VärderingsMaskinen

Unveiling the Predictive Power of Machine Learning
in the Swedish Housing Market

Oliver Bölin
Blekinge Tekniska Högskola
Institutionen för datavetenskap
Karlskrona, Sweden

I. Abstract

This report presents the findings of a machine learning project that aimed at developing a housing marked valuator based on online data. The objective of this was to create a tool capable of providing accurate predictions on properties around Sweden. The project uses a Random Forest Regressor model trained on a database comprising more than 600,000 property listings after being manipulated, achieving a mean error rate of 9% relative to the mean price of properties.

II. Introduction

The housing market has always been a challenge to predict. In 2023 house prices decreased by 6%, but in general apartment prices increased. Housing in Sweden in the last 12 years alone has risen an astounding 85% (adjusted to inflation) to the top of 2021 but has fallen a bit since then. With housing prices of a median of 2.3 million Swedish krona, buying residency is often the biggest investment a person makes in their life.

The best way to sell a house has always, throughout time, been to valuate it with a broker. The broker checks the price similarity in the region, area, and street. Calculates the average price per square meter and then takes the seller's apartment size and calculates accordingly. The amount of rooms, balcony, and the age of the apartment may vary the listing price, but it is usually the bidding price that is harder to predict, since time, economy and many other factors can change it alot. [1], [2]

But how can a buyer valuate a house listing? The current landscape lacks simple solutions for non-experienced buyers, and navigating though real estate data can be a perplexing task and the buyer could lose their interest. Relying entirely on a broker's valuation poses its own challenges, as these valuations might fluctuate. There is a need for a simple solution that only relies on data, and is easy to use. That is why VärderingsMaskingen (*The Valuation Machine*) will valuate almost any housing, even if the sort of housing doesn't even exist. It uses Random Forest Regressor to learn from a database including 600,000 different listings and gets a mean error of 9% of the mean price.

III. Interview from a valuator

The valuation of residence depends on what type of housing it is. There are different methods of housing valuation. The first one is *Comparison Method*. This method is based on what was explained in the introduction, whereas you find similar housing according to the area.
But valuations are subjective and can change according to the valuator, seller, buyer and market value. A study from Switzerland says that a third of all valuations are under or overvalued by 10%. Other studies say that sometimes over and undervaluation can exceed up to 30%. The reason for the large margin of error is because of new market trends and the valuator's incomplete information about them. [3], [4]

IV. Dataset

The dataset was sourced from ***bostadsbussen.se***, where it was extracted through web scraping of ***hemnet.se***. Note that the data is in its raw form, exactly as retrieved when it was web scraped. This raw dataset contains a variety of information, including inaccuracies resulting from miswritten details by publishers. The dataset is extensive, amounting to a 1 GB JSON file, and has the following categories:

Street, Property Type, Build Year, Ownership Type, Construction Date, Association, Broker, Sold At, URL, Price Change, Floor, Story, Housing Form, Operating Cost, Fee, Wanted Price, Final Price, longitude, latitude, County, Area.

This dataset provides insights into various aspects of residential properties, though it's crucial to be mindful of potential inaccuracies arising from misreported information by property publishers.

V. Method

The idea was to have a Jupyter Notebook create and train the algorithm, and then use that in the main software in pair with the GUI.

A. Data manipulation

After loading the data, the removal of any unnecessary parameters started, these included all prices that were 0 and if the building year was 0. The apartments that were higher than 30,000,000 were also excluded to lower the extreme values. A lot of unimportant categories were removed such as url, broker and ownership type. These were often empty or would not affect the price of real estate.

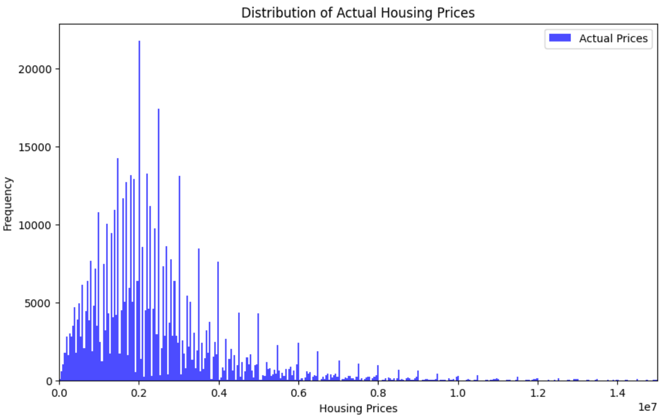


Fig. 1. Distribution of Housing prices

The date of the listing had a bad format which was fixed into an age parameter instead, meaning an age of 0 is from this year and an age of 5 is from 2019. Some textual problems were also fixed such as multiareas that included two towns were split into just one.

