My project proposal starts on the page 3

<https://docs.google.com/document/d/1pyEEX06eflB_a1HarJQYiNxUcyQ0nxp0ZKaH3EuUerc/edit>

**Rules for the development of individual project steps**

Written project proposal **DEADLINE March 1st**

* 1. Title of the proposed topic
  2. Introductory description of the problem DONE
     1. What problem are you trying to solve?DONE
     2. What dataset are you using?DONE
  3. Aim and hypotheses of problem research
  4. Review of previous research EXTEND
     1. What are the existing methods by which the problem was solved and what are the disadvantages of the existing methods?
  5. Materials, methodology and research plan
     1. In what way (by what approach) will you try to solve the problem? DONE
     2. How will you collect the data? DONE
     3. What methods / algorithms / techniques / tools do you use? DONE
     4. How do you intend to evaluate the success of the results of your project (interpretation)? DONE
  6. Expected results of the proposed project
     1. What do you expect as a final result of the project?
  7. Literature

Professor emails:

[bijelizeko@gmail.com](mailto:bijelizeko@gmail.com)

[tlipic@irb.hr](mailto:tlipic@irb.hr)

# Professors Comments:

If you're reading my paper and would like to help me out, please paste the links here or as a comment related to the particular part. Thank you! :D

Related work is still weak, better research on that aspect is needed form both theoretical and application/methodology perspective, please find following references:

* AI and Machine Learning in Real Estate Investment, <https://jpm.pm-research.com/content/45/7/43>
* <https://github.com/jiveshs98/Python-Scraping-Real-Estate-Data>

[(great for the TRY-except implementation for scraper)](https://github.com/jiveshs98/Python-Scraping-Real-Estate-Data)

* <https://oxylabs.io/blog/data-scraping-for-real-estate> ADDED
* <https://emerj.com/ai-sector-overviews/machine-learning-in-real-estate-trends-and-applications/>
* Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms, <https://arxiv.org/abs/1809.04559>
* Competitive Analysis of the Top Gradient Boosting Machine Learning Algorithms, <https://ieeexplore.ieee.org/abstract/document/9362840>
* Comprehensible Classification Models – a position paper , <https://dl.acm.org/doi/pdf/10.1145/2594473.2594475?casa_token=0CfEvgEP9aYAAAAA:knMPKND_f9zbSjxTS_hEPVISE8EAnMJV1yBu6v81o1HvpElE1bp43q8_ez436RmIqzhuhpVBPsGtUg>

Application perspective (just to name a few works):

Identifying Real Estate Opportunities Using Machine Learning, <https://www.mdpi.com/2076-3417/8/11/2321> **used, briefly**

Multi-source urban data fusion for property value assessment: A case study in Philadelphia, <https://www.sciencedirect.com/science/article/abs/pii/S0925231220308080> CANNOT ACCESS

Ho, Winky KO, Bo-Sin Tang, and Siu Wai Wong. "Predicting property prices with machine learning algorithms." Journal of Property Research 38.1 (2021): 48-70.

<https://siev.org/wp-content/uploads/2020/06/12_VALIER-MICELLI.pdf>

Bergadano, Francesco, et al. "Learning real estate automated valuation models from heterogeneous data sources." arXiv preprint arXiv:1909.00704 (2019).

<https://arxiv.org/pdf/1909.00704.pdf>

Mohd, Thuraiya, et al. "An Overview of Real Estate Modelling Techniques for House Price Prediction." Charting a Sustainable Future of ASEAN in Business and Social Sciences (2020): 321-338.

https://link.springer.com/chapter/10.1007/978-981-15-3859-9\_28

Peter, Nkolika J., et al. "Review on the Application of Artificial Neural Networks in Real Estate Valuation." International Journal 9.3 (2020).

