Radboud Universiteit

Data Mining

Project report

Authors:
Vadim Kuzmin
Erzhena Tsyrendylykova

Student Numbers:
s4844890
s4846974

March 11, 2017

Contents

1 Abstract

In this project we use the data set from Kaggle competition "House Prices: Advanced Regression Techniques" which provides a regression problem with the goal to predict the price of each house.
We've described different regressors (linear and regularised linear models, decision tree, random forest, gradient boosting) and found a model that provides the most accurate results. We've used feature selection methods (correlation coefficient and feature significance in random forest model) and PCA to reduce the dimensionality of data.
The results suggest that the most efficient prediction can be achieved by using gradient boosting regressor that produces competitive and highly robust results.

2 Application Domain and the Research Problem

The housing market has always been a topic of national attention that represents the current economics situation of a country. Urban housing markets are becoming increasingly important in interest of buyers and owners. Buying a house is one of the most significant investment for many people. That's why our analysis becomes extremely valuable for decision making.
Pricing is one of the biggest decision to a successful home sale. If you set a price too high, you run the risk of turning off potential buyers. On the other hand, if price of your home is too low, it might move quickly and your finances will be negatively impacted.

3 Data set and data collection

The given data consists of train and test sets with 1460 and 1459 samples respectively. The target variable **SalePrice**, given only for the train data set, is used to fit our model and estimate its performance. Test data set is given for testing the model through submitted file of predicted prices.
All the houses are defined by 79 variables that describe almost every aspect of the property and assess home values. It can be suggested that representative features are referred to area, quality and location of the house. These are examples of the most important attributes after manual exploration:

- LotArea: Lot size in square feet
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- BsmtCond: General condition of the basement
- HeatingQC: Heating quality and condition
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- GrLivArea: Above grade (ground) living area square feet
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- GarageCars: Size of garage in car capacity
- PoolArea: Pool area in square feet

Different types of attributes are presented in the data set, for example, YearBuild BsmtCond are ordinal and discrete, LotArea and 1stFlrS are ratio and discrete, Neighbourhood is nominal. We have both numerical and non-numerical variables. There are also missed values that could appear for different reasons (just absence, losses or confidentiality) and we had to deal with it.
Thus, the variety of attributes and sufficient amount of samples provides an opportunity for comprehensive data analysis and feature selection.

4 Preprocessing methods

Data quality
Real data sets can contain noisy, missing and inconsistent data. Low-quality data will lead to low-quality mining results. Data mining focuses on the detection and correction of data quality problems and the use of algorithms that can deal with this data. These problems may be derived from human error, flaws in the data collection or limitations of measuring devices.
We have performed data cleaning to handle missing values. A simple and effective strategy we followed was elimination objects with missing values. Another approach

we have used (and the most common) was filling with nulls and estimation the missing values by using the remaining values. If the attribute is continuous, then the average mean value was used. If it is categorical variable, then the missing values were filled with median.
It was necessary to look at the target variable and its distribution (fig.1). It is quite close to the normal distribution with a shift to the left.
Mean = 180921.195890
Min = 34900, Max = 755000

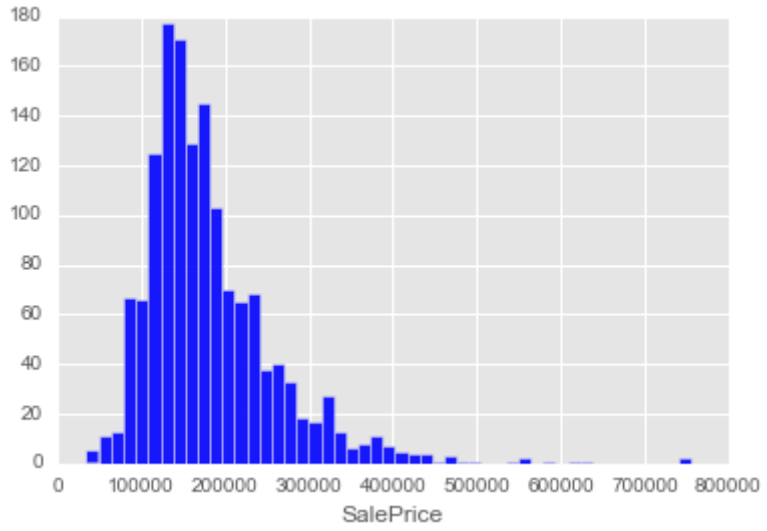

Figure 1: Distribution of target variable

For the next steps, we have concatenated train and test data set into one, because all the changes must be made for the whole data set for an appropriate work of a future model on the test data.

