Ames Housing Price Prediction

Group 2
Wan Xian - Janet - Soon Poh

AGENDA

1 Introduction

- Who are we?
- Problem statement

2 Exploratory Data Analysis

- Overview of Data
- Methodology
- Treatment of Missing Data/Outliers
- Top 5 Key features
- Multicollinearity

3 Model Preparation

- Data Pre-processing
- Feature Engineering
- Scaling

4 Model Evaluation

- Model Performance

5 Conclusion

- Recommendation
- Limitations/Opportunities
- What's next?

PROBLEM STATEMENT

Client = A property agency (IowaGuru) based in Ames, Iowa

To develop a suitable regression model capable of predicting the prices of property in Ames accurately

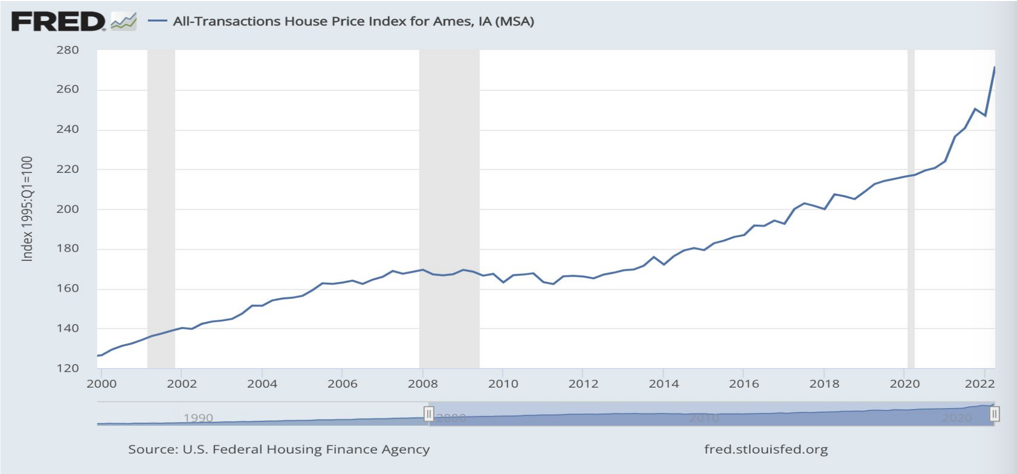
To figure out the key features of the property that are strong predictors to the sale price

BACKGROUND

- Housing market in Ames is highly competitive in recent years
  - Exponential surge in prices driven by rising demand

- IowaGuru wants to predict prices based on the property's features only
  - Dataset was taken from property sales data for 2006 to 2010 (Normal volatility in prices)

- The model will be evaluated based on regression model metrics & scores in Kaggle



OUR DATA

Upgrade Now
Features
Location
Neighborhood
Zoning
Street/Alley
Landscape
Land Contour
Land Slope
Sale Details
Sale Price
Sale Type
Month/Year of sale
Building
Overall
Type/Class
Functionality
Age
Roof
Style
Material
Foundation
Exterior
Covering
Masonry Veneer
Condition
Above Grade
Area
Count of rooms
Count of Bathrooms
Lot
Basement
Exposure to the surface
Count of Bathrooms
Finishing
Heating
Fireplace
Central Air Conditioning
Misc Features
Value
Garage
Condition
Finishing
Quality
Type
Age
Size
Utilities
Electricity
Fence Quality
Porch
Pool
Quality
Area

