University of Tartu
School of Economics and Business Administration

PREDICTING HOME PRICES:
Predicting home prices in Ames, Iowa using advanced regression techniques

Group members:
Lars Bosgraaf
Bill Sendewicz
Stas Sochynskyi

Supervisors:
Meelis Kull,
PhD, Associate Professor in Machine Learning
Laura Ruusmann,
Master's student, Teaching Assistant

Task 1. Business understanding

Business goals

The motivation for this project stems from our interest in real estate investing, and completing a project of this magnitude could give us insight into the drivers of home sale prices and the demand for certain types of real estate. We also felt that predicting the sale price of homes would be the perfect problem on which to practice our newfound regression analysis, predictive modeling and data cleaning skills, and if our model performs well, perhaps it could be useful to future users outside of this class. Possible users may include realtors, home buyers and sellers and government assessors.

The task is straightforward: to develop the best regression model possible for predicting house sale prices based on the Ames, Iowa housing dataset. However, we faced two opposing criteria in deciding what type of model to build: on the one hand, we could build an exhaustive model that considers the impact of all the available features in the dataset and their impact on home prices. On the other hand, we have been contemplating whether we can make our model more useful if it follows Occam's Razor: specifically, whether is it possible to capture most of the predictiveness of a larger model but with many fewer features (perhaps 10-20 in total). A simpler prediction model could be especially useful for nontechnical users or if one were to do "back of the envelope" calculations to compute the sale price of a home. Regardless of whether we opt for a leaner "Occam's"-type model or an exhaustive model, we will determine our model's effectiveness by how well it predicts the sale prices of homes in the test dataset when judged by some objectively-measured evaluation criteria such as MSE, RMSE, RMSLE or mean absolute error.

Inventory of resources: Lars's, Bill's and Stas's time and personal laptops combined with their ingenuity, know-how and problem solving skills; original Kaggle dataset; Python, Numpy and Pandas programming languages and libraries.

Requirements, assumptions and constraints: The only requirements or constraints are that we each spend at least 30 hours on the project and that we complete and submit the project by Monday, December 16, 2019 at 12 PM. Our main assumption is that the dataset is freely available on Kaggle (it is).

Risks and contingencies: Possible risks include not having enough time to finish the modeling or the presentation file; our model not performing to our expectations; having some unforeseen problem with the data. We will mitigate these risks by beginning our work early, and being regimented and disciplined in our scheduling and in meeting our deliverables. As of writing this document (Sunday, December 1, 2019), we believe the project is more than 60% completed.

Terminology: The only terminology we will need is the terminology specific to American real estate prices and measurements (including Imperial units) and mathematical regression and data science terminology. An explanation of variables in the dataset is available here: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Costs: Aside from the opportunity cost of Lars's, Bill's and Stas's time, there are no direct costs associated with this project. However, possible benefits include an improved understanding of the variables that impact home sale prices in Ames, Iowa and beyond, as well as an increased understanding of home sale price models and their accuracy; reduced time to compute predicted sale price; greater ease of explaining predictive models and their variables that predict home prices to laypeople.

Data mining goals

Goals and Success Criteria: The Kaggle leaderboard is based on the smallest possible Root Mean Square Log Error (RMSLE) on the test data. In that perspective, our primary goal is to place in the **top 20 of the Kaggle competition** when we submit our predicted sale prices on the test dataset.

Another goal that we will consider and strive toward, but not hold ourselves to, is the possibility of **achieving a simplified model** (simplified from the 79 original features in the dataset) that still **has significant predictive power**. As mentioned before, we will be investigating whether it is possible to create a simpler model that is just as, if not more, predictive than a model using all the original features in the dataset.

Additional goal is to submit the project document **by Monday, December 16 at 12 PM** and to earn maximum points on the document, as well as to deliver an outstanding presentation at the poster presentation session.

Success criteria include whether the project was submitted on time; whether we achieved maximum points on the project; whether we placed in the top 20 of all teams at the time we submit our test predictions to Kaggle.

Task 2. Data understanding

Gathering data

For this project we started out looking for an interesting dataset to work with. We started looking for datasets on **Kaggle.com**. We thought this would be a good place to start, as it has a lot of different datasets available. We wanted the dataset to not be too big so that we can try out different models on the data without having to wait too long to train the model.
We found a dataset about predicting the **sale price of houses**. This seemed interesting to us, as we are all interested in real estate. The dataset was not very

big. Additionally, as this is Kaggle competition it adds extra interest to work as it will be possible to compare our results against other participants in the competition.

Describing data

The dataset that was chosen for this project is the **Ames, Iowa housing dataset**. This dataset describes features of **residential houses** that have been **sold** in Ames, Iowa **between 2006** and **2010**. The dataset can be found at https://www.kaggle.com/c/house-prices-advanced-regression-techniques

The dataset was created by **Dean de Cock** from Truman State University as a practice dataset for regressions for students. The dataset is an alternative for the commonly used the Boston Housing Data Set.