Feature selection
One way to reduce the dimensionality of the data set is to use only a subset of features. Redundant features can duplicate the information regarding the property of houses. Irrelevant features can have no useful information for the data mining task. Both of them can decrease SSE and quality of created regression model.
Different types of attributes require different type of analysis and visualisation method, so it was reasonable to separate all the data into two parts: with numeric and non-numeric features.

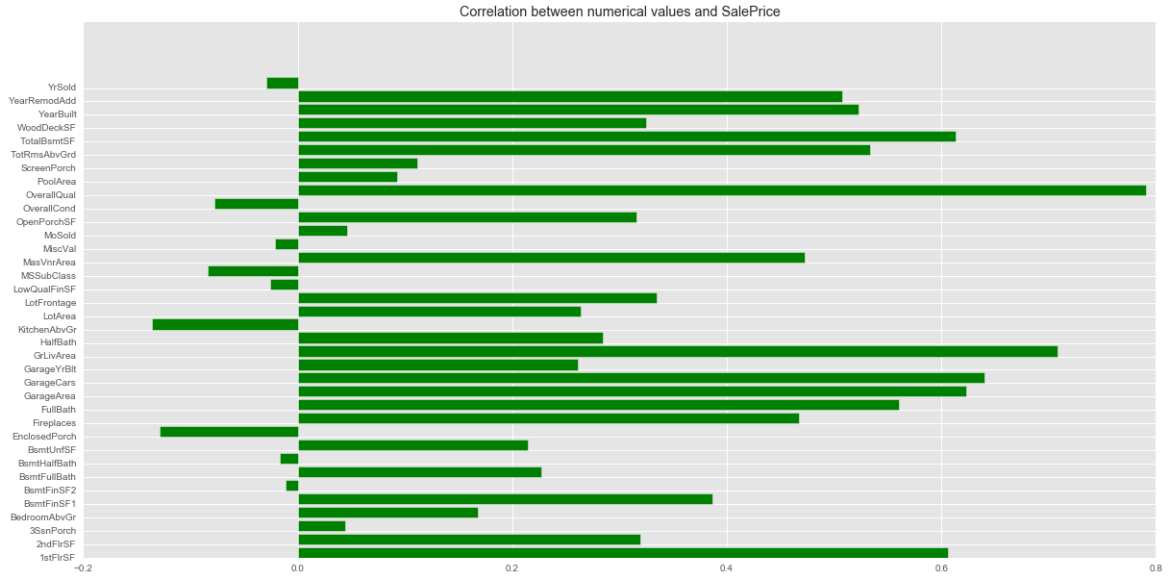

Figure 2: Correlation between numerical features and SalePrice

We calculated correlation coefficients between all the numeric attributes and the target variable. Approximately 10 features correlation coefficient are more or equal to 0.5 which tell that they contribute to the price significantly.
The most correlated coefficient = 0.79, is OveralQual. It rates the overall material and finish of the house. The relation can be seen in scatter plot. It represents how significantly the price changes with the change in quality.

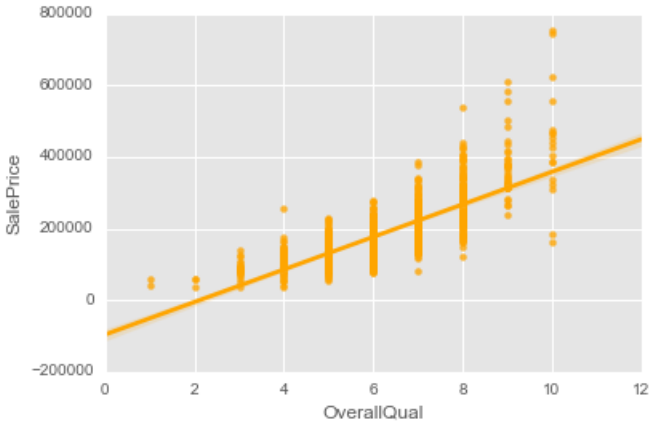

Figure 3: Repationship between OverallQual and SalePrice

Scatter plots for the other 6 most correlated numeric features help to see the relation to the target variable.