METHODOLOGY

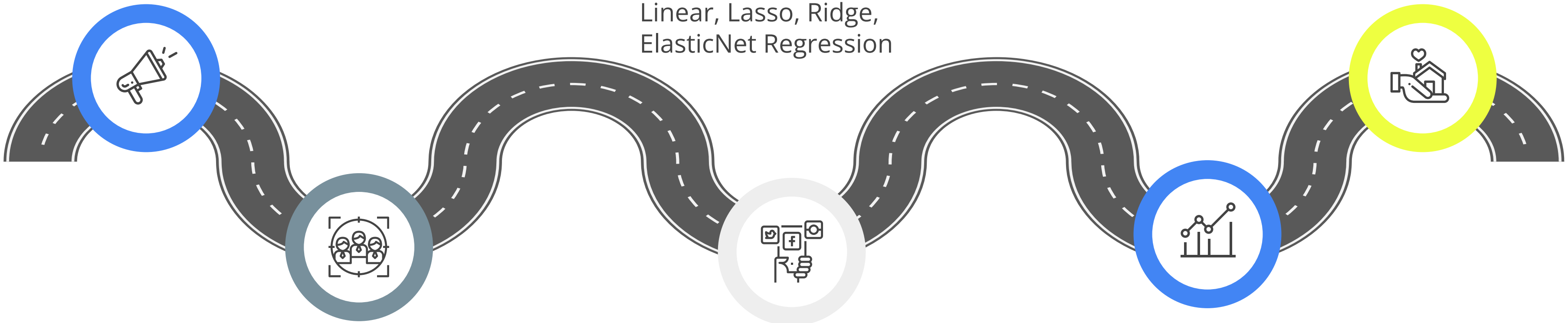
DATA CLEANING & ENCODING

Identify & address null value
Categorize Features to Ordinal, Nominal & Numeric
Encode Ordinal & Nominal Features

BASELINE MODEL

Linear, Lasso, Ridge, ElasticNet Regression

CONCLUSION & RECOMMENDATIONS



VISUALISATION & DATA PRE-PROCESSING

Plot graphs to identify & address outliers
Train-Test Split
Standard Scaler

FINAL MODEL

Feature Engineering - Polynomial
Lasso & Ridge Regression

FEATURES CATEGORY

NOMINAL FEATURE

Neighborhood, House Style, Sale Type, etc

These variables are to be encoded using One Hot Encoding [0, 1]

