Kaggle: House Prices - Advanced Regression Techniques

Team members: Maj-Annika Tammisto

Business understanding

Background

I´m preparing myself for PhD studies which will hopefully begin in the University of Tartu sometimes in 2023. The first part of my PhD project concentrates mainly on gathering and analyzing data and finding useful patterns. Considering that working with data in that sense is a new task for me, I decided to take the LTAT.02.002 course (Introduction to Data Science) in order to get a deeper understanding of how to work with data and especially, to learn some methodologies that would be useful for my PhD project.

One prerequisite for passing LTAT.02.002 is to prepare and present a data science project on a topic chosen from a good variety of options. Out of these topics I chose the one where the risk of losing time due to not being able to download data, the instructions being unclear or any other reason is as low as possible so that I could really focus on practicing and learning the huge amount of concepts and principles that were included in the lectures and practice sessions. Therefore I´m working on the Kaggle House Prices - Advanced Regression Techniques competition which is, according to the description, a good choice for expanding one´s skill set before trying a featured competition or in my case, before diving into my PhD data.

Business goals

The project is no meant to benefit a business but I myself will benefit from it and the Kaggle community hopefully as well, as they will gain a new competitor and some additional entries.

People and organizations of Estonia will benefit from the skills learned during this project, as they will be used for developing e-Estonia.

The same competition was chosen by team D1 in 2020. The reason for choosing it was similar to mine, as the team also wanted to practice their skills before moving on to any more challenging projects. In order to add additional value to this course by choosing the same topic, I have defined different data-mining success criteria for my project.

Business success criteria

1. The final Kaggle entry with house price predictions is submitted before December 12th 2022.
2. The project poster PDF is submitted by December 12th 2022.

Inventory of resources

1. Project description, tutorials, FAQ and additional information on https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview
2. Project data: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data
3. Project Kernels: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/kernels
4. LTAT.02.002 lectures and slides: https://courses.cs.ut.ee/2022/ids/fall/Main/Lectures
5. LTAT.02.002 practice session extra material: https://courses.cs.ut.ee/2022/ids/fall/Main/PracticeSessions
6. Tutorials: https://courses.cs.ut.ee/2022/ids/fall/Main/Tutorials
7. Consultation for projects on December 6th 2022.
8. Lecturer and teaching assistants in Campuswire.

Requirements, assumptions, and constraints

Project instructions are available in Google docs: https://docs.google.com/presentation/d/1wNL5wSGq1mtP9YzIsYMb-0a_XN9h_TM2y5XkX9mQYm4/edit#slide=id.g2b7c3c841b_0_193.

Important dates are:

December 12th 2022: submission of the poster PDF.

December 15th 2022 at 14:00-17:00: Poster session in Delta.

Risks and contingencies

RISK	CONTINGENCY
Underestimating the scope and complexity of this data science project that leads to an insufficient result.	Getting started with actual practical work as soon as possible in order to get an understanding about the size of the challenge as early and possible so that there would be time for finding solutions.
Getting "stuck" with the project and not knowing how to proceed.	Reaching out to teaching assistants and asking for advice. Trying to find helpful discussion topics or Kernels.
Failing to submit the poster PDF on time due to unexpected reasons (for example Internet or electricity outage).	Finalizing the poster PDF at least two hours before deadline so that there would be time to find a solution for submitting it, if necessary.
Not being able to attend the poster session due to illness or other valid reasons.	Providing the reason and proof of absence and agreeing on an alternative date.

Terminology

SalePrice - the property's sale price in dollars. This is the target variable (label) that will be predicted in this project.[1]

Costs and benefits

The most significant cost of this project is the estimated cost of time, approximately 60 working hours.

In addition, there will be travel costs from Tallinn to Tartu for the poster session on December 15th.

It´s a very good investment in education and skills which is a very important benefit.

