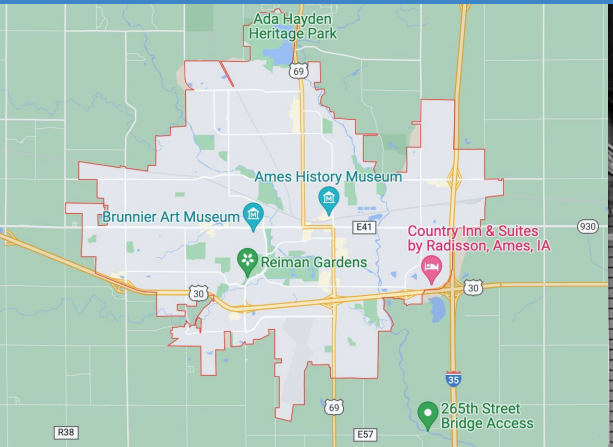
PROJECT 2

Getting Ahead of the Market: Modeling Ames Estate Sale Price based on Different House Features Available.



Overview

Problem Statement

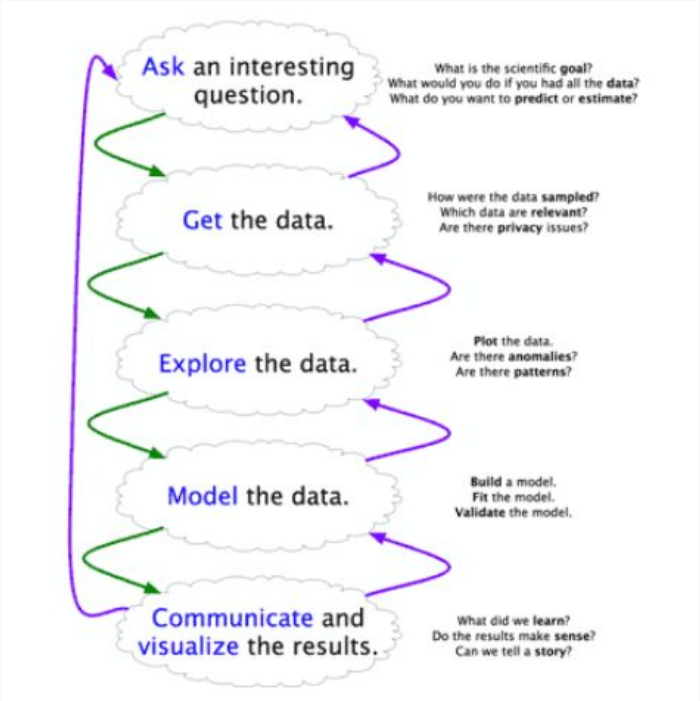
As data scientists engaged by Propnex, we are tasked to develop an accurate model to predict housing price in Ames, Iowa. In this study, we will attempt to use machine-learning model to forecast the house price in Ames, Lowas based on the given fixed house characteristics

Content

- Methodology
- Exploratory Data Analysis (EDA)
- Base Model
- Feature Engineering
- Choosing our Final Model
- Conclusion
- Recommendation

METHODOLOGY

Blitzstein & Pfister's Workflow



Step 1: Ask an interesting question

Step 2: Get the data

Step 3: Explore the data

Step 4: Model the data

Step 5: Communicate and visualize the results

EXPLORATORY DATA ANALYSIS

- Dataset Overview
- Dealing with Null Values
- Data Transformation
    - Numeric Variables
    - Categorical Features
    - Outliers

EXPLORATORY DATA ANALYSIS (EDA)

Dataset Overview

- It is observed that there are both numerical (continuous and discrete) and categorical (nominal and ordinal) features in our dataset, as such there is a need to split and categorized the dataset.
- The NA values reflected in the categorical features mostly represent the lack of the feature in the attribute rather than an actual null/blank value. Therefore, it is filled with NA for categorical features.
- Missing data are also observed for numerical data and based on the assumption that the amount of null values is insignificant so we replace it with mean value or linear regression prediction for analysis purpose.
- Some outliers are observed in high correlated variables and it will be dropped for better analysis.

