
DSI Kaggle Competition Entry

Estimating Home Prices in Ames Iowa

Owen Curtis

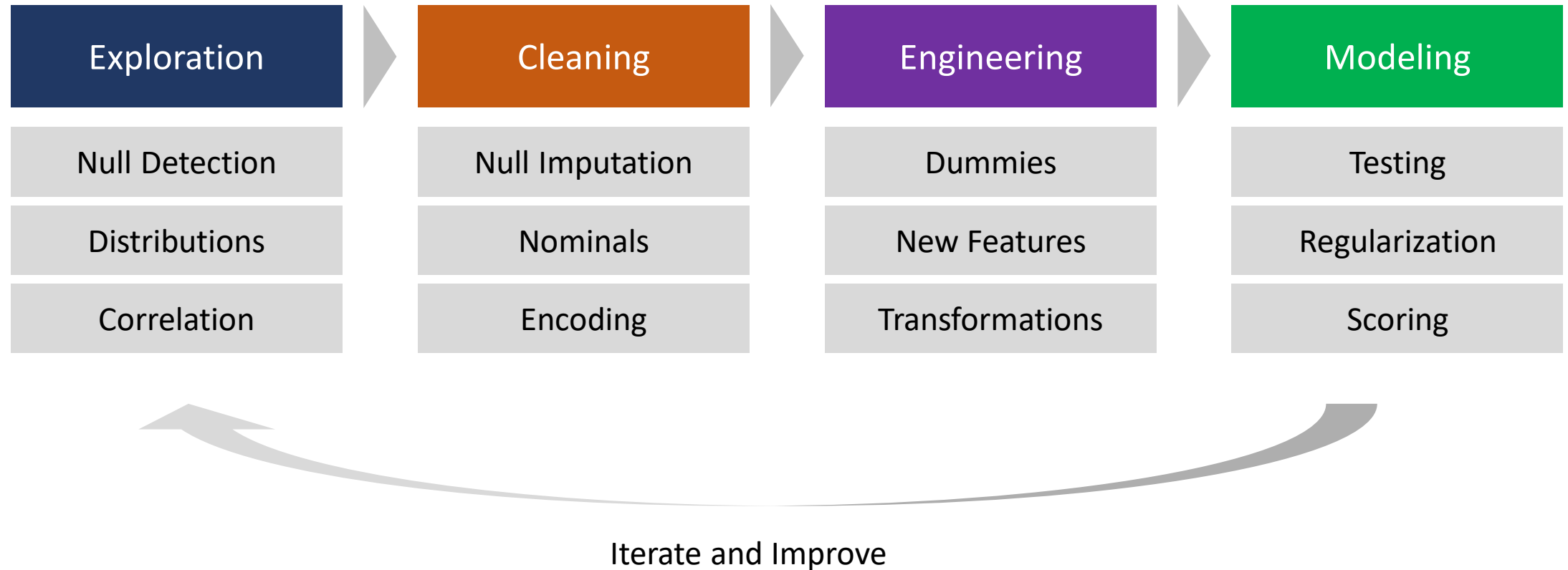
General Assembly

Background



- **Assessor's Office Data**
 - 80 data points per home
 - 2930 homes (2150 w/ price)
 - Timeframe: 2006-2010
 - Ex: Quality Scoring, Sq Ft, Garage
- **Task:**
 - Predict SalesPrice
 - Understand what home features matter
- **Scoring:**
 - RMSE
 - R2

