variable), it was chosen as a suitable alternative to ANOVA.

The null hypothesis of the Kruskal-Wallis test states that the medians of all groups are equal, while the alternative hypothesis suggests that at least one group's median is different.

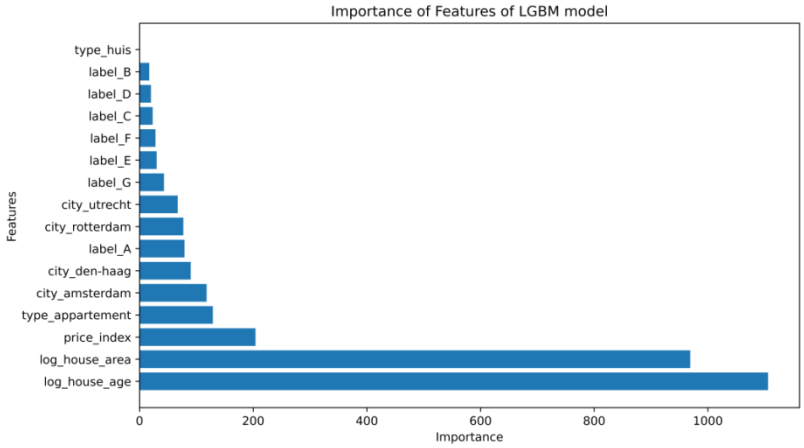
Results and Discussion

a) Assessing the influence of energy labels on house value using Machine Learning model

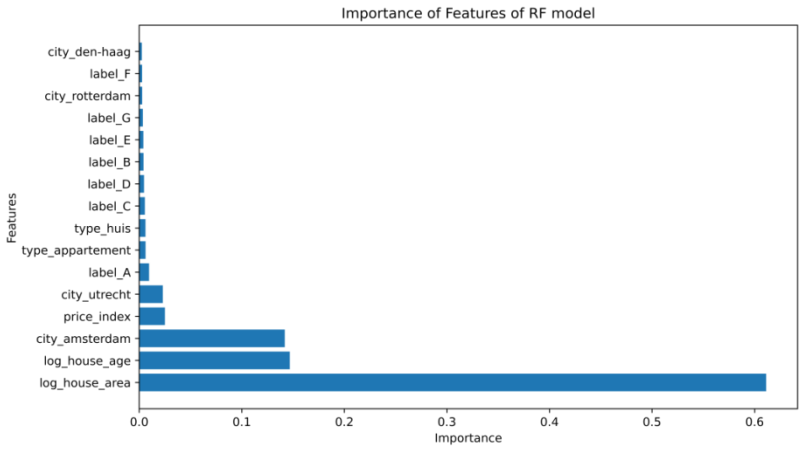
All of the trained models in this study provided evidence of a relationship between energy labels and house prices, indicating that the presence of energy labels adds value to the houses. This conclusion was based on the significant p-values and positive coefficients observed for the energy label variables in the regression models. Furthermore, the feature importance analysis conducted using Random Forest and LightGBM models highlighted the relative contributions of each input variable in the models' decision-making process, showing that energy labels played a role in predicting house prices to some extent.

However, It is crucial to highlight that the presence of energy labels in the dataset was not balanced, meaning that there are significantly more instances of certain energy labels compared to others, which introduces a bias in the analysis. This imbalance can affect the interpretation of higher coefficients or feature importance values associated with specific energy labels. The higher values may not necessarily indicate a higher added value, but rather reflect the larger representation of those energy labels in the dataset.

Additionally, it's important to note that these values are model-specific and depend on the algorithms and methodologies employed. Therefore, comparing feature importance values between different models, such as Random Forest and LightGBM, is not appropriate. Each model may have different criteria and considerations for determining feature importance, and their results should be interpreted within the context of their respective models.



