"Predicting House Prices" project report

Group C6
Link of the repository: https://github.com/laurakoldekivi/DS-2021-project

Task 2. Business understanding

1. Identifying business goals

For this project we are using the "House Prices" dataset from Kaggle. Kaggle is a platform for data science enthusiasts, where it is possible to expand your skills by competing with others in competitions on machine learning tasks. Many companies can post real problems to the site, provide data and let data scientists work on them. This particular dataset that we are using is not provided by a company, but was compiled specifically for the use in data science education. This is meant to be a suitable competition for new data science students who wish to expand their skill set. This is why our project does not benefit directly a business and does not have any specific business background or goals. For this project, the people who will benefit from it are the people who are conducting the project. This is why we will analyze the "business understanding" task from the perspective of our group. However, in addition to that, we will also analyze in what other business contexts potentially our project results could be used.

For our group, the "business" goal is to expand our data science skill sets and put into practice what we have learned from the course so far. We will measure the success of our project and its results by our position on the Kaggle leaderboard, as we aim to be in the top 50%. We believe that this is a good criterion for success, because in order to train a model with high accuracy and be in the top 50%, we need to successfully apply much of what we have learned during the course.

We also recognize that potentially this type of project and its results can benefit various stakeholders, such as house owners, house buyers, real estate companies, investors and creditors. For all of them, it is important that the value of a property is accurately estimated, so that they can make rational decisions in property transactions.

2. Assessing situation

To complete the project and achieve the desired results, we have a team consisting of three people. We also have a teacher and several teaching assistants who can help and guide us if necessary. These are our human resources. We also have other resources, such as course materials in video and slide format, the Kaggle dataset, computers and Jupyter notebook. The only requirement we have for the project is a set deadline for project completion, as we need to present the project on the 16th of December. We also have some potential risks, such as that we will not be able to finish the project by the deadline or that we will not be able to achieve our goal of reaching the top 50%. Both of these risks can be reduced by starting to work on the project early on, so that we will have time to meet the deadline and adjust our models to improve accuracy. As for the terminology of the project, it matches the terminology used in this course, and since all team members are familiar with it and on the same page, we deemed it not necessary to create a glossary with the definitions. The only cost for this project is our time, but the benefits are increased knowledge in data science and a good grade for the course. For us, the benefits outweigh the costs.

3. Defining data-mining goals

Our data mining goal is to identify the most important features that affect house prices, then build a model that would be able to predict house prices with good accuracy, and finally report our findings in the form of a poster and video. The success criteria for our data mining goal is the Kaggle evaluation, which is based on Root-Mean-Squared-Error. Since our goal was to be in the top 50%, our model's RMSE can at most be 0,14501.

Task 3. Data understanding

Our data science project uses data from a Kaggle competition. Dean De Cock gathered and prepared the data set. He gathered this data set because there weren't any other good recent home evaluation data sets and, although he found several potential data sets, they were rather limited in the number of observations ($n \leq 100$). The data set contains property sales that had occurred in Ames, Iowa, between 2006 and 2010. The original data was in Excel and given to him by the City Assessor's Office after a brief meeting.

After Dean De Cock removed extraneous variables, about 80 variables remained that were directly related to property sales. The variables focus on the quality and quantity of physical attributes of the property. Most of them are the type of information a home buyer would want to know about the property. The final edited data set includes 2930 observations and 79 variables for analysis, leaving 1 to be the actual price of the property. There are 20 continuous, 23 nominal, 23 ordinal, and 14 discrete explanatory variables. The continuous variables relate to various area dimensions for each observation. The 14 discrete variables quantify the number of items occurring within the house. Most focused on the number of bedrooms, bathrooms and kitchens. Data about the garage capacity and construction/remodeling dates are also recorded. The categorical variables range from 2 to 28 classes, with the smallest being street and largest neighborhood. The nominal variables point out various types of dwellings, garages, materials and environmental conditions, and ordinal variables rate various items within the property. This final data set includes all expected fields and sufficient cases for our project analysis to predict house prices and make different correlations between fields and property prices.

In Kaggle, they have divided this data set into two, test and train. This means that we have to perform data cleaning on both to make them ready for training, analysis and building our models. Both data sets have missing values we have to replace. It won't be a big problem, because a lot of the missing values stand there for missing items, which we can replace with another new class. Other than that, there aren't quality problems with this data set. After training, we can test our train dataset on the test data set to assess the performance of our models.

This data set by Dean De Cock is enough to support our goals. It has enough data to find correlations between observations, their variables, and price. The only problem might be that the data is quite old by now since it's about property sales between 2006 and 2010 and house prices fluctuate a lot. We believe that even if the prices are different, criteria for buying a house remains mostly the same and this dataset will help us analyze how prices differ based on different factors and which factors change the potential house buyer's opinion the most.

Task 4. Planning your project

Steps to follow for the project:

1. Data preprocessing is the first and most important task:
We have 79 variables for analysis, affecting the price of the house. But some of them have null values and some of them can't be replaced with similar values. So we need to clean the data, figure out the most important variables and deal with them by removing duplicates, incorrect, incomplete, or corrupted data.

2. Data visualization for the high-level view of data:
Visual analysis through plots, bar graphs, histograms, correlation matrices gives us an overview of how the data is distributed. In our case, like the year of sales vs number of units sold, neighborhood vs sales price, heatmap for correlation and heatmap of categorical and numerical variables concerning sales price.

3. Choosing a model
After filtering insignificant variables, we split our dataset into training and test set (70% vs 30%). Starting with regression and then trying different algorithms mostly Random Forests and XGBoost.

4. Training the Model
Post-training model on different algorithms, it's time to see the effectiveness of our models and run it on a test set split from original data. This will give us an idea of how the model may perform on the actual test set. We will evaluate the performance of the model using metrics like accuracy, confusion matrix, F-measure.

5. Parameter Tuning and Prediction
For improvising the training, we go back to initial assumptions and make some changes, like changing the correlation threshold from 40% to 30%. Finally, we run the algorithm on the actual test set and predict the values for our target houses and make a conclusive statement.

We all plan to spend at least 20 hours per person individually, other than group meetups and discussions. Since it's a group task we all are working together and helping each other with their allocated tasks, and we plan to work only with Python using Jupyter notebook. We can use Tableau for visualization, but using python makes more sense since the entire project is done on python.