Graduate thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

Olav Andreas Lystrup
Lars Nore Moene-Omholt
Mats Esperum Nielsen
Sebastian Alexander Smith
Patrick André Eckbo Vikøren

Predicting Trondheim Housing Prices by Means of Supervised Learning

Graduate thesis in Business Administration, Business Analytics
Supervisor: Denis Becker
April 2024

NTNU
Norwegian University of
Science and Technology

Olav Andreas Lystrup
Lars Nore Moene-Omholt
Mats Esperum Nielsen
Sebastian Alexander Smith
Patrick André Eckbo Vikøren

Predicting Trondheim Housing Prices by Means of Supervised Learning

Graduate thesis in Business Administration, Business Analytics
Supervisor: Denis Becker
April 2024

Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

NTNU
Norwegian University of
Science and Technology

Preface

This bachelor thesis in the field of business analytics marks the end of a three-year bachelor education in Business Administration at NTNU Business School. Throughout our years in Trondheim our interest in the intersection between economics and data science has grown significantly. The spark was lit already in the third semester, when we took the introductory subject in information technology. Thanks to the course “Essentials of business analytics” we acquired knowledge in how to implement data science to financial problems. This subject was for us the most intriguing during our time at NTNU Business School. Therefore, most of us will pursue a master's degree in business analytics.

We would like to thank our supervisor, Denis Becker, for great support and help before and during the writing of our bachelor thesis. In addition, we would like to thank him for introducing us to the subject of business analytics. With his interesting lectures and great availability, our preceding year in Trondheim has been quite informative.

In addition, we would like to express our gratitude towards “Heimdal Eiendomsmegling” and “Eiendomsverdi” for giving us access to data. This is the most crucial aspect of our thesis, and without them we would not have been able to write about this topic.

Finally, we will include a disclaimer that everything in this thesis is the authors' interpretation and may differ from others' interpretations. We also stand by what is written in the thesis.

Abstract

For young adults about to enter the housing market, being able to predict housing prices can be very valuable. Attempts have been made to do this in many different cities. In this text we attempt to create the best possible model for predicting housing prices in Trondheim based on several variables, and whether a newly graduated MBA student would be able to afford a property that meets some important criteria. The models we have created are singular and multiple linear regression models, as well as decision tree regression and random forest regression models. They were trained on housing data from Trondheim, for example variables regarding different features of the properties. As well as other macroeconomic figures, for example housing index and policy rate, from the last three years. A number of different models were utilized, and a number of different hyperparameters were used as well to tune the decision tree models to attempt to achieve a better outcome. After trying to make a more effective model and combating overfitting, while still aiming to keep the accuracy high, our findings were that using random forest regression with pre pruning gave the best results and that usable area was by far the most important variable when assessing price. Finally, we approximated what a newly graduated MBA student would have to save to afford an example property.

Sammendrag

For unge voksne på vei inn på boligmarkedet, kan det være svært verdifullt å kunne predikere boligpriser. Det er gjort forsøk på dette i mange forskjellige byer. I denne teksten forsøker vi å skape modeller for å predikere boligpriser i Trondheim basert på en rekke variabler, og hvorvidt en nyutdannet siviløkonom vil ha råd til en eiendom som møter visse kriterier. Modellene vi har laget er en enkel og en multippel lineær regresjonsmodell, samt en simpel beslutningstre-regresjonsmodell og en Random Forest regresjonsmodell. Disse var trent på data fra boligmarkedet i Trondheim, som inkluderer en rekke variabler knyttet til diverse aspekter ved eiendommen, samt annen makroøkonomisk data, som for eksempel boligprisindeks og foliorente, fra de tre siste årene. En rekke forskjellige modeller ble brukt, og en rekke forskjellige hyperparametere ble også brukt for å justere beslutningstremodellene for å forsøke å oppnå et bedre resultat. Vi forsøkte å lage en mer effektiv modell og bekjempe overtilpasning, mens vi fortsatt ville holde nøyaktigheten høy. Da fant vi ut at forhåndsbeskjært Random Tree-regresjon ga det beste resultatet og at bruksareal var den desidert viktigste faktoren for pris. Til slutt fant vi ut hvor mye en nyutdannet siviløkonom må spare opp for å ha råd til en eksempeleiendom.

Table of Contents

1 Introduction 1
2 Theory and Background 2
2.1 Literature 2
2.2 The housing market 3
2.3 Lending regulations 5
3 Data and Method 7
3.1 The data 7
3.1.1 Data Gathering 7
3.1.2 Exploratory and descriptive analysis 10
3.2 Method 13
3.2.1 Linear regression 13
3.2.2 Decision tree regression 14
4 Results and Discussion 24
4.1 Results 24
4.1.1 Linear regression 24
4.1.2 Decision Tree Regression 25
4.2 Discussion of results 32
5 Conclusion 35
Reference List 36

List of Figures

Figure 2.1: Graph of the policy rate in Norway over time (Norges Bank, 2024-2) 4
Figure 2.2: Graph of Housing index over time (SSB, 2024) 4
Figure 3.1: number of data points at each respective variable value 10
Figure 3.2: Amount of residents of a given type. 11
Figure 3.3: Box plot of the distribution of housing prices. 12
Figure 3.4: Area of where the residents are in Trondheim. The bigger the circle, the more houses are sold within that area. 12
Figure 3.5: Example of an input space and corresponding decision tree for data with two features: X_1 and X_2. R represents the regions in the input space and the leaves of the tree. 16
Figure 3.6: A visualization of the score between the training and test set for each model given an alpha. 19
Figure 3.7: Mean accuracy from cross validation for each value of alpha. 20
Figure 3.8: A visualization of the MCCP Decision Tree based on model 0. 21
Figure 3.9: A visualization of the MCCP Decision Tree based on model 1. 21
Figure 4.1: A scatterplot with a regression line of the Total price at last sale and the Usable area variable. 24
Figure 4.2: Scatter plot of actual and predicted prices. Model score 0,69. 25
Figure 4.3: Finding the most influential variables for predicting the price of a resident in Trondheim. 25
Figure 4.4: New most influential variables for predicting the price of a resident in Trondheim. 26
Figure 4.5: A scatter plot of actual and predicted prices of a resident in Trondheim. 27
Figure 4.6: Plot showing the distribution of residuals in our model. 27
Figure 4.7: New most influential variables for predicting the price of a resident in Trondheim. 28
Figure 4.8: A scatter plot of actual and predicted prices of a resident in Trondheim. 29
Figure 4.9: Plot showing the distribution of residuals in our model. 29
Figure 4.10: A scatter plot of actual and predicted price of a resident in Trondheim. Model score 0,8479. 31
Figure 4.11: Plot showing the distribution of residuals of the RFR model. 32

List of Tables

Table 3.1: Information within the data set used. 10
Table 4.1: Model score of five different RFR models. 30

1 Introduction

Purchasing a house or an apartment is one of, if not the biggest purchases people make in their life. It is especially important for young people trying to get into the real estate market. In 2022, almost 80% of the Norwegian population owned their own residence, while 90% owned a home in the course of their life (Norges Eiendomsmeglerforbund, 2022). In addition to making a model, we are comparing the starting salary for a newly graduated MBA student with a prediction from the best model. This will give an MBA student a pointer on how much they will need in equity.

In this thesis we are going to develop several different models to predict housing prices in Trondheim, and then compare the price with what a newly graduated MBA student would be able to afford. Firstly, we will include a simple and a multiple linear regression model which we will use as grounds for comparison with the machine learning model we will focus on in this thesis, decision trees. Our approach will include training of the models using our data to make the models accurate and our predictions about the housing market as good as possible. Models like these can be very useful for first time home buyers, real estate investors, and real estate brokers. Therefore, our research question is as follows:

"Predict property prices in Trondheim using regression, and determine how much a newly graduated MBA student would have to save to be able to afford a desired property."

Our initial focus will be going through the literature, to map out the methods we are going to use and explain our choices. Then we will present the theory about the Norwegian housing market as a whole, and more specifically the Trondheim housing market. We will also include calculations at which price level a newly graduated MBA student can afford a residence.

Furthermore, we will explain our data and how we prepared it for analysis. Moreover, we will give a methodical introduction on the topics linear regression and decision tree regression. The results of every model will be presented and discussed to determine which explains our data the best. Finally, we will provide our conclusion.

2 Theory and Background

2.1 Literature

Machine learning has become one of the more popular forms of how the creation of AI and data analyses are done in recent years (Karjian, 2023). It is a way to handle large datasets to perform analyses and estimate possible outcomes. There are many different types of machine learning methods that are available and prudent for different types of data. We made a choice that we wanted to use supervised learning. Methods considered were neural networks and various forms of regression. We took a look at several scientific papers to determine the most suitable method to predict the housing prices in Trondheim. Hoxha (2024) takes different machine learning methods and attempts to find out which one is the superior model when predicting the housing prices in the town of Prishtina. They used four different types of machine learning, linear regression, decision tree, k- nearest neighbours and support vector regression. The decision tree regressor showed the lowest root mean squared error (RMSE) and mean absolute error (MAE), as well as the best coefficient of determination (R-squared).

A paper published by Cornell university has applied several different machine learning models to the housing market in search of a machine learning model that is able to predict housing prices based on specific input parameters. The models used were linear regression, decision trees, random forest, supporting vector regression, as well as gradient boosting. Their results were similar to the findings in the other papers that we have reviewed. The Random Forest model performed very well, with the joint highest R-squared score and one of the lowest RMSE and MAE scores. (Jha, S.B. et al, 2006, p.18).

Given the findings in these papers, we have chosen to use machine learning in our bachelor thesis, specifically linear regression, decision tree regression and random forest regression to predict property prices in Trondheim.

2.2 The housing market

The housing market in general is affected by several different factors. In this analysis we will take a closer look at the housing market both in Norway and Trondheim, and the factors that may affect them. Trondheim is the city in Norway with the highest percentage of students compared to the total population. With a total of 37 000 students, it ranks second in terms of student population and has a student percentage of 17,24% (Trondheim kommune, 2022).
In the last 10 years the housing market, both in Norway as a whole and Trondheim, has seen substantial growth and improvement. The price per square meter in Norway the past 10 years has gone from 32 995 to 52 591, showing an increase of 59,39%, while the price per square meter in Trondheim as grown 33 404 to 47 875 which is a growth of 43, 3% (Krogsveen, 2024). The housing price in Trondheim has comparatively had a slower growth than the rest of the country. Why has it changed so much in the last 10 years? One major reason could be the change in the policy rate, which is set by the Bank of Norway (Norges Bank, 2024-1).

The housing price index serves as a reliable metric to compare the different types of houses in Norway and Trondheim. Over the past decade, the average increase in the house price index in Norway has been at 4,9. Apartments showed the highest increase with a growth of 5,93, followed by single family homes with an increase of 4,47 and complex housing with an increase of 4,86. There have been some fluctuations, but in general the growth has been steady with notable exceptions like 2021 where the general index had a 12% growth, more than double the average increase. This coincides with the central bank's choice to lower the policy rate to 0, the lowest it has ever been as we can see in figure 2.1. The reason was to minimize the negative effect the corona pandemic had on the economy (Norges Bank, 2020). This made it a lot easier for house buyers to get loans, as banks were able to borrow more money from the central bank at a lower rate and subsequently were able to give lower rates to the homebuyers. On the other hand in 2023, the policy rate was the highest it has been in the last 10 years, coinciding with a period of low or even negative growth in the housing price index from 2022 to 2023 as seen in *Figure 2.2*.