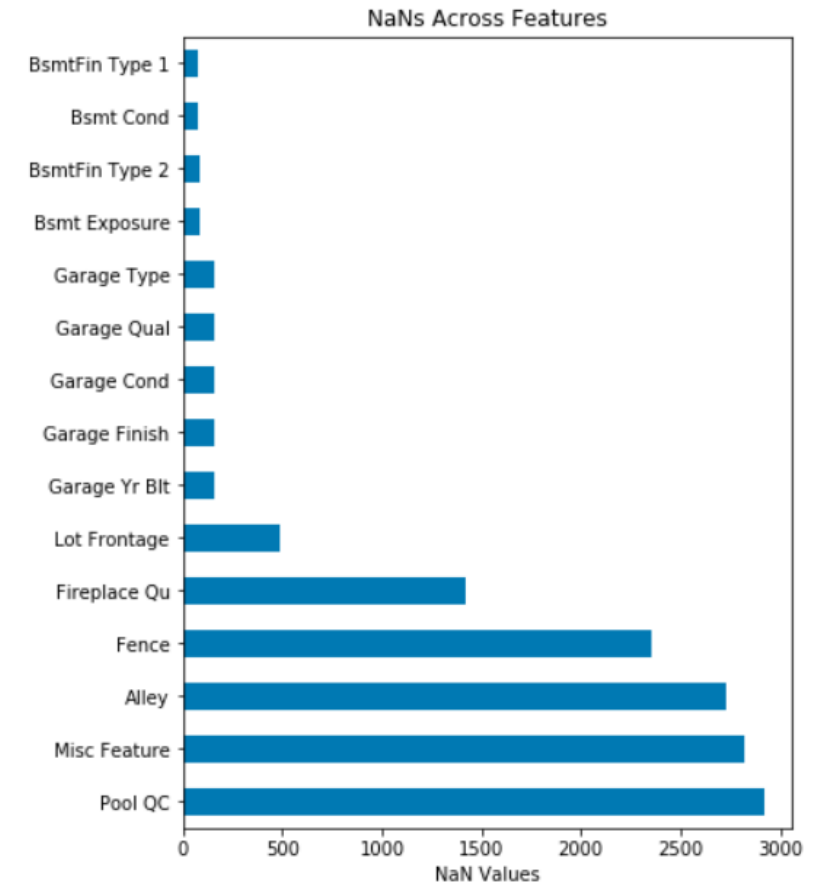
Approach



Exploration

Null Detection

- **Features:** 51 categorical, 28 continuous
- **Large amount of missing/NaN datapoints (~14,000)**
- **Major contributors:** rare house features, lot frontage, basement and garage features

