A Predictive & Prescriptive Analytics Solution to Home Improvement

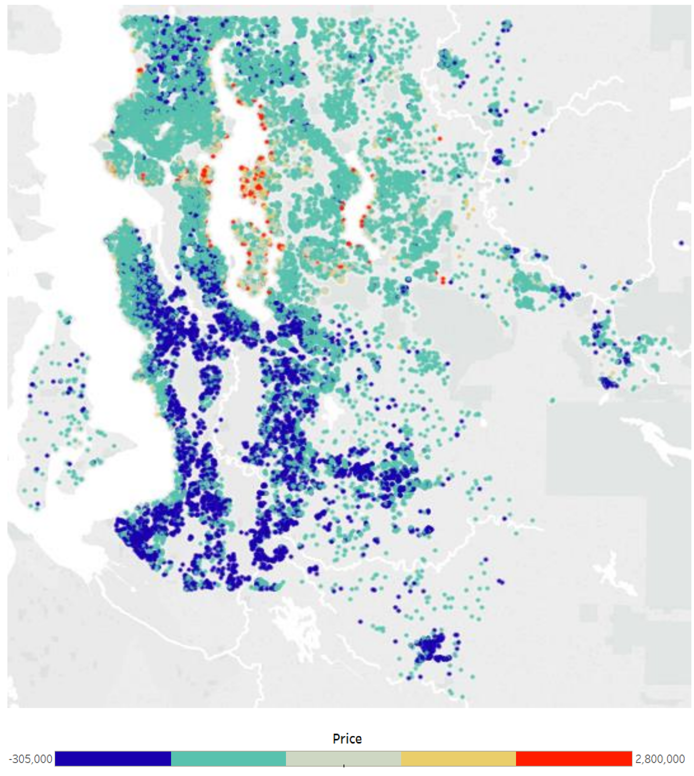
Abhishek Gupta, Komal Suresh, Akshay Ahuja, Matthew A. Lanham
Purdue University Krannert School of Management
gupta362@purdue.edu; suresh19@purdue.edu; ahuja11@purdue.edu; lanhamm@purdue.edu

Abstract

This study builds and assesses various predictive models to understand the effects that certain housing features have on the price of a home in many different markets. Using these predictive models, we formulate an optimization model that allows home owners (or better yet future home sellers) which features about their home they should invest in so as to maximize the value of their home. This study is important because purchasing a home is one the greatest investments an individual or family will make in their lifetimes. Many people make investments into their home over time, but many do not have a good idea of what their return on investment (ROI) is for their market. We provide a solution to this problem and discuss how this approach to interfacing predictive and prescriptive models can be effective for many other types of problems.

Introduction

The aim of this poster is to provide data-driven recommendations for home improvement projects, keeping in mind the return on investment. Our target decision-maker is that of the homeowner. However, this analytics solution could be very useful in real estate as realtors often provide advice to their clients in how to fix up their home to optimize the home's selling price and time on the market. We try to provide strategic guidance based on analytics to support this problem.



Given an interval response, we used regression techniques to understand the effect of different characteristics of houses on its selling price such as the size, number of bedrooms and bathrooms, number of floors, city, condition. We then investigate the costs of different home improvement projects such as adding more square feet, remodelling the kitchen, etc. to identify possible ways to increase the selling price. The prescriptive model then considers these costs and potential benefits to recommend the optimal course(s) of action for the home-owner to take.

We took a sample of house sales data for 2015 for houses in King County, WA. This served as a fairly diverse sample and contained data for house prices in different markets, from a tiny 370 sq. ft. apartment in downtown Seattle to a spacious 7,350 sq. ft. home in a more rural setting. The figure above shows the variation in the price of these homes.