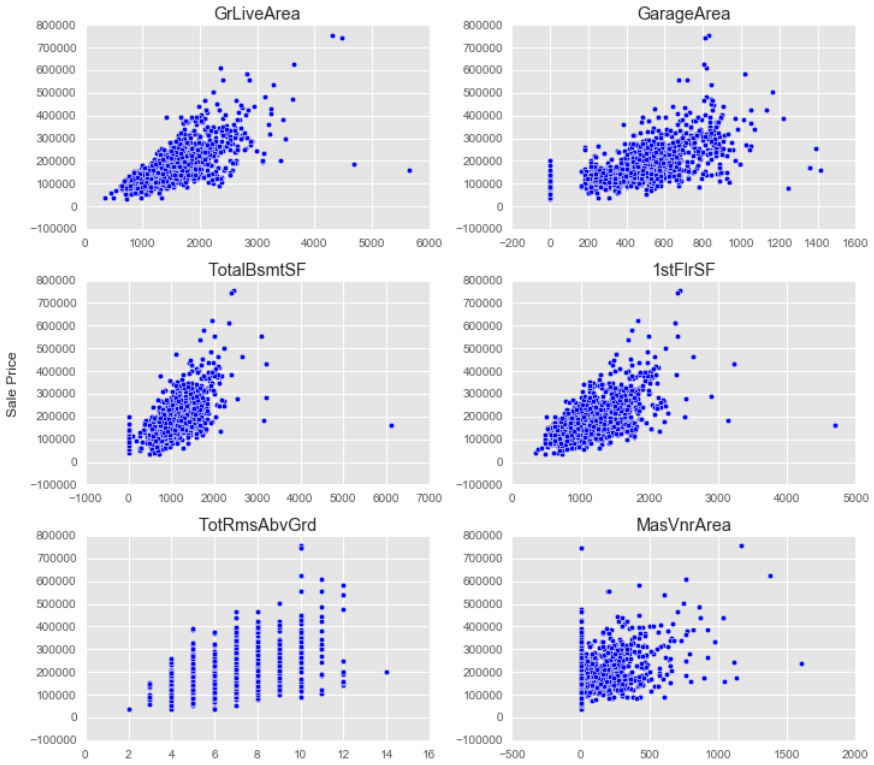

Figure 4: Relationship between the most correlated features and SalePrice

The least correlated features can be excluded without any possible damages to future models as they either will not be used by a model(in case of a tree or lasso) or elaborate an overfitting model. Features with correlation coefficient less then 0.05 were deleted.
Some features can be correlated between each other and become unnecessary to keep all of them. The heatmap was used to choose the most correlated features (GarageArea and GarageCars, TotRmsAbvGrd and GrLivArea). GarageArea and TotRmsAbvGrd were deleted as they had smaller correlation coefficient with the sale price.

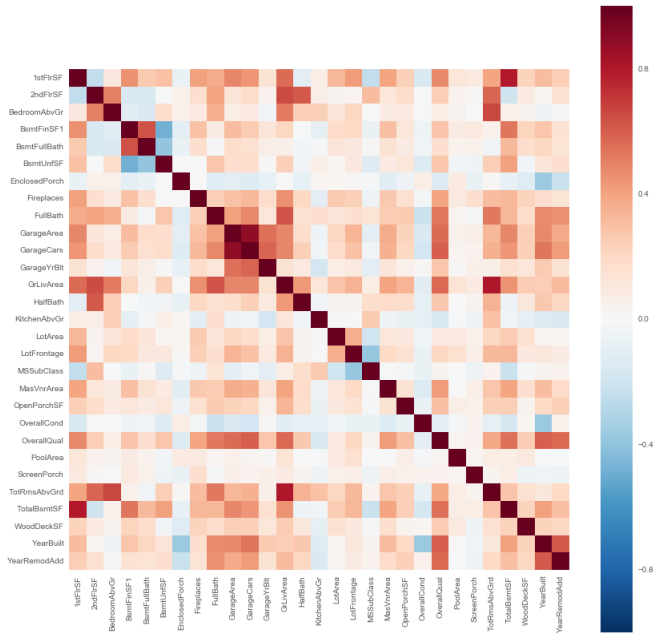

Figure 5: Correlation between features

Feature creation

It can be helpful to create new features that capture important information much more effectively, so we have created a new feature 1stFlr_2ndFlr_Sf as a sum of 1stFlrSF and 2ndFlrSF as it seemed logical and the correlation coefficient for the new attribute is 0.72 (bigger then it was before merging).