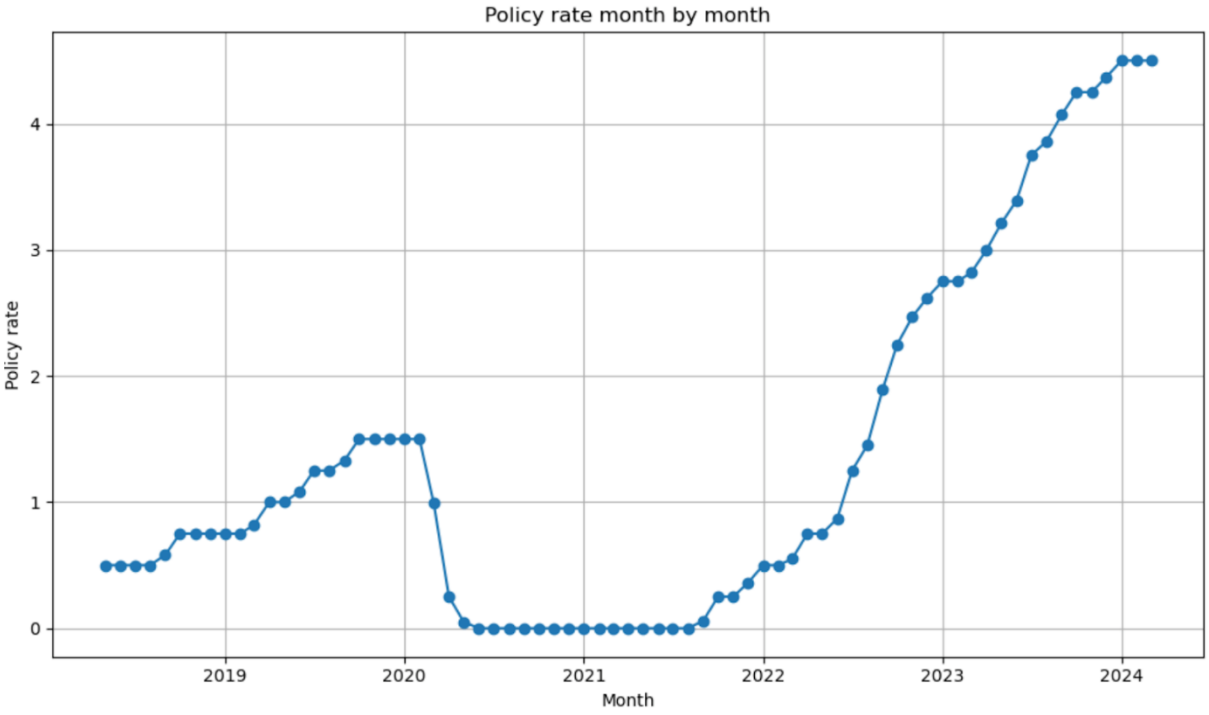


Figure 2.1: Graph of the policy rate in Norway over time (Norges Bank, 2024-2)

Figure 2.2: Graph of Housing index over time (SSB, 2024)

The housing price index in Trondheim can give us a similar impression of the housing market. The index is usually following the index of the rest of the country, it has a lower than average growth at 4,38 and some differences when it comes to the housing types. In Trondheim, single-family homes have the biggest average increase, while complex housing is second with

apartments having the smallest average increase. Similarly, Trondheim experienced a huge increase to 12,26 in 2021. As well as a very low, or even negative growth in 2023.

In general, we can say that the housing market in Norway and Trondheim are following the same trajectory, however the housing price index in Trondheim is generally a bit lower. Nevertheless, we can see some differences when it comes to the changes in the index with specific types of houses. Compared to Trondheim apartments have the biggest average increase in Oslo, in reference to the house price index. When changes happen on a macroeconomic-level, Norway as a whole will always be affected, which in turn affects Trondheim, one of the major factors being the policy rate, set by the central bank (SSB, 2011). This affects the interest rates of the bank loans, and subsequently how affordable mortgages are. If people cannot afford to borrow money, they probably cannot afford a property. Supply will then outgrow the demand, and housing prices will likely drop (Norges Bank, 2024-1).

2.3 Lending regulations

When buying a residence, the majority of people need to apply for a loan (SSB, 2022). Most banks go by the same rules when deciding on the loan amount. This is due to the lending regulations from 2021. First of all, you need 15% of the total value of the residence in equity, in other words you can only lend 85% of the value. Secondly you cannot have a debt ratio higher than 500% (5 times your yearly salary). This includes all previous debts. The final step is to do an interest rate stress test. Here you divide your ability to pay with whichever is highest of 7% interest rate or the borrowers interest rate added three percent points. The meaning of the stress test is to make sure that the borrower has a robust liquidity. (Gundekjøn and Kristensen, 2021, p. 351).

According to Econa the start salary for a newly graduated MBA student is 591 000, this means that the debt ratio is 500% when you borrow 591 000 * 5 = 2 955 000 NOK (Econa, 2024). Assuming 5 years of student loans is 410 000 NOK, you are able to borrow 2 545 000 NOK (Lånekassen, 2023). In many cases the first residence is bought with a partner, to make the equation simple we double the maximal loan amount, such that the limit is 4 890 000 NOK.

When it comes to equity the bank demands 15%. With a loan of 4 890 000 NOK the residence cannot be bought for more than 4 890 000 / 0,85 = 5 750 000. 15% of this sum is around 860 000 NOK. If you are planning on buying a residence alone you can afford 2 445 000 / 0,85 = 2 870 000, which means you need 430 000 NOK in equity. You can of course use more equity and afford a more expensive residence, but we assume that most students have a smaller amount of equity.

3 Data and Method

Within machine learning and predictive modeling we use linear regression as well as decision tree regression (DTR) and random forest regression (RFR). Each of these methods offer distinctive approaches to anticipate results based on input data. Linear regression looks at linear relationships between two variables to predict one based on the other. DTR makes sequential binary choices to split the data into distinct regions which consequently assigns a learned continuous value. RFR has the capability of improving prediction accuracy by combining the predictions of several decision trees which are trained on a random subset of the data.

First, we will describe how we gathered and processed it to make it applicable for data analysis, for then to visualize and illustrate relationships within the data. Further on we are going to take a deeper look into the world of machine learning and DTR to give a more detailed look at how it and the different models work, as well as looking at how we have applied these methods.

3.1 The data

3.1.1 Data Gathering

We acquired our data from “Heimdal Eiendomsmegling” who in turn uses “Eiendomsverdi” as their data source. The company has collected data for over 20 years, all over Norway. They only deliver their services to businesses, therefore we had to use “Heimdal Eiendomsmegling” as a third party to gather the data. The data we got includes information about every house and apartment sold in the Trondheim region from the 1st of March 2021 to the 1st of March 2024. All the variables, with description, are included in *Table 3.1* below.

In addition to the housing data, we included macroeconomic data from Norges Bank (2024-2) and SSB (2024), such as consumer price index (CPI) and policy rate from the last three years. These variables will expectantly have an influence on the prediction. We built a separate file with these macroeconomic values manually, which we then imported and merged to our final cleaned dataset. Our source, SSB, was missing housing price index and borrowing rate data for January and February 2024. To mitigate this we extrapolated these future values based on the values in December 2023 when we set up the macroeconomic data file.

The data had to be cleaned of some columns we did not want, translating the column names we did want from Norwegian to English, as well as merging the policy rate, housing loan rate, housing price index and the consumer price index of every given month. We excluded some data-points we considered irrelevant, e.g. non pertinent housing types, and in addition data-points which had too many missing features.

Later we applied the corresponding coordinates to the postal codes, which we will put into a map later to show where the residences are located in the city of Trondheim.
Lastly, we created dummy-variables of the binary categorical variables, and label encoded the rest. Now we have one column for each of these two variables where every value is represented by a number. One means "yes" and zero means "no" on the dummies, while the Broker and Housing Type variables are label encoded as shown below.

Broker; 0 = Aktiv, 1 = DNB Eiendom, 2 = EIE Eiendomsmegling, 3 = Eiendomsmegler 1, 4 = Heimdal Eiendomsmegling, 5 = Krogsveen, 6 = Lokalmegleren, 7 = Meglerhuset Nylander, 8 = Others, 9 = Privatmegleren, 10 = Proaktiv, 11 = Proper.
Housing Type; 0 = Borettslag Enebolig, 1 = Borettslag Rekkehus, 2 = Borettslag Tomannsbolig, 3 = Borettslagsleilighet, 4 = Selveier Enebolig, 5 = Selveier Rekkehus, 6 = Selveier Tomannsbolig, 7 = Selveierleilighet

The values we are left with after cleaning are:

Variable name:	**Explanation:**	**Type:**
Matrikkel / Org anr	Explains the geographical location of the residential plot	String
Address	The address of the residence	String
Primary room	Represents the area of the residence people spend most of their time in. All living spaces	Integer
Usable area	Combination of primary rooms and secondary rooms	Integer

Gross area	The total area of the residence	Integer
Year of construction	Which year the building was constructed	Integer
Floor	On what floor the residence is located	Integer
Plot size	Area of residential plot in square meters	Float
Last sold	When the residence was last sold	Datetime
Price	Price of the residence at time of sale	Float
Joint Debt at last sale	Remaining debt for a given housing association	Float
Total Price at last sale	Price added with Joint Debt	Float
Number of Rooms	Total rooms in the residence	Integer
Number of Bedrooms	Number of bedrooms	Integer
Registered date	When the residence was put up for sale	Datetime
Turnover rate	How many days the residence was on the market	Integer
Postal code	Postal code of residence	Integer
Broker encoded	What broker where used to sell the residence	Integer
Housing type encoded	Type of housing	Integer
Balcony_Yes	Is there a balcony	Boolean
Parking_Yes	Access to a parking spot	Boolean
Where and when		
Lat	This is the latitude of the respective resident, part of the coordinates used	Float
Lon	This is the longitude of the respective resident, part of the coordinates used	Float
Year	Year sold	Integer
Month	Month sold	Integer
Day	Day sold	Integer
Macroeconomic		
Policy rate	The given policy rate in Norway at the time of selling	Float

CPI	Consumer price index. how much buy power does the consumer possess	Float
change CPI	Change in CPI from the last day	Float
HPI Norway	Housing price index for the whole of Norway	Float
HPI Trondheim	Housing price index for Trondheim	Float
Borrowing rate %	The interest rate at which every household has to pay for their mortgage.	Float

Table 3.1: Information within the data set used.

One problem with the data is that formatting in a way that leaves year, month, and day as three separate variables, makes the date relatively useless, however, having the date as one lone variable has not been fruitful, and it was done in this way. On the other hand, missing the date in itself may not really matter because of all the macroeconomic variables that have been included that are connected to the date.

3.1.2 Exploratory and descriptive analysis

To get a better view of our dataset, we performed an exploratory analysis. Firstly, we wanted to understand the distribution of the variables price, year of construction and primary room. As seen in *Figure 3.1*, the histogram of price, shows us a skewed distribution with most of the variables around 3 million NOK. There are some residences that got sold for over 20 million NOK. Onto the next histogram we see that there is a large amount of buildings from the 20th and 21st century. The primary room histogram shows us that most homes are under 100 square meters, which is reasonable since apartments constitute a large part of the housing market in Trondheim.

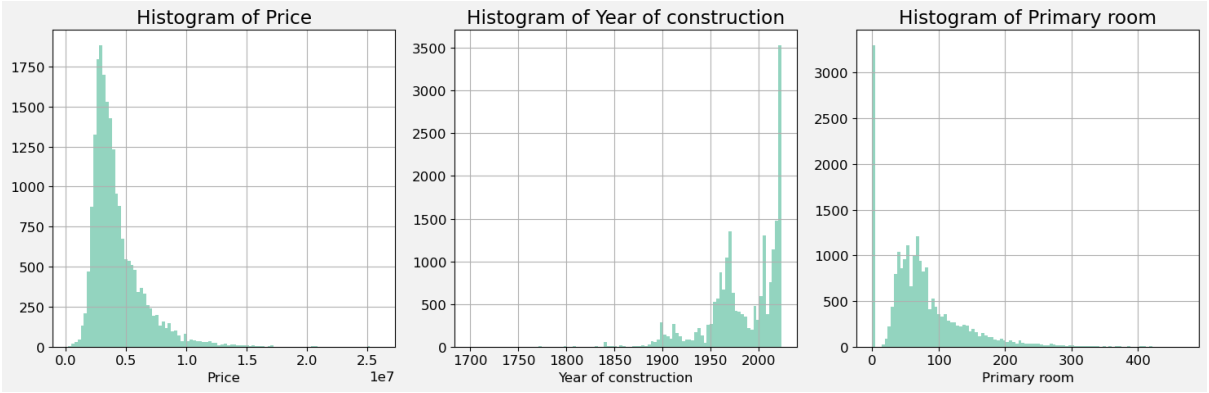


Figure 3.1: number of data points at each respective variable value

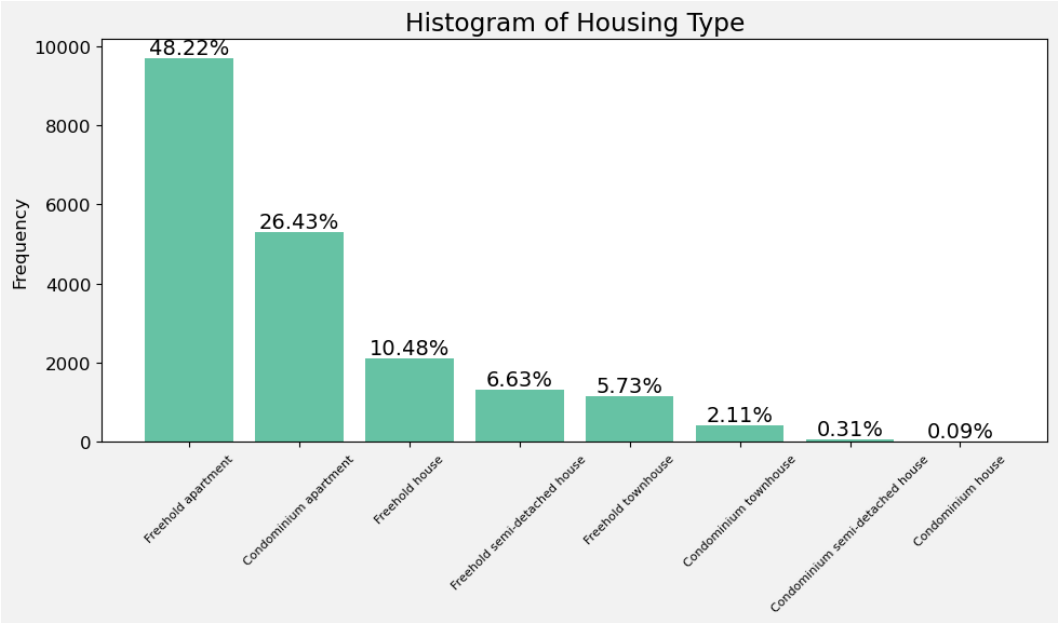


Figure 3.2: Amount of residents of a given type.

