Data Science Final Project Report

Linqi Nie, Mei Zhang, Joanna Sun, Stella Zhou, Dominique Vidjanagni

Business Understanding

In May 2017, one of the biggest real estate information companies Zillow surveyed a panel of more than 100 real estate economists and analysts for a house price prediction, which interests our group to analyze the real estate information and predict the house price. Before doing that, we define our business question as "How can real estate agents use data analytics to predict housing prices and maximize their revenue?"

To answer this question, we began with several simple assumptions:

(1) All customers are rational, which means they will choose lower price given same house condition.

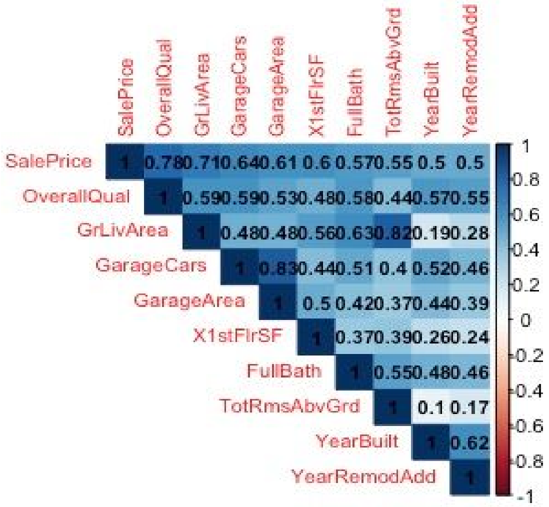
(2) The market is stable, because it is hard to include the effects of any potential economic crisis into our model.

We assume for new houses that are on sale, we can still collect all required features. For an agent, given he gets 5% commission on every successful transaction, his expected revenue function would be E(Revenue) = Sales Price * 5% * Probability (Sales Price < Market Price). To maximize the expected revenue, given 5% constant, we are essentially trying to maximize the product of sales price and the probability of sales price lower than actual market price. However, it's clear that we can't increase one of these two factors . So, an accurate understanding of Market Price becomes crucial to help agents find the optimal price, and that's why we began our study on the market pricing prediction model. With our final model, we can provide pricing suggestions to local real estate agents at Ames, Iowa and help them to maximize revenue.

Data Understanding

The dataset from Zillow Co. contains almost influential factors of the real estate price in Ames, Iowa. It contains 81 variables, which generally come from three dimensions -- the features of house itself, the location and neighbourhood, and the selling information. From condition of house perspective, we have variables regarding to building year, roof , rooms, pools and garages. For location, we know about the neighbourhood and its road access. Regards to selling information, MoSold and YrSold allows us to think about the market condition when the house was sold. In order to better address our main business problem, we set the sale price of house (SalePrice in dataset) as the target variable. With such a detailed dataset, we can describe the overall picture of real estate price in Ames, Ohio and recommend the most feasible price for Zillow Co.' agents.

In order to better understanding data, we have different plots to explain the data from different aspects. Especially, we draw some visualizations to better understand the distribution of target variable, Exhibit 1 is the correlation plot among specific variables,it shows the correlation between different numerical variables. To get the most insightful idea from our data visualization and to better prove our idea, we just choose to plot the correlation of numerical variables, which will be selected in the next step.



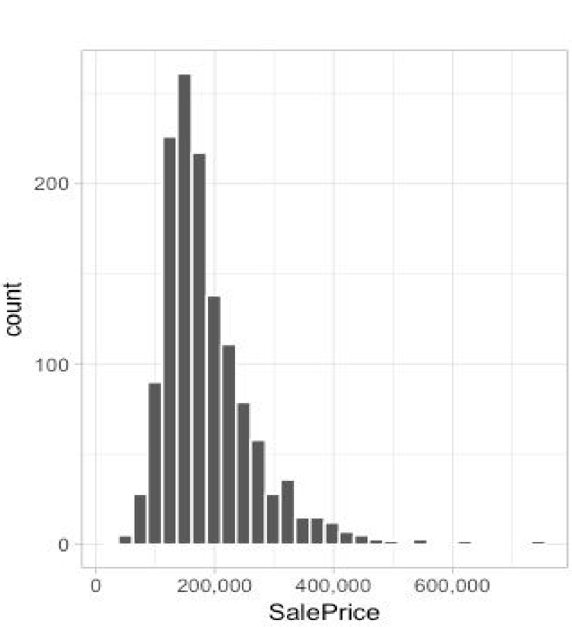


Exhibit 1

Exhibit 2

Exhibit 2 is the sale price distribution, from this plot, we can clearly see the sale price data is right skewered and a lot of the data are concentrated in $180,000, which means house commonly sell around $180,000. And we double checked our train and test dataset, the mean and median are almost the same. .