To get more info on why the apartments are more expensive, the population density was used from ***Statistiska Centralbyrån***. This was paired with the county of the listing by fixing some of the counties such as *Stockholms* or ending with *kommun* into just the county name.

Some things that were tried here were, using *openstreetmap* to convert the longitude and latitude into postal codes, for a better user experience in the end. This did not work as intended since it took a long time and there were too many requests so the database timed VärderingsMaskinen out.

Another manipulation part that was tried was considering inflation and markup of the prices. An apartment sold in 2012 and 2022 will have a very different price even though they are the same apartment. This was ultimately scrapped and the age factor was taken into account instead. With the age parameter perhaps Random Forest could learn that a lesser age gives a better price.

B. Labeling and Scaling

All categorical features, such as property type, county, area, and balcony were labelled with *LabelEncoder*. The wanted price was dropped from the data frame and it was scaled using *StandardScaler*.

VI. Algorithm

A. Random Forest Regressor

Random Forest is an algorithm that uses ensemble learning family methods. Random Forest is also built with decision trees which are flowchart structures where each internal node in the model represents tests on attributes and each branch the outcomes. Simplified, Random Forest is a massive multipath decision tree. When using it as a regressor each branch will represent the price of the housing and the final prediction will likely be the mean price. Random Forest is a good algorithm for predicting housing because of its robust handling of a larger amount of features and is less likely to overfit. [5]

B. Hyper-Parameters

To fine-tune the parameters for best optimization is called a hyper-parameter. For this purpose grid search was used for 15 hours until it was canceled due to time. Grid search takes in different hyper-parameters to be tested and then evaluates the result. Grid search was used for a very long time and the result was not better than randomly picking a parameter.

The research done by Isak and Alan was instead used as a hyper-parameter which increased the R^2 from 0.90 to 0.93 and decreased the learning time from 30 minutes to 5 minutes. [6]

C. Training

The training was done on a split database of 90/10. Random Forest Regressor takes around 5 minutes with a processor of i7.6700k at 4.0 GHz and 16 GB RAM at 2400MHz.

D. Evaluation metrics

Some good evaluation metrics for Random Forest Regressor include but are not limited to, *Mean Squared Error (MSE)*, R^2 and *Mean Absolute Error (MAE)*.

1) Mean Absolute Error: Mean absolute error is a measurement that looks at the errors in the prediction versus the actual result. By taking the mean of all these errors via,

$$MAE = \frac{\sum_{\text{i=0}}^{n} |y_i - \hat{y}_i|}{n}$$

A zero MAE means there was a perfect prediction, which would likely show that the problem at hand is not complex. [7]

2) Mean Squared Error: Mean Squared Error is by definition similar to MAE, but quadratically larger.

$$MSE = \frac{\sum_{\text{i=0}}^{n} (y_i - \hat{y}_i)^2}{n}$$

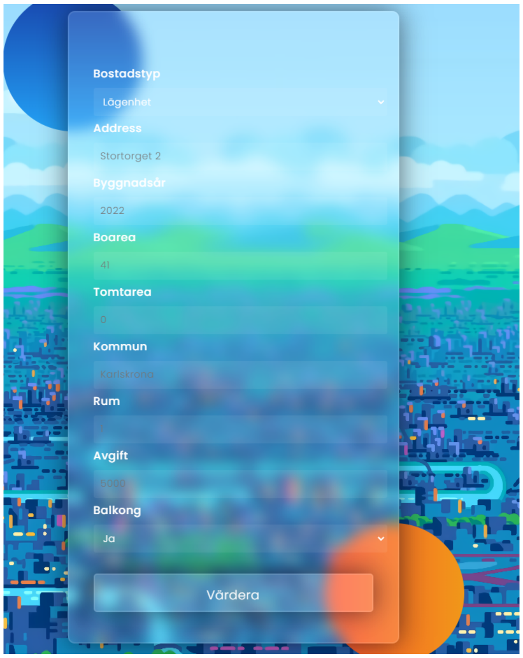
3) R^2: R^2 is the ratio between explained variance and the total variance. It is the statistical measure that represents how fit a regression model is. A R^2 of 1 is the best possible outcome and there is no difference between the prediction and the data. The R^2 explains how much the model learns about the relationship between all variables. [8]

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{\text{i=0}}^{n_{\textbf{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{\text{i=0}}^{n_{\textbf{samples}}-1} (y_i - \overline{y}_i)^2}$$

Where $\hat{y}_I$ is the prediction of the *i*th sample and y is the value of n_{samples}. Meaning,

$$\overline{y} = \frac{1}{n_{\text{samples}}} \sum_{i}^{n} y_i$$

The result of evaluation showed a R^2 of around 0.9337, a MSE of 222,373,048,290 and a MAE of 237,971 The mean price of the data is 2,463,747 SEK, which gives VärderingsMaskinen a mean error of 9.7%.