The two most frequent types of residence, observed from *Figure 3.2*, are freehold apartments and condominium apartments, which make up over 70% of all homes. In addition, our dataset is centered around the city center, with an expected overrepresentation of apartments.
To get a better understanding about how the prices in Trondheim vary, we made a box plot. Our visualization includes the maximum and minimum selling price in our dataset, which are respectively 26 million NOK and 240 000 NOK. The mean price is 4,2 million NOK and the median price is 3,62 million NOK. The data points above 9 million represent the outliers. 95% of the data points are in the interval between the black horizontal lines. The green box represents 50% of the data points, and we can see that 50% of all the residences have a price between approximately 3 and 5 million NOK.

Earlier we calculated the maximum price a newly graduated MBA student together with a partner can pay for a residence in Trondheim. From *Figure 3.3*, the amount of homes under the price of 5,75 million is 82,24%. This indicates that finding a residence probably would be a small problem, and that they have a wide selection to choose from.

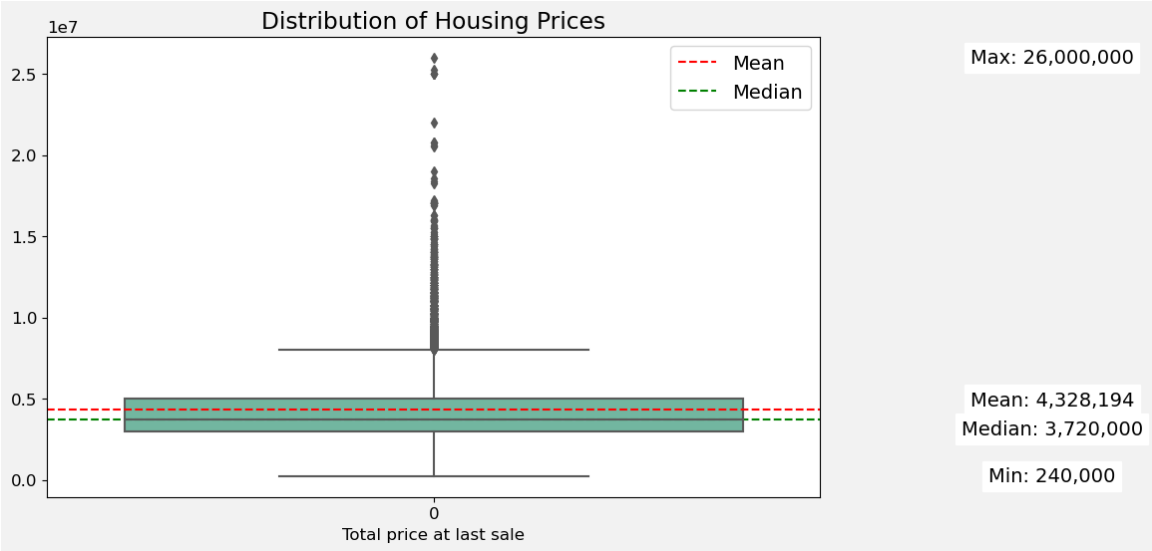


Figure 3.3: Box plot of the distribution of housing prices.

Finally, we plotted a heatmap over the map of Trondheim, *Figure 3.4*. Each circle represents a postal code, and the center of the circle is located in the center of the postal code. The size of the circle depends on how many residences that were sold in the area from 2021 to 2024. The color of the circle represents the mean price of every sold residence within the postal code. Most of these circles have a purple color which indicates a mean price of about 4 million NOK. There are also a few yellow circles around the city center and towards the forest at “Byåsen”. The cheapest areas are located towards the southern part of Trondheim, more precisely at “Tiller”.

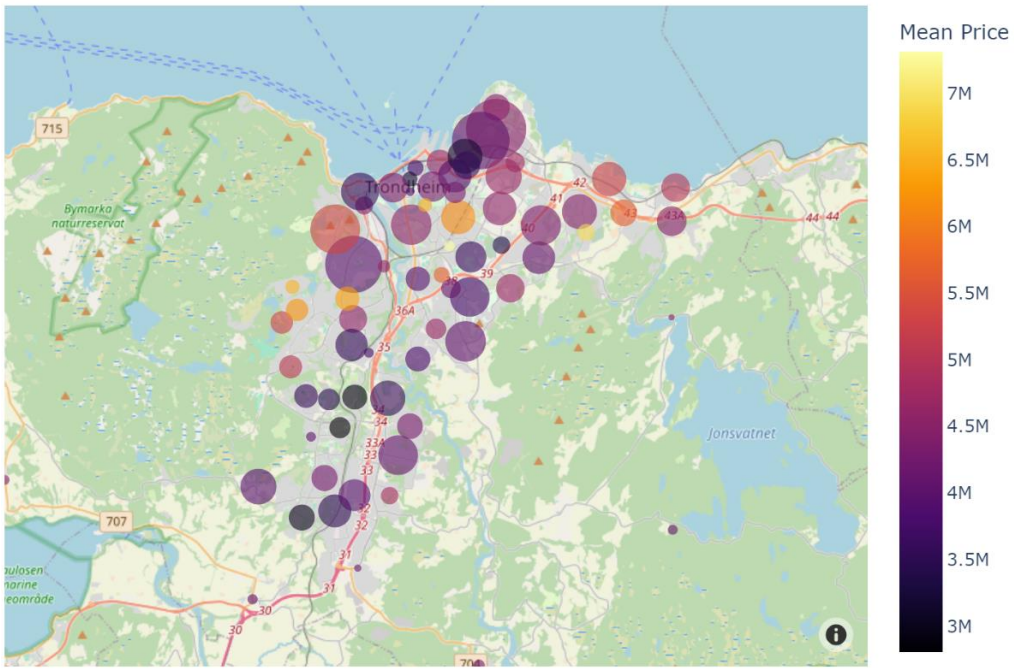


Figure 3.4: Area of where the residents are in Trondheim. The bigger the circle, the more houses are sold within that area.

3.2 Method

3.2.1 Linear regression

3.2.1.1 Theory

When analysing sets of data, linear regression is one of the easier and more popular methods, and its usefulness should not be overlooked. The fact that it is so simple means that it is not always the best fit for more complex data with correlations that are not linear. It does, however, give some indication of correlations and can be used as a starting point and as a benchmark for more complex models (James et. al., 2023, p.69). If they are not more accurate than a simple linear regression, there is either a flaw in the execution of the model or the specific model chosen.

Simple linear regression in itself is an attempt to find a linear correlation between two sets of data og values/variables, often represented on an X-axis and a Y-axis. There are a lot of different values that can be extrapolated from a linear regression analysis, but the most important values are the degrees of correlation (Pearson's R and R squared) and the slope of the regression line. A regression line is a line that represents the relationship between the two sets of data. If one assumes the correlation is completely linear, the line is a perfect representation. Pearson's R and R-squared are values that go from -1 to 1 and 0 to 1 respectively and show the degree of correlation. The closer the numbers are to zero, the less correlation there is. The slope of the correlation line represents the correlation. If the value of X goes one up, the value of Y goes down or up as much as the slope rises or falls (James et. al., 2023, p. 70-71). The slope is often calculated as the line that has the shortest distance to the actual values. This is the method of least squares. The slope is determined by the coefficients. The coefficients represent the change in price given a change of one in each variable.

When there are more variables than one that influence the values on the Y-axis, there is a need for multiple regression. This is because some of the variables may to some extent interfere with each other or cancel each other out. If this is the case, using singular regression analysis on the different predictor variables may give a misleading result. Two variables may on their own have a high correlation with the response variable, however, they may very well not have a unique explainability, making multiple regression analysis crucial to achieve a

result that reflects reality to a higher degree. A part of multiple regression analysis is then finding out which variables have the most influence, as well as their significance value (James et. al., 2023, p.80-81). A normal approach is to, one by one, remove the variables that are deemed insignificant until all that is left are significant variables (James et. al., 2023, p. 86-87).

Linear regression is a useful tool for finding simpler correlations, however, if the data is more complex, other prediction models can be more useful.

3.2.1.2 Methodology

Utilizing this theory in practice, two linear regression analyses were made, one singular and one multiple. The singular regression was done to see the correlation between total price at last sale and usable area. Firstly, the regression line was plotted on a scatter plot that shows the correlation between the two variables. Then, the summary of the regression was presented.

For the multiple regression analysis, the process was slightly more complicated. Firstly, the variables that were not integers or floats needed to be removed from the analysis. Secondly, the data was split into a training set and a test set at 70% of the data and 30% of the data, respectively. Our dependent "y"-variable is set to "Total price at last sale". The independent variables were every other variable in our data set except the dependent, as well as the variable "price". A multiple linear regression analysis was then executed with OpenAI's (2024) ChatGPT being asked to help with providing a framework for the code. After this, the goal was to remove the most insignificant variable, and then reevaluate the significance of the remaining variables. The process was repeated until the last insignificant feature had been removed. The insignificant variables were year, policy rate, turnover rate, change CPI, month, borrowing rate, gross area and day. The summary of the analysis was printed, interpreted, and plotted.

3.2.2 Decision tree regression

3.2.2.1 Theory

Decision trees are fundamental components of supervised machine learning algorithms which excel in breaking down complex prediction problems. A decision tree comprises hierarchical structures with nodes representing specific features, guiding the analysis through a series of sequential binary choices. Decision trees are versatile, serving both classification and

regression purposes, but because of the continuous nature of housing prices, we will be utilizing the decision trees' regression capabilities. (Breiman et al., 1984).

The basic idea behind decision tree regression is to recursively partition the input space into regions, and assigning a constant value to each region. Because the number of features quickly becomes more than three, the most efficient way of visualizing the partitioning of the input space is a hierarchical, top-down structure, not unlike an upside-down tree, hence the name. The tree consists of two components, internal nodes, and terminal nodes. The internal nodes split the data based on one of the features and a splitting criterion, they can be seen as creating the branches in the tree. The terminal nodes, or leaves of the tree, represent the region defined in the input space by the splitting criteria and contain the predicted continuous value. The predicted value is usually the mean or median of the target variable within each region in the partitioned input space. (Breiman et al., 1984).

Starting at the root node, the decision tree grows by selecting the feature that best partitions the data, using some splitting criterion, usually mean squared error (Scikit-Learn, 2024). The tree evaluates potential splitting points and compares the splitting criteria at each alternative, to find the minimum mean squared error. The chosen feature and corresponding threshold create branches, leading to the internal nodes. This process is repeated, generating the tree structure until a possible stopping criterion is met, culminating in a leaf node. (Breiman et al., 1984).

When making predictions, the decision tree guides the traversal from the root to a leaf node. At each internal node, a decision is made based on the feature value of the instance, directing it to the left or right child node, until the child is a leaf. The constant value associated with that leaf is used as the final prediction for the continuous outcome (Breiman et al., 1984). This process is illustrated in *Figure 3.5.*

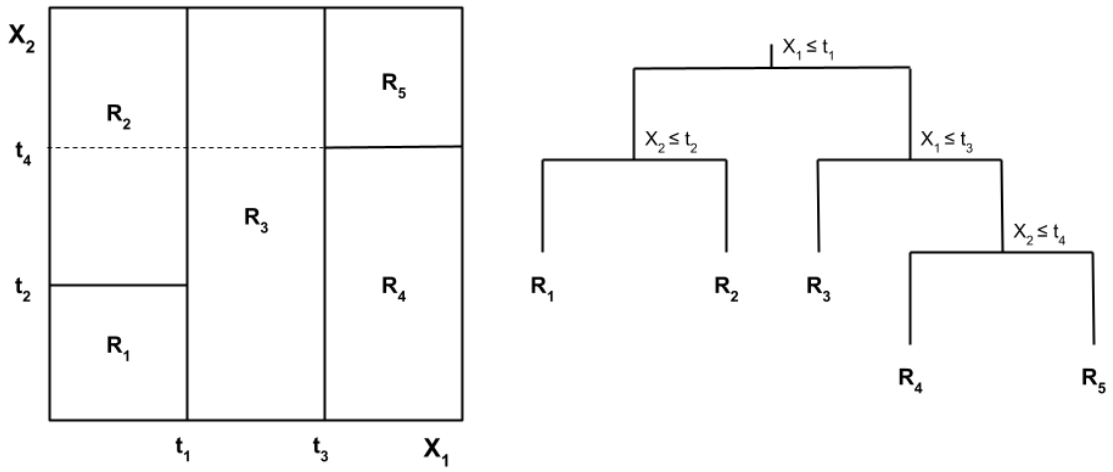


Figure 3.5: Example of an input space and corresponding decision tree for data with two features: X_1 and X_2. R represents the regions in the input space and the leaves of the tree.