Features Importance of LightGBM (Figure 5)



Features Importance of Random Forest Model (Figure 6)

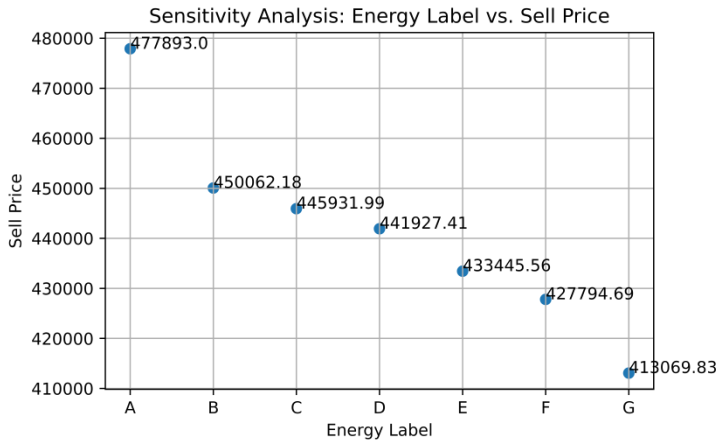
Through a comparison of the coefficients of all energy labels with those of other variables in the linear regression model, as well as an assessment of their importance in the Random Forest (RF) and LightGBM (LGBM) models, it is suggested that energy labels may not have a substantial overall impact on house prices compared to factors such as house area (size), and age that play a more significant role in determining house prices.

Based on the above results, it can be concluded that energy labels have an impact on the value of residential properties. However, this impact is observed to be relatively low. In other words, even though energy labels do affect the overall value of a property, their impact is not as significant as other factors.

b) Study the effect of changes/upgrades in the energy label

According to the above results, further analysis is required to gain a deeper understanding of the specific impact of each energy label on house prices. This was achieved through the implementation of sensitivity analysis, which allowed for a detailed examination of the effect of changing energy labels on the corresponding house prices.

During the sensitivity analysis, six perturbations were applied to the baseline dataset by modifying the energy label from D to A, B, C, E, F, and G, respectively. New predictions were then generated for each modification. The following are the mean prediction prices for each energy label:



Mean Homes Prices per Energy Label (Figure 7)

Figure 7 indicates that the mean prices of houses vary slightly among different energy labels, with an increasing trend as the energy label rating goes up.

By comparing the mean sell prices for each energy label, we can assess the average effect of changing the energy label on house prices. These findings support the hypothesis that changes in the energy label variable indeed influence the sell price, with higher energy label ratings being associated with higher sell prices.

The below results show the percentage change in house prices when transitioning from the respective energy labels to energy label A

- Residential properties with Energy label B: Approximately 6.18% increase in price when its energy label changed to A.
- Residential properties with Energy label C: Approximately 7.16% increase in price when its energy label changed to A.
- Residential properties with Energy label D: Approximately 8.13% increase in price when its energy label changed to A.
- Residential properties with Energy label E: Approximately 10.25% increase in price when its energy label changed to A.
- Residential properties with Energy label F: Approximately 11.71% increase in price when its energy label changed to A.
- Residential properties with Energy label G: A significant increase in price of approximately 15.69% when its energy label changed to A.

This information provides insights into the impact of energy label improvements on house prices and can assist in understanding the potential financial benefits of upgrading to higher energy efficiency ratings.

c) Evaluate whether the observed changes in the sensitivity analysis are statistically significant

The Kruskal-Wallis test was used to determine if there were differences in sell prices among the energy label groups. The high test statistic value suggests a substantial difference among the medians of the groups, indicating significant variation in sell prices. The low p-value provides strong evidence against the null hypothesis of equal medians, indicating that the observed differences are unlikely to occur by chance. In simpler terms, the test results robustly support the finding that sell prices vary significantly depending on the energy label, with a very low probability of the observed differences being due to random variation alone.

d) Results Evaluation

Brounen and Kok (2010) found that, when compared to similar homes with a D-label, homes with an A-label were sold at a price that was 10.2 percent higher. In contrast, homes with a G-label were sold at a discount of approximately 5.1 percent compared to similar homes. Comparing our results with their findings, we observe that they align closely. we found that homes that changed their energy label from D to A experienced an increase in price by 8.13 percent. Similarly, compared to D-labeled homes, houses with a G-label showed a decrease in sale price by 6.52 percent.