1

NUMERIC FEATURE

Year Built, Bathrooms, Floor SF, etc

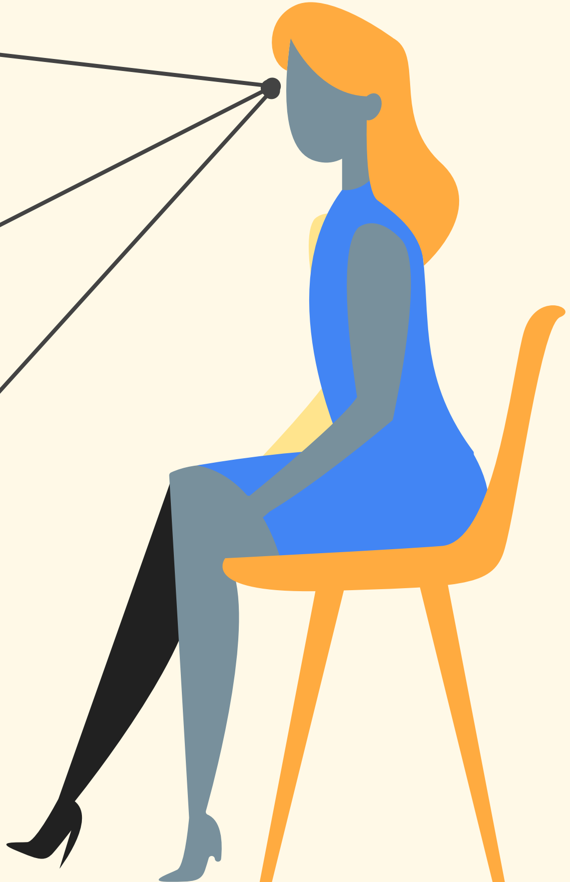
2

ORDINAL FEATURE

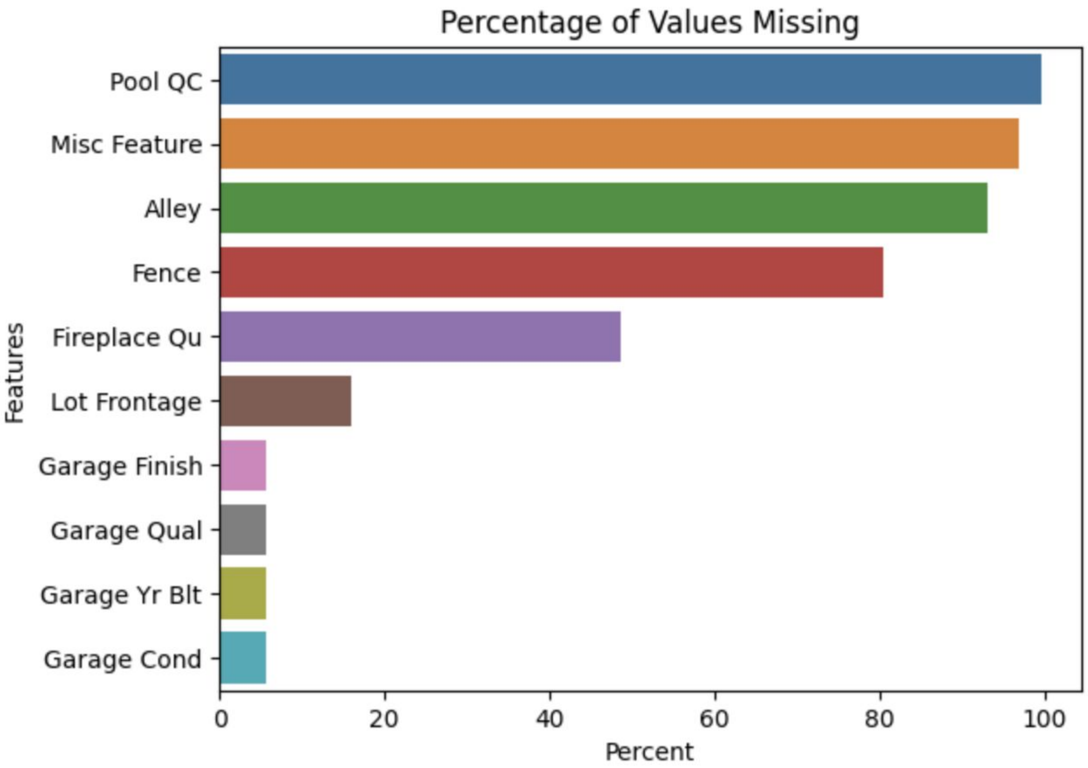
External Qualities, Basement Conditions, etc.

These variables are to be encoded as per ordinal manner stated in data dictionary by integer mapping