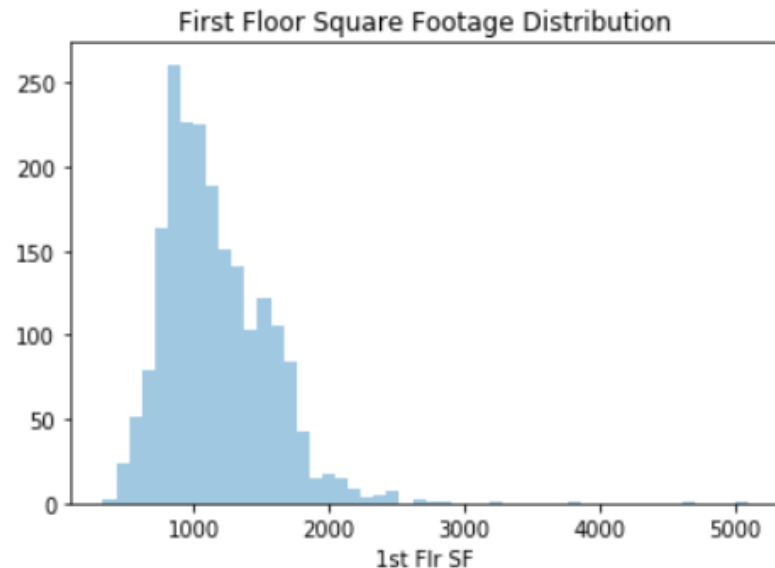


Exploration

Distributions

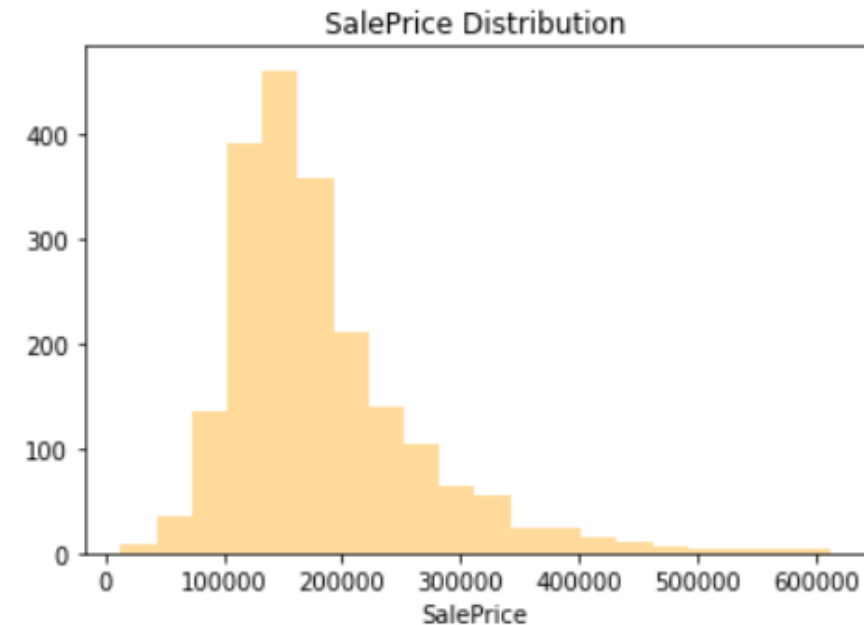
Our Features

- **Positive/Right Skew** for ~15 features (predominately SF, Area related)



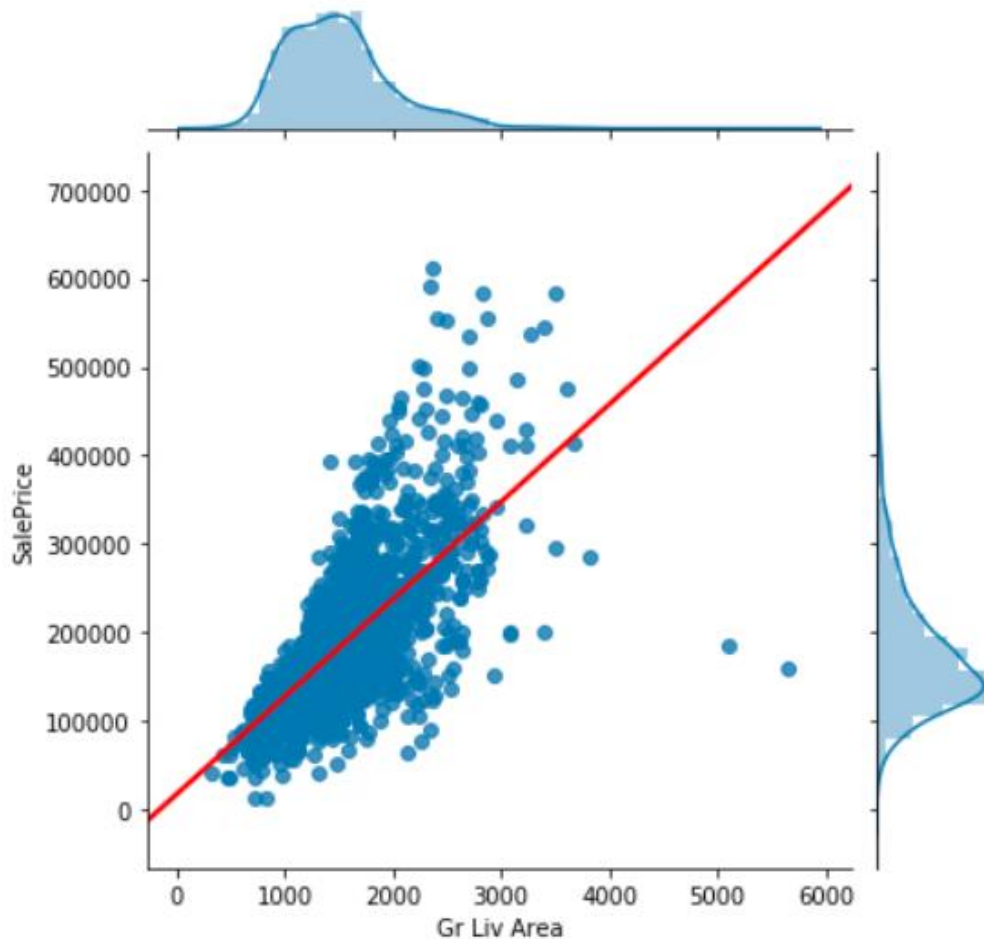
Our Target

- **Positive/Right Skew** driven by a small amount of expensive homes



Exploration

Correlation



- Feature v Target Correlations: ***Gr Liv Area, Overall Qual, Exterior Qual*** among highest
- **Extreme outliers** in General Living Area:
 - High Influence
 - Suggest in home family sales
- **Slight curve** to scatterplot suggests nonlinear relationship

