Introduction to Data Science: Course Project
Real Estate Prices

Jüri Jõul, Sten Salumets, and Tarvi Tepandi

University of Tartu

1. INTRODUCTION

The goal of our project is to create a real estate appraisal model. The project is available at https://github.com/Jyrijoul/real_estate_prediction. This report is a part of the 10th homework for the Introduction to Data Science course in University of Tartu and will give an overview of the business understanding, data understanding, and planning of our project.

2. BUSINESS UNDERSTANDING

2.1 Identifying our business goals

2.1.1 Background

There are several real estate sites in Estonia, one of the most popular being www.kv.ee. At the time of writing (28. November 2021), the site has over 5000 active advertisements all over Estonia. Some works similar to ours have already been done, most notably as a project in the same course done last year by Mattias Sokk, Marko Sasi, and Gervin Ilus. An important difference is that while they tried to predict the prices of apartments, we are trying to predict the prices of all real estate. Also, we will try to incorporate other data, such as the population density, into our models.

2.1.2 Business goals

We want to predict the prices of real estate. In addition, it would be useful to know how various factor, such as location and the floor number, affect the prices. The result of our work could be later sold as a service to companies and private citizens who want to benefit from a quick estimate on the price of their real estate.

A similar end result in the form of a built-in real estate price calculator already exists (https://www.kv.ee/kalkulaator), but this is targeted only at apartments. Our business goal is to target every kind of real estate: apartments, houses, premises, etc. Moreover, since the built-in calculator requests only a few parameters, it gives a very wide estimate. As a test, we got an estimate between 111k to 152k euros for a real apartment which is definitely not precise enough. By having more features to analyze, our goal is to give a narrower estimate than that.

2.1.3 Business success criteria

Our training model will be evaluated by its accuracy and precision to predict the prices for a test set of real estate advertisements. We aim to have a more narrow estimate than the calculator built-in to the real estate website. This way, our implementation could become more preferable to the free option.

2.2 Assessing our situation

2.2.1 Inventory of resources

For the project, there are a variety of resources available. Starting with the www.kv.ee advertisements and the data within. Also, we have access to their free-to-use price calculator, which we can use to make comparisons to our model. We shall also use other data sources to add features, such as the population density of the area. As for the tools, we have access to many free Python libraries for web scraping, data science, and machine learning.

2.2.2 Requirements, assumptions, and constraints

In order to train our model, we require a set of advertisements. We have to assume that there is a correlation between the features of an advertisement and the listed price. A constraint is that we may have to limit the number of features based on how much data we will be able to gather. The models' performances may also be constrained by the amount of advertisements currently available, as this will determine our training set size. Should we add additional features, such as the population density at the location of the real estate, we would also need access to that data.

2.2.3 Risks and contingencies

The two major risks are time management and projects in other courses. The solution for them requires setting up a precise plan for completing this particular project. An unlikely and unfavored alternative is to create everything just briefly before the deadline. As a possible contingency, we could skip some of the data processing and use the data with less features than planned, since the data we have gathered already has some very useful features even without additional processing.

2.2.4 Terminology

We could not identify any specific terminology that should be brought to attention here. All terms are related to either data science or real estate in general.

2.2.5 Costs and benefits

As the data gathered is from free to use sources, there are no direct costs in data gathering. However, that and data processing will have indirect costs in computational power and time.

The benefit of such a project is that it will create another potential source for a real estate appraisal.

2.3 Defining our data-mining goals

2.3.1 Data-mining goals

For this project the data to be mined is clear. Advertisements from `www.kv.ee` are scraped to provide us with training and test data. The pages are formatted in a uniform way so extracting specific features is relatively easy.

Some difficulty may still be ahead if we wish to add features from external sources based on preexisting features. The main such feature we would like to include is the average population density of the area for advertisements.

2.3.2 Data-mining success criteria

Data-mining is considered successful upon creation of a model that can take a variety of inputs describing a potential real estate advertisement and predict an appraisal with relative accuracy and precision. A direct comparison could be made with the preexisting free to use `www.kv.ee`'s own online appraisal tool.

3. DATA UNDERSTANDING

3.1 Gathering data

Our data needs to be enough for reliably predicting the price of real-estate. Our dataset will have to include features that numerically or categorically influence the price of real-estate the most. Also, the goal feature of price would have to be included as a label to be trained for.

The data features available to us are all the sub-articles of a real-estate advertisement. This includes properties like total area, number of rooms, year of construction, energy mark etc. Unfortunately, no public dataset of real-estate advertisements exists. This means that before doing anything related to machine learning, we will have to scrape the data together. Luckily the search page URL of `www.kv.ee` comfortably includes all the search parameters. We have already had some success with Python library called Beautiful Soup. We have not yet managed getting data from the population density database. The database is however public and should not show any real difficulty.

When we finally have scraped all the data of available advertisements, we would have to select which features are relevant to our cause. This would be done by finding out which features most strongly correlate with a lower or higher price of real-estate.

