Advanced Regression techniques to predict House prices

Bharat Mallala[*], Karthik Anbazhagan[*] and Subramanian Sivaraman[*]

Abstract

In recent times, the Ames housing data set [1], the result of an initiative by Professor Dean de Cock from Truman State University, has become a standard playground for data scientists to implement myriads of algorithms in an effort to accurately predict the resale value of a house. This paper involves an extensive analysis of this comprehensive data set after optimizing the attributes in tune to real-world constraints and transforming them into quantifiable or qualitative assets to apply different data modeling techniques, all in an effort to understand the limits and capabilities of those methods. The analysis starts off with basic linear regression, uses it as a as our baseline. What follows is an implementation of a random forest model an efficient version of XG Boost, progressively achieving increased accuracy in prediction and error rates well below the acceptable threshold thereby wrapping up the exploration. Although a handful of variables manage to account for over 60% of the variation in the sales price, practical approaches towards processing the data that help improve the performance of the model exponentially, irrespective of the algorithms in use are explored. Future work with this data set would involve using neural networks, of which a simple execution is already underway.

Keywords

ames - housing - regression - sales

[*] *Data Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*

[*]**Authors**: bmallala@umail.iu.edu, kartanba@iu.edu, ssivaram@umail.iu.edu

Contents

Introduction

For a while now, real estate in the United States been a vortex of confusion and unpredictability. Increasing demand for homes due to increasing incomes (even in the suburbs) has often led to marked up prices. It is only natural to try and understand the factors that affect the sales price of a home the neighbourhood. That fundamental aim of this paper is to provide an organized and understanding of what factors surrounding a home or what features within one contribute to its value. In regard to the same, the computational efficacy of various modeling techniques with respect to their prediction prowess are explored. The models are first trained using the train data available and then checked for performance over a part of the training data which we isolate before running the model. This data is used solely for the purpose of validation. At the final stage, the model is retrained including this validation data and is used for prediction over the test set. An exhaustive process of data cleaning helps realize that it is a key factor in improving the overall outcome of a model. Missing values are accounted for, outliers are weeded out and variable transformations are performed wherever necessary to improve their usability and significance. It is evident from those methods and the results that domain knowledge is only useful when one knows how to transform data into meaningful information for further processing. For example, some basic analysis reveal that the Sales Price is linearly associated with variables like the living room area and number of storeys in the house. However, there seems to be no real association between the price and the number of bedrooms. More insights are discussed in the findings. In the modeling paradigm, a simple linear regression model is introduced with the processed data that returns satisfactory baseline values to build up from. A lean version of random forest is then constructed, one that ranks attributes based on their significance to the model with respect to the target (Sales Price). Based on the learning from previous model iterations, a few more attributes are imputed and transformed to come up with a final list of significant factors. Finally, Extreme gradient boosting (XG Boost) is applied to arrive at the final results which are discussed at length. Also explored are ways of implementing a robust model using Artificial neural networks for future development.

1. Background

The vast feature set in this data just shows us how the value of a home is intricately dependent on factors that do not usually come up when buying homes. However obscure the dependencies may be, there are certain statistical methods than can be used to scrutinize their value addition to the home. Many techniques have already been identified for the data, thanks to the Boston housing data set [2] that was prevalently used before. Our initial study for this project primarily involved understanding the various attributes, their structure and their impact on value. In the paper, we successfully identify the different types of variables and come up with ways to transform them into useful predictors. As part of initial research, we also looked into various form of regression techniques and identified weaknesses in them compared to models like Random forest. Later into the project is when we identified gradient boosting as a valid technique, one that seemed to equal the performance of a random forest. The following sections provide an overview of the different techniques and concepts that appear in this paper.