Discretization and Binarization

Data mining algorithms require data to be in a specified form. It is necessary to transform categorical variables into binary attributes that can be used by any regression model. For non-numeric features we filled missing values with 'Na' and used one-hot coding, when attribute with n different values turns into n binary attributes.

Feature importance

As one-hot coding created a lot of binary attributes the dimensionality of our data raised rapidly and it is known as the curse of dimensionality and can lead to time-consuming overfitting models. To avoid such problems, we have reduced the size of our dataset to the 10 dimensional space with PCA. Besides we have estimated the usefulness of features during the model preparation process. We determined the

variable importance while constricting Random Forest model.

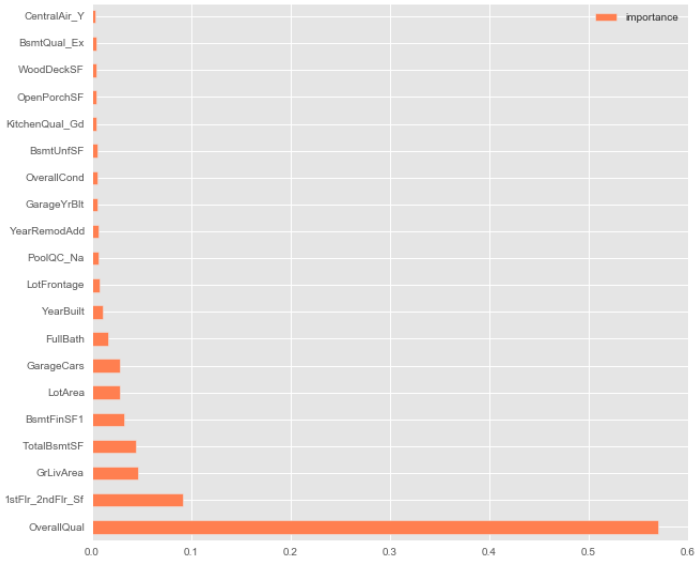

Figure 6: Features importance

As a consequence, we have created two new data sets from PCA and Random Forest with much lower dimensionality. Before fitting the models, the data was standardized to a zero mean and one standard deviation. After all these steps our data is ready to be used to fit a model.
Basically, the sum of squared errors is our way to estimate the accuracy of our regression models. Since we have an outcome variable, we can determine how well our model can predict the SalePrice.

5 Models

5.1 Linear Regression

As a first step in building the models, Linear Regression was trained and the Root Mean Squared Error(RMSE) was found to be around $9.6 * 10^{12}$. Linear Regression performed very badly, because some points in the training data have excessively large or small values compared to the rest of data. Our model has become overfitted and are not able to predict correct prices.