In their study, Fuerst et al. (2015) estimated the price differences between different energy labels. Compared to label D, property with label A and B have higher prices by 5% while properties with label C have higher prices by 1.8%. on the other hand, properties with E and F label sold by 1% less and that has label G sold by nearly 7 % less. Comparing these results to our study, we found that

properties with G-labels showed a decrease in sale price by 6.52% compared to D-labeled properties, which aligns closely with the decrease reported by Fuerst et al (2015). Additionally, properties with C-labels had a higher value by almost 1%. Properties with E-label showed a decrease of 1.9% in price compared to D-labeled properties. Furthermore, In comparison to D-labeled properties, our results showed an increase of 1.8% in price for properties that upgraded their energy label from D to B. Additionally, for houses with an F-label, we observed a decrease of 3.19% in price compared to D-labeled properties.

Conclusion

Based on the analysis conducted, we investigated the effect of energy labels on the sell prices of houses. Several Machine Learning Models were trained to discover the Influence of Energy labels on residential properties value. The LGBM model was utilized as a foundation for Sensitivity Analysis to observe the impact of changes in the energy label variable. Sensitivity analysis was performed by comparing the baseline predictions with perturbed predictions for different energy labels.

To assess the statistical significance of these results, we initially planned to use ANOVA. However, due to violations of the normality assumption, we employed the Kruskal-Wallis test instead. The Kruskal-Wallis test confirmed significant differences in sell prices across the energy label groups, supporting the notion that changes in the energy label variable influence sell prices.

The used Model confirmed the relationship between the energy label and the property's sell price and showed, that energy label of a house has an effect on its sell price. However the Extent of this effect is not so high compared to other factors, which significantly influence the price.

The conducted Analysis presents that the mean sell prices of the homes varied slightly across different energy labels, with higher energy labels generally associated with higher sell prices. Specifically, as the energy label increased from any other energy label to A, there was a progressive increase in the mean sell prices. This implies that A-label is the optimal category among other labels.

The statistical test that was applied proved that there is a statistically significant difference in sell prices among different energy label groups.

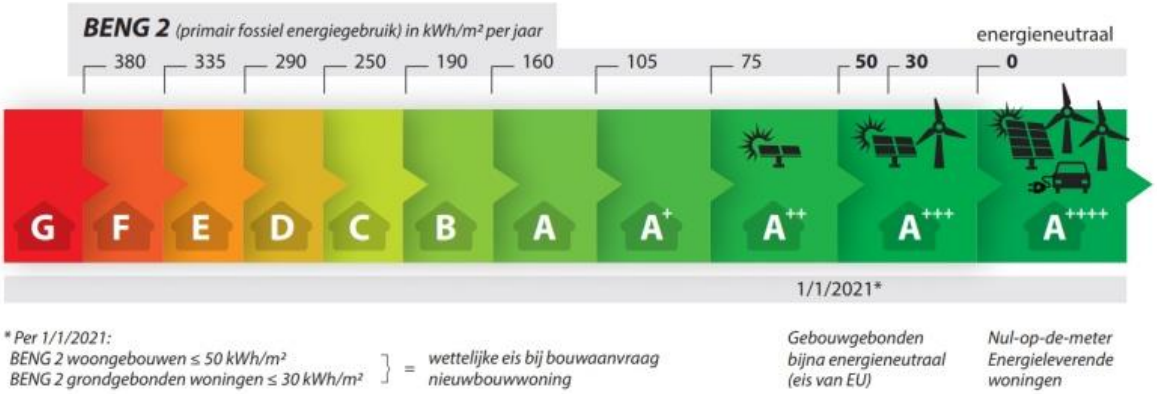
These findings emphasize the importance of considering energy efficiency and labeling when determining the value of residential properties. They also highlight the significance of energy labels in the real estate market and the potential benefits for various stakeholders. Taking energy label into account in the decision making process is crucial due to the effect it has on the property value, Investors may increase their portfolio value, mortgages providers may offer better terms to homeowners with higher energy labels, and buyers may be willing to pay money for properties with higher energy efficiency rather than buying properties with low energy efficiency .