1.1 History

The Boston housing data set [3] introduced in 1978 was considered a standard for all kinds of model testing since it involved parameters that did not require specific domain knowledge; all it had were features of a home and its neighbourhood that any layman would understand. This led to its widespread usage in academia to practise modeling on. But the data does not hold up to the world we now live in. That is where Professor Dean De Cock, who has had experience with the Boston data set, decided to develop another one that befit the modern world. In collaboration with the Ames City Assessor's Office, was able to obtain sales of houses in the Ames County between the years 2006 and 2010. Dean worked extensively on the data, thereby creating a set that held so many opportunities for the budding data scientist. The sheer number of features available that defined a home and justified its sale price, was also easily accessible and understandable from a practical standpoint. It is this data set that we source in this paper to analyze and predict the prices of homes, given the key parameters that hold monetary value.

1.2 Concepts and Understanding

Data Cleaning This describes the steps taken to transform raw logged or unmaintained data sets into information that are tangible and quantitative or qualitative in nature with regard to the premise. This process usually involves checking for errors and missing values. Also part of the procedure is identifying unusual data that do not make sense in reality and either adjusting or removing them in order to hold the integrity of the data. Once corrections are made, we perform data validation to verify them and make sure their significance is unchanged.

Simple Linear Regression Regression is a statistical procedure that involves predicting the values of a certain measure based on the independent observations of another variable's past values. Linear regression works when these two variables are linear with respect to each other and follow an approximate normal distribution. Regression also assumes homoscedasticity wherein the variance of both the variables are considered equal. Highly correlated bi variate data work best with linear regression as they have almost equal variance and as their spread is linear with respect to each other. This method is one of the most straightforward procedures for any prediction. The regression for a bi variate set of data can be visualized in terms of a line across the scatter plot of the variables, one that provides the closest approximation to an expected value.

Decision Trees: Random forest Random forest [4] is another classification algorithm that basically involves multiple iterations of decision trees. It randomly selects a small group of features among the training set and builds a decision tree. This process iterated for a user specified number of times creating just as many trees. The features are selected at random and usually has constraints on the variance. The result of a random forest is that each feature is ranked based on its overall significance in all the decision trees that were generated. This helps in identifying the most important factors that affect a target variable.

Extreme Gradient Boosting Extreme gradient boosting [5] combines weak models together and estimates an output that is the best version of all the previous models. It is basically another classification technique that that proceeds linearly and stepwise. It is an optimization algorithm that focuses on a given cost function returning the best prediction for that function with all the data provided [6].

2. Algorithms and Methodology

2.1 Software and System Specifications

All computations and analysis were performed in a 64-bit version of Windows 10 powered by an Intel(R) Core(TM) i7-6560U CPU @ 2.20Ghz with 16 GB of RAM. The primary scripting language we used was R version 3.3.1. The different packages made use of in this paper have been listed in the References section.

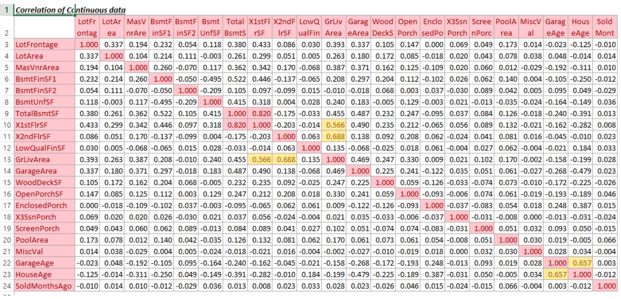
2.2 Data

The Ames housing dataset provided on Kaggle [2]lists 1460 houses as the training set with nearly 80 variables covering all aspects of a house like the living area,the garage area, the house's exterior finish et cetera. The test data set has 1459 rows of houses with just as many variables except for the Sales price attribute. Overall, the features provided for each house describe the property quantitatively and qualitatively.

Description Among the 1460 rows of training data, there are 46 categorical variables which can be broken down to 23 nominal and 23 ordinal variables. These variables provide categorical information like the year the house was built, the year when any kind of remodeling or renovation was done, the street of the house and its general neighbourhood within the Ames county. Apart from these, there are 14 discrete variables that provide factual information about the property, like the number of kitchens, bedrooms and bathrooms within the property. The 20 continuous variables provide estimates of square footage of the living room and garage or the total lot size.