3



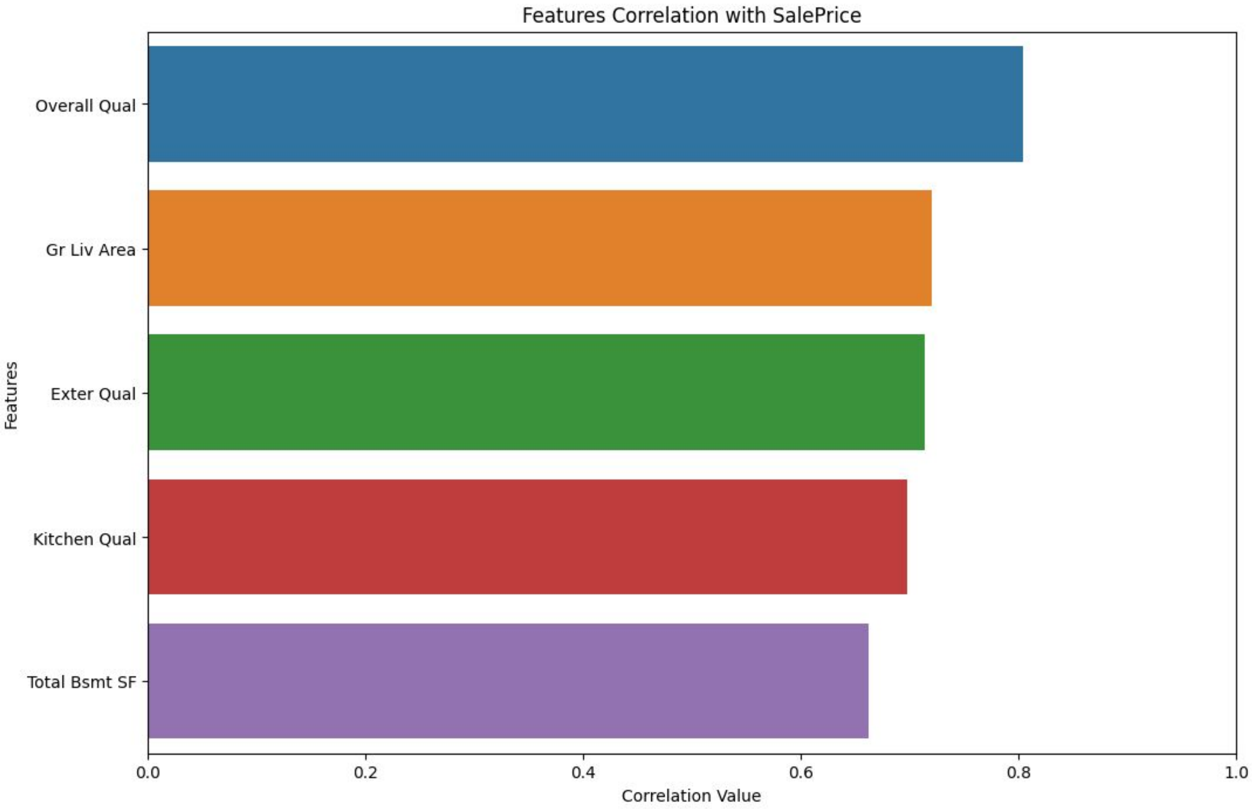
MISSING VALUES



These values should be “NA” / 0.0, not genuine missing

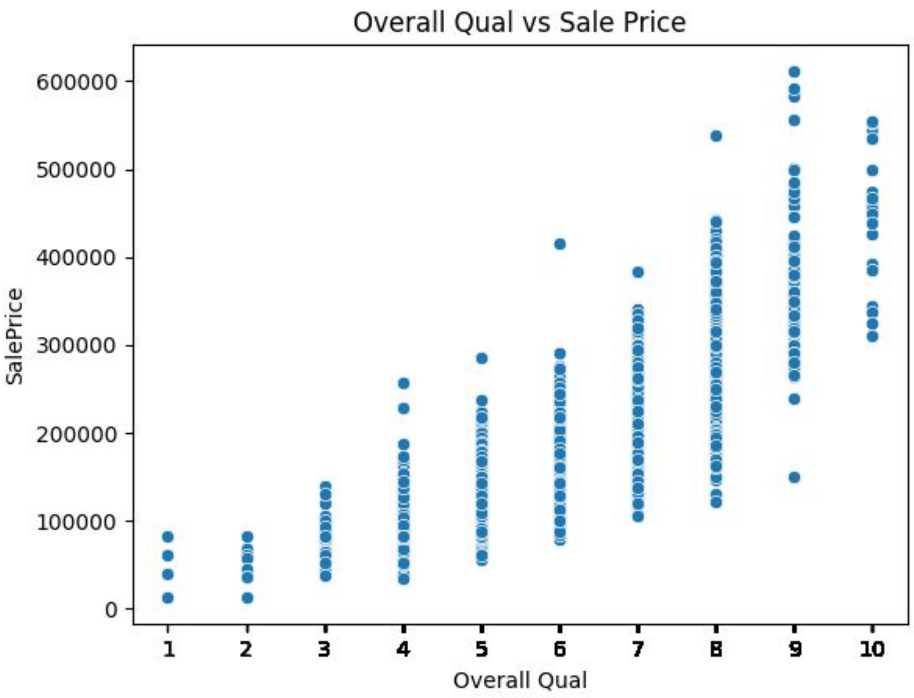
These features do not exist in the property