Truong, Quang, et al. "[Housing Price Prediction via Improved Machine Learning Technique](https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050920X00123/1-s2.0-S1877050920316318/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEAgaCXVzLWVhc3QtMSJGMEQCIFefVwGprKESZzmNivKL4rRJb7nd3JN93XbhPg%2FJAoELAiBfbj5FnYky9geB8d8rbJzMTvtidmasseAlsFUTwWidxyq0AwhAEAMaDDA1OTAwMzU0Njg2NSIMa%2FNsPsJUePSD%2FFkwKpED6fyPa2s%2BacxlgzD66vJW7SxSXu8q%2BxANri6nOKNZs6yi5P7RpvSKrj4BejwIf5X4ojo5%2Ftl9ha2GS6FQV6Y93%2BaQZQTD7UCVlfqujvXCB4wgctuzC%2FLM3fJxEb%2F%2FgFOV4ExnQW7B8rYb5EEkzvqvhW2z4y84r5B5pMdw82bd0e6goIKVJe9DahRxDiZPo7vMUpZ8j%2BLGAlD%2BH5q19QZRexx%2Bm3%2FyypiCQij3Eti3PP6p%2B2%2FFIKI1rjCLRVbg3GBqIcNk5WiDrnbVpJXkFun2EU%2B42TD8FnVCvybqPrIvtvQSQcjzTgXzAgEXrZZpj3d2NXR9O%2FwCVGNa8U9sz8jiJPXX0WrTJtFCY9q8xnNzXUo8lBLPd%2FBNBuOzo1iW4ooCUAb%2B4s8KtaMLW%2Fwe927ztT8GbG5stQI0sumVFzzA7koRC2F%2FBjJFBY3z1BHMbkfKtWeLnJ9Sahoxs7bRBmHGGfpap8nDwU22AQLNqDL7at247DNrKe7CiJ083MF%2BnNTZNWxM0edQ07YpP7VgLRHb5hMw0%2FzLggY67AHW9pCpPdg8ERASjJQbPjfXp%2FHcdZ1WUN7sn2nu%2F5qcbPb31EF%2F0UA5%2B%2By6AcUBI8majLp4IKAa6y5n42yAC0Q0ysMZH%2FDr7iGC4ku%2BpZ%2BZDVOxy6a17NWBKpuhgaMgF0dy9ya6fTM7oe8VnpEqsgKXS0dH5rIC%2FupouQgyLvx2HIUNy8tmQXn9kJVTr3RecVipJ0USpoyfPTcjQCy2oKtgyp4VHx3U3H4mrfDm3Hhnai%2FwF2FAFNI9gr5fJY3f%2BQSjPzsgOjlq3uRLWWwlt%2FlVmbXyVav%2B2qDx14DsSplPt4k53mOtnbg4jN7HAQ%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20210318T084021Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYTSVF7JVM%2F20210318%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=f3c657cbc563536768cf1a8935aaf49279dad2bb85664d1de45d8b31860528f0&hash=6ec1250dc589b0857e114c7c47590e581c47e8ded2feb1fe6eac2085941781b5&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050920316318&tid=spdf-1c6bd2f2-187a-4334-b276-bec63a02040b&sid=5499ea211807724bd96be211849cac33179egxrqb&type=client)s." Procedia Computer Science 174 (2020): 433-442.

Fan, Chenchen, Zechen Cui, and Xiaofeng Zhong. "[House prices prediction with machine learning algorithms](https://dl.acm.org/doi/10.1145/3195106.3195133)." Proceedings of the 2018 10th International Conference on Machine Learning and Computing. 2018.

Jha, Shashi Bhushan, et al. "Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study." arXiv preprint arXiv:2008.09922 (2020).

<https://arxiv.org/pdf/2008.09922.pdf>

Master thesis examples:

<https://inspire.redlands.edu/gis_gradproj/291/>

Another important aspect is feature engineering:

* You should explicitly state what are considered features for now (and how these relate to data collection and processing schemes in the project)
* <https://link.springer.com/chapter/10.1007/978-3-030-43823-4_10>

Model selection:

* <https://machinelearningmastery.com/a-gentle-introduction-to-model-selection-for-machine-learning/>
* <https://www.scikit-yb.org/en/latest/tutorial.html>

The comments during the lectures:

What you write in the proposal, is basically your paper, just add the practical part and conclusions.

**Intro**: grab attention. **Previous research**: What has been done, what you will do differently. **Methodology:** how you will do it. **Expectations**: What you expect, what conclusions?

<https://github.com/ai4eu/tutorials/tree/master/House_Price_Prediction>

<https://github.com/ageron/handson-ml2/blob/master/02_end_to_end_machine_learning_project.ipynb>

*To identify underpriced properties for sale and build a recommendation algorithm based on the selected inputs.*

<https://github.com/ageron/handson-ml2>

<https://github.com/slundberg/shap>

<https://docs.seldon.io/projects/alibi/en/stable/>

<https://catboost.ai/>

Look into:

Representation learning

End-to-end learning

Real Estate Investing Decisions Simplified

(With The Help Of Decision Tree Algorithms)

**Jernej Brlek**

JBRLEK@STUDENT.ZSEM.HR

Business Mathematics and Economics student, Zagreb school of Economics and Management, 110 Jordanovac, 10000 Zagreb, Croatia

# Abstract

This paper gives insight into the usefulness of decision tree based algorithms in the space of investing decision making in real estate. The issues covered include gathering the data, preprocessing, exploratory data analysis, different tree algorithms and their predictive power, as well as comments on the usefulness of the predictions.