Data Preparation

We began with detailed data filtering and cleaning to better prepare for the following model process. It is especially crucial when we have 81 variables.

The first step is to clean the dataset by finding problematic variables using our business sense and data understanding, the problems including:

1. Irrelevant variables such as ID (which is only the order of our observation);
2. Missing variables (which have almost all N/A in that columns)

3. Redundant variables such as garge and garage (garge and garage data have same value for each row, we believe garage is some typo or backup or update for the dataset, and we should only keep one of them);
4. Variables which has less than 8 observations in each one specific level (since when we ran 10-folder cross validation in train set, 8 is too small to be divided into train and test sets and then into 10 folders and such variables will always cause error.) ;

For problem 1 and problem 2, we dropped those columns. For problem 3, we dropped every garge column, and left garage columns for the following model validation. For problem 4, it is the most difficult issue in our data. And after observing data, we deleted specific rows because in that observation, the value will damage our model building.

Modeling

- Model Selection

From unsupervised perspective, we use PCA to get more data insights and explore the dataset before building models, helping us to find patterns and to understand our data. PCA is specifically suitable for marketing and finance area and can provide a fundamental dimension reduction. We got exhibit 3 and exhibit 4. Exhibit 3 tells us that the variance explained by factor, and in exhibition 4, PCA reduced the variables from 81 to less.

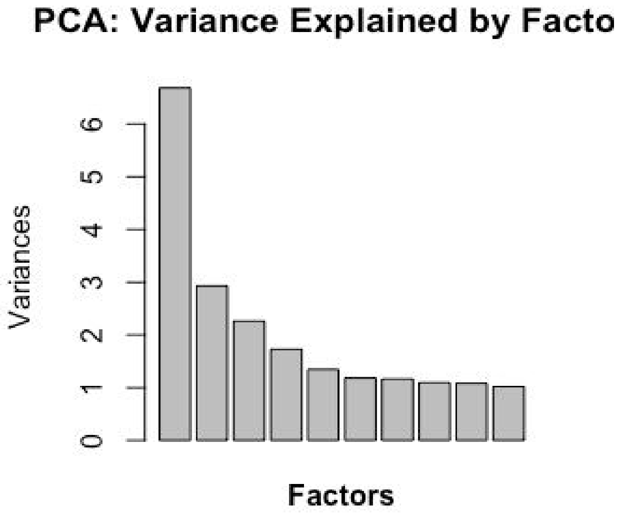


Exhibit 3

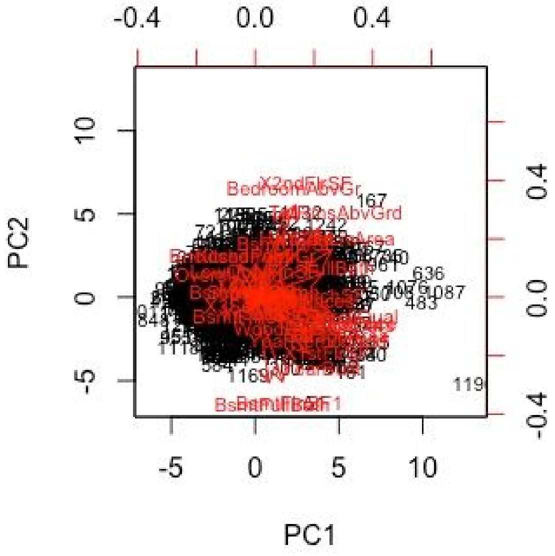


Exhibit 4

From supervised perspective, with the core task of predicting real estate sales price, we identified sales price as our target variable and focused most of our work on supervised learning. The models that we recognized as relevant are linear regression, linear regression with interaction, and classification tree. We also used a null model as a benchmark to evaluate different models' performances. However, along the process of model building, we took linear regression with interaction out of our consideration because it is not feasible considering the large amount of variables contained in the dataset.

- Variable Selection

After data preparation process,. there are still more than 60 variables, and many of them have high strong correlations with each other, such as Bedroom (Number of bedrooms above basement level) and TotRmsAbvGrd (Total Rooms Above Ground). So, we decided to further eliminate variables to avoid multicollinearity issue. For this consideration, we use VIF (Variance

Inflation Factor) to assess the severity of multicollinearity. The rule of thumb is that if VIF) > 10, then multicollinearity is high, so we used 10 as a cut off and removed all variables with a VIF score higher than 10.