Cleaning

Encoding

- Assessor's data includes a number of **quality/condition** metrics on various scales

Exter Cond

Value	Description	Encoded
Ex	Excellent	5
Gd	Good	4
TA	Average	3
Fa	Fair	2
Po	Poor	1

Nominal Data

- Alley, Misc Feature** were two nominal features with minimal information

Misc Features

Value	Descrip
Elev	Elevator
Gar2	Garage
Other	Other
Shed	Shed
TenC	Tennis Court



Feature:Shed
1
0

Null Imputation

- Majority of nulls mean feature non-existent. Imputed with **zero values**

Garage Type	Garage Cars	Garage SF
NaN	NaN	0.0

- There are some edge cases, however...



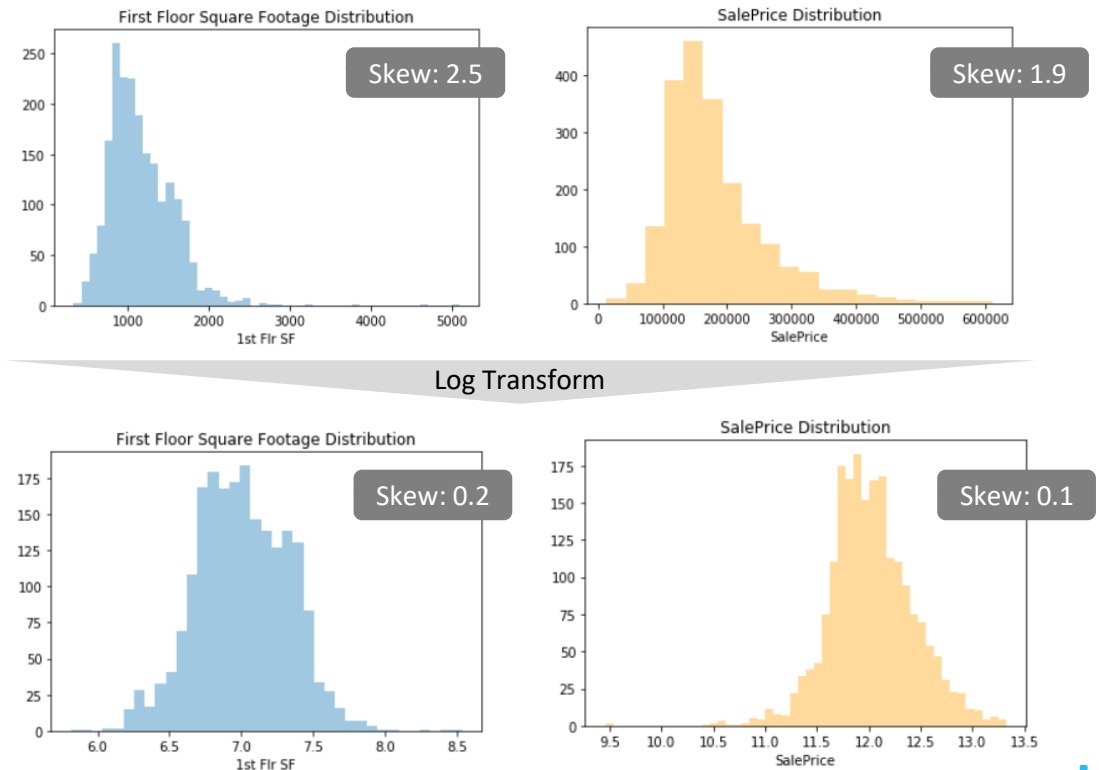
Dummy Features

140 dummy
features created;
~60 removed



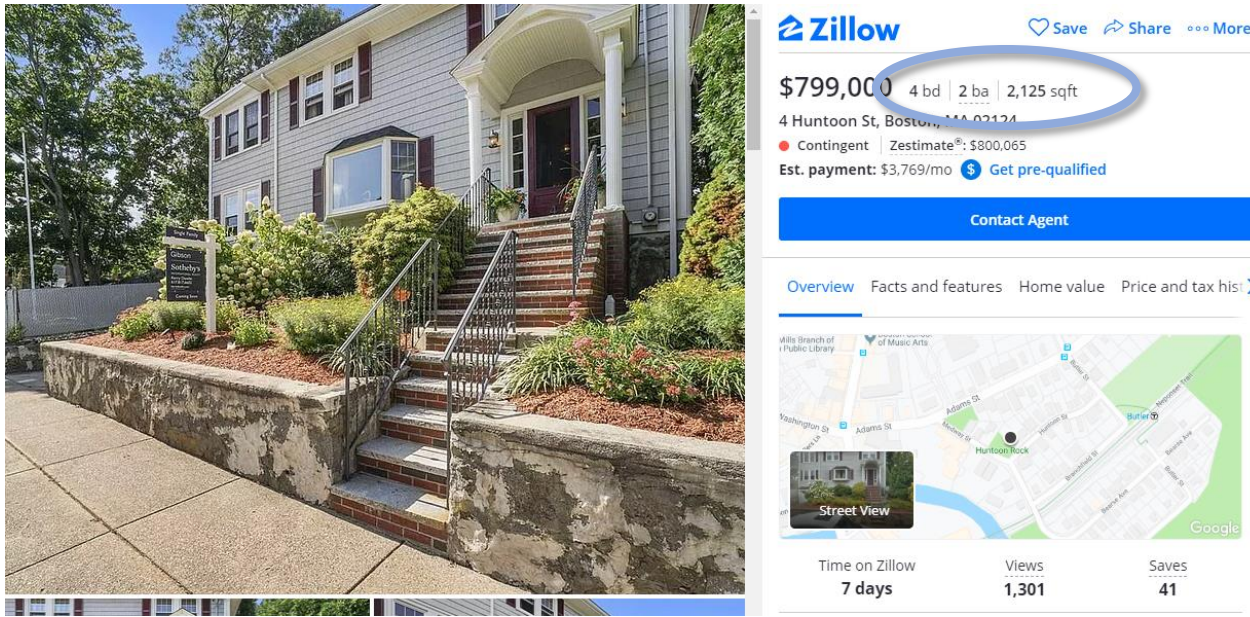
Transformation

- Select SF/Area features, Target log transformed



New Features - Manual

- Real Estate Aggregator research conducted; **Total SF, Total Baths, Age** among most important features
- New feature inclusion validated through correlation analysis



Zillow Save Share More

\$799,000 4 bd | 2 ba | 2,125 sqft

4 Huntoon St, Boston, MA 02124

Contingent Zestimate®: \$800,065

Est. payment: \$3,769/mo Get pre-qualified

Contact Agent

Overview Facts and features Home value Price and tax history

Street View

Time on Zillow: 7 days Views: 1,301 Saves: 41

Feature	SalePrice Correlation
Total Area	0.80
Gr Liv Area	0.72
Kitchen Qual	0.69