Methodology

Data Sources

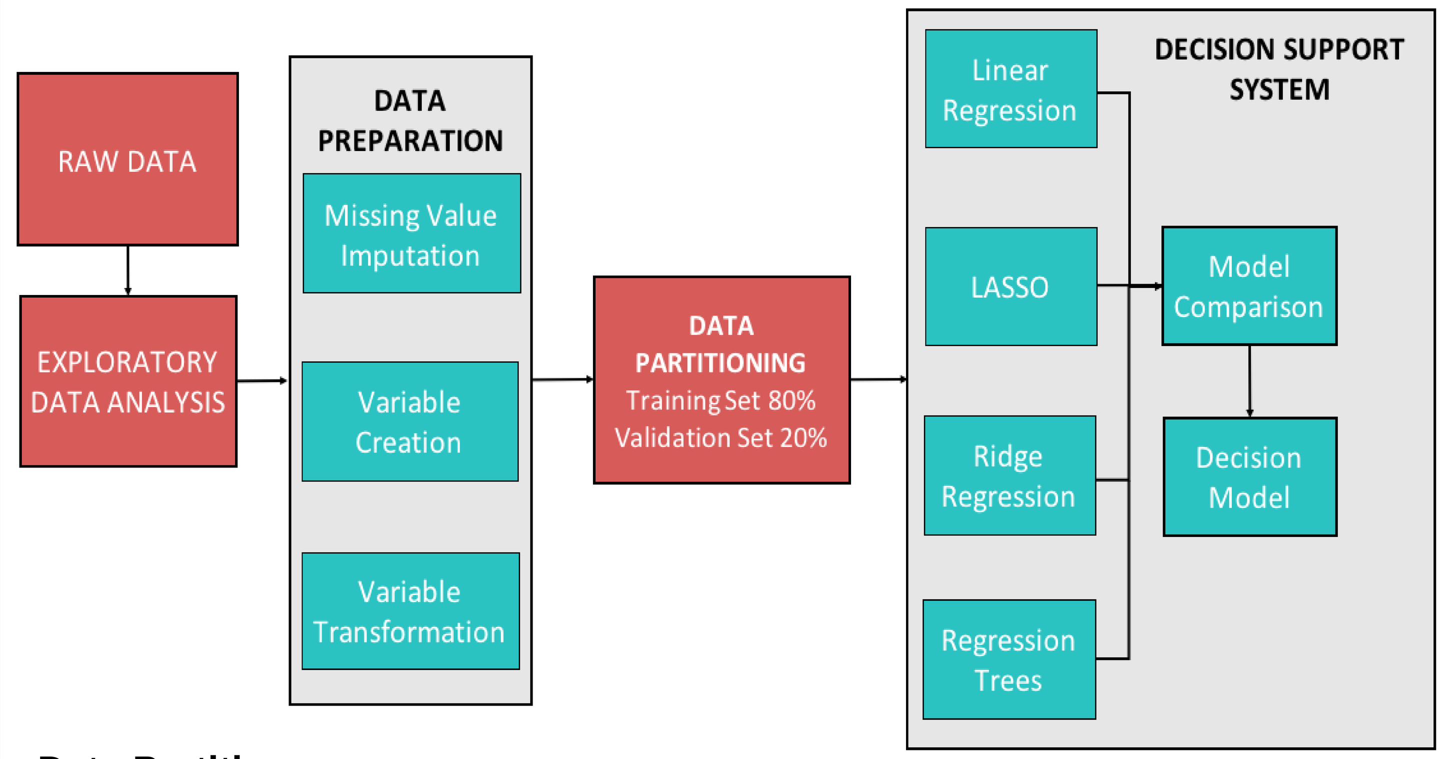
Housing data used for the development for this model was downloaded from a Kaggle data set as a csv file containing 21,613 rows and 21 columns. The economic data was then gathered from a public federal government zip code data base as a csv file with 81,831 rows and 20 columns. Finally, cities and their respective zip codes were gathered from the capital impact government gateway site. These various data sources were then merged together for analysis.

Exploratory Data Analysis

We started our analysis by looking at some of the descriptive statistics of the variables in the dataset to gain some understanding of the dataset. We then decided to look at the scatter and box plots between independent and dependent variables to see if any trends were apparent. This helped us identify the variables that were able to explain the difference in house prices. Using a correlation matrix, we were able to determine that bathrooms, grade, sqft_above, and sqft_living15 had a correlation of over 0.70 with another predictor variable and therefore could be removed from the final model.

Data Preparation

The next step was to see distribution of the price variable to decide if any transformation was necessary. After looking at the histogram of the price variable we observed that it was right skewed. With this information we decided to apply a log transformation to the variable in order to make the distribution more Gaussian.