Bostadstyp

Lägenhet

Address

Stortorget 2

Byggnadsår

2022

Boarea

41

Tomtarea

0

Kommun

Karlskrona

Rum

1

Avgift

5000

Balkong

Ja

Fig. 2. Input page

E. User-Interface

Python Flask was used for the web interface with a combination of HTML and CSS markup languages. The input for the user contains a variety of different parameters, but longitude and latitude were excluded because of user experience. Instead of coordinates the user can input an address and county and it will be geocoded by *Geopy Nomatims* API. This then results in longitude, latitude and area. This simplifies the input for the user. In the results for the prediction, a low, predicted, and high price rounded to the nearest 5,000 SEK are shown. Accordingly to,

$$\text{Low price} = \text{Predicted price - MAE}$$

$$\text{Medium price} = \text{Predicted price}$$

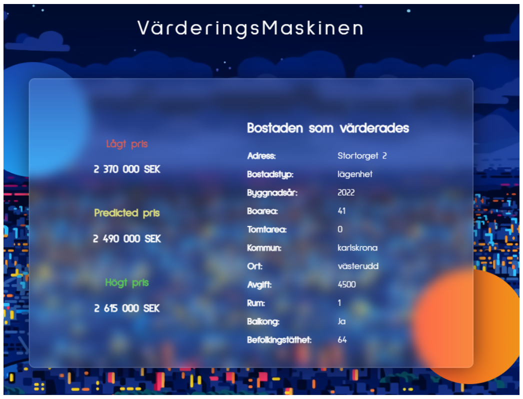
$$\text{High price} = \text{Predicted price} + \text{MAE}$$

F. Saving the algorithm

Instead of relearning the algorithm on every run, *joblib* was used to save the model. *joblib* is a tool for pipelining Python jobs, and for Random Forest regressor it first converts it to a savable format, which is a .pkl or .joblib extension. This contains all kinds of different information about the model such as parameters, feature importance and other necessary information to create the model. The model is later loaded in when the *vm.py* file is started.

VII. Analysis

A. Feature importance



Lågt pris

2 370 000 SEK

Predicted pris

2 490 000 SEK

Högt pris

2 615 000 SEK

Bostaden som värderades

Adress:	Stortorget 2
Bostadstyp:	lägenhet
Byggnadsår:	2022
Boarea:	41
Tomtarea:	0
Kommun:	karlskrona
Ort:	västerudd
Avgift:	4500
Rum:	1
Balkong:	Ja
Befolkingstäthet:	64

Fig. 3. Result page

Feature importance is a technique which calculates a score of all the features used for the model. This was used for the Random Forest model, whereas a higher score in the table gives that specific feature a greater effect on the prediction. [9] After the feature importance was evaluated, it showed that the model was affected by noise, irrelevant information or overfitting. An investigation was conducted into this with correlation and removal of features.

Feature	Importance
living_area	0.321219
population_density	0.280876
build_year	0.135090
longitude	0.079351
latitude	0.073264
age	0.052360
fee	0.016140
area	0.015667
rooms	0.011253
county	0.005369
land_area	0.004756
property_type	0.002488
balcony	0.002165

TABLE I
Feature Importance

B. Correlation

Correlation will explain how the features are related to each other. A positive correlation means the feature increases the valuation, while a negative decreases the valuation. This means a correlation of zero does not do much to the valuation. [10]

C. Feature Removal

To fix the overfitting issue, a test was made to ensure none of the features affected the valuation negatively. This test yielded no good result as the *Mean Price Ratio* was not affected positively by removing any feature.

D. Results

Even though the feature importance and correlation implied that *age*, *balcony* and *build year* were features that affected the prediction negatively, a separate run without them yielded

Feature	Correlation
population_density	0.471403
living_area	0.410599
rooms	0.369930
longitude	0.212303
county	0.146396
property_type	0.133796
area	0.126844
latitude	0.037127
land_area	0.021031
fee	0.002190
build_year	-0.078951
balcony	-0.131165
age	-0.209563

