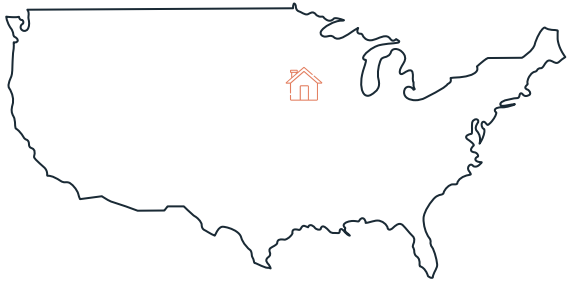
Predicting & Maximizing Home Value

NYCDSA Machine Learning Project

by Lucas Kim, Ryan Park, Sita Thomas

September 2020

Project Overview



Audience

- Fictional Data Mining Company

Dataset

- 1460 houses in Ames, Iowa
- 80 features

Objective

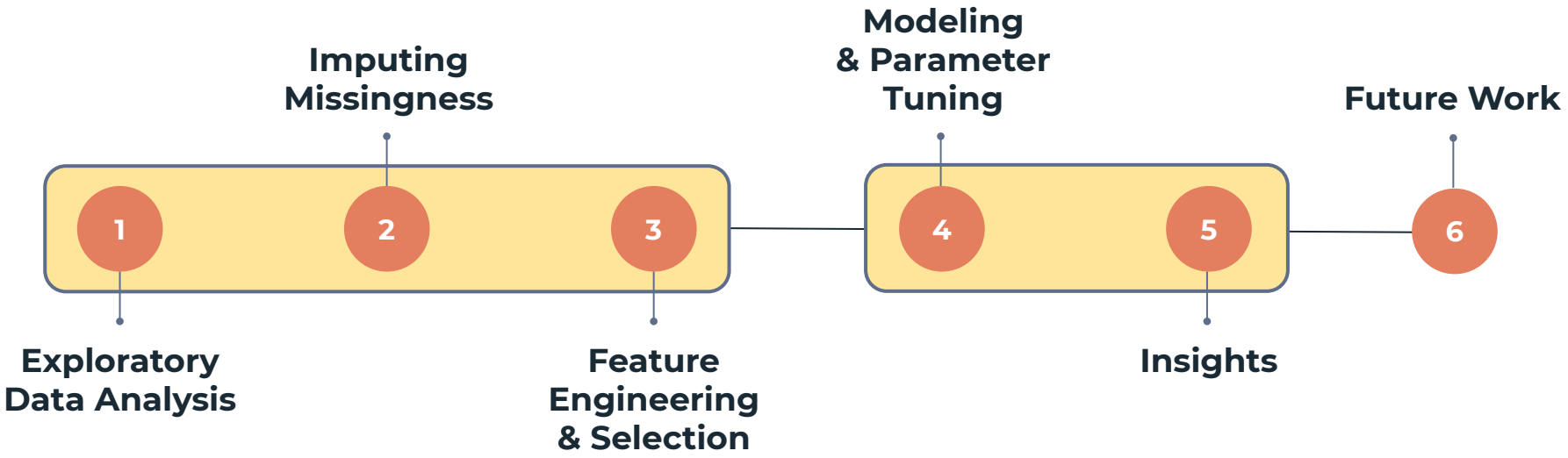
- Predict home sale prices
- Describe feature relationships

Project Overview

For this project, we imagined that we work for a fictional data mining company that has given us a database including 1,460 observations, each representing a single home, and 80 features describing different aspects of the house, including the exterior, interior, the lot, the home's surroundings, as well as information about the last sale of the house.

The fictional company wants to sell a pricing prediction model and the insights from this dataset to as many different target audiences as possible. So our tasks included building a predictive model and collecting useful descriptive information from the features and their relationships to each other. We were additionally tasked with documenting our modeling process so it can be used in production and adjusted for better performance or to be used with new data.

