PREDICTIVE MODELLING: AMES HOUSING DATA

AGENDA

- Problem Statement
- Data Cleaning
- Exploratory Data Analysis(EDA)
- Pre-processing/Feature Engineering
- Model
- Evaluation(Prediction & Inference)
- Recommendation

PROBLEM STATEMENT

- To create a regression model based on Ames housing data to predict sales price
- Analyse data to identify top features of a house that help or hurt sale prices
- U.S. homeowners spend more than $400 billion each year on remodeling
- 56 percent return of their renovation investments
- Make recommendations to increase value of houses to:
  - House owners
  - Investors and Real Estate Companies

Sources:
"2019 Remodeling Impact Report" - https://remodelingdoneright.nari.org/RemodelingDoneRight/media/Assets/Remodeling-Impact-Survey-2019.pdf
"Home renovation: highest ROI remodeling projects" - https://themortgagereports.com/39910/home-renovation-highest-roi-remodeling-projects#:~:text=The%20Remodeling%20report%20found%20that,the%20average%20was%2064%20percent.

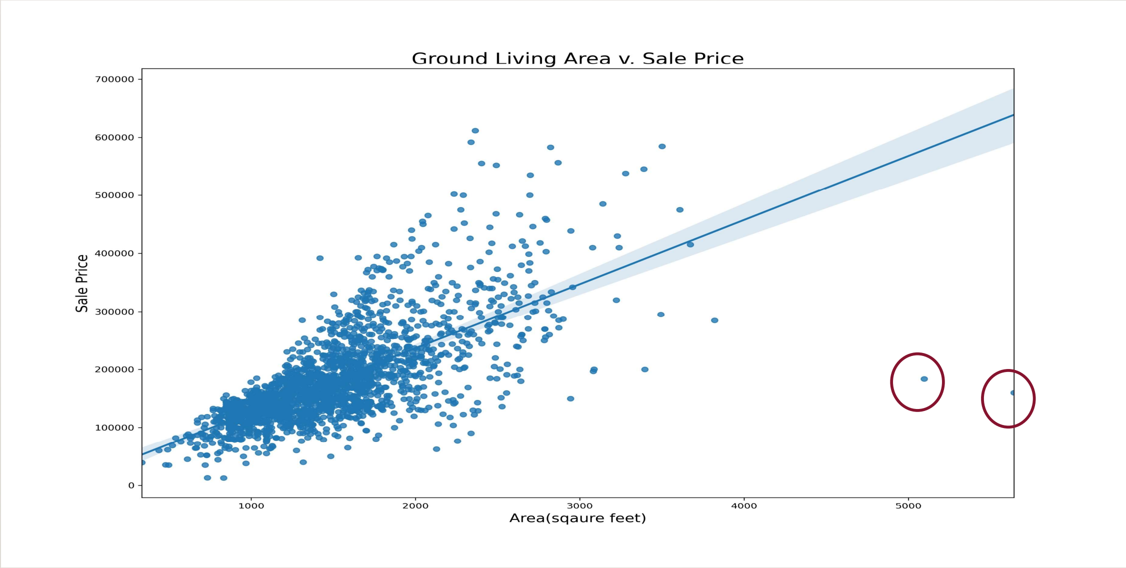
SOURCE: AMES HOUSING DATA

- Data of about 2050 homes sold in Ames, Iowa between 2006 to 2010
- Target: Sale Price
- About 80 features of a house that affects sale price:
  - Numerical and Categorical
    - Nominal – 23
    - Ordinal – 23
    - Discrete – 14
    - Continuous – 20
- Examples
  - Neighbourhood(Nominal)
  - Overall Quality(Ordinal)
  - Fireplace(Discrete)
  - Ground Living Area(Continuous)