Cleaning The different types of variables - Nominal, Ordinal, Continuous and Discrete are grouped together as part of the data cleaning process. Missing values are accounted for by either using the mean, median or the mode or by transforming them into dummy variables that convey more meaning and are usable by the models. Since various categorical variables have more than 10 values, they are approximated to less than three by imputing values for all. This allows us to have minimum number of beta coefficients. This is important as all the classification models do not work well if the categorical variables have many values or classifiers. Some values in the data set marked NA are actually not missing values but have meaning in the form of a string, ”NA”. These values are made differentiable to R by changing them to meaningful terms. For example, under the Alley attribute, NA stands for No access. Correspondingly, the value for that particular classification was changed to ”No”. As part of the cleaning and pre processing, multivariate analysis was performed to check for variables that are closely related to each other. We also calculated the correlation coefficients for several numerical variables in order to remove those that are similar in their distribution across the data set thereby influencing the model in the same way. There were some variables that has values that spread over 90% of the data. These values are known to have very little impact on the target variable as there is minimal variation in distribution. So these variables were also removed. The data cleaning process transformed existing categorical variables into either Yes/No attributes or with ranges of numbers like 1 to 5. Values for attributes related to quality, for example, as assigned a range between 1 and 5 where 1 is poor and 5 is excellent. Ultimately, the processed data contained nearly 56 variables down from the initial 80.

Partitioning The training data set itself is partitioned randomly using R in order to generate a 70-30 split. This split provides a validation set that is used to check the performance of the model. This split serves the purpose of evaluating the model with known data, as a form of supervised learning.