In order to get a final set of variables, we ran a backward stepwise regression with the 61 variables. This process leaves us with 37 variables. Reducing variables can help prevent overfitting issues, but as a tradeoff, we will lose a certain amount of information contained in the eliminated variables. We examined the rest 37 variables and economically they all seemed relevant, so we proceeded to cross validation with this set of variables.

We used 10-fold cross validation and out-of-sample R-square to choose between linear regression and classification tree. However, in the process of cross validation, we discovered that for certain categorical variables, there are levels in the testing set but not in the train set. It creates a problem because our model does not know how to predict these properties' prices. This is actually due to some properties' special attributes. For example, only two houses in the dataset are made of wood. As a result, we just removed all houses with special features out of the dataset, and if a categorial variable contains too many unique levels, we just threw the variable out from the model. Concerned with overfitting issues, we then ran a Lasso function to punish having too many variables, and used the variables returned as our final choice. Using the linear regression model with the final set of variables, we successfully predicted the test set with an out of sample R-square of 0.85.

- Business Application

In the real estate industry, pricing is always challenging because if the price is too low, they are not maximizing their commission, and if the price is set too high, the selling cycle can become unnecessarily long and may incur additional cost due to more house showings. However, if they have a model to predict the market price given the features of the house, agents can then strategically price the properties to maximize their commissions.

Evaluation

- Discuss how the result of the data mining is/should be evaluated. Provide good measures of the performance of predictive models. How should a business case be developed to project expected improvement? ROI? If this is impossible/very difficult, explain why and identify any viable alternatives.

In training data, we have a linear regression model obtained an out-of-sample R^2 value of 0.87. In testing data, we have an out-of-sample R^2 of about 0.85, which means that we have a good predictive model. However, we may also need to evaluate our model further. Live testing can be a solution. We could implement the model and use it to price houses for a part of our sample while not using it for the other group. This instance of A-B testing could help us gather more data about the accuracy of our model. By comparing revenue generated from A and B separately, we can know how much value our model adds to the business. To develop a business case for expected improvements, we need to look at the current margins on sales for the firm and we need to demonstrate expected gains in revenue from using our model.

Deployment

- Discuss how the result of the data mining will be deployed.

The linear model will be used to predict future market prices of houses. This information will be used by real estate agents to price their houses more competitively, which will result in an increase in both profits and customers.

- Discuss any issues the firm should be aware of regarding deployment.

Some houses with features that are not typical have been removed from the dataset and this is something /the firm should be mindful of when deploying the model. In addition, the model is built on data from Ames, Iowa. therefore, the model should only be used at first in areas similar to Ames.

- Are there important ethical considerations?

Some possible issues about fair pricing could arise and the firm should not overcharge its customers. They should always keep in mind that it's important to keep the market well-functioning and sustainable. In addition, The firm must also be very mindful to not engage in any real estate fraud. For example, since the result of the data mining is manipulatable, the firm is able to interpret the data in an unethical way to both its shareholders and customers to attract more investments.

- Identify the risks associated with your proposed plan and how you would mitigate them.

The key risk in our plan is the unpredictability of real estate market conditions, because market does have a significant influence on real estate sales price. In addition, the data points in this dataset represent housing transactions from 2006 to 2010, so we need to recognize that because of the 2008 financial crisis, which greatly affected the housing market, the data might be biased (biased high before 2008 and biased low after 2008). One way to mitigate the risk is to use more recent data to fit the model.

Appendix 1 - Team Members' Contribution

Linqi Nie: Data integration, data selection, data cleaning, data transformation, data mining, pattern evaluation, deliverable

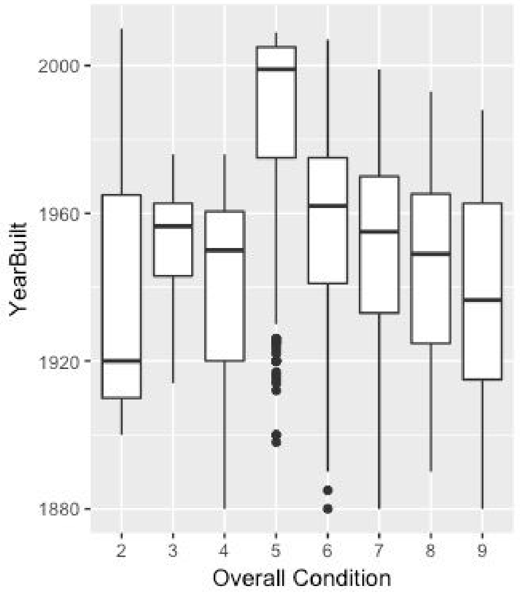
Mei Zhang: Data integration, data selection, data cleaning, data transformation, data mining, deliverable