# Introduction

Real estate is an often overlooked option for investing due high requirements for entry. However, it is an option that everyone is considering, at least at some point in life. Moving out or sticking with parents? Renting or buying? House or a flat? City or suburbs? What is the right price? Options are vast. Which one is the right one?

Everyone dislikes the feeling of making a bad decision and regret that comes from it. That is only partially why many of those decisions are difficult to make. The other reasons include lack of awareness of one's true preferences and the sheer number of possibilities as a result of the market size.

Since the focus of this paper is on machine learning algorithms and their predictive power in a real estate setting the questions we would like to answer are as follows: Is there a place in real estate pricing for machine learning? Can an ML model help make a decision whether to rent or buy?

What data is available online for the real estate listings in Slovenia through the advertising agencies websites and is it sufficient for the use of ML?

# Gathering the Data

Two main advertising platforms for real estate in Slovenia are [nepremicnine.net](https://www.nepremicnine.net/) and [bolha.com.](https://www.bolha.com/) For the ease of aggregating the information from individual advertisements, the latter has been chosen, with possibility of extending the research onto the former.

Dataset will be generated using web scraping algorithms with the use of Python packages Selenium and Beautiful soup. Given the size of the aforementioned site and its status as one of the two main Slovenian real estate advertising platforms, the goal of the research is to create a ‘useful’ model, which can be used to identify investing opportunities.

# The Goal

The goal of this project is to learn about ML and expand the knowledge of more advanced algorithms and processes. That already sets up the first category of goals, which are related to my personal progress in the field. That includes going more in depth on every stage of the process, mainly in data acquisition, preprocessing, exploratory data analysis (EDA), model optimisation and interpretation.

The second part of the goals relates to the content of the project itself. It consists of designing multiple models and selecting for the best predictor of the price of properties in a regression setting. Based on the price estimates, a tool will be designed that can effectively and reliably provide aid in judgement for whether to rent or buy based on some input parameters and detect whether newly published properties are under or overpriced.

# What Has Been Done

## Web Scraping

In the Github repository from Luka Tašler[[1]](#footnote-0), an implementation of such a method for the purpose of tracking the activity is implemented. It provides the user with information about newly published listings with an automated scanning feature.

I would like to take it a step further and provide insight on whether the new listing is actually a ‘good deal’ or ‘not worth the bill’.

## 

## Real Estate

A renowned ongoing Kaggle competition with the name House Prices - Advanced Regression Techniques[[2]](#footnote-1), intended for advancing ML knowledge and sharing it deals with the problem of predicting real estate prices based on various features (81 to be exact). The dataset consists of roughly 3000 samples split equally between train and test. With more than 30 thousand submissions of various skill levels, different approaches are displayed, plenty of them based on tree models with varying degrees of success.

Predicting Housing Prices: Simple Approach[[3]](#footnote-2) uses multiple decision tree based algorithms, making it a great resource for this project.

However, due to the vastly different dataset, we cannot expect the models to transfer without some additional optimization and oversight.

A paper Identifying Real Estate Opportunities Using Machine Learning[[4]](#footnote-3), written by Alejandro Baldominos et al. details a search for the best performing regression model for predicting real estate prices in the section of Madrid, Spain. It considers 4 categories of models, namely k-nearest neighbors regression, support vector machines, ensembles of decision trees and multi-layered perception. Out of those, the decision tree based model is performing the best, based on several metrics. Authors explain feature selection and transformation using one-hot encoding for non-continuous variables.

The paper is a great example of a project proposal with very in-depth explanations of the procedure. Decision tree based models will be further examined and compared among themselves.

The paper ‘Housing Price Prediction via Improved Machine Learning Techniques’[[5]](#footnote-4) by Quang Truong et al. deals with the Kaggle dataset ‘Housing prices in Beijing’ and explores Random Forest, LightGBM, XGBoost, Hybrid regression and Stacked Generalization. The metric used for evaluation or the models is root mean squared logarithmic error (RMSLE), which offers an advantage over the RMSE as the effect of the difference between predicted and actual price is the same regardless of the price level. Performance of the methods is very closely matched on the test dataset, the best being Stacked Generalization regression with a score of 0.16350 and the worst LightGBM with 0.16944. The paper states that although Random Forest does well on the train set, it is due to overfitting and it does not perform as well on the test.