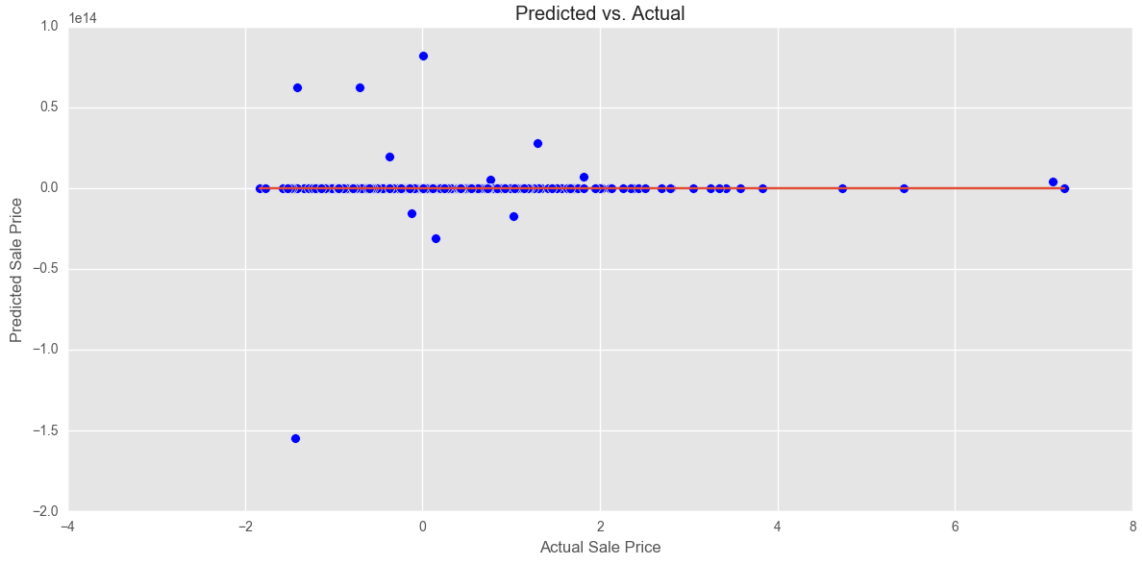

Figure 7: Relationship between predicted and actual values using linear regression

5.2 Lasso

Outliers can lead to significant regression coefficients, so we used Lasso and Ridge techniques to penalize these coefficients, and hence tried to retain the good features of both feature selection and lasso regularisation. RMSE is equal to 0.556. It is significantly better if we compare with the previous result.
Lasso regularisation use an additional parameter `alpha`, which we tuned by cross-validation `Lasso-CV`. The best result was obtained with `alpha = 0.013` and `RMSE = 0.38`

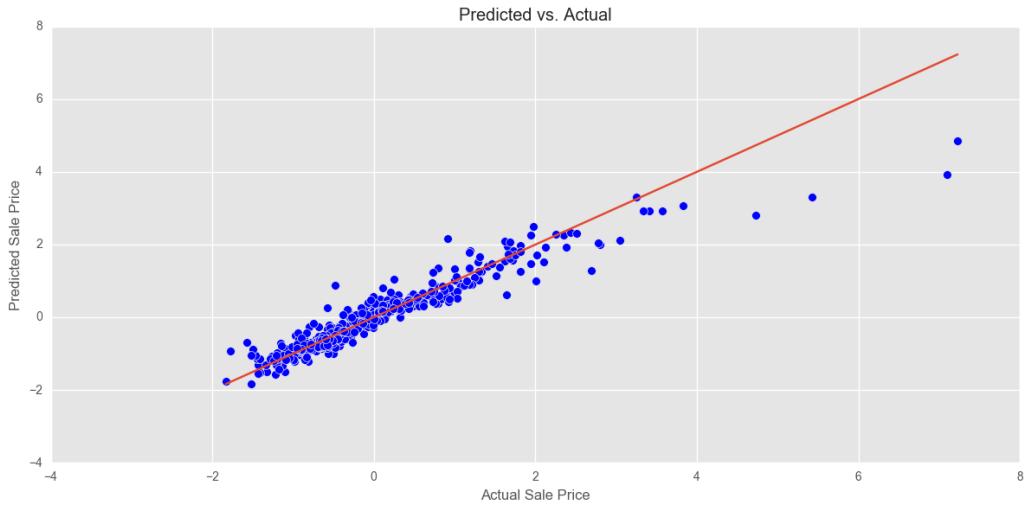

Figure 8: Relationship between actual and predicted values with Lasso regularization

5.3 Ridge

RMSE is equal to 0.421. RidgeCV has improved RMSE not significantly: 0.416 with alpha = 4.4306