Workflow


Imputing Missingness
Modeling & Parameter Tuning
Future Work
1
2
3
4
5
6
Exploratory Data Analysis
Feature Engineering & Selection
Insights

Workflow

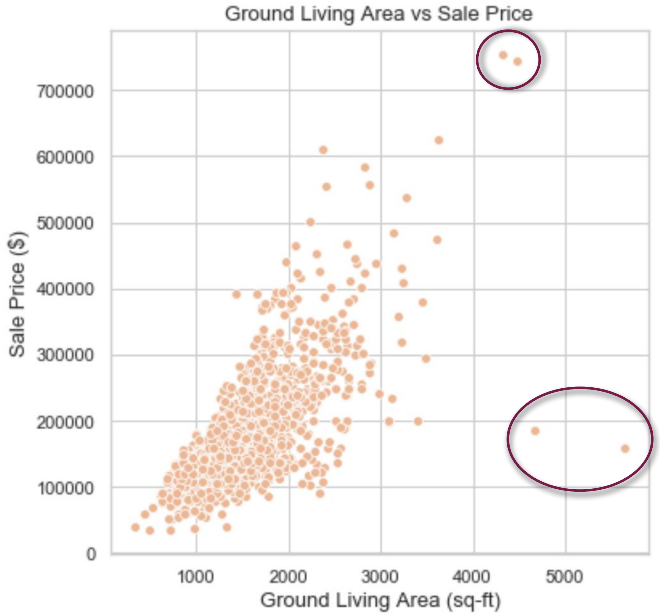
This project naturally divided itself into sections. First we pre-processed the data in three stages. Then we ran several models on our data, adjusted their parameters, and modeled again until we reached an error threshold of 10% or less. Finally, we examined the modeled data for insights, with consideration to how this project could be improved in the future.

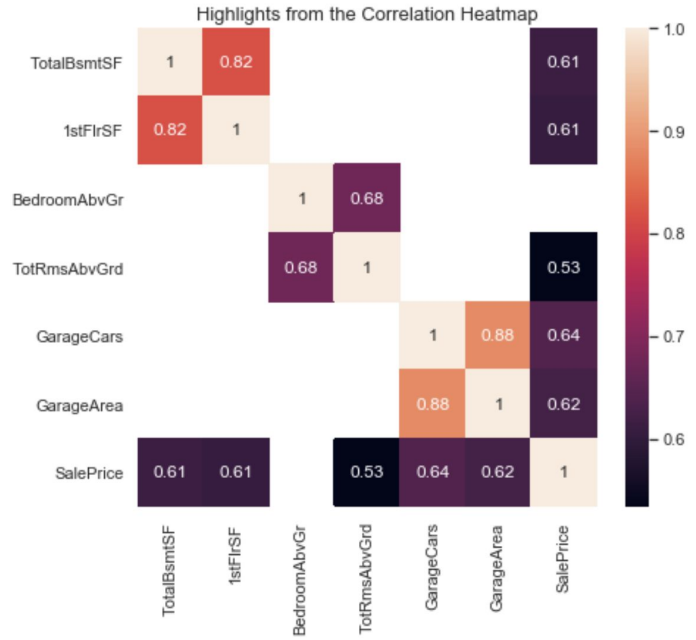
01.

Pre-Processing

Exploratory Data Analysis, Imputing Missingness, and Feature Engineering & Selection

Exploratory Data Analysis





Exploratory Data Analysis

We engaged in graphical and numeric exploratory data analysis to gain a deep understanding of the dataset, as well as the relationships between the features and the target variable, Sale Price. We removed outliers that may affect correlations and modeling, such as the ones circled for Above Ground Living Area on the previous slide.

A handful of variables were highly correlated with Sale Price and some of the highlights are shown in the correlation heatmap on the right side, such as 1st Floor Square Feet and Basement Square Feet.

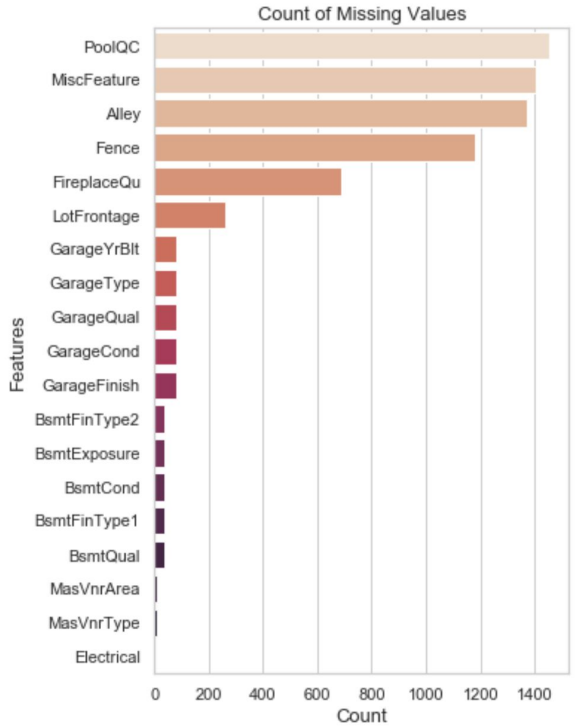
Another interesting observation we made was that Total Rooms Above Ground is highly correlated to Sale Price but Bedrooms Above Ground is not. These observations and others formed the basis for the feature engineering we'll discuss soon.