1 *Correlation of Continuous data*

		LotFrontag	LotArea	MasVnrAre	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtS	X1stFlrSF	X2ndFlrSF	LowQualFin	GrLivArea	GarageArea	WoodDeckS	OpenPorch	EnclosedPo	X3SsnPorch	ScreenPorc	PoolArea	MiscVal	GarageAge	HouseAge	SoldMont
3	LotFrontage	1.000	0.337	0.194	0.232	0.054	0.118	0.380	0.433	0.086	0.030	0.393	0.337	0.105	0.147	0.000	0.069	0.049	0.173	0.014	-0.023	-0.125	-0.010
4	LotArea	0.337	1.000	0.104	0.214	0.111	-0.003	0.261	0.299	0.051	0.005	0.263	0.180	0.172	0.085	-0.018	0.020	0.043	0.078	0.038	0.048	-0.014	0.014
5	MasVnrArea	0.194	0.104	1.000	0.260	-0.070	0.117	0.362	0.342	0.170	-0.068	0.387	0.371	0.162	0.125	-0.109	0.020	0.060	0.012	-0.029	-0.192	-0.311	0.010
6	BsmtFinSF1	0.232	0.214	0.260	1.000	-0.050	-0.495	0.522	0.446	-0.137	-0.065	0.208	0.297	0.204	0.112	-0.102	0.026	0.062	0.140	0.004	-0.105	-0.250	-0.012
7	BsmtFinSF2	0.054	0.111	-0.070	-0.050	1.000	-0.209	0.105	0.097	-0.099	0.015	-0.010	-0.018	0.068	0.003	0.037	-0.030	0.089	0.042	0.005	0.095	0.049	-0.029
8	BsmtUnfSF	0.118	-0.003	0.117	-0.495	-0.209	1.000	0.415	0.318	0.004	0.028	0.240	0.183	-0.005	0.129	-0.003	0.021	-0.013	-0.035	-0.024	-0.164	-0.149	0.036
9	TotalBsmtSF	0.380	0.261	0.362	0.522	0.105	0.415	1.000	0.820	-0.175	-0.033	0.455	0.487	0.232	0.247	-0.095	0.037	0.084	0.126	-0.018	-0.240	-0.391	0.013
10	X1stFlrSF	0.433	0.299	0.342	0.446	0.097	0.318	0.820	1.000	-0.203	-0.014	0.566	0.490	0.235	0.212	-0.065	0.056	0.089	0.132	-0.021	-0.162	-0.282	0.008
11	X2ndFlrSF	0.086	0.051	0.170	-0.137	-0.099	0.004	-0.175	-0.203	1.000	0.063	0.688	0.138	0.092	0.208	0.062	-0.024	0.041	0.081	0.016	-0.045	-0.010	0.023
12	LowQualFinSF	0.030	0.005	-0.068	-0.065	0.015	0.028	-0.033	-0.014	0.063	1.000	0.135	-0.068	-0.025	0.018	0.061	-0.004	0.027	0.062	-0.004	-0.021	0.184	0.033
13	GrLivArea	0.393	0.263	0.387	0.208	-0.010	0.240	0.455	0.566	0.688	0.135	1.000	0.469	0.247	0.330	0.009	0.021	0.102	0.170	-0.002	-0.158	-0.199	0.028
14	GarageArea	0.337	0.180	0.371	0.297	-0.018	0.183	0.487	0.490	0.138	-0.068	0.469	1.000	0.225	0.241	-0.122	0.035	0.051	0.061	-0.027	-0.268	-0.479	0.023
15	WoodDeckSF	0.105	0.172	0.162	0.204	0.068	-0.005	0.232	0.235	0.092	-0.025	0.247	0.225	1.000	0.059	-0.126	-0.033	-0.074	0.073	-0.010	-0.172	-0.225	-0.026
16	OpenPorchSF	0.147	0.085	0.125	0.112	0.003	0.129	0.247	0.212	0.208	0.018	0.330	0.241	0.059	1.000	-0.093	-0.006	0.074	0.061	-0.019	-0.193	-0.189	0.046
17	EnclosedPorch	0.000	-0.018	-0.109	-0.102	0.037	-0.003	-0.095	-0.065	0.062	0.061	0.009	-0.122	-0.126	-0.093	1.000	-0.037	-0.083	0.054	0.018	0.248	0.387	0.015
18	X3SsnPorch	0.069	0.020	0.020	0.026	-0.030	0.021	0.037	0.056	-0.024	-0.004	0.021	0.035	-0.033	-0.006	-0.037	1.000	-0.031	-0.008	0.000	-0.013	-0.031	-0.024
19	ScreenPorch	0.049	0.043	0.060	0.062	0.089	-0.013	0.084	0.089	0.041	0.027	0.102	0.051	-0.074	0.074	-0.083	-0.031	1.000	0.051	0.032	0.093	0.050	-0.015
20	PoolArea	0.173	0.078	0.012	0.140	0.042	-0.035	0.126	0.132	0.081	0.062	0.170	0.061	0.073	0.061	0.054	-0.008	0.051	1.000	0.030	0.019	-0.005	0.066
21	MiscVal	0.014	0.038	-0.029	0.004	0.005	-0.024	-0.018	-0.021	0.016	-0.004	-0.002	-0.027	-0.010	-0.019	0.018	0.000	0.032	0.030	1.000	0.028	0.034	-0.004
22	GarageAge	-0.023	0.048	-0.192	-0.105	0.095	-0.164	-0.240	-0.162	-0.045	-0.021	-0.158	-0.268	-0.172	-0.193	0.248	-0.013	0.093	0.019	0.028	1.000	0.657	0.003
23	HouseAge	-0.125	-0.014	-0.311	-0.250	0.049	-0.149	-0.391	-0.282	-0.010	0.184	-0.199	-0.479	-0.225	-0.189	0.387	-0.031	0.050	-0.005	0.034	0.657	1.000	-0.012
24	SoldMonthsAgo	-0.010	0.014	0.010	-0.012	-0.029	0.036	0.013	0.008	0.023	0.033	0.028	0.023	-0.026	0.046	0.015	-0.024	-0.015	0.066	-0.004	0.003	-0.012	1.000

Figure 1. Random Forest

2.3 Data Modeling

Once the data has been processed and the significant contributors have been selected, the various discussed models are executed. Linear regression sets an initial baseline while random forest helps strong prediction using multiple decision trees. Finally, XGboosting is performed using selected attributes for an even more compelling prediction.