Total Bsmnt SF	0.67
1 st Flr SF	0.65
Total Baths	0.57

Engineering

New Features - Polynomial

- Programmatic generation of interaction terms returned strong correlation
- ...but multi-collinearity and overfitting a concern

Polynomial Term	SalePrice Correlation
Overall Qual Total Area	0.92
Exter Qual Total Area	0.91
Kitchen Qual Total Area	0.89
Yr Sold Total Area	0.87
Overall Qual 1 st Flr SF	0.84
Pave Drive Total Area	0.88

Polynomial Term	SalePrice Correlation
Overall Qual Total Area	0.92
Exter Qual Total Area	0.91
Kitchen Qual Total Area	0.89
Yr Sold Total Area	0.87
Overall Qual 1st Flr SF	0.84
Pave Drive Total Area	0.88

R2 improved
from **.88** to **.89**
with reduction of
polynomials
from 50, to 2

Data Processing Summary

Nulls Identified and Imputed



Ordinals Encoded



New Features



Interaction Terms



Log Transformations



Modeling – Process

What We Want

Low RMSE / High R2

Bias – Variance Balance

Residuals: Homoscedasticity

Residuals: Normal Distribution

What We Did

Test, Train, Split

Linear Reg Model Tuning (CV)

Lasso, Ridge, ElasticNet

Train to Full Data

Modeling – Base Model (Linear)

Linear Regression Model (Training)

Rank	Adjustments	RMSE (Est)	R2 (CV)
1	Selective X Transformations	19.8K	.901
2	Outlier Removal	20.8K	.889
3	-48 Polynomials	22.5K	.881
4	Log Transform X,y	23.5K	.870
5	+50 Polynomials	25.2K	.860
6	All Numericals	27.3K	.830