Dealing with Null Values (Numerical Data)

Number of Null values in both the Train and Test dataset

Train Dataset

```
check_numeric_cols(X_train)
```

```
Lot Frontage has 330 NaNs, this represents 16.09% of the data
Mas Vnr Area has 22 NaNs, this represents 1.07% of the data
BsmtFin SF 1 has 1 NaNs, this represents 0.05% of the data
BsmtFin SF 2 has 1 NaNs, this represents 0.05% of the data
Bsmt Unf SF has 1 NaNs, this represents 0.05% of the data
Total Bsmt SF has 1 NaNs, this represents 0.05% of the data
Bsmt Full Bath has 2 NaNs, this represents 0.1% of the data
Bsmt Half Bath has 2 NaNs, this represents 0.1% of the data
Garage Yr Blt has 114 NaNs, this represents 5.56% of the data
Garage Cars has 1 NaNs, this represents 0.05% of the data
Garage Area has 1 NaNs, this represents 0.05% of the data
```

Test Dataset

```
check_numeric_cols(X_test)
```

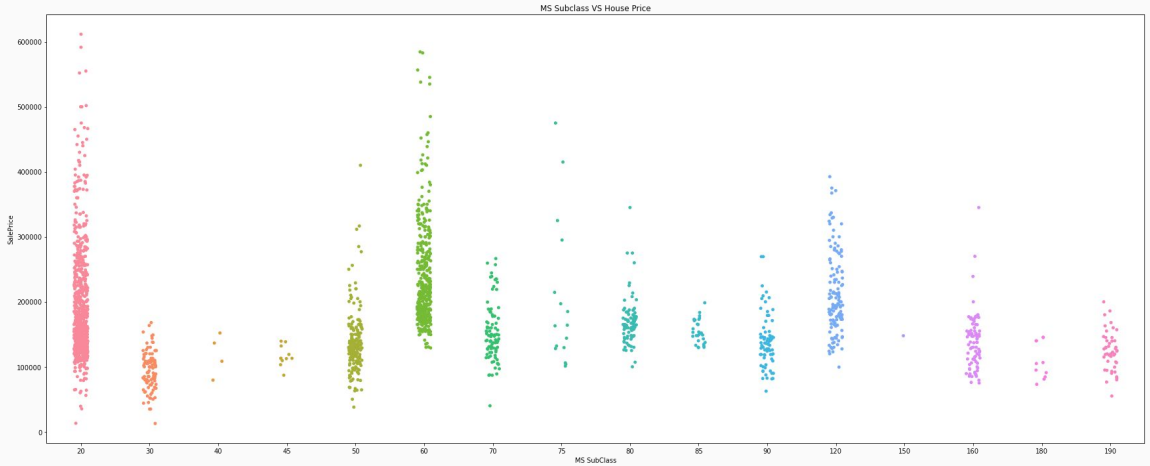
```
Lot Frontage has 160 NaNs, this represents 18.22% of the data
Mas Vnr Area has 1 NaNs, this represents 0.11% of the data
Garage Yr Blt has 45 NaNs, this represents 5.13% of the data
```

- For this analysis, we will fill in the numeric features with less than 1.5% of missing data with mean value,
- Use linear regression to fill in the NaN values for 'Garage Yr Blt' feature,
- And mean value to fill in NaN values for 'Lot Frontage' feature.