DATA CLEANING

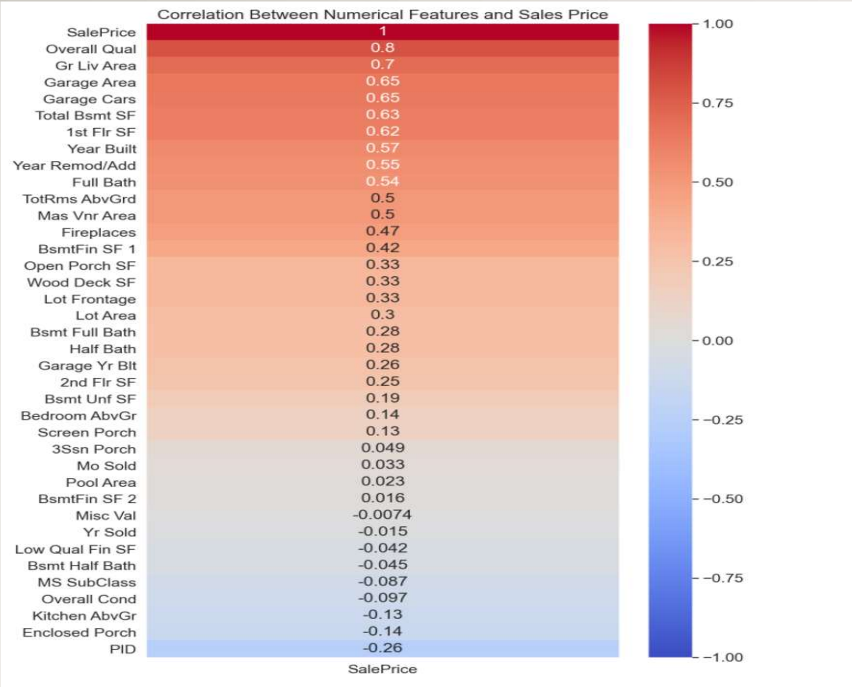
- Fill in null values
  - Minimise data lost
  - Data Description
  - Used "None" or "0" for features that are non existent
  - Mode, mean or median for truly missing data