3.2 Describing data

We have had success scraping together data from www.kv.ee using Python library called Beautiful Soup. The dataset includes parameters from every advertisement from the website, for example: location coordinates, number of rooms, number of bedrooms, total area, year of construction, number of floors etc. Total number of features from advertisements is 20. The feature to be added in the future is population density.

The format of the data is the same as it appears in the website and will need further processing. For example the total area feature has the unit m^2 added to every field and will have to be removed. Some features like for example 'extra features' is actually a list of real-estate properties and will have to be split.

After preprocessing our generated data seems very promising for the task.

3.3 Exploring data

Name of feature	Description	Distributions	Signs of problems
Description	Short description of the advertisement. Includes type and size of the estate	2093 unique descriptions with repeating keywords	Needs to be split
Location	Coordinates of the real-estate. Can be useful for adding population density to the machine learning algorithm	2220 unique locations	None
Rooms	Number of rooms	Values are in range of 1 to 47. Mean value is 5.	None
Bedrooms	Number of bedrooms	values are in range of 1 to 33. Mean value is 3.	None
Total area	Total area of the real-estate in square meters	Values are in range of 1 to 169900. Mean value is 570.	None
Floors	Number of floors	Values are in range of 1 to 5. Mean value is 2.	None
Year of construction	Year of construction	Values are in range of 1 to 11934. Mean value is 1981.	Unrealistic values will have to be filtered out.
Ownership	Type of property	Feature has six unique values, like private property and co-property.	None
Ground area	Ground area of the property in square meters.	Values are in range of 2.6 to 679037. Mean value is 11102.	None
Energy mark	Energy mark of the building	The feature has 9 unique values. Seven are for energy marks from A to H. Two are for missing energy marks.	Two missing energy mark values should be combined into one.

Additional information	Extra information about the advertised real-estate.	Various keywords about the real-estate like the existence of balcony, basement, type of roof.	Should be split into separate features.
Kitchen	Description of kitchen properties.	Different keywords about the state of the kitchen like wood stove and the existence of kitchen furniture.	Should be split into separate features.
Sanitary arrangements	Description of the sanitary arrangements.	Different keywords about the state of the sanitary arrangements like sauna and shower	Should be split into separate features.
Heating and ventilation	Description of the heating and ventilation state of the building.	Different keywords about the state of heating and ventilation aspects like heated floors and conditioner.	Should be split into separate features.
Neighbourhood	Description of surroundings of the real estate	Different keywords of the surroundings like proximity to a lake or a forest	Should be split into separate features.
Condition	Condition of the building.	Different categories of the condition	None
Communications and security	What communication ways and security measures are available.	Different keywords like internet and neighbourhood watch.	Should be split into separate features.
Readiness	How liveable is the real-estate in its current state.	Six unique keywords like ready and foundation.	None
This floor/ Number of floors	What floor is the real-estate being sold at and what is the total number of floors of the building.	The floors range between 1 to 12.	Should be split into separate features.
Expenses in the summer/winter	Cost of utilites in summer and winter in euros.	Costs range between 30 to 500 euros.	A lot of advertisements have not added this information. Might not use it because of this.

3.4 Verifying data quality

Our data is good enough for supporting our goals. It needs some work like separating features, but the information is all there.

4. PLANNING

The tasks needed to be done for successful completion of the project (with the estimated contributions in hours) are the following:

- Scrape together data from all the advertisements of www.kv.ee
  Jüri - 5 h
  Sten - 3 h
  Tarvi - 2 h

- Find a way to add population density feature, using the location feature from www.kv.ee
  Jüri - 3 h
  Sten - 2 h
  Tarvi - 5 h

- Choose the features that seem promising for machine learning
  Jüri - 3 h
  Sten - 5 h
  Tarvi - 1h

- Clean data
  Jüri - 5 h
  Sten - 5 h
  Tarvi - 1 h

- Properly format the data
  Jüri - 1 h
  Sten - 1 h
  Tarvi - 1 h

- Research suitable machine learning algorithms
  Jüri - 3 h
  Sten - 3 h
  Tarvi - 2 h

- Train models
  Jüri - 7 h
  Sten - 3 h
  Tarvi - 5 h

- Evaluate
  Jüri - 1 h
  Sten - 1 h
  Tarvi - 2 h

- Making the introductory video of the project
  Jüri - 1 h
  Sten - 6 h
  Tarvi - 8 h

- Making the poster of the project
  Jüri - 1 h
  Sten - 1 h
  Tarvi - 3 h

Here is the list of methods and tools we plan to use:

- Python and Jupyter Notebook
  We plan to do most of our programming in Python. We may also use Google Colaboratory to aid collaboration.

- Sklearn machine learning library
  We plan on using more traditional machine learning algorithms. For models like that, Sklearn is one of the best and easiest libraries to use.

- Beautiful Soup and Requests Python libraries
  Beautiful Soup is a web-scraping library and Requests is a library for making the queries to the websites. Both are Python libraries and are essential for our project.

- Adobe software for presenting the project
  Adobe Premier Pro 2020 could be used for making the video and Adobe Illustrator 2020 will be used for the making of the poster.