TABLE II
CORRELATION WITH WANTED PRICE

Removed Features	MAE	R-squared	Mean Price Ratio
None	237,971	0.9337	0.9034
build_year	251,820	0.9285	0.8977
land_area	237,870	0.9334	0.9034
rooms	240,074	0.9328	0.9025
population_density	237,883	0.9343	0.9034
fee	239,127	0.9335	0.9029
balcony	238,879	0.9331	0.9030
age	313,245	0.9090	0.8728
longitude	251,848	0.9262	0.8977
latitude	255,646	0.9239	0.8962
property_type	238,035	0.9335	0.9033
county	238,224	0.9335	0.9033
area	236,957	0.9342	0.9038
build_year, age	338,185	0.8953	0.8627
build_year, balcony, age	348,139	0.8904	0.8586
land_area, fee, latitude	257,078	0.9238	0.8956
age, balcony, build_year, property_type, land_area, fee	382,324	0.8764	0.8448

TABLE III
MODEL PERFORMANCE WITH REMOVED FEATURES

a result of a mean prediction error of 0.85%, which is a 5% decrease from with them. Removal of the low-importance features and near zero correlated features also yielded no better results, which leads to believe that the original features for the model fit it well enough to yield the best possible result with the data available.

VIII. CONCLUSIONS

As demonstrated in this report, features such as living area, population density and rooms play a crucial role in predicting housing prices. The process involved a dataset containing over 600,000 property listings after data manipulation and labelling techniques. Evaluation metrics such as Mean Squared Error, R^2 and Mean Absolute error showcased the accuracy of the model.

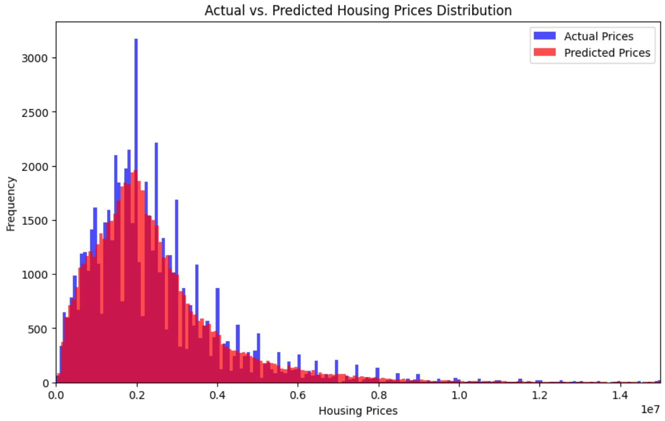


Fig. 4. Actual prices vs. Predicted prices

As G. Bölin said, "Valuation can only be as good as the data which they are based on" which fits with machine learning as well as a normal valuation done by brokers. If there was data of a thousand different real estate per area,
of pictures for each apartment and where valuators could build an AI that uses image processing to rate each apartment,
or perhaps even just more information such as when the estate was renovated, then the predictions would get closer and closer to real life. With a Mean Error of 9.7%, around 250,000 SEK the prediction is relatively close to the real price. The model is good at predicting common, housing in well-known areas. But so are normal broker valuators. What normal valuators can't do is predict on-demand housing that does not exist yet and do that without any bias. With a model like VärderingsMaskinen, anyone can valuate a future apartment they might dream of, and eventually buy. [11]

IX. CONTRIBUTIONS

This report and project were done by me, Oliver Bölin. A special thanks to my brother Gustav Bölin who studies real estate economics at Karlstad University.

X. REFERENCES

REFERENCES

[1] "Prognos bostadspriser 2023 och 2024," www.lendo.se. https://www.lendo.se/bolanebloggen/prognos-bostadspriser-23-och-24-trendbrott-pa-vag (accessed Jan. 06, 2024).
[2] See subsection Valuation
[3] G. Bölin, Karlstad University, Jan. 06, 2024
[4] E. Meins, H. Wallbaum, R. Hardziewski, and A. Feige, "Sustainability and property valuation: a risk-based approach," Building Research & Information, vol. 38, no. 3, pp. 280–300, Jun. 2010, doi: https://doi.org/10.1080/09613211003693879.
[5] A. Ihre, I. Engström, "Predicting house prices with machine learning methods", pp. 5, 2019.
[6] A. Ihre, I. Engström, "Predicting house prices with machine learning methods", pp. 13, 2019.
[7] Wikipedia Contributors, "Mean absolute error," Wikipedia,https://en.wikipedia.org/wiki/Mean_absolute_error Aug. 16, 2019.
[8] D. Jain, "ML — R-squared in Regression Analysis - GeeksforGeeks," GeeksforGeeks, May 07, 2019. https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/
[9] "Understanding Feature Importance in Machine Learning — Built In," builtin.com. https://builtin.com/data-science/feature-importance
[10] A. Upadhyay, "What Is Correlation?," Medium, Aug. 09, 2020. https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47
[11] G. Bölin, Karlstad University, Jan. 06, 2024