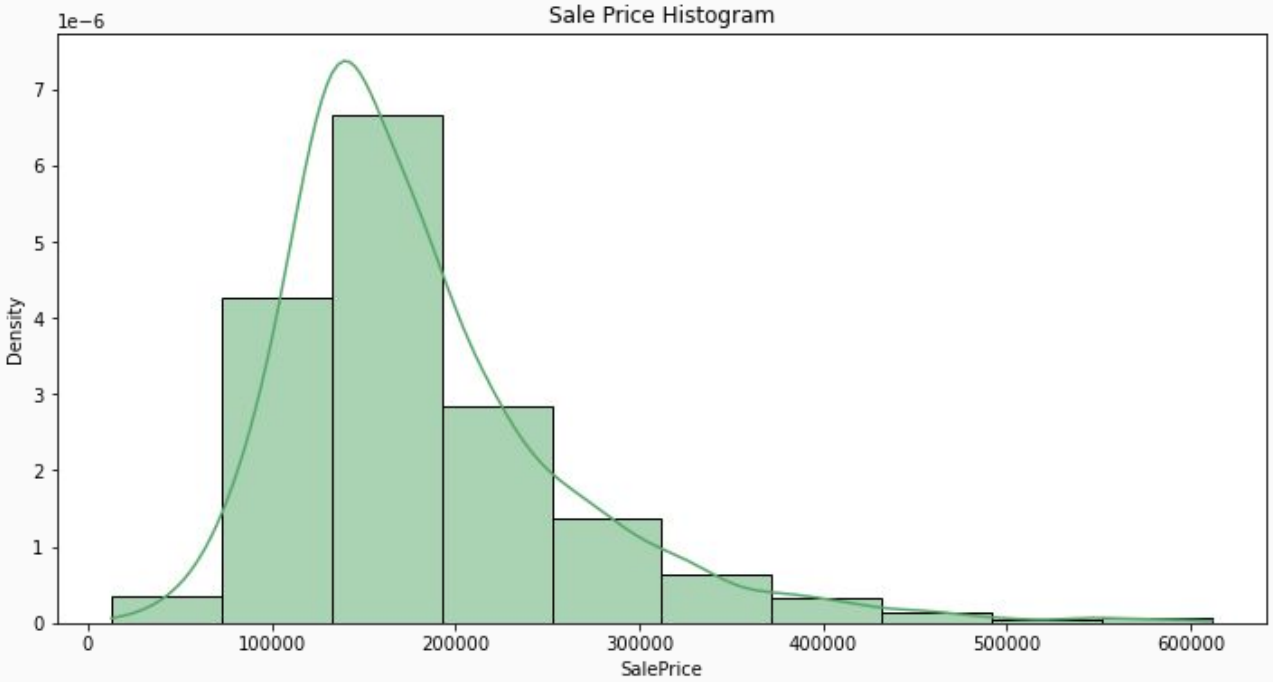
Data Transformation (Categorical Features)

Nominal Features

- We noticed ‘Neighbourhood’ feature may contribute to the overfit during modeling.
- To reduce the chance of overfitting, we ranked the feature according to the median price, 3 being Excellent (>75% of the sale price), 2 being Good (50% to 75% of the sale price), 1 being Fair (25% to 50% of the sale price), 0 being Cheap (<25% of the sale price).