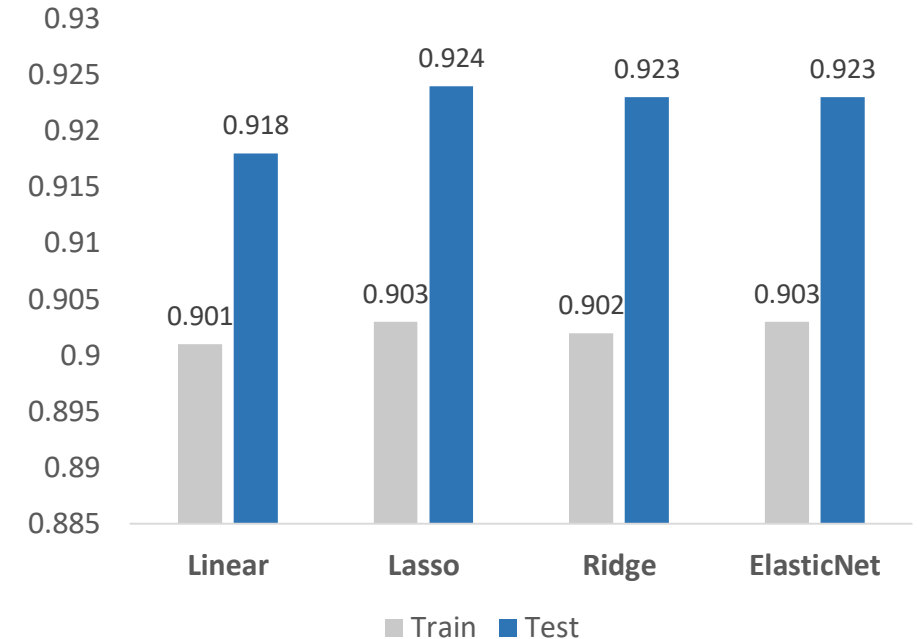
- **Major Gains**
 - Transformations
 - Outliers
- **Winning Recipe**
 - **94** Total Features
 - **2** Polynomials
 - **Selective X** Transformations
 - **Outlier** Removal

Modeling – Regularization

Model Ranking			
Rank	Description	RMSE	R2 (CV)
1	Lasso	19.6K	.903
2	Elastic	19.6K	.903
3	Ridge	19.7K	.902
4	Linear	19.8K	.901