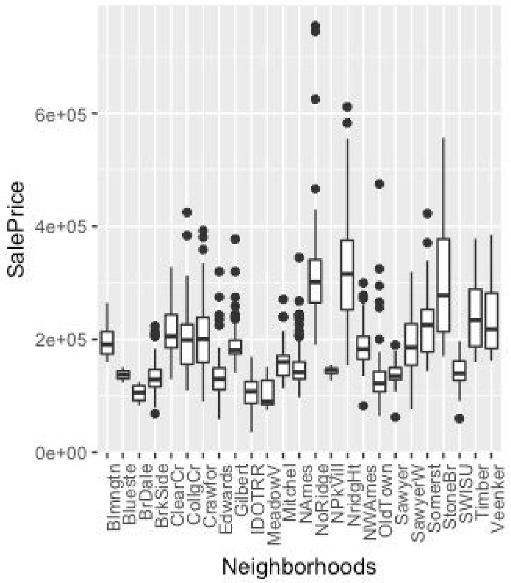
Joanna Sun: Data transformation, pattern evaluation and knowledge presentation, deliverable

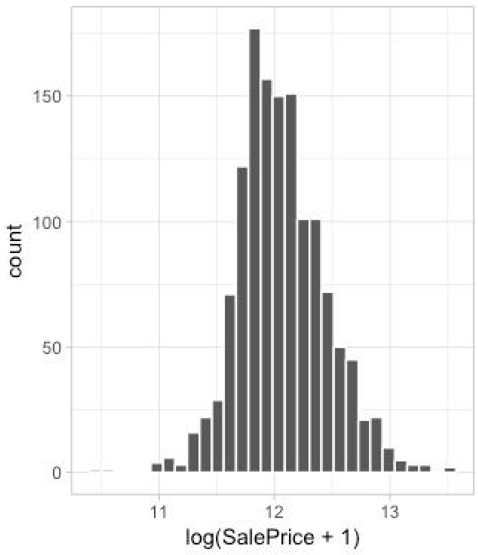
Stella Zhou: Data Integration, data selection, data cleaning, data mining, data transformation, deliverable

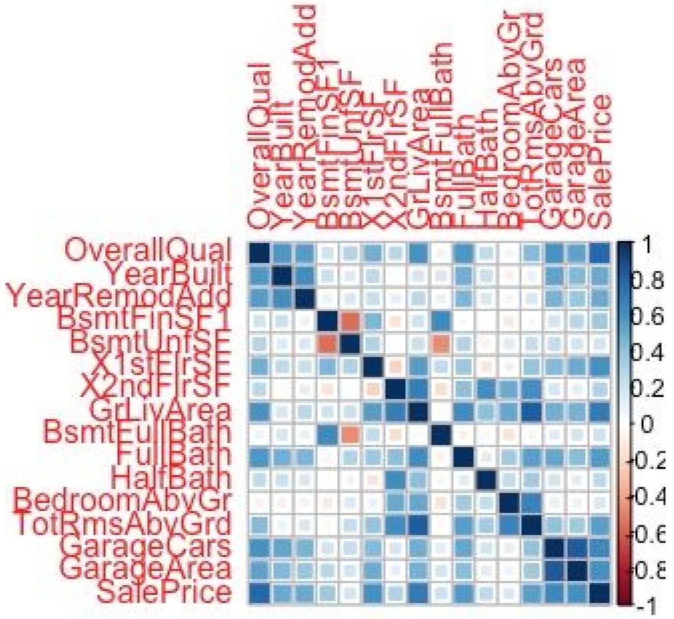
Dominique Vidjanagni: Data Integration, data selection, deliverable

Appendix 2 - Additional Visualization


YearBuilt
Overall Condition


SalePrice
6e+05
4e+05
2e+05
0e+00
Blmngtn
Blueste
BrDale
BrkSide
ClearCr
CollgCr
Crawfor
Edwards
Gilbert
IDOTRR
MeadowV
Mitchel
NAmes
NoRidge
NPkVill
NridgHt
NWAmes
OldTown
Sawyer
SawyerW
Somerst
StoneBr
SWISU
Timber
Veenker
Neighborhoods


count
log(SalePrice + 1)


OverallQual
YearBuilt
YearRemodAdd
BsmtFinSF1
BsmtUnfSF
X1stFlrSF
X2ndFlrSF
GrLivArea
BsmtFullBath
FullBath
HalfBath
BedroomAbvGr
TotRmsAbvGrd
GarageCars
GarageArea
SalePrice