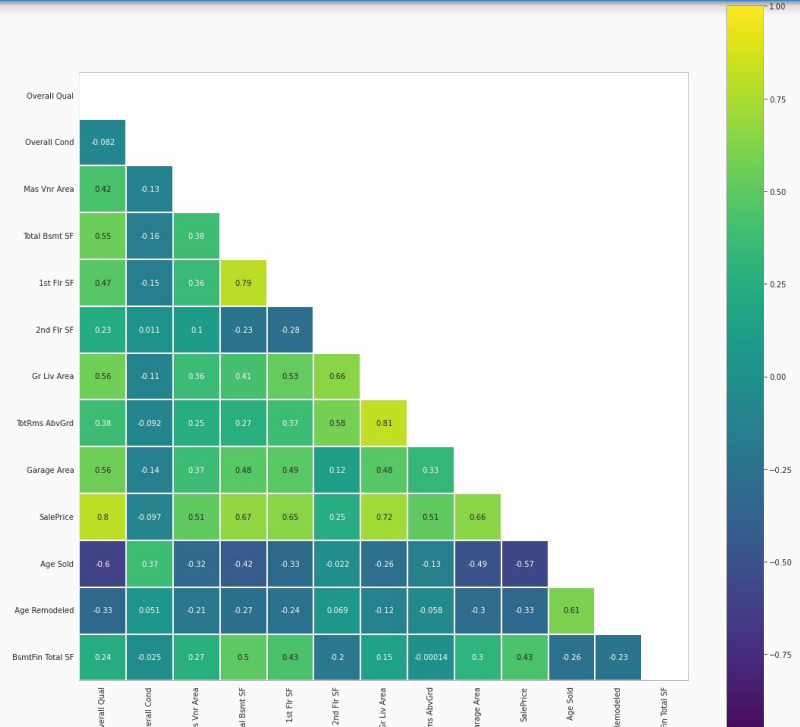


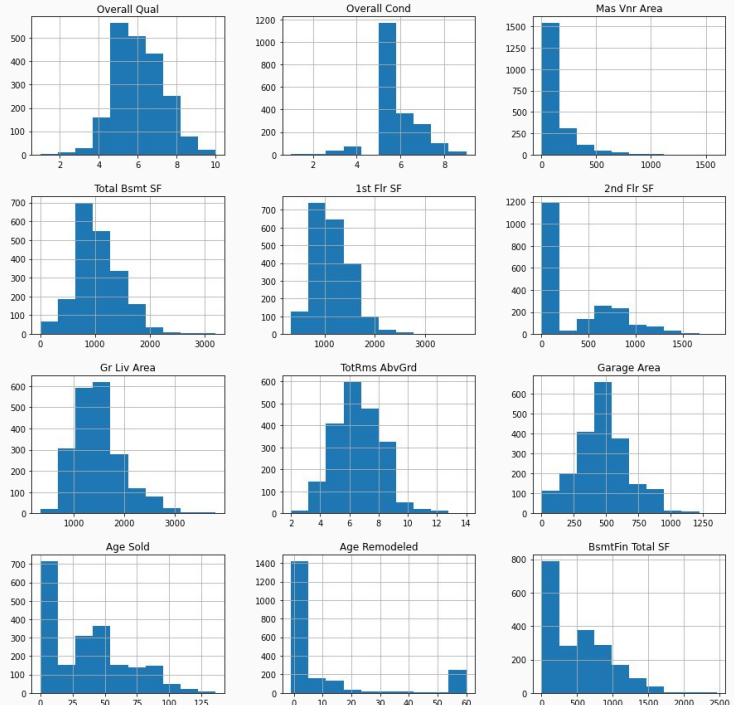
Data Transformation (Outliers)



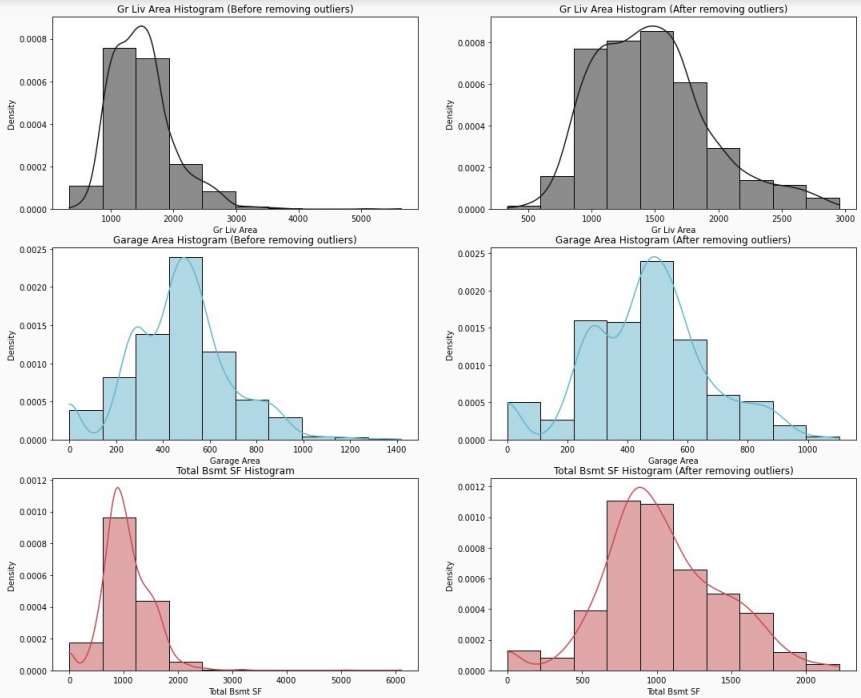
- The housing price is a right-skewed histogram, indicating that the data contained a lot of outliers with extremely high prices.

Data Transformation
(Top 5 Variable Correlated to Sale Price)





Data Transformation (Outliers)