Data Partition

The data was divided into two different groups, the training and validation sets which comprised of 80% and 20% of the total number of observations respectively. The training set was used to build the models. The validation set was used to assess, fine-tune, and compare the models.

Model Building and Comparison/Selection

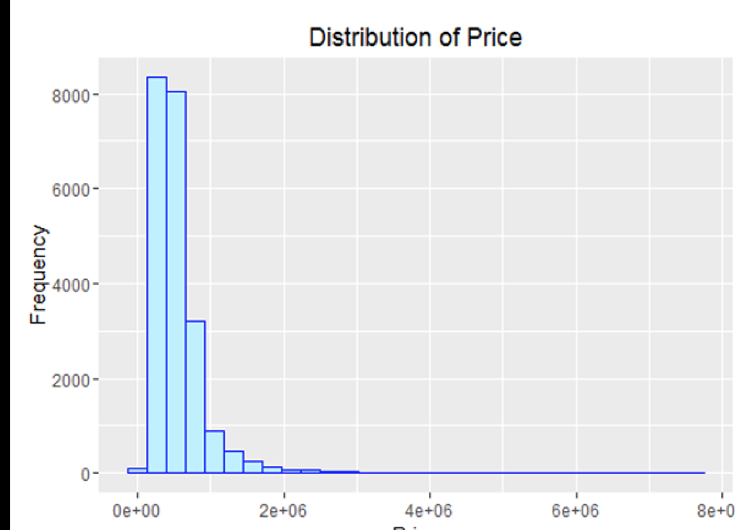
To estimate the house prices, predictive models were built on the training set using machine learning techniques, namely, linear and ridge regression, LASSO and regression trees (CART). The models were compared and selected for prediction on the basis of Mean Square Error (MSE).

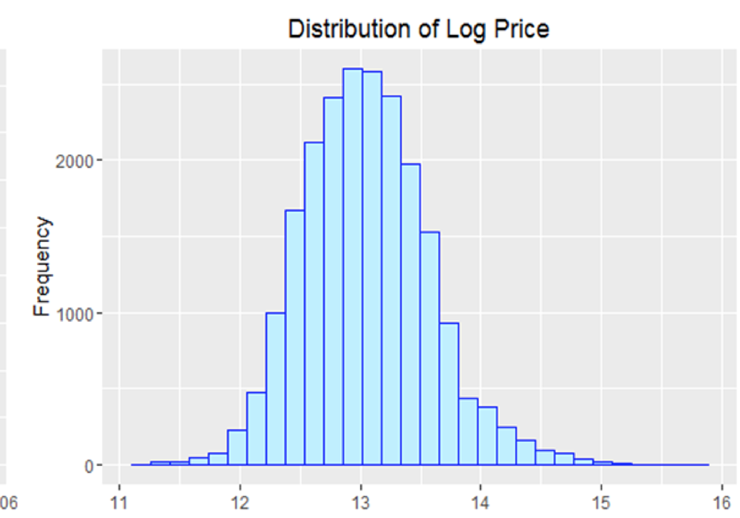
Decision Model

Since the coefficients of linear regression are more interpretable than those obtained from the rest of the techniques, results from linear regression were used to build the decision model using optimization techniques. The decision model is used to make recommendations of house improvements given constraints such as budget for the fixes and increase in sale price of the house after remodeling or additions. Features not available in our predictive models were researched for their expected costs and ROI in home value (e.g. bathroom remodel).