A valuable area for further investigation would be to explore the energy transition cost and its impact on potential savings associated with improving the energy efficiency of residential properties. This could involve analyzing the costs associated with implementing energy-efficient measures, such as insulation upgrades, energy-efficient appliances, or renewable energy systems, and assessing the potential long-term savings in terms of reduced energy consumption and utility bills. This information would be beneficial for homeowners, mortgages providers, investors and other stakeholders.

Appendix

Appendix A: Energy Labels

The below image illustrates the energy labels classes for all building types (residential or not) based on the new following method NTA 8800 since 1-january-2021. This method uses the total fossil fuel energy that a building needs fir heating, cooling, hot water and ventilation to determine the energy label of the property (Woonbewust.nl, 2022).



Appendix B: Data Exploration

Variable Descriptions

City	The City where a home is located.
House Type	The type of the property, which can be a house or an apartment.
Price	The price in Euro.
Price m2	The price in Euro per square meter.
Living Area	The size of the property.
Energy label	The energy label class assigned to the property indicating its energy efficiency.
zip	Postcode.
Address	The complete address.
Year Built	The year when this property was constructed
House Age	The age of the property calculated based on the current year and the year it was built.
Date List	The date when the property was listed for sale.
Term Days	The number of days it took for the property to be sold.
Date Sold	The date when it was sold
Street	The street where the property is located
House Number	The property number.
Suffix	Additional address info.
Extra	Additional address info.
Month sold	Created for merging datasets, indicating the month when the property was sold.
Year Sold	Created for merging datasets, indicating the year when the property was sold.
Quarter Sold	Created for merging datasets, indicating the quarter (three-month period) when the property was sold.
Period	The specific quarter and year when the house was sold, used in conjunction with the price index.
Region	The city for which the price index is calculated.
Price Index	The House price index for the specific region and time period.

Correlation matrix

	Price	Living Area	House age
Price	1	0.639833	0.097294
Living Area	0.639833	1	-0.01271
House Age	0.097294	-0.01271	1

The number of data points per energy labels:

A	4642
B	2179
C	4185
D	3205
E	1890
F	991
G	740

Appendix C: Modeling

Linear regression results:

The table below displays the coefficients of the independent variables for the multiple linear regression model, where the dependent variable is the logarithm of the house price.

	log_house_price
Const	4.6639 (0.0469)
city_amsterdam	1.4714 (0.0108)
city_den-haag	1.0090 (0.0134)
city_rotterdam	0.9904 (0.0163)
city_utrecht	1.1932 (0.0155)
type_appartement	2.3296 (0.0224)
type_huis	2.3344 (0.0256)
price_index	0.0468 (0.0644)
log_house_area	0.8526 (0.0120)
log_house_age	0.0614 (0.0062)
label_A	0.7842 (0.0125)
label_B	0.6873 (0.0128)
label_C	0.6568 (0.0107)
label_D	0.6596 (0.6596)
label_E	0.6338 (0.0135)
label_F	0.6529 (0.0174)
label_G	0.5893 (0.0182)
R-squared	0.7000
R-squared Adj.	0.6989

Standard errors in parentheses.