Imputing Missingness

5.96% missingness

Missing Not at Random
Garage, Bsmt, PoolQC, etc.

Missing at Random or Missing Completely At Random
LotFrontage, MiscFeature, etc.



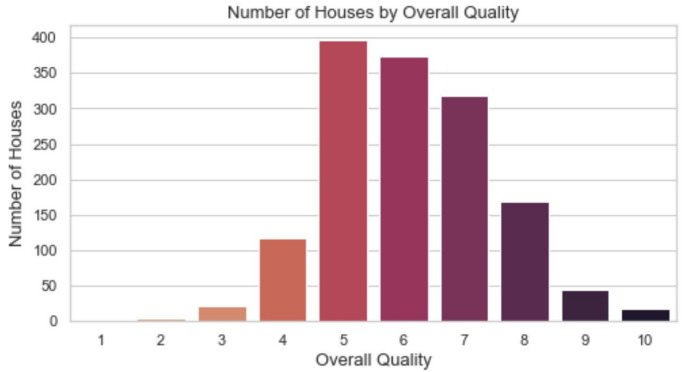
Imputing Missingness

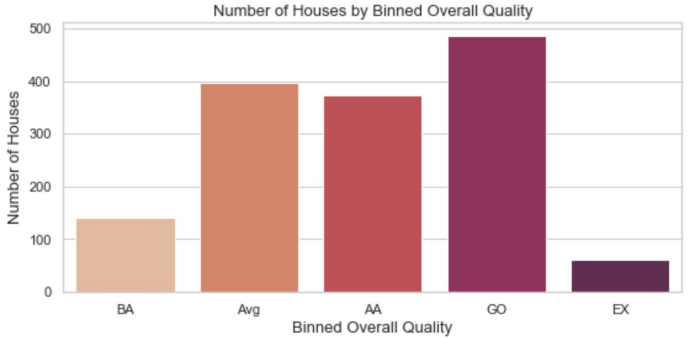
Once we had a basic understanding of the features, we looked at missing values, which was about 6% of the values and was concentrated in a small number of columns. Many of those columns actually contained values Missing Not at Random, because their missingness was due to the lack of presence of those features, such as a garage or basement. Those values were imputed with a ‘Not Applicable’ string for categorical features. There were no numeric features Missing Not at Random.

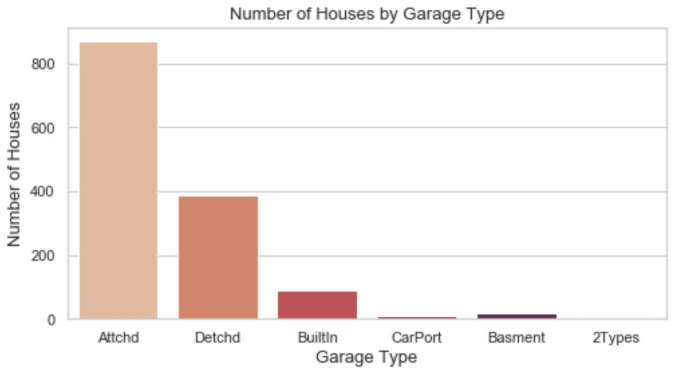
After imputing values that were missing not at random, Lot Frontage had the highest number of missing entries. So, we imputed those values with the Lot Frontage mean by Neighborhood.

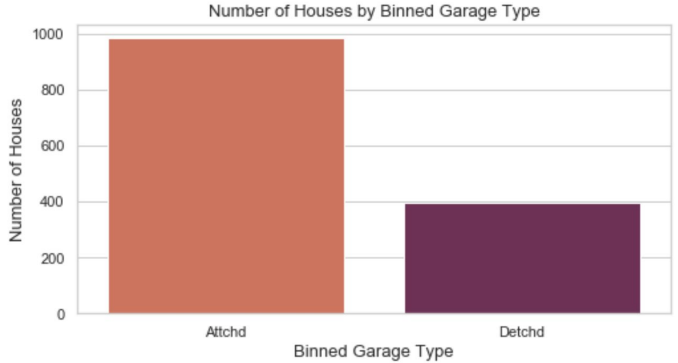
The remaining missing values were imputed with the mean or mode of their columns. Imputation in both the training and test datasets used fill values from only the training dataset to avoid data leakage.