TOP 5 FEATURES CORRELATED WITH SALE PRICE



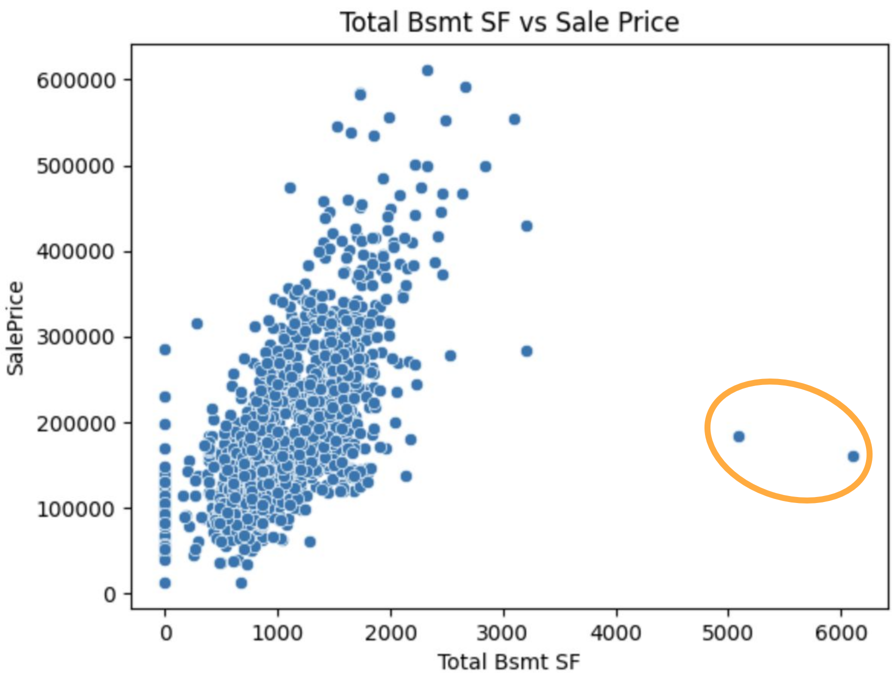
- Overall Quality (0.8)
- Exterior material quality (0.71)
- Above Grade Living Area (0.7)
- Kitchen Quality (0.7)
- Total Basement Area (0.65)

IDENTIFYING & REMOVING OUTLIERS



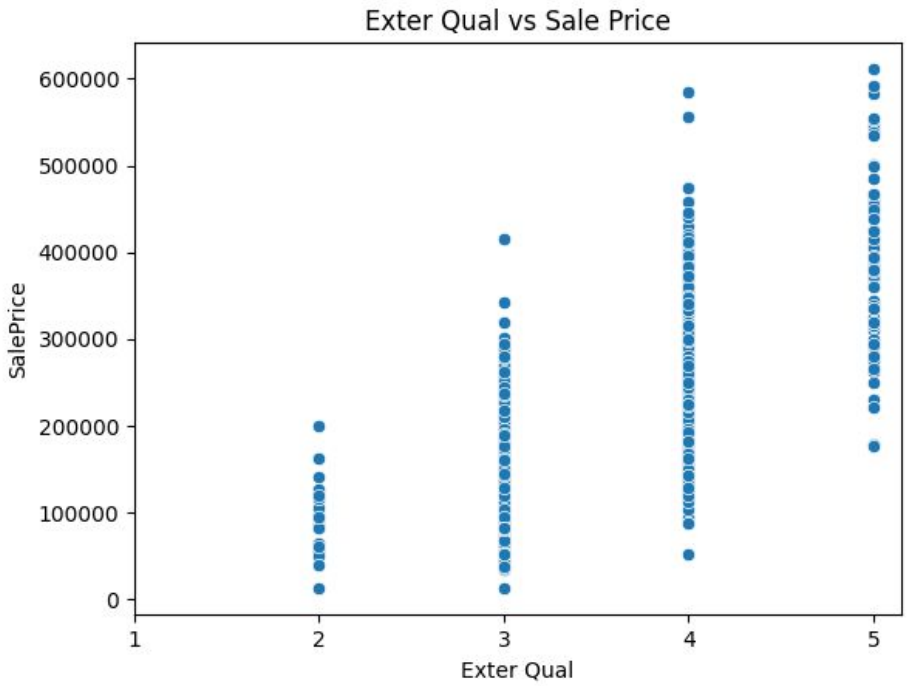
- No outliers in Overall Quality data series
- 1 outlier in Kitchen Quality data series & to be dropped