Comparison between models:

	MSE	RMSE	R2 Score
Linear Regression	0.055	0.235	0.699
Polynomial Regression	0.04	0.201	0.781
Random Forest	0.035	0.186	0.812
LightGBM	0.032	0.178	0.827

Appendix D: Statistical Tests

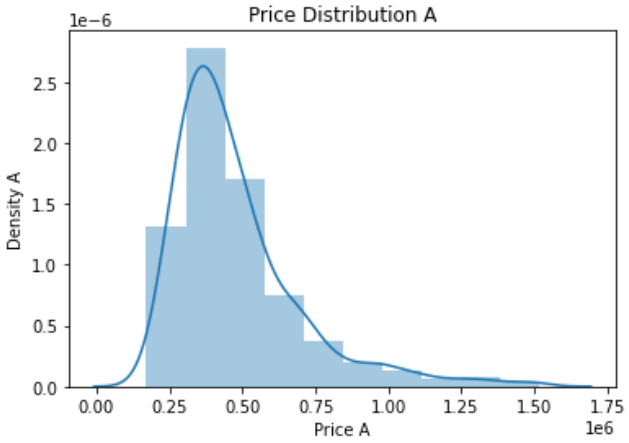
Checking the assumptions of the ANOVA test:

- We applied the Levene's test to assess the homogeneity of variance assumption:
  Based on the P-value: 0.313, the variances appear to be homogeneous across groups

- We conducted a normality test (Shapiro-Wilk test) for each group of energy label predictions, which yielded a very low p-value. As a result, we reject the null hypothesis that assumes a normal distribution

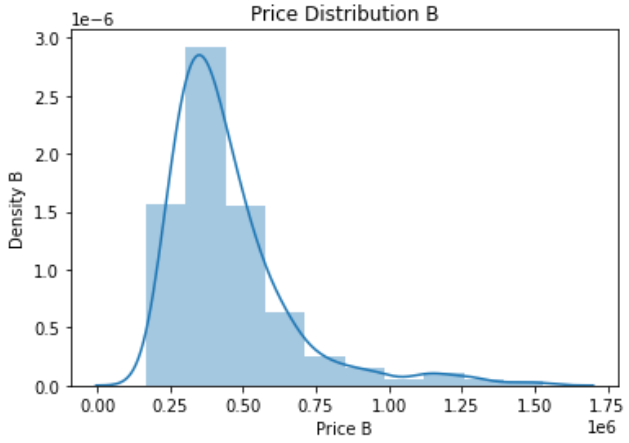
  Check normality by plotting the distribution of each group:

Group A:



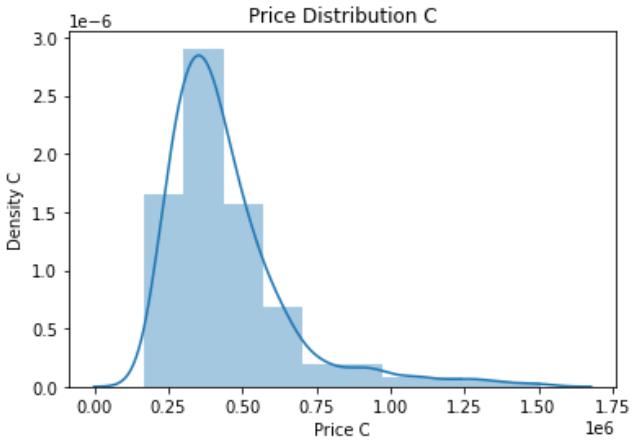
The distribution of house price predictions for Label A (Distribution 1)

Group B:



The distribution of house price predictions for Label B (Distribution 2)

Group C:



The distribution of house price predictions for Label C (Distribution 3)

The above distributions, along with the distribution of the labels D, E, F, G, have almost the same distribution shape and provide evidence for the violation of the normality assumption for the ANOVA test.