[1] https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data (25.11.2022)

Data-mining goals

1. README file, HW10 reports and project code in https://github.com/annimaj/KaggleHousePrices
2. The final submission file (.csv) for Kaggle.
3. A poster PDF.

Data-mining success criteria

1. At least three regression techniques are practiced on the project data.
2. The final Kaggle entry submission is among top 100 on the leaderboard.
3. At least 15 points are earned with the project and poster session.

Data understanding

Gathering data

- Data requirements

The Kaggle House Prices - Advanced Regression Techniques competition has been available for some years already and thousand of teams have submitted their entries. Therefore I´m quite confident that the data provided in Kaggle is sufficient for reaching the data-mining goals of this project.

- Data availability

Competition is available on https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data and it has been added to the project repository https://github.com/annimaj/KaggleHousePrices.

- Selection criteria

The following files are relevant for this project:

data_description.txt: will be used for understanding the meaning of data variables.

train.csv: will be used for training different models created during this project.

test.csv: will be used for testing a chosen classifier at the very end.[2]

[2] https://courses.cs.ut.ee/MTAT.03.183/2017_fall/uploads/Main/domingos.pdf, page 80 (26.11.2022)

Sample_submission.csv: will be used as an example when creating submission files.

Data description report

Data is provided in .csv format.

Training data file contains 1460 instances and test data file contains 1459 instances of data.

There are 79 different variables that can be used for predicting the SalePrice label.

The provided data is suitable for the data-mining goals of this project.

Data exploration report

The 79 data variables are, according to the competition description, describing (almost) every aspect of residential homes in Ames, Iowa[3]. The data_description.txt file provides a detailed explanation of possible values of every variable.

For predicting the value of the SalePrice label, I will potentially use the following variables:

1. Location information: MSZoning, Neighborhood
2. Potentially value-adding features: LotArea, Utilities, Condition1, Condition2, OverallQual, OverallCond, ExterQual, ExterCond, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinType2, HeatingQC, CentralAir, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, Bedroom, Kitchen, KitchenQual, TotRmsAbvGrd, Functional, Fireplaces, FireplaceQu, GarageCond, PoolQC, Fence, MiscFeature
3. Sales information: MiscVal, MoSold, YrSold, SaleType, SaleCondition

Data quality report

The data needed for this project is available and added to the project repository https://github.com/annimaj/KaggleHousePrices.

The preliminary quality check shows that there are several variables that contain NaN values. Otherwise the data seems to be complete and suitable for reaching the data-mining goals of this project without the need to obtain or gather any further data.

A more thorough data quality test is to be made as one of the next steps, in order to double-check if the variables that I will potentially use indeed contain only the values that are described in the

[3] https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview (28.11.2022)

data_description.txt file. Should there be any additional values present, they need to be either interpreted as valid values or excluded as typos or errors.

The code file house_prices.ipynb has been added to the project repository https://github.com/annimaj/KaggleHousePrices and it will include the code used for data cleaning.

Project Planning

There is one team-member in team A5, therefore all project tasks will be done by one person.

The estimated total time to be allocated for this project is 60h.

Project tasks

1. Data preparation according to the CRISP process (25 h)

- Selecting data
- Cleaning data
- Constructing data
- Integrating data
- Formatting data

2. Modeling, regression technique #1 / Random Forest (10 h)

- Designing test
- Building model
- Assessing model

3. Modeling, regression technique #2 / tbd (10 h)

- Selecting modeling technique
- Designing test
- Building model
- Assessing model

4. Modeling, regression technique #3 / tbd (10 h)

- Selecting modeling technique
- Designing test
- Building model
- Assessing model

5. Evaluation and poster PDF creation (5 h)

The same amount of time is allocated for steps 2,3 and 4, although there are more sub-tasks in steps 3 and 4. It has been done under the assumption that the first modelling task will require some more time due to practicing feature selection, practicing coding skills etc. and that that the following modelling tasks will go a bit faster.