Fontys Hogescholen ICT

Project Proposal

Housing Price Prediction by Zhaklin Yanakieva

Zhaklin Yanakieva

Eindhoven, 08/03/2021

Table of content

Versioning table

Time period(weeks)	Comments on changes	Version
3 — 5	Initial changes to the document — creating the structure	1.0.0
6 — 7	Filling up with the needed information, such as explaining graphs, and summarizing	2.0.0
8 — 10	Adding the conclusion and more explanations	3.0.0

Project definition

Background

House price forecasting is an important topic of real estate. Machine learning is applied to analyze the information gathered to create models for housing for tenants and landlords, and therefore, help them have more knowledge about the real estate they are renting or giving for rent. The project will be conducted by analysing value ranges, such as price, living area, type of housing, postcode, street and rooms, that are in the dataset of a website (https://www.pararius.com/apartments/eindhoven / https://www.friendlyhousing.nl/nl) after scraping the information from the websites. In this project I am going to predict the price of houses and I will verify my results by the training of the models.

Project Goal

The objective of the project is to predict the rental value of housing properties in Eindhoven. The data is refreshed every week, so the prediction will be always up to date.

Project Domain

The domain of this project is determined to be: *Prediction of a real-estate price*

Also the following research methods were applied:

- Literature study
- Stakeholder analysis

Domain research

Literature study

- **On what basis does the price of a house change?**[1]

  1. Neighborhood comps - One of the best indicators of your home's value is the sale prices of similar homes in your neighborhood that have sold recently. These comparable homes are often referred to as "comps".

  2. Location - Your current home may be the ideal location for you — close to your job or near your parent's house — but when appraisers determine how much value to assign based on the location of the house, they're looking at three primary indicators, according to Inman:

     - The quality of local schools
     - Employment opportunities
     - Proximity to shopping, entertainment, and recreational centers

  These factors can influence why some neighborhoods command steep prices, and others that are a few miles away don't. In addition, a location's proximity to highways, utility lines, and public transit can all impact a home's overall value. When it comes to calculating a home's value, location can be more important than even the size and condition of the house.

  3. Home size and usable space - When estimating your home's market value, size is an important element to consider, since a bigger home can positively impact its valuation.

  4. Age and condition - Many buyers will pay top-dollar for a move-in-ready home. This is why most buyers require an inspection contingency in their contract — they want to negotiate repairs to avoid any major expenses following the sale.

  5. Upgrades and updates - Updates and upgrades can add value to your home, especially in older homes that may have outdated features. However, not all home improvement projects

[1] https://www.opendoor.com/w/blog/factors-that-influence-home-value

are created equally. The impact of a project or upgrade varies based on the market you're in, and your existing home value.

6. The local market - Even if your home is in excellent condition, in the best location, with premium upgrades, the number of other properties for sale in your area and the number of buyers in the market can impact your home value. If there are a lot of buyers competing for fewer homes it's a seller's market. Conversely, a market with few buyers but many homes on the market is referred to as a buyer's market.

7. Economic indicators - The broader economy often impacts a person's ability to buy or sell a home, so in slower economic conditions, the housing market can struggle. For example, if employment or wage growth slows, then fewer people might be able to afford a home or there may also be less opportunity to relocate for new opportunities. It's important to keep up with the current status of home sales and home price appreciation in your area, especially when you evaluate the best time to sell your house.

8. Interest rates - Why care about interest rates? Both short-term interest rates (like what you pay on a credit card) and long-term interest rates (like what you pay on a mortgage) influence your ability to afford a home, but in different ways. A rise in short-term interest rates may increase the interest on your savings, but it also makes short-term debt more expensive. For example, if you're spending more money paying off a credit card or short-term loan, then you will likely have less money available in your budget to afford a home.

- **What do people look for in a house?**[2]
    - What the Experts Say Women Want?

      Experts say it is recommended spending the most money in the kitchen and the bathroom, which again meant new kitchen cabinets, new vanities and new flooring.