Decision trees are a robust machine learning technique, and there are several advantages to using them, however they also come with a handful of disadvantages like any other method. Our primary concern is that decision trees do not output continuous values very well; it is in their nature to give rough estimates (the mean or median) for entire partitions in the input space (Scikit-Learn, 2024). Decision trees are in addition quite prone to overfitting, meaning that they do not necessarily generalize efficiently by themselves. Instead of capturing underlying patterns in the data, decision trees may memorize noise or outliers, resulting in overly complex models that perform well on the training data but fail to extrapolate effectively to new, unseen data (Scikit-Learn, 2024).

Despite the mentioned disadvantages of decision trees for regression, there are several advantages. One being that a DTR model can handle nonlinear relationships well. Another significant upside of decision trees is their inherent interpretability, the ability to visualize them makes it easy to follow the reasoning behind the model and to find potential sources of error. The characteristic of sequentially splitting the input space using binary conditions makes it simple for computers, and even humans to assign a value to an instance based on its features. (Scikit-Learn, 2024).

Another upside of decision trees is that they require minimal data preparation. Preprocessing with decision trees is often more straightforward and less demanding compared to several other machine learning methods. Because decision trees split data based on relative

differences rather than absolute values, they do not require scaling. Furthermore, some decision tree algorithms are able to handle missing values quite well during model training, often alleviating the necessity for manual intervention. Encoding of categorical values is an essential part of preprocessing for decision trees, typically achieved through dummy-variables and label encoding. (Scikit-Learn, 2024).

3.2.2.2 Simple Decision Tree for Regression

The first step in making any decision tree regressor is to split the data into training and test sets. Like described in the section on linear regression, we split the data into the dependent and independent variables and further into training and test sets with a 70/30 split.

As a benchmark for the various decision tree regressors we will make, a model is created without any intervention. This means that the training of the model is left entirely up to the implemented algorithm and its default configuration. This model will serve as a reference point for evaluating the performance of customized models so that we can assess their robustness and effectiveness.

Decision trees without intervention do not necessarily perform very well, the accuracy of their predictions might not be entirely precise, and they may overfit to the data. These problems can be mitigated through diverse optimization methods. The methods being pre-pruning, post-pruning and ensemble methods.

3.2.2.3 Pre-Pruned Decision Tree for Regression

One way is to specify some initial hyperparameters, or rules for how the tree is allowed to grow before initializing the tree, this is called pre pruning. These hyperparameters are the depth of the tree, the minimum samples in an internal node required for a split, and the minimum samples required for a node to be a leaf. By specifying these upon the initialization of the tree, the performance can be improved drastically (Scikit-Learn, 2024). We learned from OpenAi's ChatGPT that a method for finding the best initial hyperparameters is a grid search, where you test several different combinations of parameters and compare the accuracy of each potential model (OpenAI, 2024).

By means of grid search with cross validation, we find the best values for the maximum depth of the tree, the minimum samples required for a split, and the minimum samples required for a leaf node. Iterating over different combinations of these parameters with values 5, 10, 15, 20, 25, and 30, and evaluating the performance with five different validation subsets of the training data returns the best combination. The grid search reveals that a tree with a maximum depth of 15, minimum samples for a split of 20, and minimum samples for a leaf of 5 is optimal. With the implementation of these values for the hyperparameters, a decision tree regressor can be created, and consequently pre pruned, for a more robust and accurate model.

3.2.2.4 Post-Pruned Decision Tree for Regression

Another method for optimizing the performance of a decision tree after its creation is minimal cost-complexity pruning. Minimal cost-complexity pruning, hereby referred to as MCCP, is a technique which omits branches in the tree which do not significantly improve the performance of the model, this is achieved by penalizing trees with more branches. (Breiman et al., 1984).

After a tree has been fully grown, a set of subtrees, or pruning paths is determined. For each subtree, the cost complexity is calculated by summing the impurity (squared error) of each datapoint for that tree. A penalty term is added to the cost complexity of each subtree, this is the number of leaves in the subtree times a complexity parameter named alpha. Each subtree is allotted the value of alpha which minimizes its cost complexity, meaning that every subtree has a unique corresponding alpha. (Breiman et al., 1984).

Using cross validation, the best subtree can be found by iteratively scoring each tree multiple times with different training and validation subsets of the data. The value of alpha which gives the subtree that has the highest mean score, is the alpha that best prunes the tree to generalize the data. Using the method of MCCP with cross validation, a tree can be pruned so that it has higher accuracy when working with new, unseen data. (Breiman et al., 1984).

MCCP will be employed to refine the robustness and predictive capabilities of the two decision tree regressors previously configured. The first regressor was established with default settings, capturing the patterns in the data without manual intervention (model 0). The second was pre-pruned by explicitly defining its hyperparameters (model 1), by performing

MCCP we aim to optimize and enhance its performance and generalization capabilities further. The purpose of performing MCCP on both models is to determine whether pre pruning has a substantial impact on the resulting pruned regressor.

First, the two models' subtrees are defined and consequently their values for alpha. We find that model 0 comes with 11 859 different values of alpha, model 1 has 490. To save on computational costs we limit the alphas so that we are only checking those that result in a model with a substantial coefficient of determination, and those that do not result in an overly overfit model. We find the relevant interval by trial and error, resulting in 445 alphas for model 0, and 426 alphas for model 1. The coefficient of determination for each alpha for both models on the training and test sets can be seen in *Figure 3.6* below.

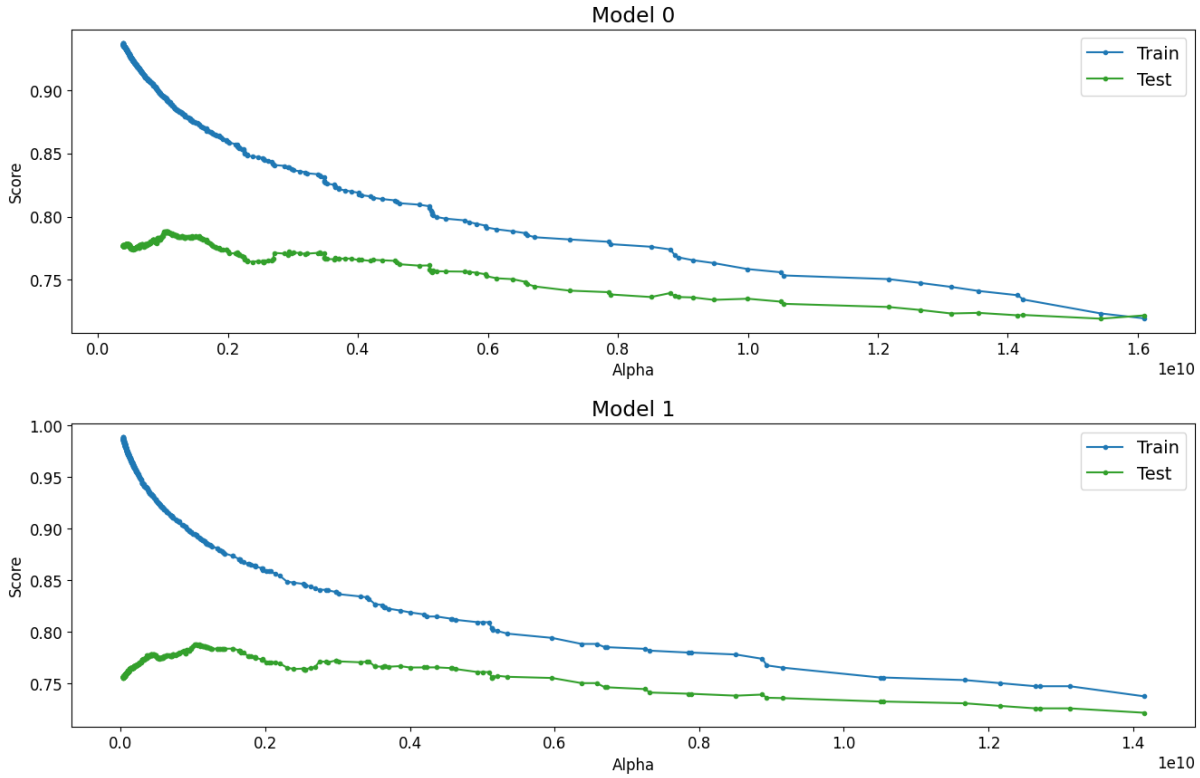


Figure 3.6: A visualization of the score between the training and test set for each model given an alpha.

Next, with the relevant ranges of alpha for both models defined, we employ four-fold cross validation to compute the average accuracy of each created model given an alpha. These accuracy scores and their corresponding alphas are put into data frames so we can find the alpha that defines a model which performs the best on average. The performance of each alpha for both models can be seen in *Figure 3.7* below.

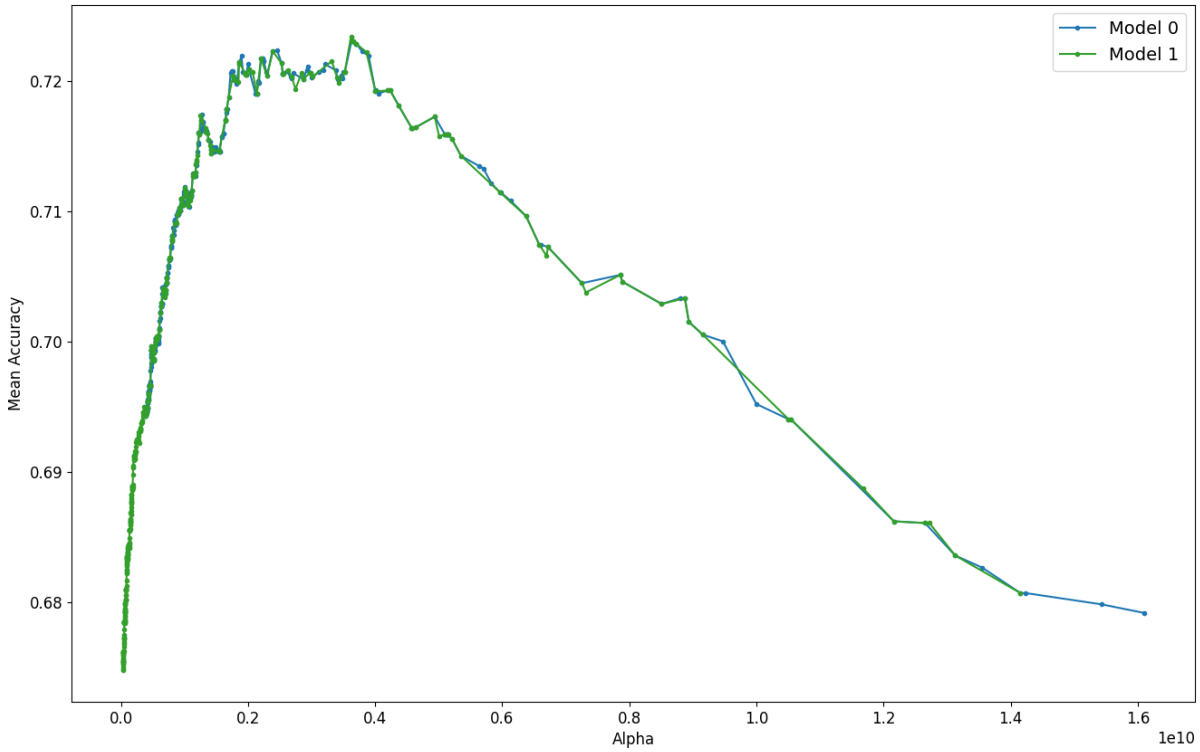


Figure 3.7: Mean accuracy from cross validation for each value of alpha.

By finding the highest mean accuracy in the data frame for each model, we can identify the alpha that produces the best regressor on average. This gives us an alpha of value 3632224380.81 for model 0, and 3620267843.29 for model 1. Using these alphas when creating decision tree regressors, generates a tree where several branches have been omitted, with the goal of generalizing and increasing model performance. The following *Figure 3.8 and 3.9,* show the resulting grown trees from MCCP for both models with their respective optimal values of alpha implemented.

