Project D16: KAGGLE-House-Prices

TEAM:
Kalmer Keerup
Arlo Tammekun
Mehis Taevere

Business understanding

Identifying your business goals

Business goal of this project is to identify the areas of improvement potential in house sales in Ames, Iowa. This enables optimizing the house sales prices and better identifying the competitors miscalculated house prices.

This project will discover the most significant factors of house sales price and the most significant fixable factors.

Assessing your situation

Our project is most limited on human resources, what we have, are three team members working on this from spare time from other university work and challenges. We will need to create "internal" deadlines for ourselves to have more discipline for completing tasks and reaching milestones on time. Our research is centered around the Kaggle competition "House Prices", which also provides us with training and testing data.

Terminology used within the project will correspond to the terminology used in the course this project is created for, "Introduction to Data Science" in the University of Tartu.

Defining your data-mining goals

We aim to produce models for predicting housing prices based on provided features, to help validate future house pricing decisions and report on the charasteristics and main factors of house pricing.

We aim for prediction models to have accuracy of at least 75% and to find three main factors for house pricing.

Data understanding

Gathering data

Data has been aquired from the Kaggle "House prices" competition.

Describing data

The Ames Housing dataset was compiled by Dean De Cock in 2011 and describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. (http://jse.amstat.org/v19n3/decock.pdf) The data consists of a training set and test set, each having 1460 instances of data and 80 features (23 nominal, 23 ordinal, 14 discrete, and 20 continuous). Descriptions of features can be found in data_description.txt.

As the data has a large number of features detailing various information about residential properties it is suitable for our purposes.

Exploring data

A cursory look over the data revealed no problems and showed that the data is in line with the descriptions from data_description.txt.

Verifying data quality

The data is of good quality and very suitable to our purposes.