[2] https://toughnickel.com/real-estate/What-Do-Women-Look-For-When-Buying-a-House

Also, a clean house, no reminders of the previous tenants, an easy to clean house, "neutral" as beige colors, no repairs needed.

Stakeholders

Name	**Details**
Landlords	People who have the intention to sell a property.
Tenants	People who have the intention to rent a property.
Real-estate agencies	Institutions through which people rent/purchase houses.

Planning

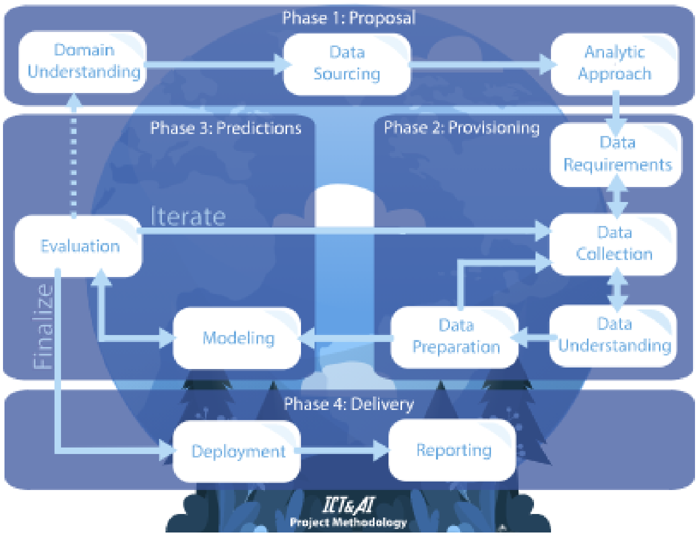
In this chapter, I am going to explain the project planning for the upcoming weeks this semester and how I am going to achieve my goal. It gives me a better understanding of the structure of the semester and also helps me to deal with deadlines. I am using my own sprints that are represented in the table (see Picture 1).

Part	**Weeks**	**Project days**	**Phase(s)**
A	3-6	12	1 and 2
B	7-10	10	2 and 3
C	11-12	10	4

(Picture 1:Phases)

From week 3 until week 6, this is the first phase (phase 1) of the project. This is where the business proposal and the exploratory data analysis report must be delivered, however Ialso decided to start with the first steps of the provisioning phase.

The next phase is week 7 until week 10. That is the time for the second part of the project (phase 2 and 3). The reason for doing phase 2 and 3 in the same sprint is because in the second iterations, the product has to be improved compared to the first one. This is the same for the final part of the project (phase 2,3 and 4). That is the final sprint where I have the final product ready.



(Picture 2: Phases of the project)

As you can see on Picture 3, the 4 weeks will be spent on the Project Proposal, the EDA (Exploratory Data Analysis), Data requirement and Data collection ledger. After the 6th week I am going to spend the time on the Provisioning and Prediction phases. At week 10 to 12 , there will be the deployment of the project presented.

Step	Weeks	Phase	Deliverables
1	3 – 6	Proposal	Project Proposal, Exploratory Data Analysis Report, Data requirement, Data collection ledger

2	7 – 10	Provisioning	Data Analysis Report, Preparation notebook
3	14 – 15	Prediction, Deployment	Modelling notebook attached, Deployment report

(Picture 3: Planning of the project)

Ethical considerations & Impact assessment

Artificial intelligence can dramatically improve the efficiencies of our workplaces and can augment the work humans can do. When AI takes over repetitive or dangerous tasks, it frees up the human workforce to do work they are better equipped for—tasks that involve creativity and empathy among others. There are both negative and positive aspects of the impact of AI on society. I use TICT as a tool to create the impact on society document, where I state why this project will affect people positively. As a result, I state that there is not going to be any personal data exploited and the project will only be beneficial to society.

For a deeper look into the Project Impact document, which is submitted together with the Project Proposal document.