- Outliers
  - Removed outliers that may distort model prediction

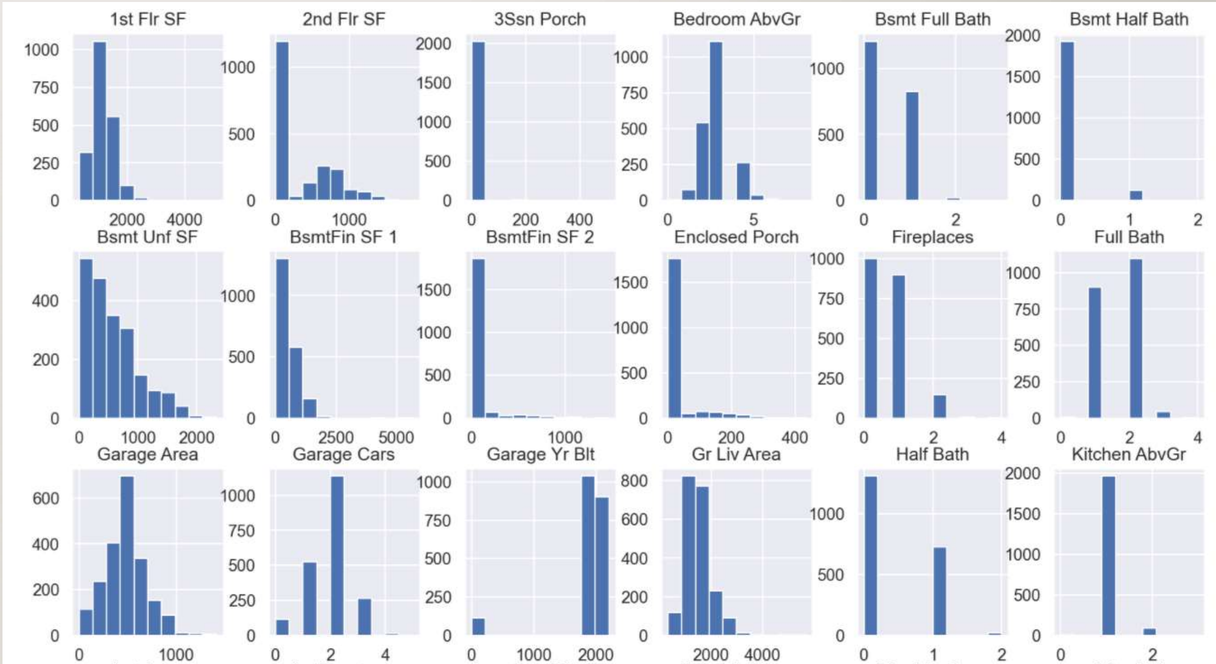


Exploratory Data Analysis(EDA)

- Heatmap of correlation of features with target sale price



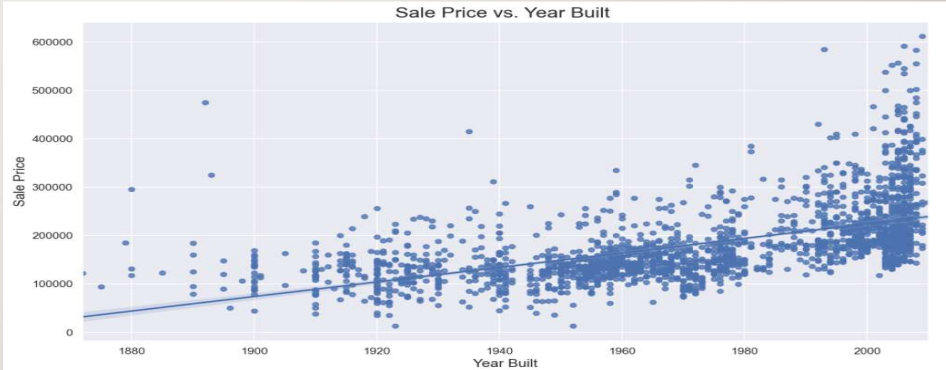
- Histograms of features

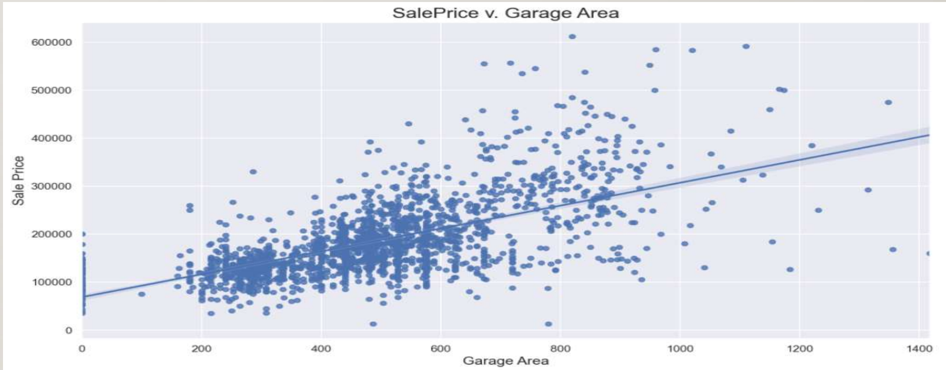


Exploratory Data Analysis(EDA)

- Box plots

- Scatter plots



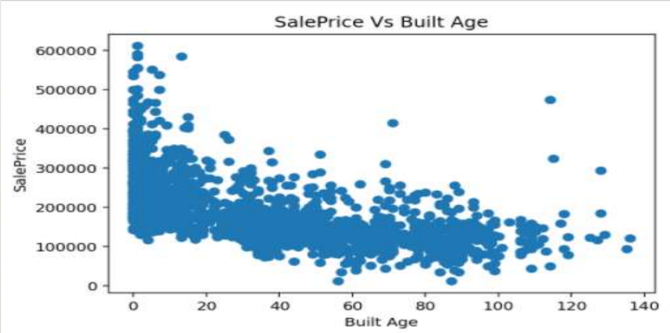


Pre-processing/Feature Engineering

- Quantify Categorical Data
  - Nominal
    - Dummy encoding for column features
  - Ordinal
    - Map numerical values in order of quality

