Predicting house prices

Project A11: Raul Tölp, Hendrik Šuvalov

Task 1. Setting up

Link to the repository: https://github.com/hendriksuvalov/housepriceskaggle

Task 2. Business understanding

Identifying business goals

The people who benefit from this project are people who are either seeking to buy real estate or already have a property and wish to sell it. For both of those people, it would be useful for them to know what factors influence the house price. Those who wish to sell would know how to set the price of their real estate and those who wish to buy would know if the price was fair.

Our goal is to create a model that helps the users do this and analyze what factors influence the price of the real estate. With the latter information, the user can take into account how their preferences (e.g. they are insistent that they want their house to have a balcony) influence the price of the house.

Due to the fact that the data is linked with a Kaggle competition, our success criteria can be directly measured by how accurate our results are once we submit them. The higher we place on the leaderboard, the more successful we can consider our project.

Assessing the situation

Our team consists of two students with some experience in the field of data science. We have standard university laptops and will be using Jupyter Notebook with Python. Our data is the Ames Housing dataset, which was compiled by Dean De Cock. It consists of about 3000 observations which include 79 features along with the price, describing the sale of individual residential property in Ames, Iowa from 2006 to 2010.

The project has to be completed by December 16, 2021, when we have to present the results in a poster session. By that time, we will have to have developed a working model along with analysis of the data. We will also be required to make a poster, describing the results and visualizing our findings.

We do not foresee any great risks during this project, with the exception of perhaps managing our time poorly. Should we misassess how long any given step takes us, we will have to reschedule individual tasks on-the-fly and reprioritize them in order to ensure completion of the project.

We expect that everyone involved in this project has a clear understanding of the terminology involved, as it is relatively simple and does not contain anything new domain-specific.

There are no costs involved in this project, other than the time resource of the members, which is expected to be 30 hours per member. The expected benefit is a generous assessment of our results in the form of a good grade.

Defining data-mining goals

Our goal in the context of data-mining is to gather the relevant features in the given dataset that influence the house prices and use them to develop a model that predicts the price of the house. The actual deliverable we have to present is predictions of test data, which we have to submit to Kaggle. Then Kaggle assesses how accurate our predictions are by scoring the submission using root mean squared logarithmic error.

Task 3. Data understanding

Gathering data

Gathering data is not in the scope of this project, as it has already been gathered by Dean De Cock. The data initially came to him directly from the Assessor's Office in the form of a data dump from their records system and initially had more features, however, he removed some, most of which were related to weighting and adjustment factors used in the city's modeling system.

Describing data

The training data consists of 1460 rows which describe a property sale that took place in Ames, Iowa between 2006 and 2010. The features range from being very specific, such as the type of exterior covering on the house, to quite generic, such as a 1-10 rating of the overall condition of the house. We expect all the relevant price-influencing features to be included in this dataset, though we suspect that some features are also irrelevant and thorough analysis has to be done to separate the two.

Exploring data

Each row has 23 nominal, 23 ordinal, 14 discrete and 20 continuous variables (also including the price itself), though, in some cases, certain features might be not available. For example, the feature describing the type of alley access to property has 93% missing values, which already tells us that it is not going to be very useful for the prediction. We also suspect that some very specific features, such as Exterior 2nd, which is the second exterior covering on a house should the exterior be of more than one material, will likely not be the best indicators of price, as they do not seem very relevant in pricing the house. What we hypothesize to be very important factors, however, are the year in which the house was built, which we associate strongly with the quality of the house, the lot size in square feet, as the bigger the house, the higher the price and any quality indicating features. There is also a feature that describes the condition of the sale (normal sale, sale between family members e.t.c), which we think might have a strong influence on the pricing of it, as it describes the motivations for selling the house.

Verifying data quality

The source of this data seems reputable enough, as it is used widely in a popular Kaggle competition and no major complaints have been made. There seems to be a sufficient amount of it to make a relatively accurate model, as the results in the leaderboard seem very strong. The most severe issues that we seem to be facing are the missing values for many features and some strong outliers (for example, 30 houses built in 1899 - 1913, whereas the mean is around 1970. Apart from that, the data seems relatively clean and ready to be used.

Task 4. Planning your project

Detailed plan of tasks

All the tasks are followed by how many hours we suspect they will take for both team members combined.

- Importing data into work environment: ~ 1 hour
- Initial exploration: ~ 1 hour
- Pre-processing (removing outliers, NA-s, e.t.c): ~ 2-3 hours
- Analysis, feature correlations, e.t.c: ~ 4-6 hours
- Feature engineering: ~ 4-6 hours
- Implementing different models (we suspect 3-4, including Random Forest & Linear regression): ~ 6-12 hours
- Tuning parameters and testing the models: ~ 6-12 hours
- Poster & presentation: ~ 10 hours

Methods and tools

The data will be imported to an environment containing a Jupyter Notebook file, in which we will write all the necessary code for the analysis, models and outputting the predictions using

Python. We hypothesize that we will mostly be using Pandas and NumPy for handling the data and scikit-learn for training the relevant models.