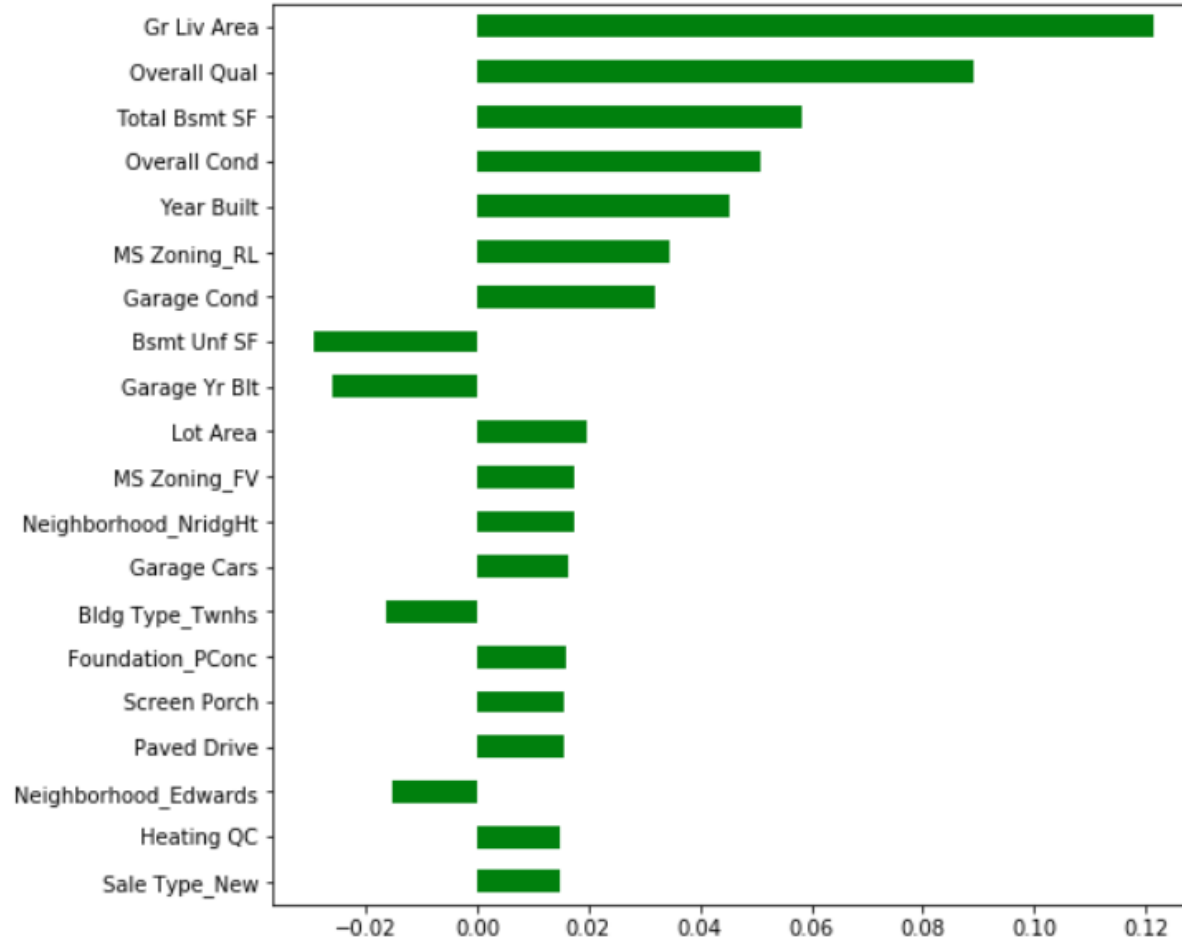
Final Kaggle Submission: 20K RMSE

R2 (Train v Test)