Feature Transformation









Feature Transformation

In terms of feature transformation, to reduce the number of categorical features that have almost no variance, and to prevent the Curse of Dimensionality caused by dummifying the dataset for linear models, several features were binned into fewer groups, such as Overall Quality on the left, which was ranked from 1 - 10, including the oddly quantified “Very Excellent” category, and Garage Type on the right, which had six categories, although the vast majority were contained in only two of them.

Some other features whose information we were confident was contained elsewhere, were dropped altogether in order to prevent multicollinearity.

Feature Engineering

ExtraRoom

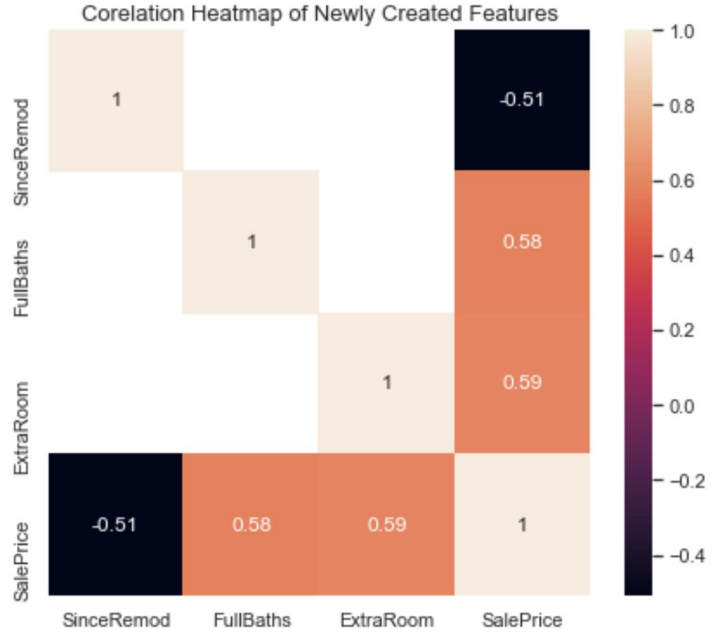
Total Rooms - Bedrooms

FullBaths

Above Ground Bathrooms + Basement Bathrooms

SinceRemod

Year Sold - Year of Remodel



Feature Engineering

Based on what we observed in the Exploratory Data Analysis stage, we created 6 new features. This choice was to reduce dimensionality, as well as to better explain and predict Sale Price.

For example, the new ExtraRoom feature calculates the difference between Total Rooms and Bedrooms, allowing TotalRooms to be dropped without losing much of its information. The new FullBaths predictor combines main living area and basement bathroom quantities, and HalfBaths were treated likewise. The new SinceRemod input contains the duration between the year the home was sold and the last time it was known to have been remodeled.

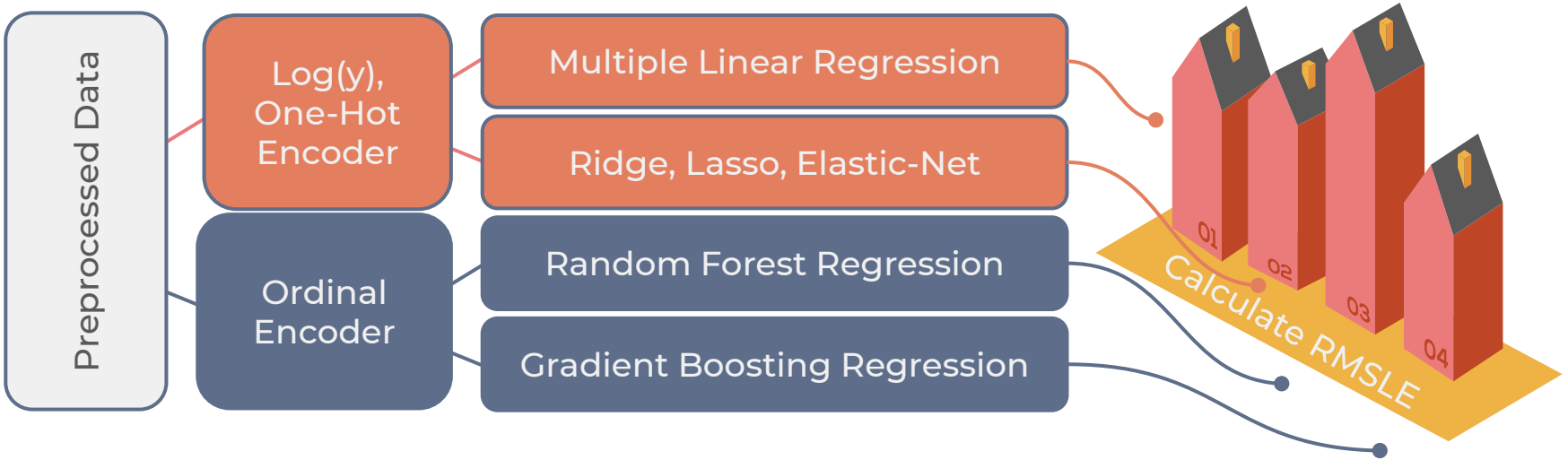
A a result, you can see in the correlation heatmap on the previous slide that these newly created features are highly correlated with Sale Price. In total, we used 49 out of 79 predictors for our models.