Kitchen quality	Value	Mapped Value
Excellent	Ex	5
Good	Gd	4
Typical/Average	TA	3
Fair	Fa	2
Poor	Po	1

- Feature Engineering
  - Creating our own features-Property Age



- Scaling of features
  - Standard Scaler to rescale data that can reduce the effect of features with different units of measurement

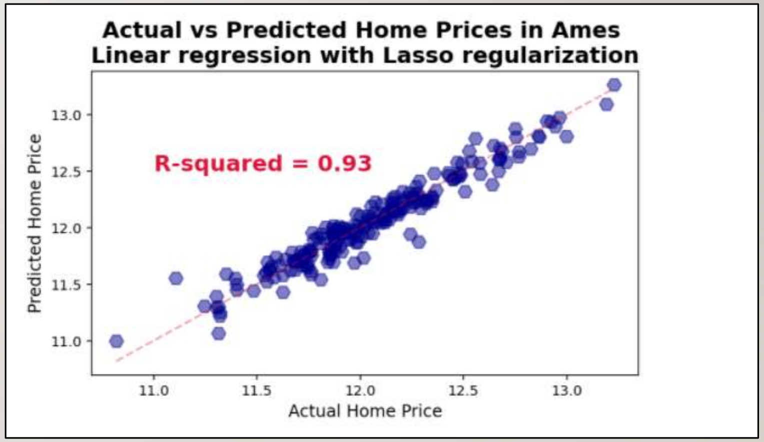
MODELING

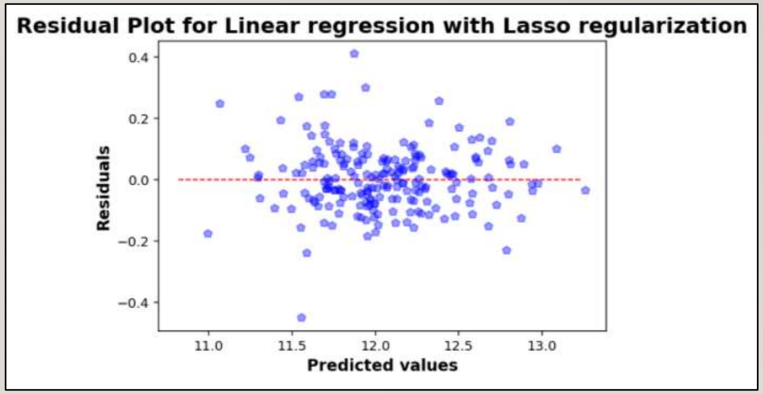
MODEL	OPTIMAL VALUE	SCORES	
	Hyperparameters	R-Squared	RMSE
Simple Linear Regression(LR)	-	-1.82	55294755457.6
LR with Lasso Regularization	**Alpha=527.01105**	**0.93**	**21946**
LR with Ridge Regularization	Alpha=182.51834	0.89	21456
LR with Elastic Net Regularization	Alpha=514.41383 L1_ratio=0.5	0.90	21920

The selected model to predict sale price in test data is **LR with Lasso regularization**