2.3.1 Algorithms

After the data is cleaned and processed, a simple linear regression to predict values for the target variable (Sales Price) is performed. The reason to choose a linear model is because the target here is a numerical variable. Some variables are lost due to multi-col-linearity but the impact is not problematic. The primary measure of performance for linear regression is the value of r^2 which is indicative of how much impact the model has had on the outcome or how well the model describes the data and predicts the target. The features supplied for the linear model are factors like the Lot area, Garage space, the price level in the neighbourhood, basement quality, the condition of the pool and the quality of the exterior. The model gives a decent r^2 value of about 80% which is a good result for this method even if some over fitting may happen. Proceeding to Random forest, this algorithm worked by building simple decision trees using randomly sampled attributes and finding out the most important factors for the model. This step when iterated several hundred times, yields a robust combination of decision trees that consider those values that contribute to the model in the majority of the randomly sampled trees. Using R, a random forest for all the transformed variables is performed.

3. Experiments and Results

Once the models were trained with the training set, they were applied on the validation data set to check for their prediction accuracy. XGboosting provided the most accurate results and was chosen as the model for the test data set. Initially, linear regression worked with an RMSE of 80% which was a good start. Random forest gave a list of significant predictors that influenced the Sales prices the most. XGBoosting performed the best as the predictive trees that it built stage-wise performed better overall, compared to random forest, thereby resulting in an RMSE of over 87%. Random forest is very effective for feature selection while XGboosting as a model

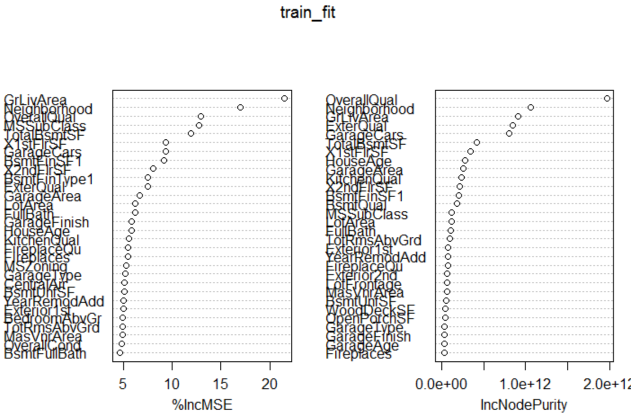


Figure 2. Random Forest

performs better at prediction as the decision trees are built stage wise and not in random like random forest.

4. Summary and Conclusions

It is clear from the results that Extreme gradient boosting offers significant leaps in performance when compared to random forest. But it is worthy to note that there are more complicated and improved versions of random forest that may provide results that are as good as xgboosting's. The RMSE value of 0.87 for gradient boosting earned a rank of 900 in the Kaggle competition. It is clear from the method that the Sales Price estimates provided after boosting are accurate but can be improved further. This paper has helped understand the difference between the various prediction algorithms

5. Future Work

Artificial neural networks (ANN) have been known for their efficiency in prediction under various situations. The next step would be to implement ANN and compare its performance to gradient boosting. A simple version of neural networks using features that were selected from the earlier models has already been implemented. It resulted in an r^2 value of 82% which can be improved upon if the feature selection is performed exclusively for the model. Interpreting the performance of a model using the RMSE (Root Mean Square Model) may lead to misleading results as this only conveys the model's effects on the results that have been produced, rather than the accuracy provided. So exploring another method of evaluating the models that are executed is also part of the road map for future development.

References

[1] https://www.kaggle.com/c/house-prices-advanced-regression-techniques/

[2] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3), 2011.

[3] Boston housing dataset: Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics Management, vol.5, 81-102, 1978.

[4] Classification and Regression by randomForest by Andy Liaw and Matthew Wiener. vol 2, p18-22, 2002.

[5] xgboost: Extreme Gradient Boosting by Tianqi Chen and Tong He and Michael Benesty, R package version 0.4-4, 2016

[6] https://en.wikipedia.org/wiki/Gradient_boosting#Algorithm

References