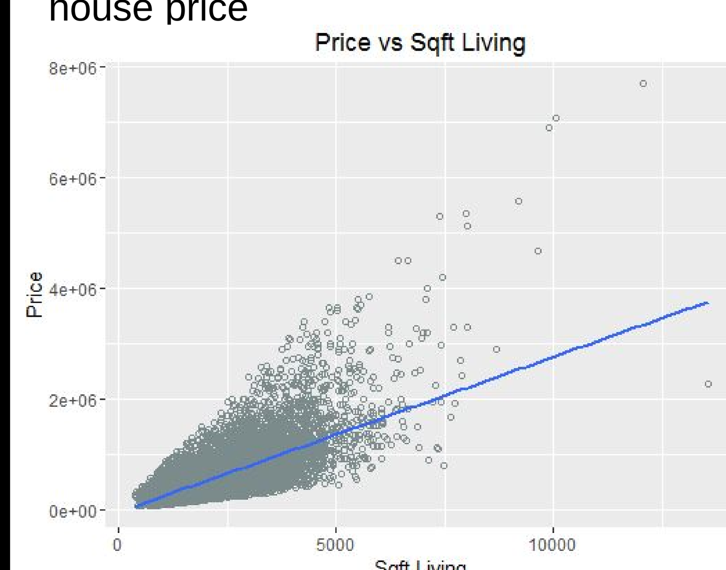
Results

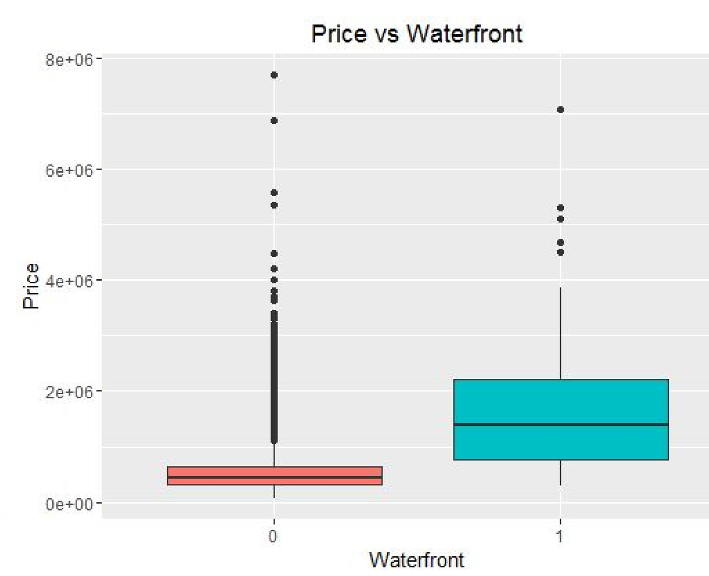
The response was log transformed to meet the normality assumption of regression