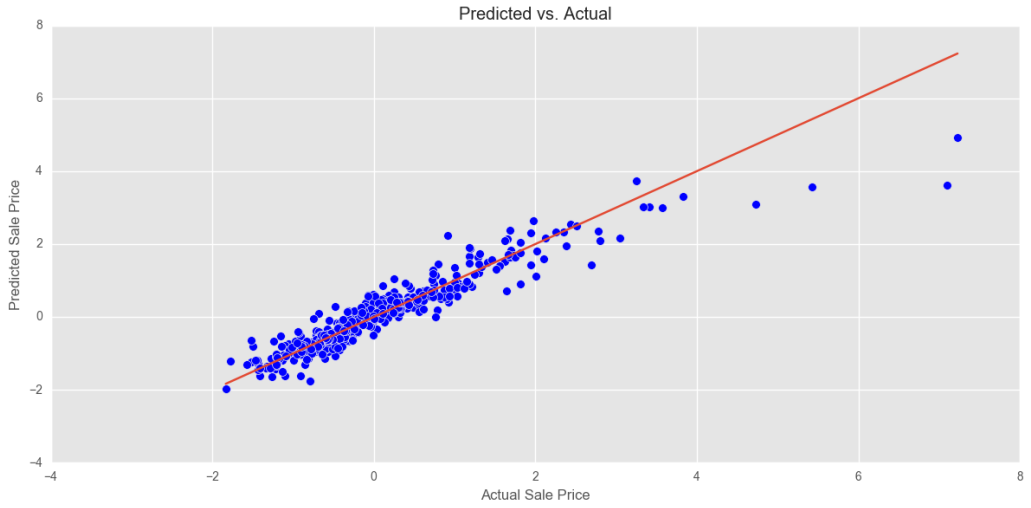

Figure 9: Relationship between actual and predicted values with Ridge regularization

5.4 Decision Tree Regressor

We have constructed a regression tree. To achieve the best performance of a predicting model, we need to tune the depth of our algorithm.
As we can see from the plot, the best choice of depth is 19. With a specified depth we constructed regression tree with RMSE = 0.5291.

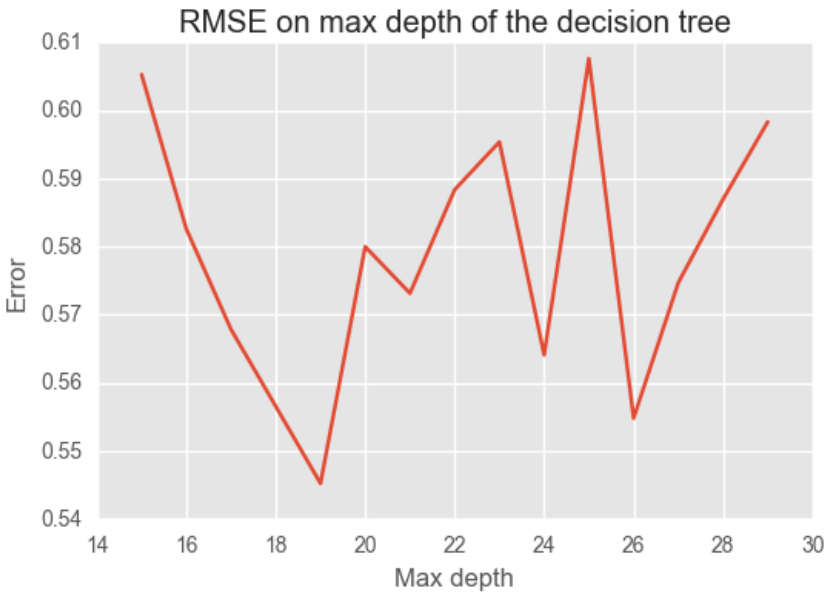

Figure 10: Error rate as a function of tree depth

5.5 Random Forest

Then we created an ensemble of decision trees, where each tree is constructed using a random subset of trained data. We obtained RMSE is equal to 0.404. It performs often well, since it is not sensitive to the usage of specific hyper-parameters. Also because of randomness, we obtained an overall better model.

5.6 Gradient Boosting

We know that boosting algorithms are widely used in machine learning competi-

tions. Boosting gives power to prediction models and improve accuracy. In gradient boosting method, models are trained sequentially. Each new model gradually minimizes the error in order to focus on cases with the highest error.
As we expected, we achieved the smallest RMSE = 0.332.