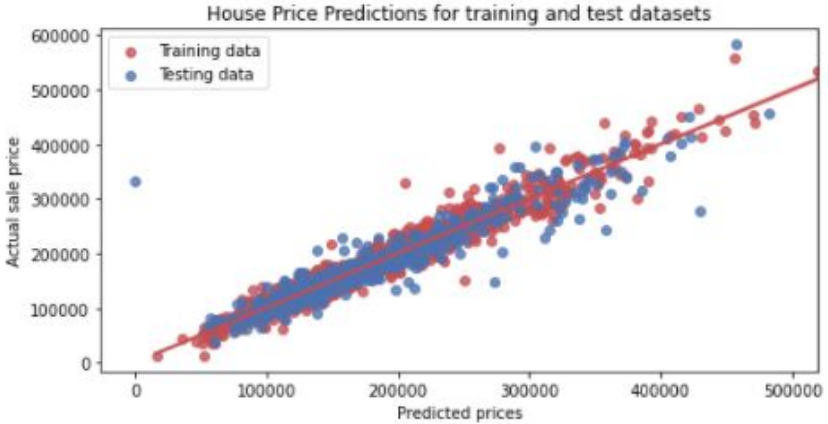


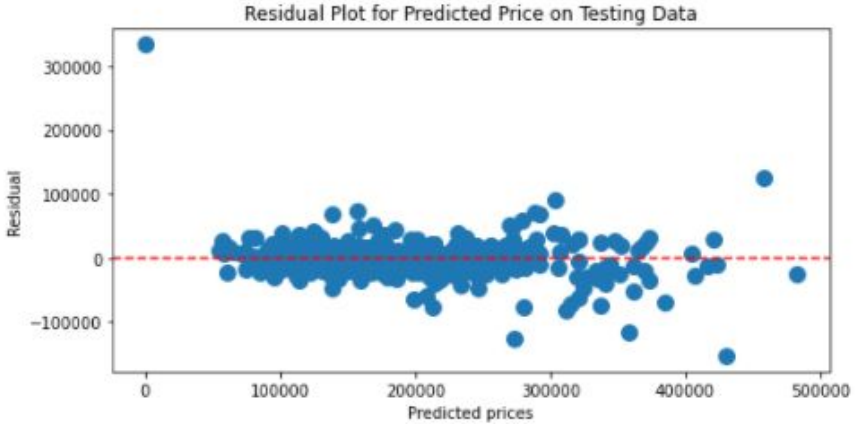
- Top 3 features that are highly correlated to the sale price. They are ‘Gr Liv Area’, ‘Garage Area’, ‘Total Bsmt SF. All 3 features also follow the pattern of a right-skewed histogram, indicating that the data contained a lot of outliers with extremely high values

- After removing the outliers with 3 STD away from the mean values, the histogram of the 3 features definitely appeared more symmetrical, with less extreme values.

BASE MODEL

```
R2 for train dataset: 0.9309609131910173
R2 for test dataset: -3.5047116753419224e+21
RMSE for train dataset: 0.10408964776545655
RMSE for test dataset: 23443445554.032436
```





- Obviously Extremely Overfitted Base Model
- Base Model Perform Worst than Null Model

Feature Engineering

Narrowing Down Independent Variables

Feature Engineering
(Narrowing down independent variables)

Method 1: Feature Selection based on corr() to the sales price

- 18 features with a minimum threshold of +/- 0.5 correlation score with the sale price

Method 2: Feature Selection based on Lasso regression

- 50 unique features extracted with minimum threshold coeff of +/- 0.005 score with the dependent variable

Method 3: Hybrid - Identifying Features from Lasso regression & Corr()

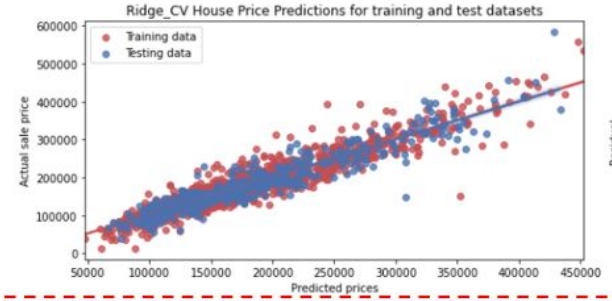
- 56 unique features extracted from the hybrid method

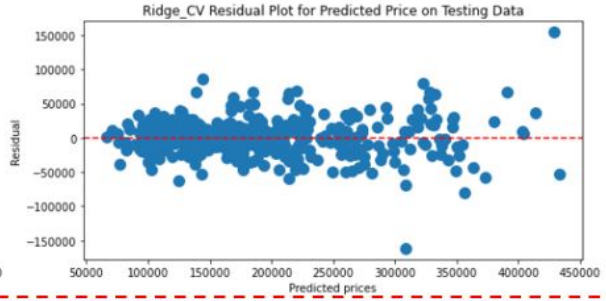
Removing collinear features

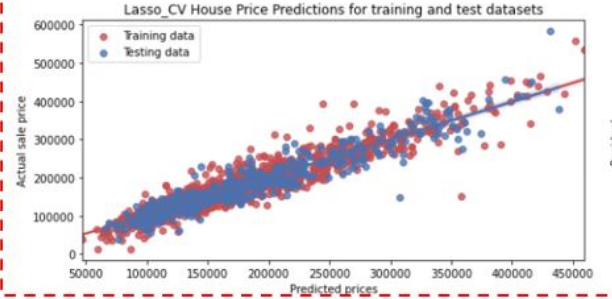
- For this analysis we dropped 'Garage Cars', 'Garage Yr Blt','1st Flr SF','Exter Qual Rank_3
- 52 unique features for the final model