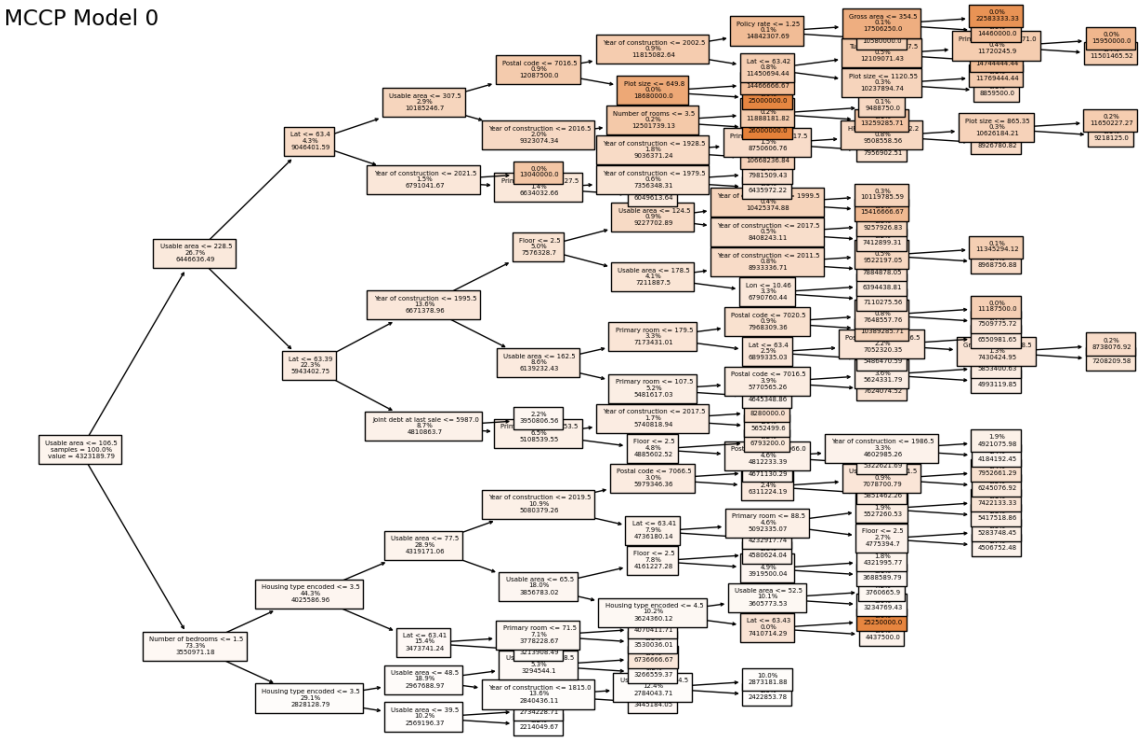


Figure 3.8: A visualization of the MCCP Decision Tree based on model 0.

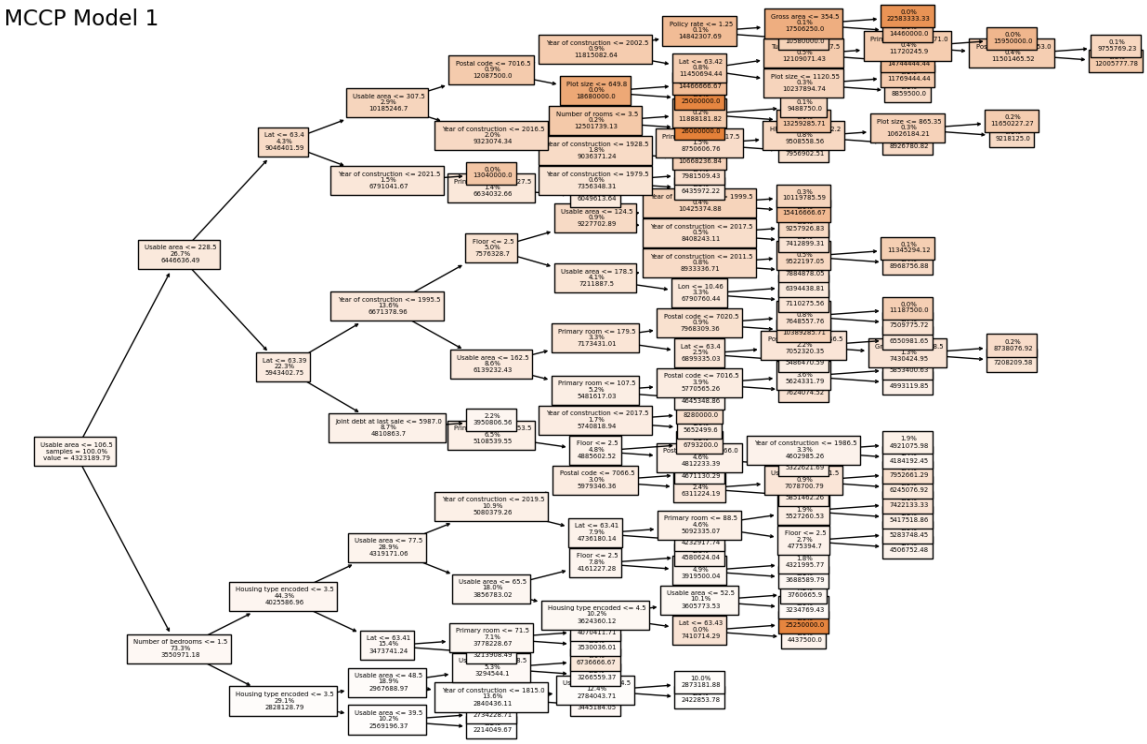


Figure 3.9: A visualization of the MCCP Decision Tree based on model 1.

Left un-pruned, the decision tree for model 0 would have been far too chaotic for us to understand and interpret. Hence, it is not visualized in this paper. There would have been too many optional branches for the model to consider and would not be effective in capturing the patterns in the data.

3.2.2.5 Random Forest Regressor

Ensemble methods are techniques which combine the predictions of multiple individual decision trees in a unified model, with the goal of improving robustness and predictive accuracy. Every decision tree in the ensemble is trained on a subset of the training data and it makes its own prediction. The final prediction of the ensemble is obtained by aggregating the predictions of all individual trees, usually by averaging or applying a weighted sum. Ensemble methods help diminish the limitations of individual decision trees, such as overfitting, by applying the predictions of multiple trees to its advantage. By their nature, ensemble methods are often capable of achieving higher predictive accuracy and generalization compared to individual decision trees. (James et al., 2023).

There are several different ensemble methods available for decision tree regression. One of the most recognized is random forest regression. Random forests combine multiple trees which are trained on a random subset of the data. The prediction for each individual tree is averaged and given as the final prediction (James et al., 2023). In our endeavor to construct the best possible model for our data, we have used this ensemble method.

To begin with, we load our already cleaned dataset the same way as the methods above. For us to be able to test our models performance and at the same time make sure it does not overfit we split the dataset into a train-set and a test-set.

Further on we tested different random forest regression models with different amounts of estimators. The amount of estimators represents the amount of trees made in the random forest. (Scikit-Learn, 2024). Our goal was to find the number of estimators where the model score did not improve much, to find the best model and to reduce the computational power needed. Therefore, we made five different models with 10, 20, 30, 40 and 50 trees which were scored and compared according to the coefficient of determination for the prediction (Scikit-Learn, 2024).

For further improvement we tested various maximum depths which sets a limit for the amount of nodes in the trees. If this parameter is at its default value, the trees will continue until each leaf node is pure or includes the minimum number of samples, which is 1 by default. We found that a maximum depth of 20 was the best.

In addition to manually pruning the model, we used the grid-search method with four-fold cross validation, with ChatGPT helping us to get started (OpenAI, 2024). The parameters we experimented with in this method was maximum depth (5, 7, 10), the amount of trees in the forest (60, 80, 100), the minimum number of samples required to split a node (3, 5, 7), the minimum number of samples to be a leaf node (3, 5, 7), and the maximum number of features to consider when finding the optimal split (14, 15, 16) (Scikit-Learn, 2024). From performing the grid search we found that the best estimator has a maximum depth of 10, 100 trees in the forest, the minimum samples for a split was 3, the minimum samples for a leaf was 3, and finally the maximum number of features was 15.

An attempt was also made to perform MCCP on the RFR. We manually set several values for alpha and evaluated cross validated model scores with three validation sets against each other. The best value for alpha turned out to be 0, which is the default value.

Finally, we created a scatter plot between actual and predicted price. This makes it easy to observe any outliers. In addition, we can observe if the model shows any sign of a pattern which can indicate bias. When the random forest method does not capture the actual relationship between the variables, we call it bias (StatQuest, 2018). We will also include a histogram of the residuals to get a view if the model is over- or underestimating our values. If the graph has an expectation value above or below 0, our model is predicting a different price than the actual price.

4 Results and Discussion

Throughout our analysis we have used different forms of machine learning, in this section we are going to look at how our models worked and evaluate how they performed compared to what we expected. R-squared and mean absolute error (MAE) are metrics used to evaluate the models. Finally leveraging our findings, we aim to determine the categories of housing and prices a newly graduated MBA student can afford.

4.1 Results

4.1.1 Linear regression

4.1.1.1 Singular Linear Regression

The singular linear regression analysis is somewhat insightful; however, it is severely lacking. The R-squared and R-squared adjusted are both 0.561, there is some correlation. This is supported by the plotted regression line that shows that there is some degree of correlation present. The P value of 0 shows that it is significant. The coefficient is approximately at 24 670. The low correlation, however, makes this of little value. *Figure 4.1* shows the semi-low correlation. The observations do not follow the regression line closely.

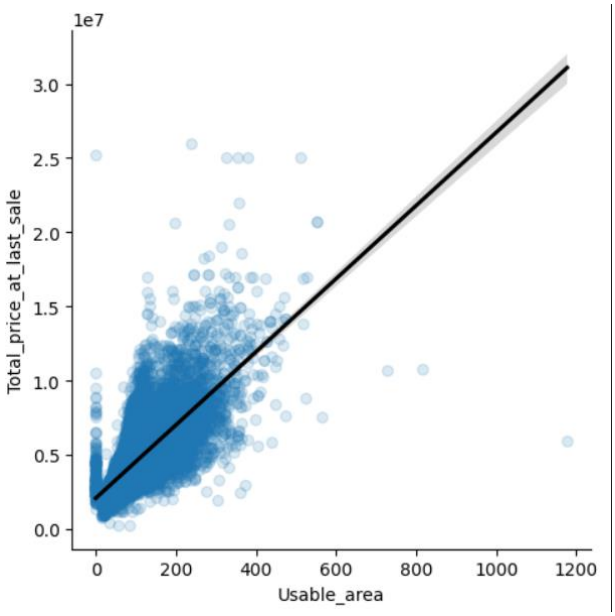


Figure 4.1: A scatterplot with a regression line of the Total price at last sale and the Usable area variable.

4.1.1.2 Multiple Linear Regression

The multiple regression analysis has more analytic value. Before removing the insignificant variables, the R-squared and the R-squared adjusted for the test set are 0.694 and 0.693. After removing the insignificant variables, the model has not improved. The R-squared and the R-squared adjusted are still 0.694 and 0.693. There is a certain degree of correlation between the variables and the total price at sale. The model is not terrible, but it is not good either, there is room for improvement. In *Figure 4.2* the observations follow the regression line more closely,

as the correlation is higher. This model will serve as a benchmark for the decision tree models we will use later on.

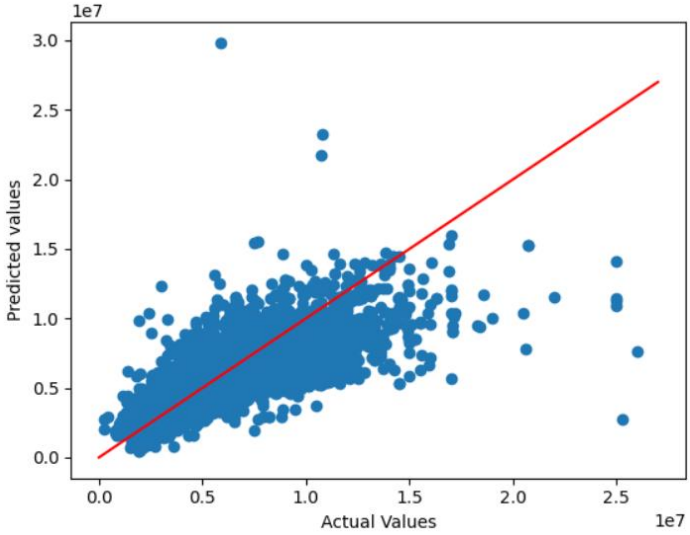


Figure 4.2: Scatter plot of actual and predicted prices. Model score 0,69.

4.1.2 Decision Tree Regression

4.1.2.1 Simple Decision Tree for Regression

Looking at the feature importances of this benchmark model, we find in *Figure 4.3*, that usable area is by far the most important variable, explaining more than half of the model decisions. Other significant variables are the latitude, number of bedrooms and year of construction of the residence. Although, most variables do not have a significant impact on how the model makes its decisions.