02.

Modeling

Linear Regression
and Tree Based Models

Modeling Pipelines



Modeling Pipelines: Regression

We explored two types of models, linear regressions and tree-based regressions.

For linear models we took the log of Sale Price and dummified the predictors, then we studied Multiple Linear Regression as well as the regularized models Ridge, Lasso, and Elastic-Net.

Modeling Pipelines: Tree-Based

For tree-based models we numerically encoded the features, then studied multiple types of decision trees, including Random Forest and Gradient Boosting.

After fitting each of the models, we compared the difference between the train and test set errors in order to measure the fit and adjusted parameters as needed. Then, we evaluated performance using R-squared and Root Mean Squared Log Error.

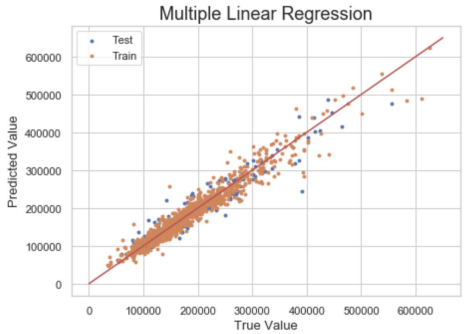
Log Transformation of Sale Price

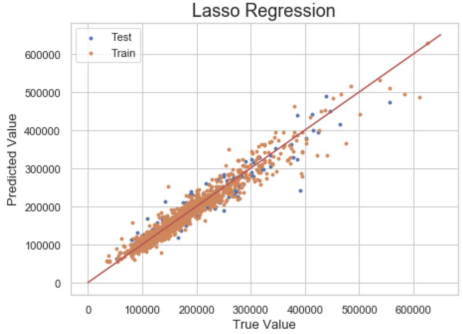
Log Transformation of Sale Price

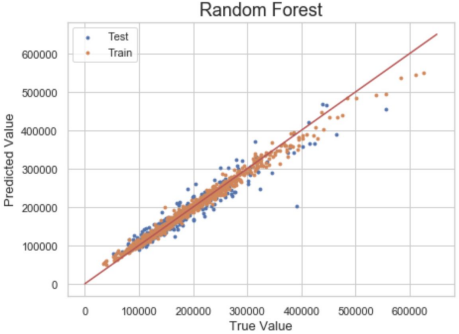
Taking the log of the output variable, Sale Price, better satisfied the Linearity assumption of regression models and improved performance of all the linear regressions we explored.

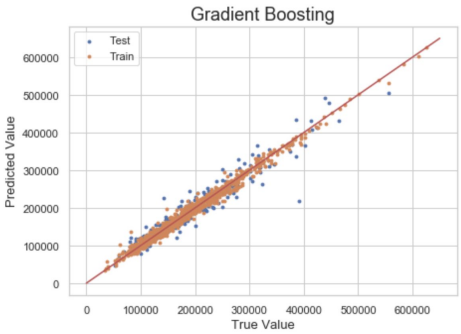
We did not apply the log transformation to the dependent variable when running tree-based models because tree-based models do not make assumptions about the relationships between the target and predictor variables.