The Ames Housing dataset consists of 81 variables and 2919 observations. The dataset has **3 types of variables**:

1. Numeric variables (35)
2. Categorical variables (23)
3. Ordinal variables (23)

For the Kaggle competition the dataset has already been split into separate train and test datasets. The **train** dataset consists of **1460** observations and the **test** dataset consists of **1459** observations. Both datasets are provided in **.csv** format.

Exploring data

We explored dataset using **Jupyter Notebook** and **Excel**. The data can be imported straight into a Jupyter Notebook using ***Pandas*** module in Python. ***Seaborn*** and Excel were used to do different visualizations of dataset. When looking at the first few observations in the data, it is immediately visible that there are NA values in the

data. We will have to deal with those values. There are also 43 variables of object type. These variables will have to be converted to numeric values, so that the regression models can fit the data and create predictions of the prices. For all object type variables we will have to see if they are categorical or ordinal variables and handle them accordingly. One resource we can use to decide this is the data_description.txt file that comes with the dataset. This file has a detailed description of all variables.

Verifying data quality

After finding the data, we looked for the original source of the data. We found the original paper that introduces this dataset. The goal of the paper was to introduce a new dataset for practicing regression analysis for students learning regression models. The paper has a very detailed description of how the data was created, how it can be used and some helpful notes about how the data can be analysed in a different way. According to the paper, the raw data came straight from the Ames City Assessor's Office in the form of a data dump, after a request from the author of the paper. The raw data has been altered by removing some variables that were too complex for people without a lot of real estate knowledge to understand. Most of these deleted variables were related to weighting and adjustment factors used in the city's current modeling system. The creator of the dataset also removed observations of non residential house sales and houses that were sold multiple times during the 2006-2010 time period. Because of the very extensive and detailed documentation of the dataset provided by the creator, we believe the data used for this project to be of good quality.

Reference

A link to the paper introducing this dataset can be found here:
http://jse.amstat.org/v19n3/decock.pdf

Task 3. Planning your project

In general, our work plan for the project is split into the following seven categories:

- Documentation **(D)**
- Data exploration **(DE)**
- Data cleaning **(DC)**
- Feature engineering **(FE)**
- Model training **(MT)**
- Model validation **(MV)**

The first management rule is “one task per person.” However, the general categories are too broad to assign to just one group member. Hence, each category is split into sub-tasks that were assigned to one (or in some cases two) group member.

The “prepare poster” task is a bit complicated to split into subtasks that will take a lot of time, thus every team member is going to contribute to the final presentation either doing visualization, formating or proofreading.

In general, we used **Python** and sometimes we used **Excel** to validate the results.

Table 1. List of tasks with hours workload

*Note #1: Category letters mentioned above.
*Note #2: Time also includes time spent for discussions and presenting results to others, exploring different options

Task name	Category	Description	Resp.	Time (H)
Documentation set up	D	Set up communication channel (FB), Github environment, Google Drive folder for documents, share with other team members	Bill	2
Proofreading	D	Poster grammar error check	Bill	2
Formating	D	Final formating of the poster	Stas	2
Prepare poster	D	Prepare final poster: describe initial goal, main results & findings, prepare the visualization of the data, provide evaluation metrics values, prepare	Bill Stas Lars	7 7 2

		structure presentation of the results		
Examine correlations	DE	Examining the correlation. Excel was used in this part. Identifying which features can be removed	Bill, Stas	6 (3 each)
Data visualization	DE	Visualization of the data. Creating figures for the poster that would be the most telling	Bill, Stas	6 (3 each)
Data transformation	DC	Identify ordinal features that can be transformed to numeric, transforming values from ordinal to numeric	Stas	4
Remove outliers	DC	Remove outliers from the dataset. Includes temporary visualization.	Stas	4
Remove features	DC	Examine whether feature has enough observations of each type (balanced or imbalanced). Feature removal.	Stas	4
Replace NAs	DC	Remove or replace NAs with numeric features	Lars	4
PCA/PCR	FE	Apply PCA/PCR for features	Bill	5
One-hot-encoding	FE	Replace categorical variables with dummy variables to allow for regression analysis	Bill	3
Apply different models	MT	Applying linear regression (lasso, ridge), XGboost, random forest, SVM algorithms	Lars	10
Model refinement	MV	Metrics evaluation. Model refinement to achieve the best RMSLE metric. Creating the list of best results.	Lars	10
Cross-validation	MV	Test dataset observation, measuring the performance of the models on the test data.	Lars, Stas, Bill	15 (5 hours each)

Total number of hours	93