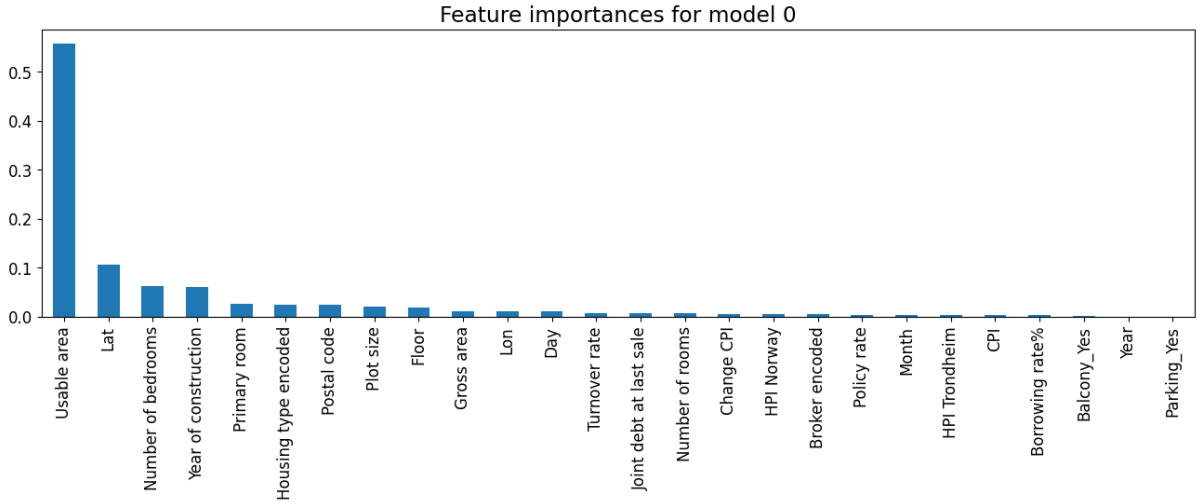


Figure 4.3: Finding the most influential variables for predicting the price of a resident in Trondheim.

When reviewing the simple benchmark model which was created without any intervention, we find that it does not perform very well. The coefficient of determination on the training set for this model is 0.997 and 0.749 on the test set, a strong indication that this model is overfitting to the training data.

Looking at the mean absolute error for the predictions of this model, we see that MAE for the training set is 8 111, and 556 060 for the test set. This tells us that this model is only off by 8 111 NOK on average when predicting the values it has been trained on, and 556 060 NOK on average when predicting unseen values. With these numbers in mind we can confidently state that this model is overfitting and is not adequate for our purposes, it will not be a good predictor for housing prices in Trondheim.

4.1.2.2 Pre-Pruned Decision Tree for Regression

For the pre-pruned regressor we find that the feature importances are in essence very similar to the benchmark model. Here the usable area is still the most explanatory variable, with the next most important being latitude, number of bedrooms, and the year of construction. One slight difference is that the features with less importance are closer to zero for this model. As shown in *Figure 4.4*.

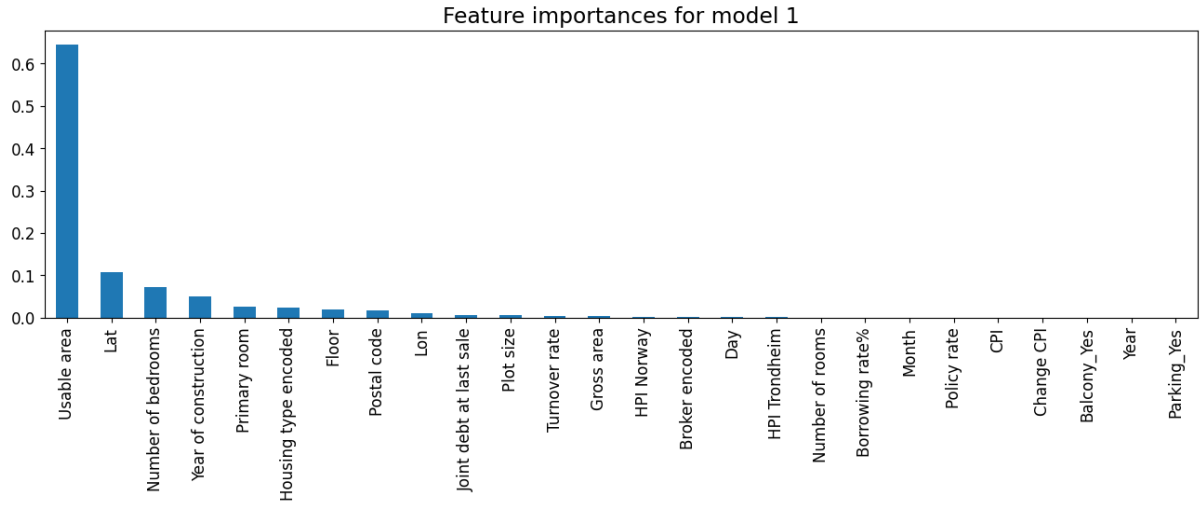


Figure 4.4: New most influential variables for predicting the price of a resident in Trondheim.

The coefficient of determination for this model on the training set is 0.849, and on the test set is 0.812. This is a welcomed improvement over the benchmark model, these values tell us that the pre pruned model probably is not overfitting. However, it does not predict with much confidence. By looking at *Figure 4.5* below, we can see the price the model predicted compared to what the price is supposed to be. The red line through the center represents what a perfect prediction would look like, the closer to this line each point in the scatterplot is, the better the prediction. As you can see, for the most part the predictions are quite close to the true values, though there is a good amount of noise and several outliers.

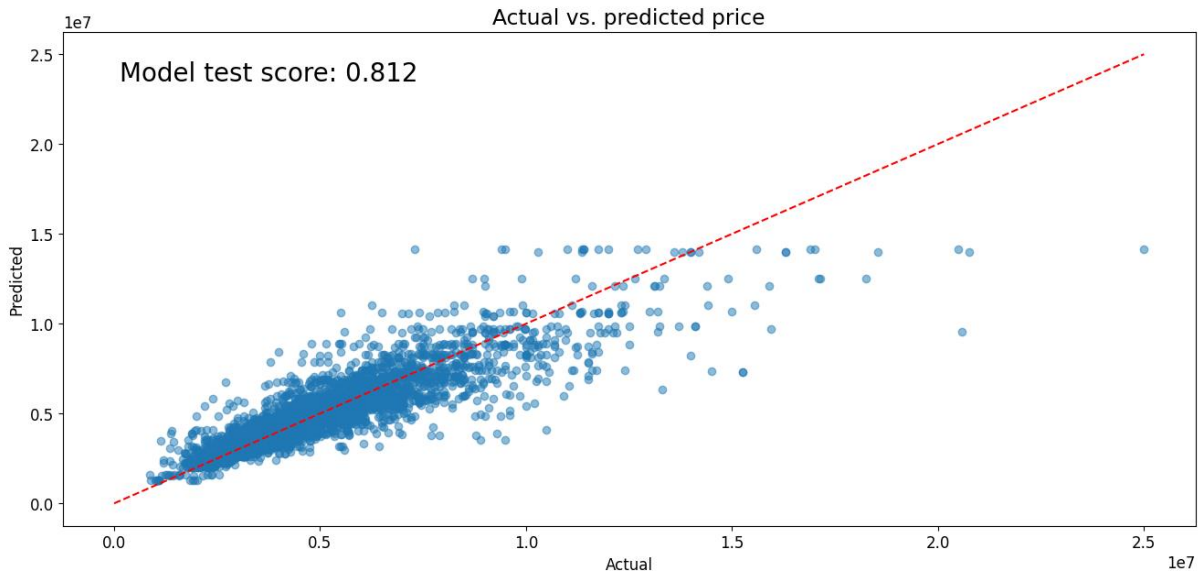


Figure 4.5: A scatter plot of actual and predicted prices of a resident in Trondheim.

In conjunction with R squared, we see that the MAE on both training and test sets are quite similar, being respectively 449 919 and 526 805. With this information we can tell that the pre pruned model is not overfitting much, but it is off by just under half a million NOK on average, which is not preferable.

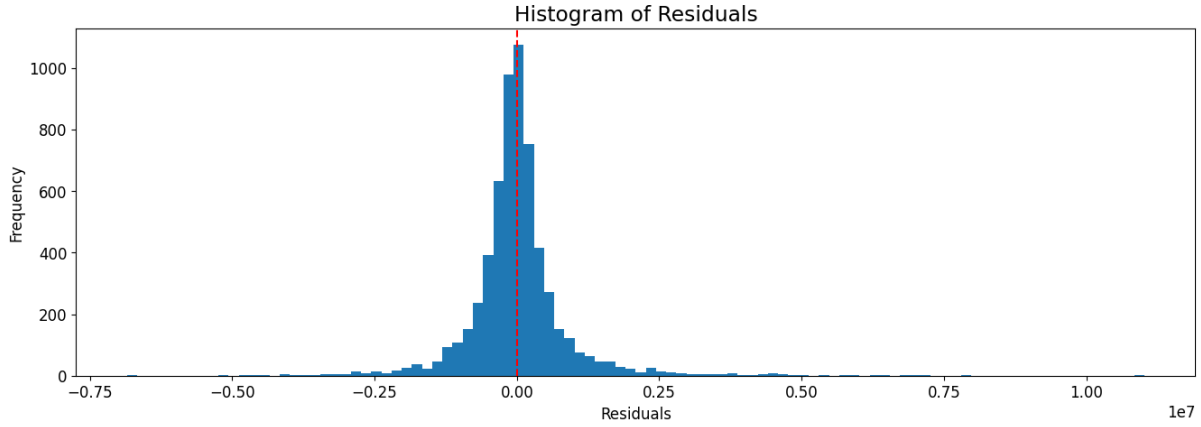


Figure 4.6: Plot showing the distribution of residuals in our model.

By looking at the histogram of residuals above, we learn that there is a substantial amount of uncertainty in the prediction of price. Although the residuals are primarily centered around zero, the distribution has a considerable amount of deviation, shown in *Figure 4.6.*

4.1.2.3 Post-Pruned Decision Tree for Regression

By looking at the feature importances for both models pruned with MCCP, we observe that the importances are in essence the same for both post pruned models, as shown in the *Figure*

4.7 below. In addition, they are quite similar to the importances for the model which was purely pre pruned.

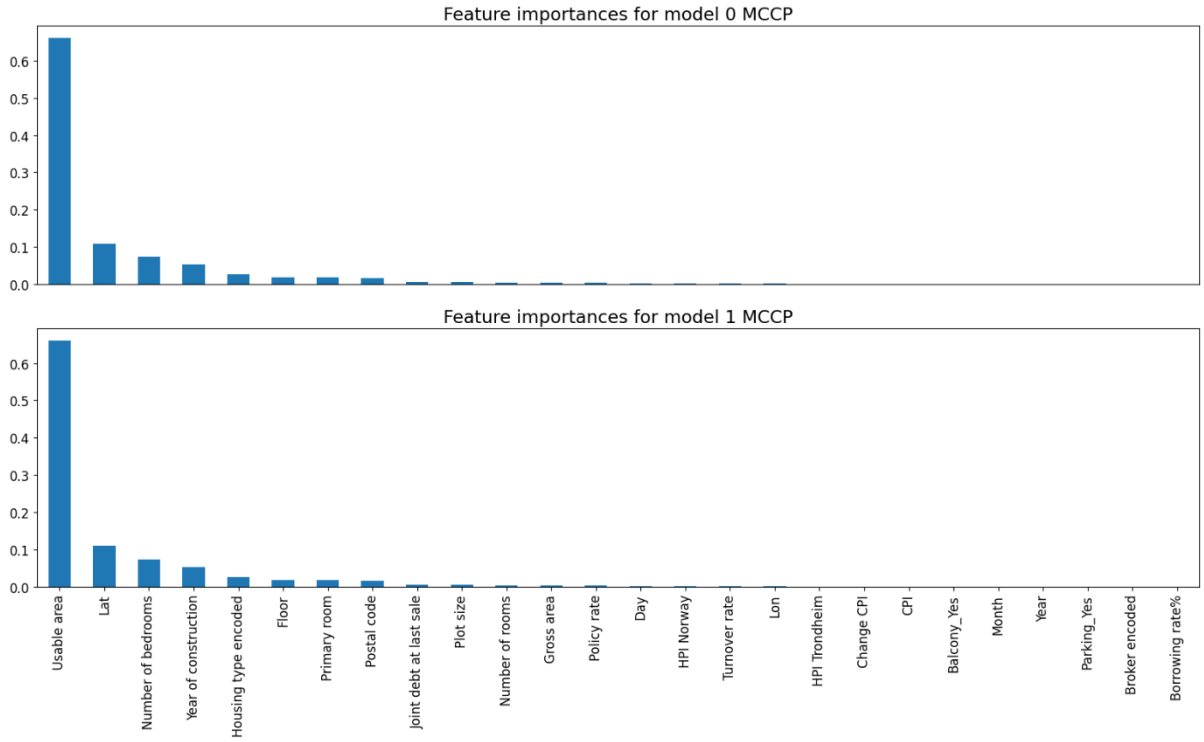


Figure 4.7: New most influential variables for predicting the price of a resident in Trondheim.