Data understanding

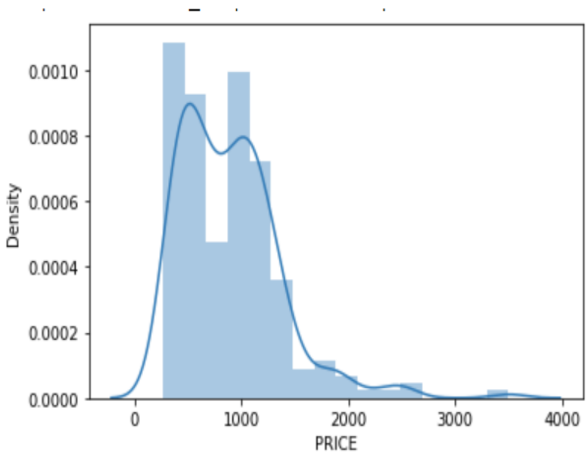
In the EDA, I decided to use data that is scraped from two housing websites. After I created a dataframe by joining the data from the two places, I saved it into a csv file in order to use it in the document.

In the EDA, I seperate the analysis into several parts. The first part is the Data collection. For this, I decided to use web scraping as a technique because it gives the opportunity to work with a data set that is up to date and therefore, makes more accurate summaries. Then, the following part is the Data preprocessing, for which I tried different types of data transforms to expose the data structure better, so the model accuracy may be improved later.

Standardizing was made to the data set so as to reduce the effects of differing distributions.

The skewness of the features was checked in order to see how distorted a data sample is from the normal distribution.

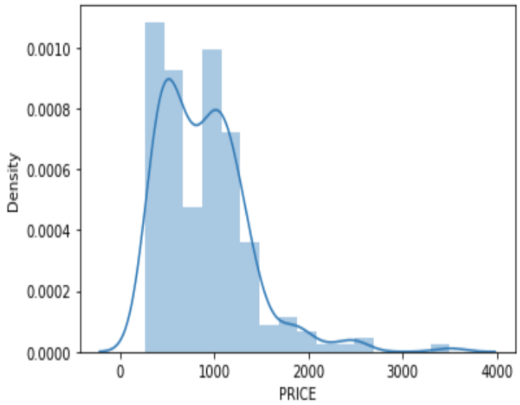
Rescaling (normalizing) the dataset was also included to reduce the effects of differing scales.



> Examining the data distributions of the features. We will start with the target variable, PRICE, to make sure it's normally distributed.

> This is important because most machine learning algorithms make the assumption that the data is normally distributed. When data fits a normal distribution, statements about the price using analytical techniques will be made.

Rescalled:



> The PRICE distribution is not skewed after the transformation, but normally distributed. The transformed data will be used in in the dataframe and remove the skewed distribution:

> **Normally distributed** means that the data is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

For a deeper look into the EDA, you can find it here or in a pdf format.

Modelling

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.[3]

The modelling stage will consist of training a Linear Regression and Random Forest model. The reason for choosing to train this machine learning model is based on the explored information - the predicting label, which is a numeric value for which the models are suitable. In this case, the predicting label is the price of a real estate. The price is measured in euros since the project data is extracted from a Dutch housing website.

1. Linear Regression:

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).[4]

2. Lasso Regression:

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).[5]

[3] https://docs.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model
[4] https://en.wikipedia.org/wiki/Linear_regression
[5] https://www.statisticshowto.com/lasso-regression/

3. Ridge Regression:

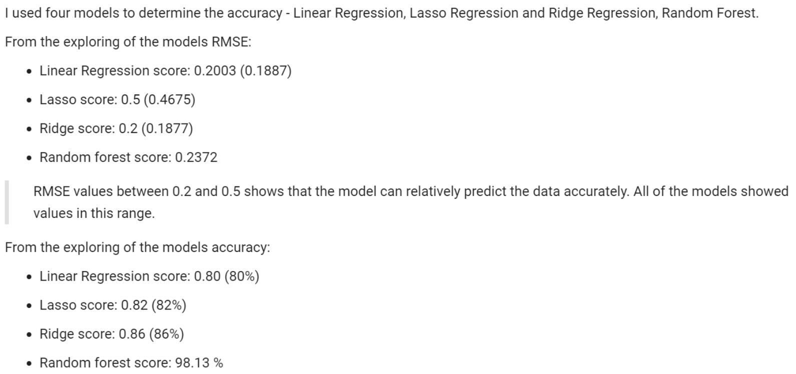
Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. It has uses in fields including econometrics, chemistry, and engineering.[6]

4. Random forest regressor:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.[7]

Evaluation

This chapter explains how I will assess the performance of the project and what features will be optimized. After exploring the data set and training the models, a vital part of the project is the evaluation of the models performance.