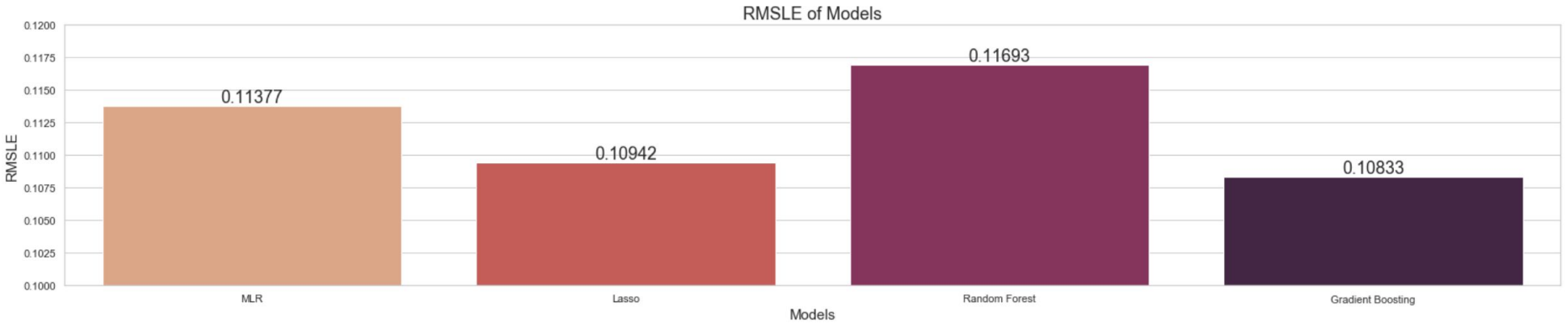
Model Performance











Model Performance

The top row of plots in the previous slide illustrate the comparison between true and predicted values. The models performed similarly, indicating that the nature of the data in combination with our pre-processing work went a long way toward building solid models, avoiding much of the garbage-in, garbage-out trap.

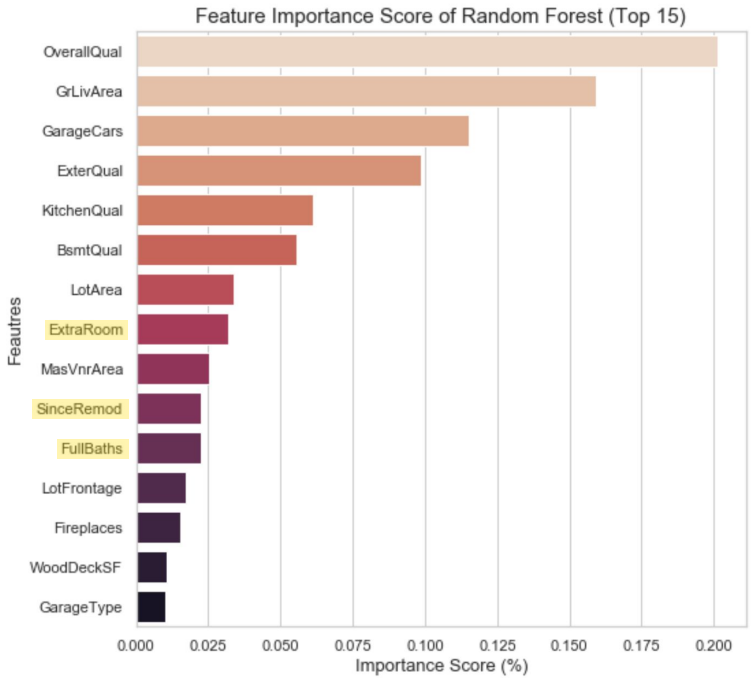
Closer observation shows the progressive tightening of values around the mean from plot to plot, especially more obvious for the more highly priced homes, where there is much less data on which to train the model.

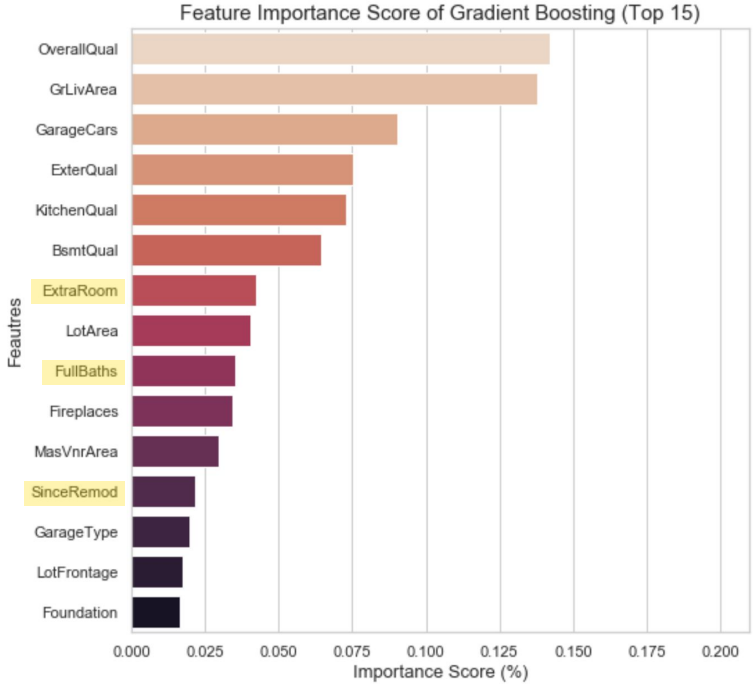
Model Performance

In the RMSLE (Root Mean Squared Logged Error) graph on the first Model Performance slide, we see that interestingly, the tree-based models had both the highest and lowest error scores, despite being in the same class.