Using MCCP on model 0 resulted in an R squared on the training set of 0.8252 and 0.7663 on the test set, and for model 1, the R squared is 0.8261 on the training set and 0.7664 on the test set. This is illustrated in *Figure 4.8.* The MAE for model 0 is 574 751 on the training set and 631 460 on the testing set. For model 1 the MAE is 573 701 on the training set and 631 043. In other words, MCCP on the benchmark model and the pre pruned model yielded about the same results, with the model which was pre- and post-pruned performing marginally better. This could have been presumed when seeing the trees plotted in *Figure 3.8 and 3.9.* They seem to be overfitting, and their predictive accuracy is not satisfactory. By looking at the predicted prices by model 1 after undergoing MCCP compared to the actual prices in the figure below, we observe that the model is unable to predict prices with much certainty.

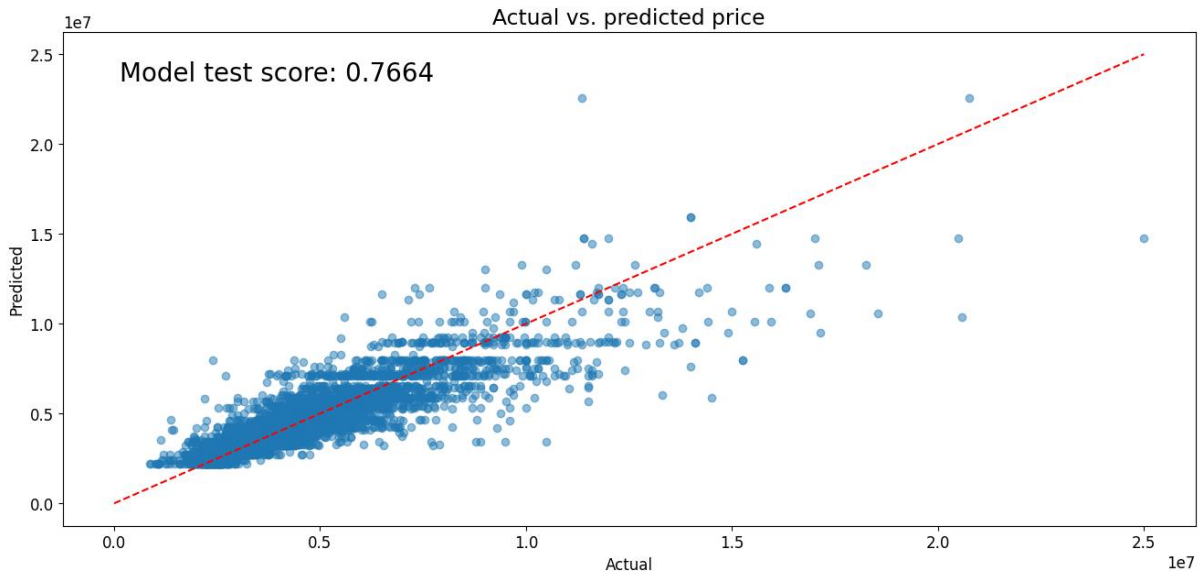


Figure 4.8: A scatter plot of actual and predicted prices of a resident in Trondheim.

From *Figure 4.9, w*e can see in the histogram of residuals for model 1, that though most of the residuals are close to zero, the deviation from the target is more significant than expected and preferred. Put differently, MCCP has not made the models perform significantly better, rather the opposite.

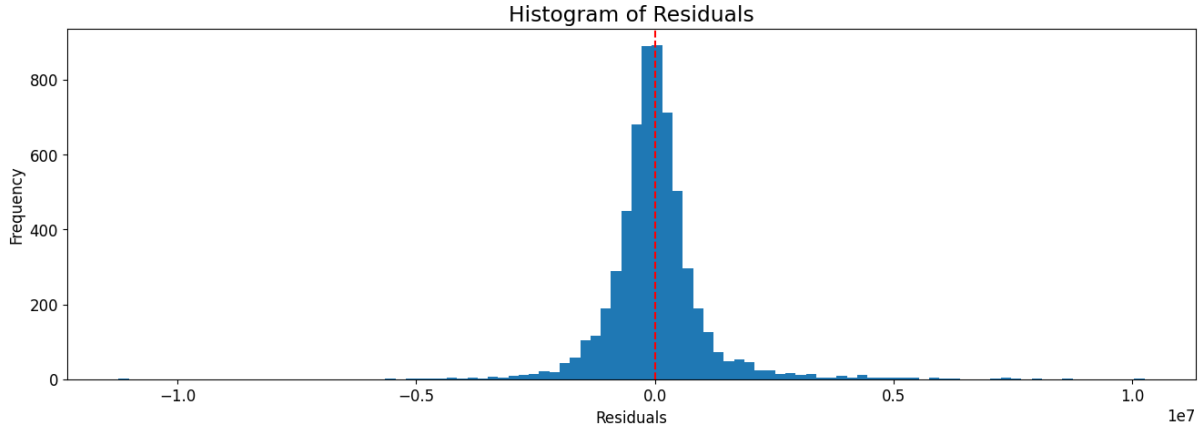


Figure 4.9: Plot showing the distribution of residuals in our model.

4.1.2.4 Random Forest Regressor

The same way as with the DTR models, we include the feature importances of the variables in every RFR model created. Here our results are almost identical to each other, and the results of the previous DTR models. The usable area is approximately five times as important as the second variable “latitude”. In other words, the usable area is most likely the most important factor for the price of a residence in general.

Next, we will discuss the R-squared on the test set for each of the five simple models with the number of estimators specified. In the *Table 4.1* below, we can see that the score does not improve drastically. Nevertheless, we see that model 4 has the best results.

Model 1 score (estimators = 10):	0.8526
Model 2 score (estimators = 20):	0.8602
Model 3 score (estimators = 30):	0.8631
Model 4 score (estimators = 40):	0.8647
Model 5 score (estimators = 50):	0.8640

Table 4.1: Model score of five different RFR models.

Looking at the score on the training score for model 4 however indicated that the model is overfitting, the score being 0.976. In an attempt to mitigate the overfitting of this model and increase accuracy, we set the maximum depth of the trees in the forest to 20. This increased the test score to 0.866, however the score on the training set remained high at 0.975. In addition, the MAE was 165 829 and 419 132 for the training and testing sets respectively. Put differently, the model still seems to be overfitting.

In an attempt to deal with the overfitting problem of the model above, we used the grid-search method. The grid-search method iterated over different combinations of five hyperparameters. These hyperparameters being the amount of estimators in the forest, the maximum depth, the minimum number of samples required to split a node and to be a leaf node, and the maximum number of features. This method yielded a coefficient of determination of 0.908 and 0.848 on the training and testing sets respectively. Also, an MAE of 374 667 and 472 812 on the training and testing sets respectively.

Our experiment with MCCP on the RFR did not yield a better model either with scores of 0.9776 and 0.865 on the training and test sets respectively, and MAE of 158 090 and 417 492. These results indicate the model is overfitting. This method also required a large amount of computing power which discouraged us from exploring it further.

Looking at the pre pruned model found by grid search which gave the best scores, we visualize the comparison of predicted price and actual price in *Figure 4.10*. From this we immediately observe outliers, the model has a problem with perceiving characteristics of expensive properties. When the price is above 15 million NOK, the prediction of our model is consistently less. If we ignore the outliers, there are few signs of patterns in the plot. Which means that the model is good at capturing the relationship between the variables, and therefore has low bias.

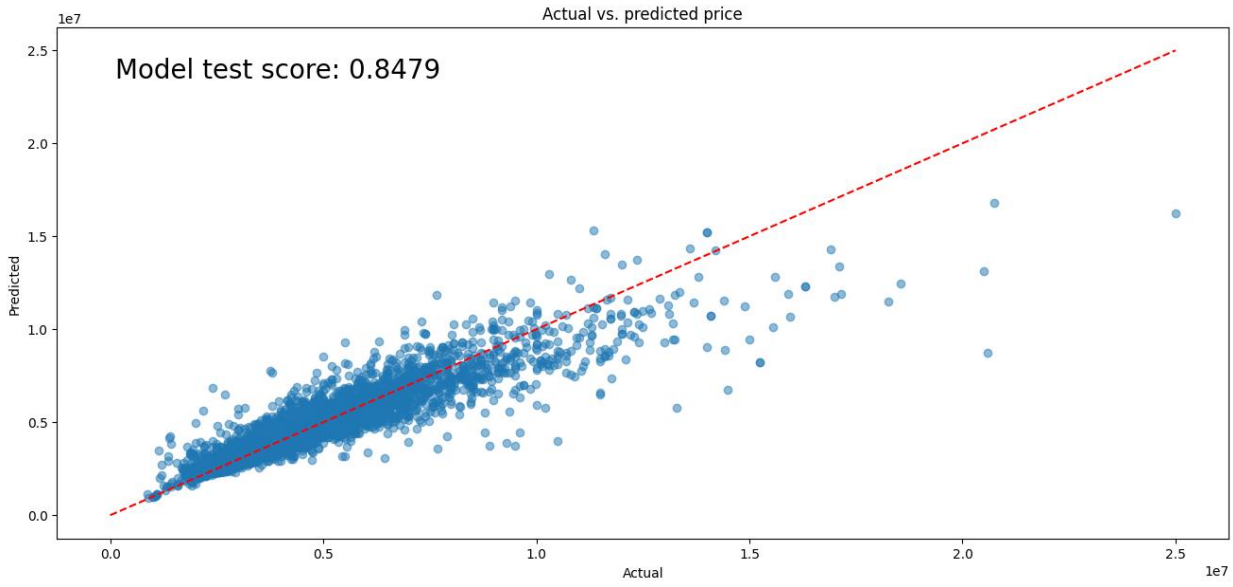


Figure 4.10: A scatter plot of actual and predicted price of a resident in Trondheim. Model score 0,8479.

When studying the histogram of residuals for the same model, we see an expected value of 0, as shown in *Figure 4.11*. This tells us that the model is not predicting the price too high, nor too low. The residuals also look normally distributed, which is a positive indicator, however, there is still some deviation.

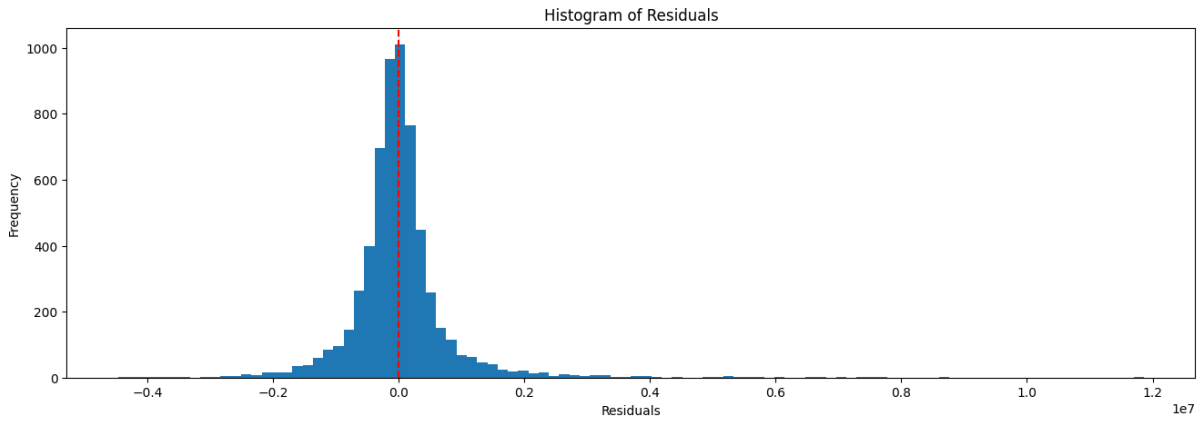


Figure 4.11: Plot showing the distribution of residuals of the RFR model.

4.2 Discussion of results

Linear regression leaves a lot to be desired. A simple linear regression with an R-squared value of 0.561 is not enough to be a useful model. The usable area seems to be a very important factor if it on its own accounts for this much, however, multiple linear regression makes for a better model. The increase of R-squared by approximately 0.127 is an improvement, but it is still not a good model. It is apparent that a linear model has its limitations when applied to a data set as complex as this. Interestingly, removing the insignificant variables did not improve the R-squared value. This suggests that not only were they insignificant, but their coefficients were also so low as to render them utterly futile.

The linear regression model failed to provide any fruitful answers, with the excessively low R-squared value. Hence, we have decided to discard it as an optional method to predicting the housing market in Trondheim, and treat it as it was intended, simply a benchmark. Therefore, we will focus on the more interesting findings using DTR and RFR.

After having built a simple DTR model, a pre-pruned DTR model, as well as the MCCP optimization of those models, we got some prudent results. Based on the R-squared score of the simple DTR model, it is clear it had a lot of overfitting interfering with its predictions. For the pre-pruned model it showed more promise, in the sense it had a lot less overfitting, and decent R-squared results. Given we restricted the pre-pruned model, more than in simple DTR, it was expected that the pre-pruned model gave better results, as proven above.