Kruskal Wallis test results:

Kruskal-Wallis Test - Statistic: 62.43362615570999
Kruskal-Wallis Test - p-value: 1.4396725420342192e-11= 0.00000000001439

The significant test statistic and extremely low p-value indicate robust evidence for the presence of significant differences in sell prices across the energy label groups.

Appendix E: The datasets and Code

The datasets that was used and the code that was written for this research can be found on GitHub using the following link:

https://github.com/DanyaMawed/Thesis-Project.git

References

Aroul, Ramya R. and Hansz, J. Andrew, The Value of 'Green:' Evidence from the First Mandatory Residential Green Building Program (December 9, 2010). Journal of Real Estate Research, Forthcoming, Available at SSRN: https://ssrn.com/abstract=1888778

Aydin, E., Correa, S.B. and Brounen, D. (2019) 'Energy performance certification and time on the market', *Journal of Environmental Economics and Management*, 98, p. 102270. doi:10.1016/j.jeem.2019.102270.

Brounen, D. and Kok, N. (2010) 'On the economics of energy labels in the Housing Market', *SSRN Electronic Journal*. doi:10.2139/ssrn.1611988.

Chien, W. (2023). Funda scraper (Version 0.0.3) [Library]. Available from https://pypi.org/project/funda-scraper/

Dielessen, M. (2023) 'Unveiling the Impact of Energy Labels on House Energy Consumption', Utrecht University, Applied Data science Master Thesis.

European Commission, Energy Performance of Buildings directive (2020) *Energy*. Available at: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en.

Energielabel Woningen (20-April-2023) *RVO.nl*. Available at: https://www.rvo.nl/onderwerpen/wetten-en-regels-gebouwen/energielabel-woningen

Fuerst, F. *et al.* (2015) 'Does energy efficiency matter to home-buyers? an investigation of EPC ratings and transaction prices in England', *Energy Economics*, 48, pp. 145–156. doi:10.1016/j.eneco.2014.12.012.

Glen, S. Sensitivity Analysis ("What-if"): Definition. Retrieved from StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/sensitivity-analysis/

Jermain, N. (2019) *Transforming skewed data for machine learning*, *Open Data Science - Your News Source for AI, Machine Learning & more*. Available at: https://opendatascience.com/transforming-skewed-data-for-machine-learning/

Ke G. et al. (2017)LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 2017:3146–54.

Kok, N. and Jennen, M. (2012) 'The impact of energy labels and accessibility on Office rents', *Energy Policy*, 46, pp. 489–497. doi:10.1016/j.enpol.2012.04.015.

Loucks, D.P. and Beek, E. van (2005) 'Model Sensitivity and Uncertainty Analysis', in *Water Resources Systems Planning and management: An introduction to methods, models and applications*. Paris, France: UNESCO, pp. 255–285.

Microsoft. LightGBM [GitHub repository]. Retrieved from https://github.com/microsoft/LightGBM

Pannell, D. (1997) ‘Sensitivity analysis of normative economic models: Theoretical Framework and practical strategies’, *Agricultural Economics*, 16(2), pp. 139–152. doi:10.1016/s0169-5150(96)01217-0.

Park, D. *et al.* (2015) ‘Analysis of a building energy efficiency certification system in Korea’, *Sustainability*, 7(12), pp. 16086–16107. doi:10.3390/su71215804.

Thabane, L. *et al.* (2013) ‘A tutorial on sensitivity analyses in clinical trials: The what, why, when and how’, *BMC Medical Research Methodology*, 13(1). doi:10.1186/1471-2288-13-92.

Truong, Q. *et al.* (2020) ‘Housing price prediction via Improved Machine Learning Techniques’, *Procedia Computer Science*, 174, pp. 433–442. doi:10.1016/j.procs.2020.06.111.

Woonbewust.nl (2022) *Het Nieuwe Energielabel voor Woningen. Wat Verandert Er in 2021?*, *Woonbewust*. Available at: https://woonbewust.nl/energielabel-woningen.