The true vs. predicted value plots indicate that the Random Forest model is performing poorly because it is biasing the data, shown by the overall reduction in slope. The Gradient Boosting model, on the other hand, does *not* affect the slope, *and* reduces the spread around the mean compared to the other models.

Feature Importances





Feature Importances

Although Random Forest and Gradient Boosting performed notably differently, they produced nearly identical Feature Importances.

The top 6 features that contribute the most to Sale Price are Overall Quality, Ground Living Area, Garage Size by Car Capacity, Exterior Quality, Kitchen Quality and Basement Quality.

Our engineered features, *ExtraRooms*, *FullBaths*, and *YearsSinceRemod* appear in the top 15 features.

This confirmed the multicollinearity we found among the individual variables and the correlations between the combined variables and Sales Price.

03.

Insights

Quality and Additions

Quality

The single most important factor in selling a home, given the context of this dataset, is "the overall material and finish of the house." A home with any set of features will likely sell better than any other house with comparable features as long as the home has higher-quality materials and finishes.

Exterior, Kitchen, and Basement Quality all closely follow behind Overall Quality. If one is prioritizing areas to remodel, outdoor finishes, followed by indoor finishes and finally basement finishes may be the best approach. That said, if remodeling over several years with plans to sell the home in the future, Exterior Quality has the advantage of staying in style many decades longer than interior finishes. Thus it may be wise to prioritize the order of interior finishes so that the most outdated areas of the home will be those that contribute less strongly to Sale Price, given that the Years Since Last Remodel also influences Sale Price.

Additions

While Above-Ground Square Footage is the second most influential on Sale Price, one must consider the costs associated with increasing this feature. It may be more worthwhile to simply replace fixtures and/or add cladding. Also to be considered are things like lot size and local regulations - some homes may be limited in their ability to grow.

The Garage Size by Car Capacity can be considered similarly. Where garage size may be most valuable is at homes that have neither a garage nor carport. Adding *some* sort of vehicle storage structure, even a poor quality one to keep the cost down, may be worthwhile.

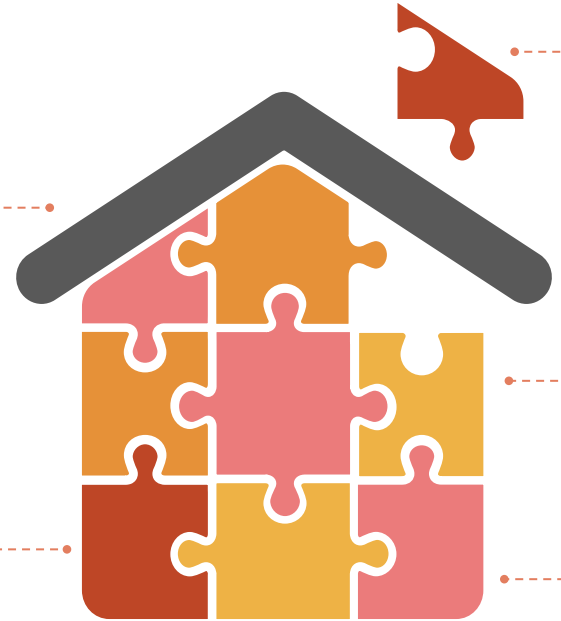
Remodeling Insights

Garage Size by Car Capacity

May be a good return on investment for homes with no garage or carport.

Above Grade Living Area

Long-term value but a significant investment.



Overall Quality

Most added value. The materials and finishes of the home are key to pricing.

Exterior Quality

May improve both home safety and utility costs.

Kitchen and Basement Quality

Relatively easy and inexpensive to update, but easily outdated.

04.

Future Work

Dataset and Workflow

Limitations

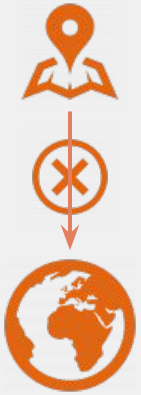
Despite providing highly useful insights, this dataset is acutely limited - only capturing about a thousand homes, specifically in Ames, Iowa, sold only between 2006 and 2010.