## Interpretation

Scott Lundberg presents numerous examples of implementation of SHAP package for explanation and understanding of the model on the individual level of a sample.

It comes as an ideal tool for this project as identifying drivers of the price of a property is directly one of the goals, as well as being a very useful insight into the fabric of the market.

# Methodology

## Collecting data

Data is around us, we just have to aggregate and use it. Web scraping is a form of automated aggregation of data from the web. It has become increasingly important in real estate for aggregating offers from multiple websites, forecasting, monitoring competition and reacting to trends[[6]](#footnote-5)

For the purpose of this project, it will be collected from the site [bolha.com/nepremicnine](https://www.bolha.com/nepremicnine), through the process known as web scraping, performed in Python using Selenium and BeautifulSoup4 libraries. It will be then manipulated and stored as Pandas data frame for the ease of use in the following phases. Given the nature of user-generated ads and the inconsistent information available, preprocessing will be necessary to prepare a useful dataset. It will include imputing missing values and manipulation of entries to prepare numeric or categorical features.

## Exploratory Data Analysis

Importance of intuition cannot be overlooked when designing and interpreting a machine learning model. Using Matplotlib and Seaborn, relationships between features will be visualized to gain the first insights.

## The Model

In the case of this paper, finding a model that has the most predictive power is desired, as in so many cases. Using various implementations of decision trees, ranging in complexity from a simple decision tree and RandomForest to LightGBM and XGBoost, each of those models will be optimized and then compared with the rest, to determine the absolute best.

## Testing and Evaluation

Models will be evaluated through root mean square logarithmic error (RMSLE[[7]](#footnote-6)) method. This method is selected to avoid bias towards the more expensive properties.

## Interpretation

A great model is worthless without anyone to interpret it. In this case, SHAP[[8]](#footnote-7) library provides a great tool for determining the relationship between features and their interaction. Based on the outstanding visualizations and its explanatory power, I intend to use it as the main tool in the final phase of the project. It will also enable visualization of key price drivers.

After gaining in-depth insight into the real estate market, the final goal remains, to provide advice on whether to buy or rent. For that, a program will be developed that given parameters such as interest rate of the mortgage calculates the return period of buying instead of renting at rent equivalent mortgage payments.

# Expected Outcomes

I expect that the research process as described above will provide the following outcomes:

* Insight into the real estate market by understanding the drivers of the price of properties (for both; houses and flats).
* Using the models predictive property to assess how well the price of new listings reflect the underlying value given the current situation.
* Evaluate the options of renting or buying based on the currently available possibilities.

# References

1. Kirjazovas, Vytautas. ‘Real Estate Web Scraping: How’s It Revolutionizing this Industry?’, Oxylabs, 11 May 2020: <https://oxylabs.io/blog/data-scraping-for-real-estate>
2. Tašler, Luka. Nepremičnine.net Scraper, Github: <https://github.com/ltasler/nepremicnine.net-scraper>
3. Kaggle dataset House prices-Advanced regression techniques <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>
4. Abhinand. Kaggle, Predicting Housing prices: Simple approach: <https://www.kaggle.com/abhinand05/predicting-housingprices-simple-approach>
5. Baldominos et al. Identifying Real Estate Opportunities Using Machine Learning MDPI: <https://www.mdpi.com/2076-3417/8/11/2321/htm>
6. Truong et al. Housing Price Prediction via Improved Machine Learning Techniques, Elsevier: <https://www.sciencedirect.com/science/article/pii/S1877050920316318>
7. Google, Machine Learning Crash Course: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>
8. Moody, James. Towards data science, What does RMSE really mean: <https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e>
9. Lundberg, Scott Github, SHAP: <https://github.com/slundberg/shap>

1. <https://github.com/ltasler/nepremicnine.net-scraper> [↑](#footnote-ref-0)
2. <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview> [↑](#footnote-ref-1)
3. <https://www.kaggle.com/abhinand05/predicting-housingprices-simple-approach> [↑](#footnote-ref-2)
4. <https://www.mdpi.com/2076-3417/8/11/2321/htm> [↑](#footnote-ref-3)
5. <https://www.sciencedirect.com/science/article/pii/S1877050920316318> [↑](#footnote-ref-4)
6. <https://oxylabs.io/blog/data-scraping-for-real-estate> [↑](#footnote-ref-5)
7. <https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e> [↑](#footnote-ref-6)
8. <https://github.com/slundberg/shap> [↑](#footnote-ref-7)