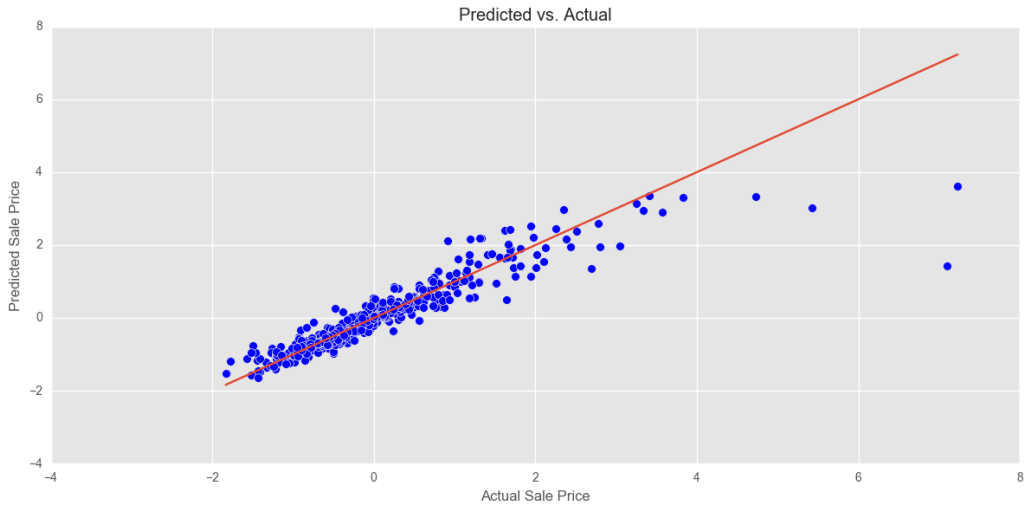

Figure 11: Relationship between actual and predicted values using Gradient Boosting

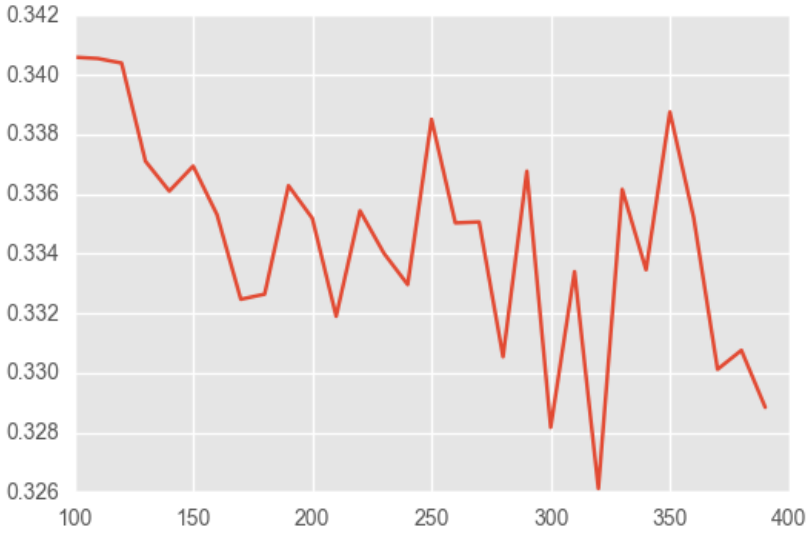

Figure 12: Error rate as a function of number of estimations

Gradient Boosting performed better than Random Forest, although there is a price for that power: we need to tune few hyper-parameters.
We adjusted a number of boosting stages to perform. As Gradient Boosting is fairly robust to overfitting, we tried to choose the best number of estimators for our model. After fitting a number of different regressions, we found Gradient Boosting regressor to be one of the cleverest algorithm, because it has the smallest error measure(RMSE) in the estimation period.

6 Conclusions

Due to the increase in the interest of buying a house, companies are investing money to the development of new technologies of selling process. Quality certification is a crucial step for both buyers and vendors. Our work is intend to help them with the prediction and building a model that can explain all the prices. The large dataset (2919 samples) with 79 attributes was considered. This case study was addressed by a regression task with a great opportunity of feature engineering.
Encouraging results were achieved with model Gradient Boosting Regressor that provides the best performance: Root Mean Squared Error = 0.3331. Which means our accuracy is almost 67%, that is acceptable.
It should be noted that our score on Kaggle is 0.13212 and you can find out our team in the first half of the ranking of competition - "House Prices: Advanced Regression Techniques".
In future work, we intend to improve our model with exploring the popular XGBoost library. We also plan to apply other DM algorithms to achieve higher accuracy and move towards the top of the Kaggle leaderboard.

References

1. PN. Tan, M.Steinbach, V.Kumar, 2006. *Introduction to Data Mining*
2. P.Cortez, A.Silva, 2008. *Using Data Mining to predict secondary school performance*
3. F. Pagnotta, M.A. Hossain, *Using Data Mining to predict secondary school alcohol consumption.*
4. Scikit-learn documentation
5. Matplotlib documentation