Modeling – Regularization

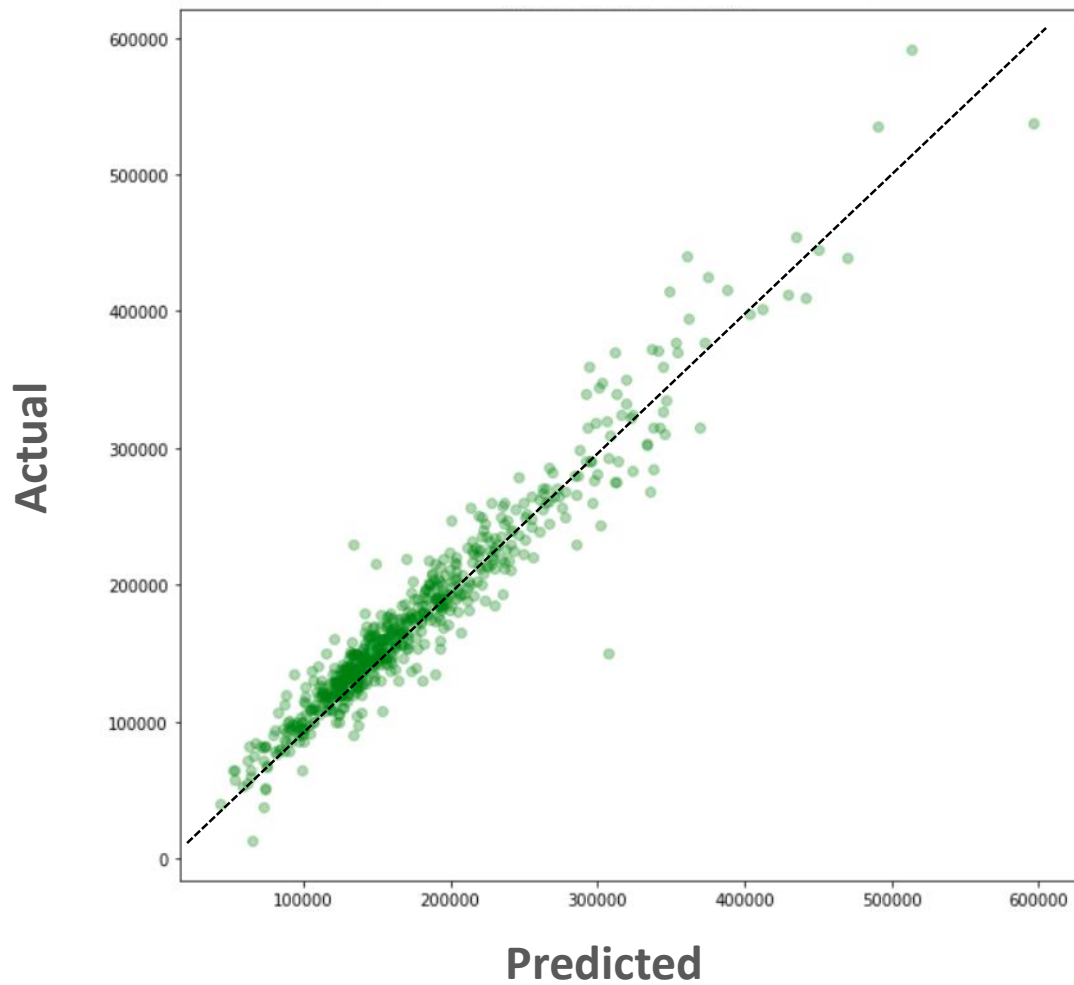
Lasso Coefficients



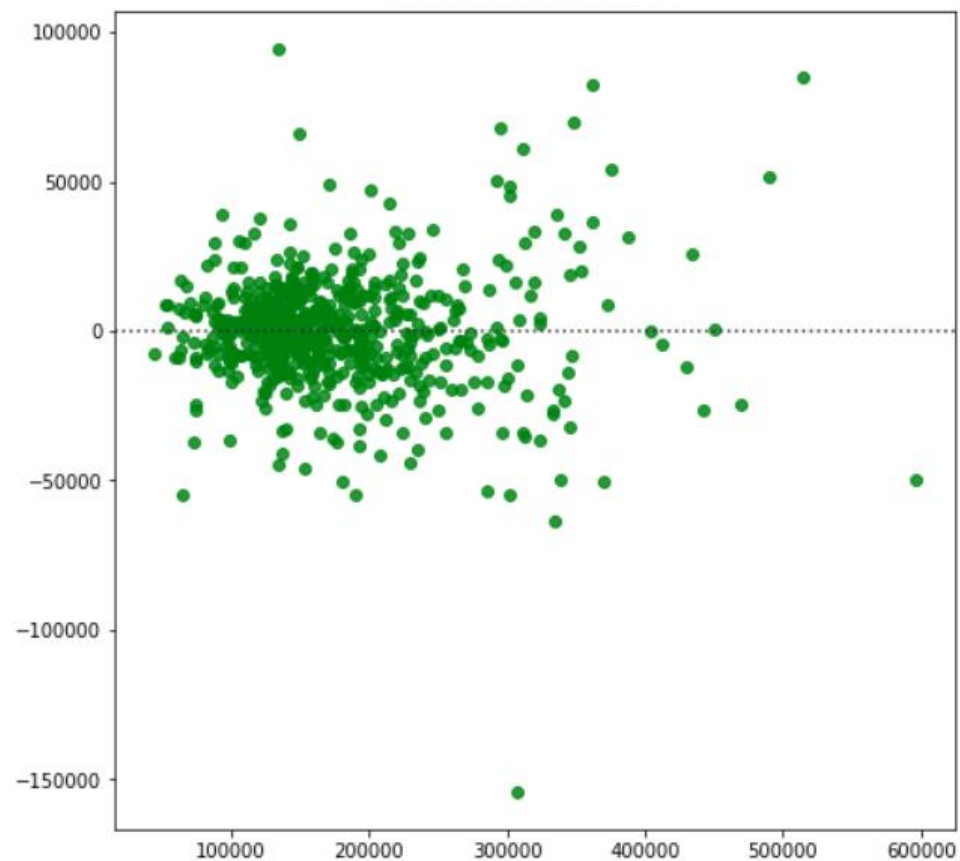
- General Living Area
- Overall Quality
- Basement SF
- Overall Condition
- MS Zoning Residential Low Density

Modeling – Residuals

Actuals v Predicted



SalePrice Residuals (Lasso)



Findings


What Drives Sale Price

“Quality” of Home

“Condition” of Home

Square Footage

Zoning/Neighborhood



Prediction

\$176K



Final RMSE

20K

THANK YOU