I used four models to determine the accuracy - Linear Regression, Lasso Regression and Ridge Regression, Random Forest.

From the exploring of the models RMSE:

- Linear Regression score: 0.2003 (0.1887)
- Lasso score: 0.5 (0.4675)
- Ridge score: 0.2 (0.1877)
- Random forest score: 0.2372

> RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately. All of the models showed values in this range.

From the exploring of the models accuracy:

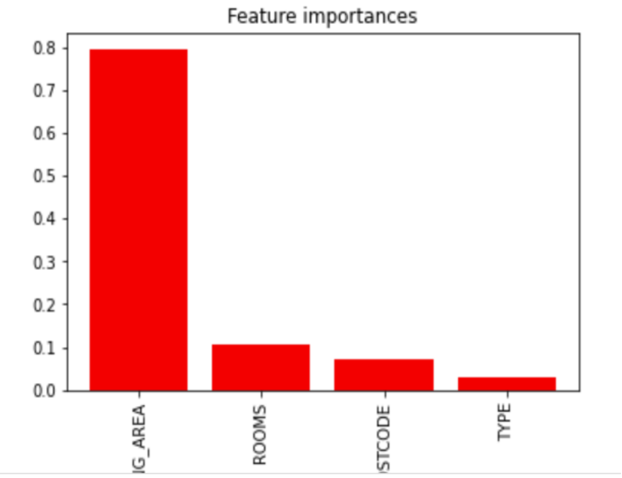
- Linear Regression score: 0.80 (80%)
- Lasso score: 0.82 (82%)
- Ridge score: 0.86 (86%)
- Random forest score: 98.13 %

[6] https://en.wikipedia.org/wiki/Ridge_regression
[7] https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

First of all, the performance of the model has to be tested with common evaluation metrics. Then, an approach, which consists of predicting a future value, to evaluate the model is accurate enough to determine the performance.

On the picture above is stated the information needed to compare the models and then, decide which is the best one to be chosen for the deployment phase. As a conclusion, the Random forest had the best performance of all because its accuracy stated 98% and the RMSE was a bit more than 0.2, which says that the model may well predict the score accurately.

1.1 Plotting the feature importance



> The idea behind the plotting of feature importance is that after evaluating the performance of the model. The feature importance (variable importance) describes which features are relevant.

> Random Forest determined that overall the living area of a home is by far the most important predictor. Following are the sizes of above rooms and postcode.

Deployment

Ideally the model will be deployed by using an API and creating a Web application. The users will have the chance to access the Web application so as to query the API, which will be open for it. More information can be found In the deployment report where it is also explained by visualizing the deployment itself.

Conclusion

From the gathered information I was able to limit down the information to make the application more productive and feasible for future orientation and other predictions. By limiting the data, it is possible to scale down the application, but making it in such a way that quality comes forth rather than a rushed product. Thereby having a more reliable product and a product that will work in the end. All planning done by the team and the minimal risk factor, will make sure the end product will not suffer any form of lack in quality.

Tools used:

1. Numpy
2. Pandas
3. Matplotlib
4. Scikit Learn
5. Google coollaboratory
6. TCIT
7. Python

References

Toughnickel, https://toughnickel.com/real-estate/What-Do-Women-Look-For-When-Buying-a-House.

Opendoor, https://www.opendoor.com/w/blog/factors-that-influence-home-value.

Canvas, https://fhict.instructure.com/.

“Impact on society and ethics.” https://www.tict.io/.