EVALUATION OF LASSO MODEL ON SALE PRICE

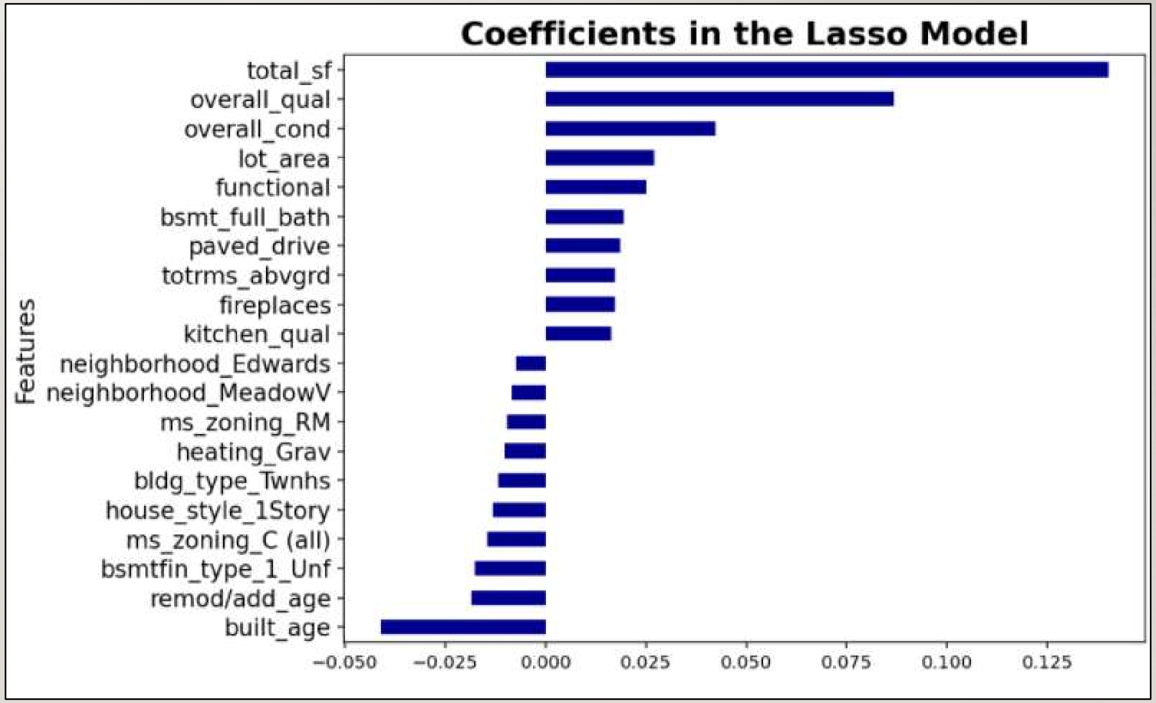
- A strong model with R^2 score of 0.93 on unseen data
- There are very few errors and no outliners to be highlighted
- No evidence of overfitting since R^2 score on training data is slightly low 0.92
- The model predicted well on unseen data





INFERENTIAL DETAILS FROM LASSO MODEL

- Lasso model helps in inferential analysis for feature selection
- Shows features that positively and negatively affects the price of house



CONCLUSION AND RECOMMENDATION

(FOR HOUSEOWNERS , INVESTORS)

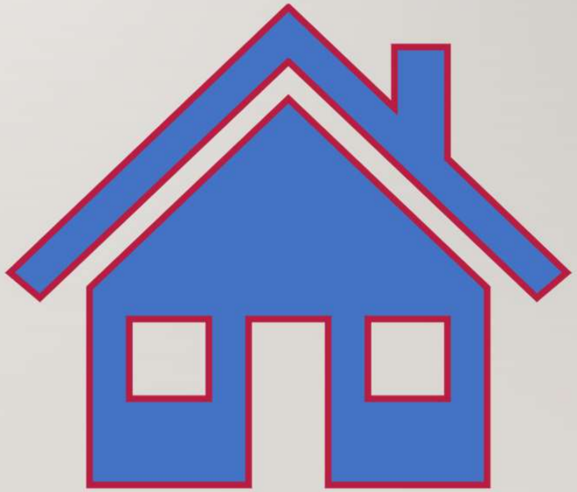
- Accurately predicted the sale price of the houses
- Helped in identifying the significant features that helps and hurts the price that guided our recommendations.

Features that <u>helps</u> increase sale price	**Features that <u>hurt</u> sale price**
Size of the house	Age of the house
Overall material and finish of the house	Remodel Date
Overall condition of the house	Unfinished basement

FUTURE RECOMMENDATIONS TO INCREASE OUR USER BASE

Improve accuracy of the model by collecting more data.

Expand scope of the model to predict price for other cities.



QUESTIONS