Sander Ruusmaa
Uku Roio
Georg Demidov
28.11.2020

House price prediction

Business understanding

We have been asked to deal with a problem. There is a copious amount of data regarding houses in Ames, Iowa with information about various features of the house and their qualities. Trying to price a house based on its many features is a cumbersome and time consuming task for a real-estate agent. There may be many unknown features that silently reduce or increase the price of real-estate. Furthermore, automatisation of this process would be greatly beneficial for various real-estate companies or individual real-estate agents.

For this reason, we would like to give our solution to this problem and increase the real-estate sale numbers. If this project turns out to be a mild or even great success, then it would be possible to implement this machine learning algorithm on other real-estate all over the US or even Europe. The assessment of the success ratio of the algorithm would be to see if houses will get sold fast with prices that we predicted.

An additional feature of our algorithm would be to give a rough estimation of a buyer's "dream home". For example, if a person were to come and say that they want to have a paved driveway, a porch and a fireplace in their home, then they would get an estimated price of such a home.

Assessing our situation

To solve this problem, we have a team of 3 versatile potential upcoming data-scientists. The data gathering task of our problem has already been taken care of by Dean De Cock, who has done the hard work. Python, along with its many tools are in our disposal to tackle our task.

The data that we have at our disposal is anonymised and open source. Since it is hosted by Kaggle and the competition is available to all, then there are no legal issues regarding this project. Our project is scheduled for completion in about 2 weeks.

Terminology:

Machine learning - Machine learning is the study of computer algorithms that improve automatically through experience.
Data - facts and statistics collected together for reference or analysis
Feature engineering - Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques.
Feature selection - In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction.
Framework - an essential supporting structure of code

Classifier - In the terminology of machine learning, classification is considered an instance of supervised learning, i.e., learning where a training set of correctly identified observations is available.
Accuracy - In measurement of a set, accuracy is closeness of the measurements to a specific value
Standard deviation - In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values.
Mean - The mean (average) of a data set.

Data-mining goals

Since we have already been given a data set, then our job would be to analyze it and start feature engineering it. We would have to refactor the data so we could later give it to a classifier, such as Random Forest Classifier, and train it with our train data. Of course we would leave a part of the data untrained, as a test data.

It would be a great success if our classifier had an accuracy of over 90%. That would be our initial goal of the classifier.

Gathering data

The data has already been gathered by Dean De Cock. Since the data set is already given to us, we do not have to gather it. Our data set should have many features since it is hard to predict a price of a house based on only a handful of features. The data should preferably be mostly numerical, although verbal and descriptive data makes it easier for the human to understand its value.

The data set has already been used in many ways to educate data science students. It is a handy alternative for the already overused Boston dataset, which has been used for educational purposes as well.

Data description

It contains 79 features of a house, like type of alley access, slope of property, heating quality, garage condition etc. The features vary from integers, doubles to verbal descriptions. This will make training the classifier harder, since computers work best with numerical values. We believe that the size of the data set is sufficient enough to train a classifier. We know nothing about the source of the data, nevertheless, Kaggle should be a reputable data source.

Exploring data

The first thing that strikes the eye is that there are missing values of course, but the “Alley” feature seems to have the most missing values - 93% of it is NA. Utilities feature seems to be quite pointless as well, 100% of it is the same value. Since the SalePrice feature seems to follow a normal distribution (this is the feature that our classifier will try to predict). The features that we found to potentially have a significant impact on the pricing of a house were

”YearBuilt”, “KicthenQual” and “MoSold”. “YearBuilt” contains the year in which the house was built. It is a numeric value ranging from 1872-2010. The most densely populated year range is from 1996 - 2010 which possibly indicates a higher sale price for a house and a population increase in Ames. “KitchenQual” feature indicates the quality of the kitchen in the house. The values are verbal descriptions ranging from “Excellent” to “Poor”. There are 5 values in total with the other 3 being “Good”, “Average” and “Fair”. “Average” and “Good” are the most frequent values. “MoSold” shows the month in which the house was sold. The values are numerical ranging from 1-12. The highest amount of houses sold was during the summer, with autumn being the time of month with the least amount of sales. We also found the “CentralAir” feature contains mismatched values since it is expected to contain boolean values but instead it is filled with string values of “y” and “n”. However, this is a problem well within our brain capacities, therefore it should be an easy fix.

Verifying the data

Since the data has already been given to us then there is no way that we can not obtain it. The data has enough features to solve our task which is to create a classifier which could predict house prices. Of course, feature engineering has to be done in order to feed the machine train data. This is due to the fact that our data has some weak points - whether some of the data is missing or a feature is completely irrelevant and may throw off our machine learning algorithm. Overall, the data is very nicely organized and completing our goal of a house price predicting algorithm is highly probable.

Planning your project

Cleaning the dataset is the first task to be tackled by us. We will deal with missing values by replacing them with the mean of the column or by deleting the entire feature if it proves to be insignificant. We will also correct any mismatched values by applying the correct data type to the column. Lastly we will identify all the outliers in the features by scatter plotting the features and removing the row with the outlier(since we have a big dataset, we can afford to remove some of the rows). We will be spending approximately 3-4 hours(per person) on this task.

The next issue on our agenda is feature selection. We will ignore features that contain 100% of only one value when training a classifier. Later we will calculate the significance scores of the features and remove features with low scores. This task will take us 5 hours.

Then we will do some feature engineering to transfer the data so that the classifiers can understand it. We will use Pandas get_dummies to get rid of non numerical values. This task will take us 6 hours.

After that we will train different classifiers like random forest with our modified data and create prediction models. We will calculate the accuracy scores of the models so that we can improve them if necessary. This will take us 5 hours.

Second to last, we’ll find the features that had the biggest impact on pricing for each classifier and analyze them. This will take us 4 hours.

Lastly, we will create the poster. We will include the results of our prediction models and graphs with important features. We will hopefully bring out various interesting facts about our data. Creating the poster will take us 6 hours.