Therefore, this pricing prediction model, while relatively accurate for Ames, Iowa 10 years ago and possibly even today with an adjustment for inflation, is unlikely to perform well on unseen data from different locations and different years. It should be trained on a substantially larger dataset with a far greater variety of observations.

Although the prediction tool needs work before production launch, the insights could be valuable to a variety of audiences. They could be sold directly to home buyers and sellers, to real estate companies, real estate investors, developers, materials manufacturers, home goods retailers, media companies, and local governments.

Limitations

Too few sample points



Limited generalisability

Valuable insights and potential for improved prediction

Model Improvements

Because this model needs so much work, we wanted to analyze our workflow and the pipeline in preparation for future iterations. We would like to explore other imputation methods and apply different dimensionality reduction techniques such as Lasso Regression and Principal Component Analysis, instead of or before manual feature selection and engineering. We also removed some outliers manually but would like to approach them more empirically.

Primarily, however, we would like to explore more advanced models such as LightGBM or XGBoost, and/or stacking multiple models. While we believe a roughly 10% error threshold is reasonable for this particular problem, we know it can be lowered considerably, which would make the model appreciably more valuable, especially in regard to high end homes where a 10% over or under-estimation of price can be tens to hundreds of thousands of dollars.

Dimension Reduction

Automate feature engineering
and selection

Outliers

Identify and remove
outliers automatically

XGBoost, etc.

Try more advanced models
to improve predictions

Stacking

Combine multiple models

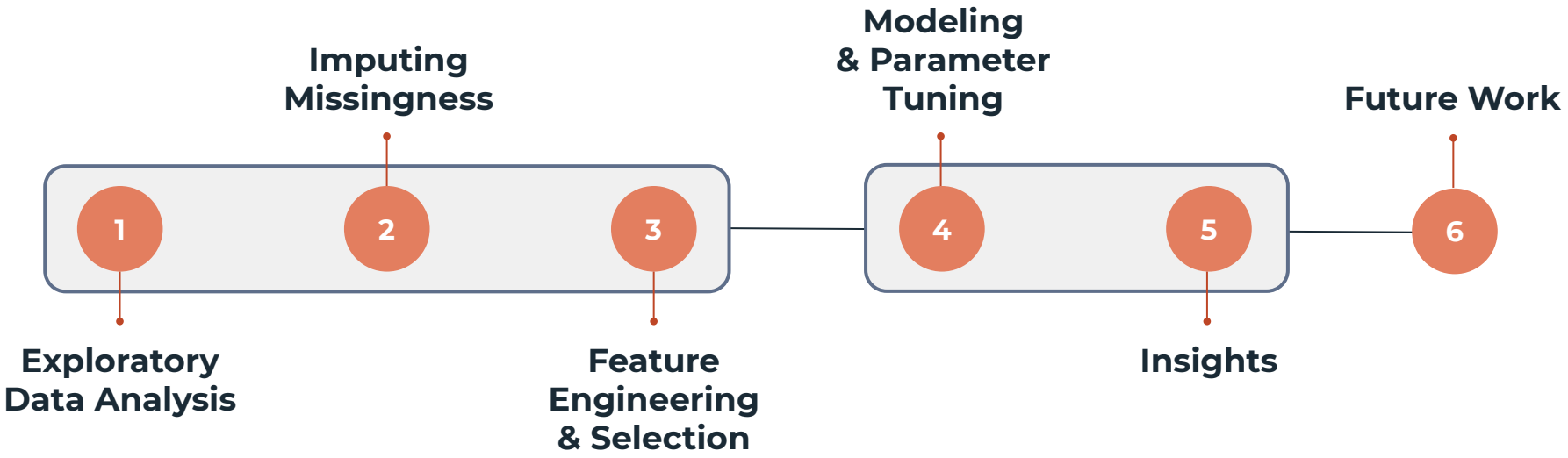
Summary

In summary, we were asked by a data mining company to design a home price prediction tool and gather insights about the factors that influence a home's value from a small dataset of observations.

We pre-processed the data by performing exploratory data analysis, imputing missing values, and manually selecting, combining, and dropping features. Then, we modeled the data with both linear and tree-based regressions until we reached a roughly 10% error rate. Using the best-performing model, we evaluated the influence of each predictor on the sale price of the home and developed several insights from the data that could be sold to a variety of audiences.

Lastly, we reviewed our work and determined the ways in which we would like to expand and improve the next iteration of this project.

Summary


Imputing Missingness
Modeling & Parameter Tuning
Future Work
1
2
3
4
5
6
Exploratory Data Analysis
Feature Engineering & Selection
Insights

Thank you for reading!