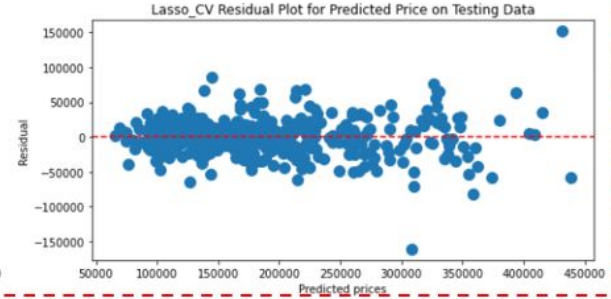
Choosing our Final Model

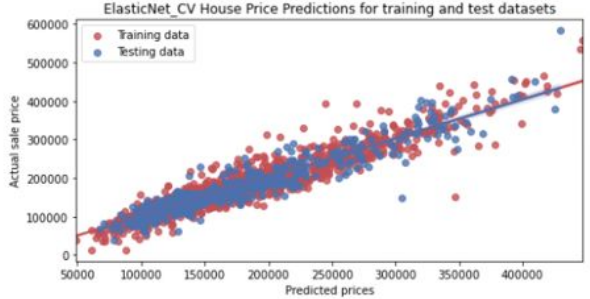
RidgeCV, LassoCV, ElasticNetCV

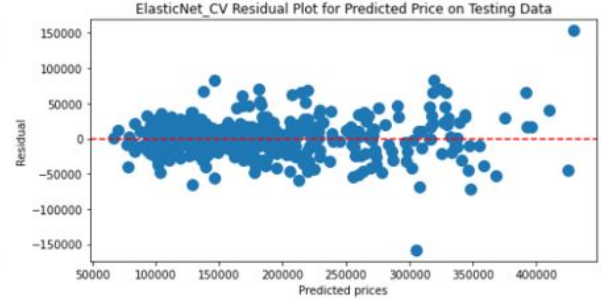












Models Statistical Summary

```
Ridge_CV Statistic Summary
R2 for train dataset: 0.9140692591330569
R2 for test dataset: 0.9127692361214845
RMSE for train dataset: 0.11612729176257494
RMSE for test dataset: 0.11695805347344425

Lasso_CV Statistic Summary
R2 for train dataset: 0.9137800352844344
R2 for test dataset: 0.9148584349486778
RMSE for train dataset: 0.11632255697242828
RMSE for test dataset: 0.1155489779533173

ElasticNet_CV Statistic Summary
R2 for train dataset: 0.9106393888942106
R2 for test dataset: 0.9142414881894058
RMSE for train dataset: 0.1184221890542645
RMSE for test dataset: 0.115966863703299
```