After performing MCCP on both models, we found that the initial tree had little to no effect on the resulting pruned tree. Their performances were roughly identical, with the pre- and post-pruned model achieving slightly better, but negligible results. As seen in *Figure 3.8 and 3.9*, their grown trees were also for the most part identical, with some slight differences. We believe the indistinguishable performance of the MCCP regressors reside in the depth of the tree, with the final maximum depth of the post pruned trees at 10. The benchmark tree has a maximum depth of 29 and the pre pruned tree has a maximum depth of 15, meaning that both trees probably had approximately the same structure down to about a depth of 10. Nevertheless, both post pruned models performed worse than the pre pruned model and can be discarded from consideration of our final model. For further comparison, we will continue discussing the best method for predicting the housing prices using the pre-pruned DTR model as a new lower bound.

In comparing the RFR models, the first model and the model where grid search was used to define the hyperparameters, we find that the basic pre-pruned model had room for improvements. These were found in the grid search model and applied, in order to reduce the amount of overfitting. The results proved beneficial in the sense we got a more reliable model, shown by the smaller difference in score between training and test sets. Although the model test score dropped by 0.2, we believe the sacrifice to be worth it. Hence, the RFR model including grid search was the best model of the two, in predicting the housing prices in Trondheim.

When comparing the pre-pruned RFR model, using grid search, with the pre-pruned DTR model, we get an accuracy score of 0,908 (train) and 0,848 (test), compared to an 0.849 (train) and 0.812 (test). Even though we valued the extra reliability when comparing the two RFR models, we now prefer the accuracy score of the RFR model to the DTR model. This is because, we deem the difference in accuracy between the training and test sets of the RFR model to be acceptable. The deviations in their respective residual plots are better in the RFR model and is another positive when compared to the DTR model. The MAE of the RFR model is about 100 000 less than the DTR models. It means, by choosing the RFR model it predicts on average about 100 000 NOK closer to the actual value than the DTR model. We deem this to be too much error to disregard. Hence, we trust RFR to be a better method with more reliable, as well as accurate results than the pre-pruned DTR model.

Even though our model has a high accuracy score, there is still room for improvement. It was expected that the usable area is important to the price, but the fact that the other variables have such a low importance in all models is surprising. This might be because of our dataset lacking information. When we cleaned the data some of the variables were estimated and some were set to 0, which could have caused the low feature importance score on some variables. This could also have affected the accuracy score in the sense, we could have created some inaccuracies.

Another reason for the low feature importance score could be variables that were absent from the data we received. For example, a farm can be worth a lot, but at the same time have a smaller usable area than other houses in the same price range. Valuable characteristics such as view, proximity to infrastructure, activities and the rights to forests and fields are not included in our dataset, and may be some of the reasons behind the outliers in all the figures of actual price against predicted price.

When predicting the price of an apartment, certain criteria are set as an example: Usable area of 85 square meters, two bedrooms, a balcony, and a parking spot. A preference for freehold properties is noted, influencing the housing type parameter. In addition, we have entered values for the rest of the parameters. These are in many ways random, but necessary for the model to be able to predict. We therefore ended up using average figures, as well as values close to the last observed values. This applies to most macroeconomic values. The model predicted a price of 5 702 386 NOK, which is below the maximum price of 5 750 000 NOK calculated earlier. When it comes to the equity requirement, for the sake of simplicity we will round up the price to 5 700 000 NOK. This means that the required equity is 15% of this sum, which corresponds to 855 000 NOK. To conclude, a newly graduated MBA student would have to save half of the amount, which is 427 500 NOK, to be able to afford this type of housing.

5 Conclusion

In this thesis our goal was to make a model to predict housing prices in Trondheim. The linear models, both singular and multiple, did not make for good predictors. Removing the insignificant variables did not improve the multiple linear model. A reason for this, might be due to the presence of nonlinearity when predicting housing prices. Decision tree regression and random forest regression are methods that can handle nonlinear relationships. After pruning the DTR model with pre-pruning and minimal cost-complexity pruning, as well as pre-pruning the RFR model based on the results of the grid search, our most accurate model is the latter. With a higher test score and a lower mean absolute error, our conclusion is that the RFR model is the best at predicting property prices in Trondheim.

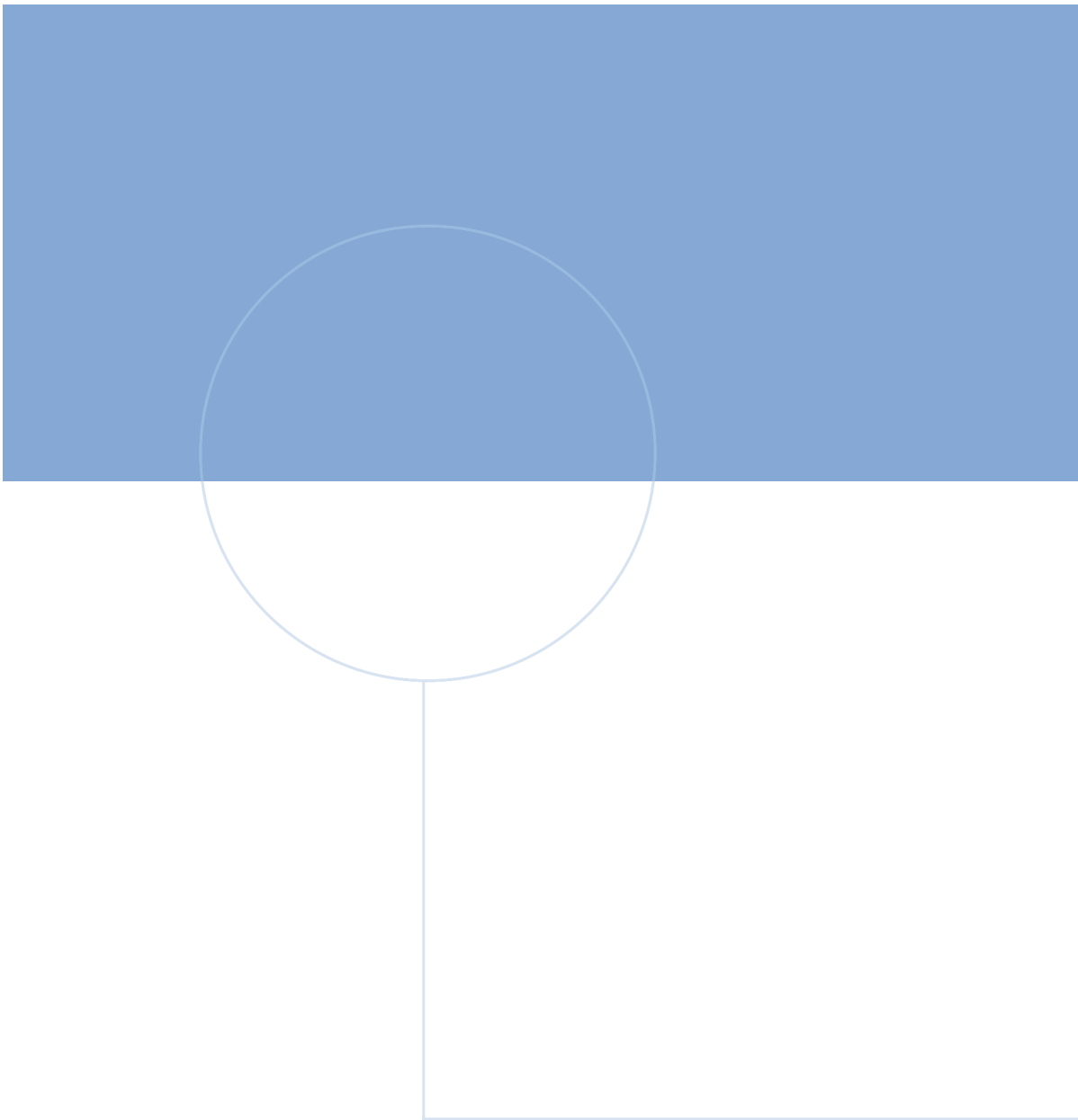
Even though the RFR model had a high R-squared score, it has some clear weaknesses. First of all the model is strictly limited by geographical area and will only provide reasonable predictions in Trondheim. In addition, our data points and variables were lacking in volume, which makes it harder to train a model to perceive relationships. The dataset also had missing values. Some of these properties were deleted, but most missing variables were replaced with estimates. This most likely had an impact on the results. Another weakness in the models is the fact that data from only the last three years were used. To make even better models, it would be preferable to have data spanning more than three years. The fact that year, month and day were split into separate columns, can also be an obstacle for the models.

As mentioned above, the models have their weaknesses. Therefore, further research is recommended. The random forest model had the highest score, but there were still signs of overfitting. A possible way to deal with this is to introduce another ensemble method like gradient boosting. In addition to experimenting with different methods, a good idea would be to collect more data. Not only in terms of observations, but also more variables that can explain the variance in the models. Whether or not the model performs well on other cities would also be interesting to research.

Reference list

- Breiman, L. et al. (1984) *Classification and Regression Trees.* Wadsworth International Group.
- Econa (2024) *Dette tjener du som nyutdannet*. Available at: https://nye.econa.no/medlemsfordeler/lonn/dette-tjener-du-som-nyutdannet/ (Accessed: 10th March 2024).
- Gundekjøn, O.T. and Kristensen, R.K. (2021) *Din privatøkonomi*. Oslo: Gyldendal akademisk forlag.
- Hoxha, V. et al. (2024) "Comparative analysis of machine learning models in predicting housing prices: a case study of Prishtina's real estate market", *International Journal of Housing Markets and Analysis*, Available at: https://www.emerald.com/insight/content/doi/10.1108/IJHMA-09-2023-0120/full/html#tbl1 (Accessed: 15th April 2024)
- James, G. et al. (2023) *An Introduction to Statistical Learning with Applications in Python*. Springer.
- Jha, S.B. et al. (2006) *Housing Market Prediction Problem using Different Machine Learning Algorithms:* A Case Study. Available at: https://arxiv.org/ftp/arxiv/papers/2006/2006.10092.pdf (Accessed: 17th April 2024)
- Karjian, J (2023) *History and evolution of machine learning: A timeline.* https://www.techtarget.com/whatis/A-Timeline-of-Machine-Learning-History (Accessed: 23rd April 2024)
- Krogsveen (2024) *Prisutvikling for Norge.* Available at: https://www.krogsveen.no/prisstatistikk. (Accessed: 25th March 2024)
- Lånekassen (2023) *Flere får over 1 million i studiegjeld.* Available at: https://lanekassen.no/nb-NO/presse-og-samfunnskontakt/nyheter/flere-far-over-1-million-i-studiegjeld/ (Accessed: 10th March 2024)
- Norges Bank (2020) *Pengepolitikkens rolle i en koronatid.* Available at: https://www.norges-bank.no/aktuelt/nyheter-og-hendelser/Foredrag-og-taler/2020/2020-10-06-cme/. (Accessed: 25th March 2024).
- Norges Bank (2024-1) *Hvordan påvirker styringsrenten andre renter?* Available at: https://www.norges-bank.no/kort-forklart/styringsrenten/hvordan-pavirker-styringsrenten-andre-renter/ (Accessed: 23rd. April 2024)

- Norges bank (2024-2) *Norges Banks dataorg.* Available at: https://app.norges-bank.no/query/#/no (Accessed: 5th March 2024)
- Norges Eiendomsmeglerforbund. (2022) *FØRSTEGANGSKJØPERE 2022 Q4,* Available at: https://nef.no/wp-content/uploads/2023/03/Forstegangskjopere_2022Q4_ny-1.pdf (Accessed: 15th February 2024)
- OpenAI (2024) ChatGPT (3.5) [Large language model]. https://chat.openai.com
- Scikit-Learn (2024) *Decision Trees.* Available at: https://scikit-learn.org/stable/modules/tree.html#tree (Accessed: 22nd April 2024)
- Scikit-Learn (2024) *sklearn.ensemble.DecisionTreeRegressor.* Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html (Accessed: 22nd April 2024)
- Scikit-Learn (2024) *sklearn.ensemble.RandomForestRegressor.* Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (Accessed: 20th April 2024)
- SSB (2011) *Hva driver utviklingen i boligprisene?* Available at: https://www.ssb.no/priser-og-prisindekser/artikler-og-publikasjoner/hva-driver-utviklingen-i-boligprisene (Accessed: 23rd April 2024)
- SSB (2022) *Statistikkbanken - Boforhold, levekårsundersøkelsen.* Available at: https://www.ssb.no/statbank/table/14066/tableViewLayout1/ (Accessed: 24th March 2024)
- SSB (2024) *Statistikkbanken - Prisindeks for brukte boliger.* Available at: https://www.ssb.no/statbank/table/07221/ (Accessed: 5th March 2024)
- StatQuest (2018) *Machine Learning Fundamentals: Bias and Variance.* Available at: https://www.youtube.com/watch?v=EuBBz3bI-aA&ab_channel=StatQuestwithJoshStarmer (Accessed: 20th April 2024)
- Trondheim Kommune (2022) *Kunnskapsbyen*. Available at: https://www.trondheim.kommune.no/aktuelt/om-kommunen/annet/kunnskapsbyen. (Accessed: 25th March 2024).



NTNU

Norwegian University of
Science and Technology