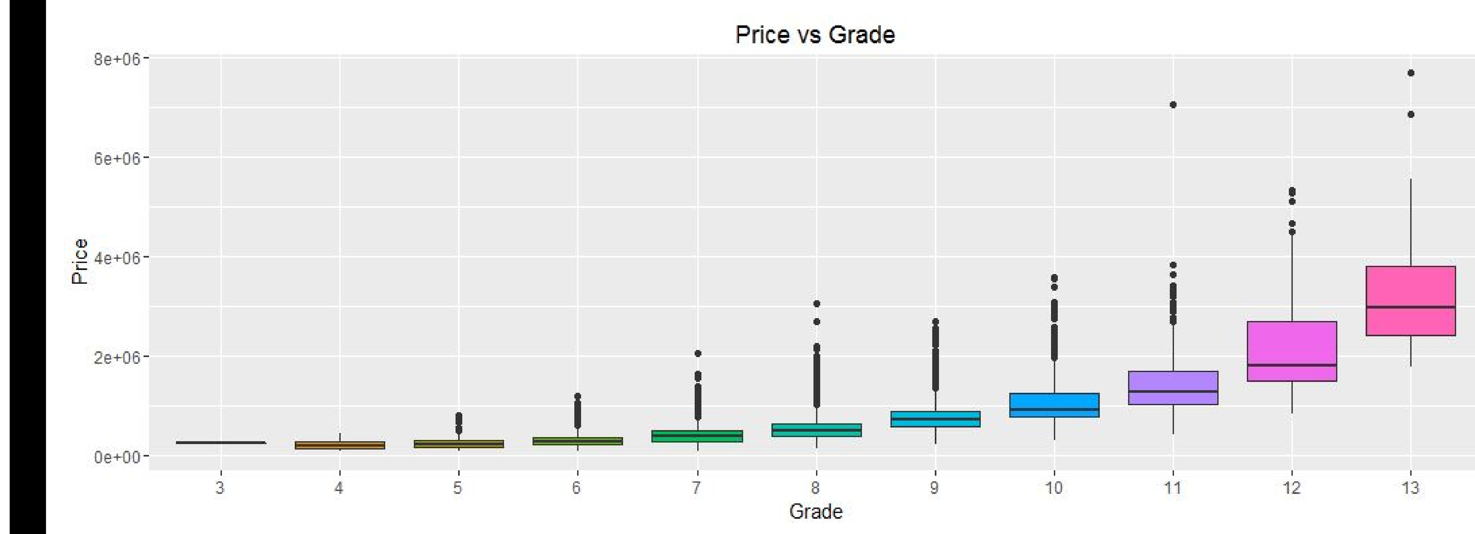




Graphical representation of relationship between different characteristics with house price







Model Comparison

Statistic\Model	Linear Regression	Regression Tree	Ridge Regression	LASSO
MSE	0.054	0.075	0.046	0.046
RMSE	0.231	0.273	0.214	0.214
R-squared	0.813	0.738	0.840	0.839

We compared non-linear and linear models and among linear models, LASSO and ridge regressions performed best on basis of Mean Square Error.

Decision Model

Decision variables

x_j = # of bedrooms to add in environment $j;\ i = 1,..,N$
s_j = # of sq. ft. to add in environment $j;\ i = 1,..,N$
m_j = # of additional finished basement sq. ft. to add in environment $j;\ i = 1,..,N$
k_j = upgrade kitchen or not in environment $j;\ i = 1,..,N;\ k_j \in \{0,1\}$
b_j = upgrade bathroom or not in environment $j;\ i = 1,..,N;\ b_j \in \{0,1\}$

Parameters

r_{ij} = expected return parameter for decision variable i in environment $j;\ j = 1,..,N$
c_{ij} = expected cost parameter for decision variable i in environment $j;\ j = 1,..,N$
$\pi_{ij} = (r_{ij} - c_{ij})$ = expected value add for decision variable i in environment $j;\ j = 1,..,N$

To define end users environment *j*

A = end users investment budget ($)
X = end users current number of bedrooms
$X_{[l,u]}$ = end users lower and upper bounds
S = end users current home sq. ft.
$S_{[u]}$ = end users upper bound
M = end users current finished basement sq. ft.
$M_{[u]}$ = end users upper bound
K = end users feedback if kitchen is already upgraded or not
B = end users feedback if bathrooms are already upgraded or not
L = end users current home location

$$K = \begin{cases} 1\ if\ already\ updated \\ 0\ if\ not\ already\ updated \end{cases} \quad B = \begin{cases} 1\ if\ already\ updated \\ 0\ if\ not\ already\ updated \end{cases}$$

Model:

$max[\pi_{1j}x_j + \pi_{2j}s_j + \pi_{3j}v_j + \pi_{4j}m_j + \pi_{5j}k_j + \pi_{6j}b_j]$ (maximize house ROI)

subject to:
$x_j \leq X_u$ (# of bedrooms that can be added)
$s_j \leq S_u$ (# of sq. ft. that can be added)
$m_j \leq M_u$ (# of finished basement sq. ft. that can be added)
$k_j + K \leq 1$ (update the kitchen or not)
$b_j + B \leq 1$ (update the bathroom or not)
$c_{1j}x_j + c_{2j}s_j + c_{4j}m_j + c_{5j}k_j + c_{6j}b_j \leq A$ (improvements within budget)
$x_j, s_j, c_j, m_j \geq 0$
$k_j \in \{0,1\},\ b_j \in \{0,1\}$

Conclusions

Maximizing the value of one's greatest investment is an important decision. We provide an analytics-based decision-support-system (DSS) that provides a solution to this problem. We posit that many business problems can be approached and supported in a similar fashion by merging multiple data sources, understanding cause-and-effect relationships (i.e. descriptive analytics), building predictive models (i.e. predictive analytics), and streamlining those predictive models into decision/optimization model (i.e. prescriptive analytics). Lastly, having a visual interface that allows a decision maker to explore these insights is something we will showing when presenting this study using built R Shiny app on a tablet.

Acknowledgements

Business Information and Analytics Center(BAIC) partially funded this project