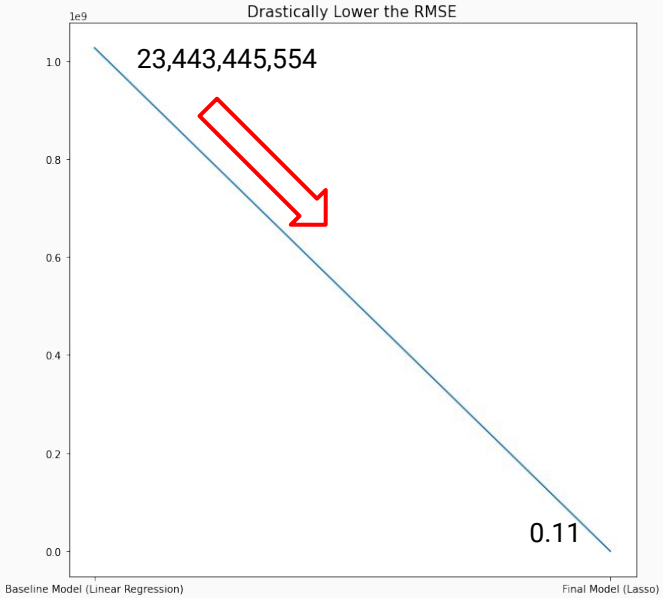
Conclusion & Recommendations

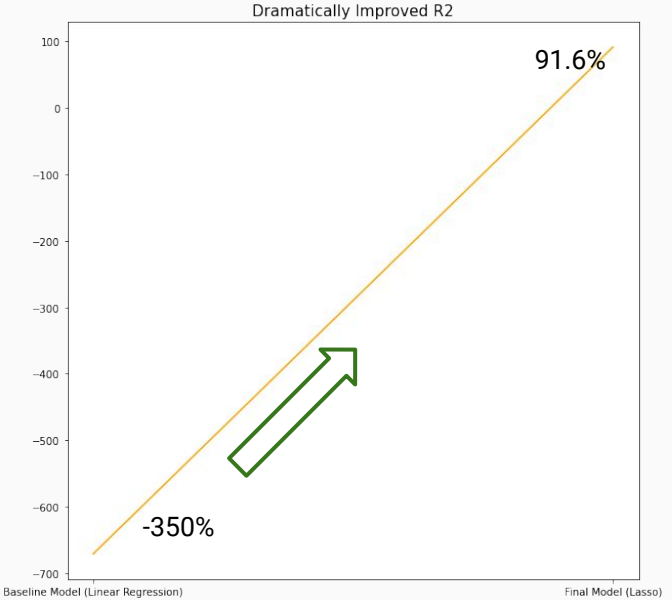
(Significant Findings and Further Improvements)

Conclusion
(Significant Findings)

We have successfully been able to develop a lassoCV model for Propnex to predict the housing price in Ames.

Using our base model as the baseline, we have make significant improvement in terms of metrics (R2 and RMSE) for our testing dataset with regularization and features selections.





Conclusion
(Significant Findings)

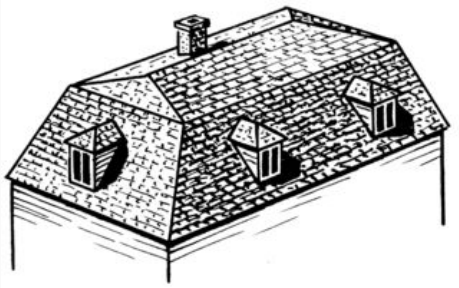
From the lasso coefficient and statistical significant analysis, we **observed having an excellent rating for exterior quality adds the most value to the property (**$41,058.73**)**. A **strong preference for brick common on the exterior covering of the house is also observed, adding** $15,360.11 to the house value. Aside from the exterior quality of the house, **having an excellent rank for Bsmt Quality helped to add** $27,702.84 to the house value

	features	coef
46	Exter Qual Rank_5	41058.729546
9	Bsmt Qual Rank_5	27702.839435
50	Sale Type_New	15658.648256
6	Exterior 1st_BrkFace	15360.117692
26	Condition 1_PosA	15174.310375
24	Sale Type_Oth	14222.120329
39	Functional Rank_7	12033.150953
29	Land Contour_HLS	10502.997671
41	Overall Qual	9093.237029
49	Condition 1_Norm	8759.179585

Conclusion
(Significant Findings)

Conversely, having a Mansard Roof Style depreciates a home's value by $28,555.19. Thus, it is best to remove/revamp this feature if it makes financial sense before selling the house.

	features	coef
12	Functional Rank_1	-54155.893819
0	Roof Style_Mansard	-28555.192622
23	Overall Cond_3	-17559.524615
31	Overall Cond_2	-11626.499347
48	Heating_Grav	-10365.582770
14	Garage Cond Rank_1	-9845.427103
36	Overall Cond_4	-9060.287303
43	Overall Cond_5	-6246.182933
45	Garage Cond Rank_3	-6064.983626
47	Garage Cond Rank_2	-4805.919195



Recommendations
(Further Improvements for Model)

- Use dimension reduction algorithms such as Principal Component Analysis (PCA) to take care of multi collinear features instead.
- Utilize other ML methods such as gradient boosting to check for improvement of performance.
- Considering reducing the number of columns (original 17) in nominal 'Exterior 1' feature by studying the correlation it has with the house sale and converting them to ordinal feature instead.
- Check for outliers among the 52 selected features for the final model.

Thank You!

Presenters:

Ivan Cho

Ng Jia Sheng

Joel Tan

GA Lead Instructor:

Divya

GA Asst. Instructors:

Ben and Shao Quan