IDENTIFYING & REMOVING OUTLIERS



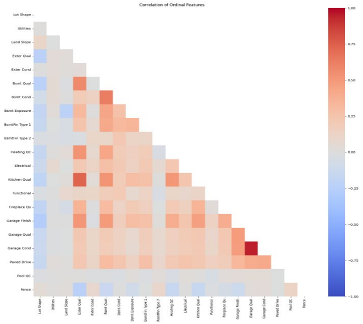
- 2 outliers spotted in Ground Living Area & Total Bsmt SF data series
- To be dropped from analysis

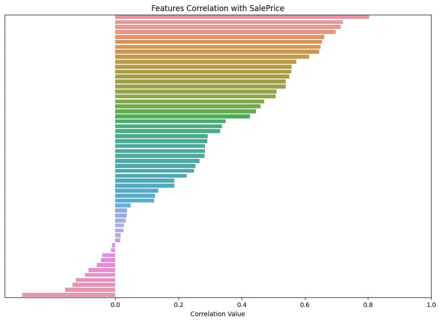
IDENTIFYING & REMOVING OUTLIERS



- No outliers for both Exterior Material Quality data series

FEATURE CORRELATION





- Identify features that has high correlation with each other (> 0.75)
- Drop features that has lower correlation with Sale Price

Feature 1 (Corr with Sale Price)	**Feature 2 (Corr with Sale Price)**	**Feature 1 & 2 Corr**
Garage Qual (0.28)	**Garage Cond (0.26)**	0.95
Garage Area (0.65)	**Garage Cars (0.64)**	0.89
Yr Blt (0.57)	**Garage Yr Blt (0.55)**	0.86
Total Bsmt SF (0.66)	**1st Flr SF (0.64)**	0.81
Gr Liv Area (0.72)	**TotRms AbvGrd (0.51)**	0.81

*In red: Feature has lower correlation with Sale Price -> Dropped

DATA PRE-PROCESSING

TRAIN-TEST SPLIT

75% of data for Training
25% of data for Testing

TRANSFORMATIONS

Standardize features using
sklearn StandardScaler

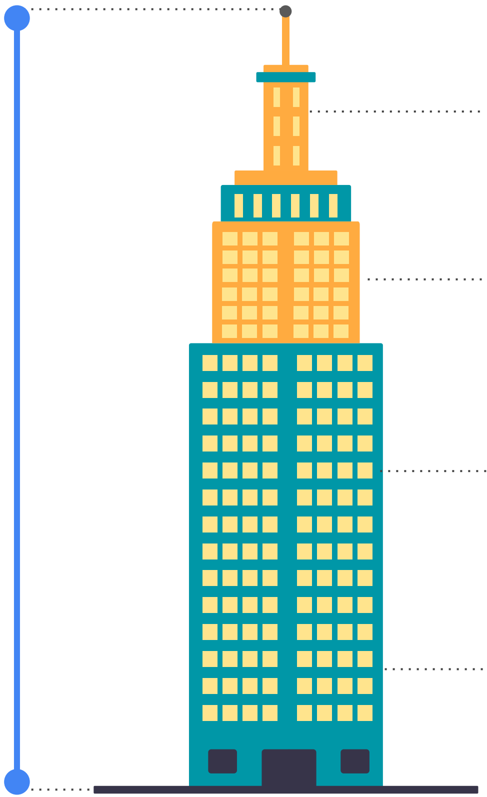
BASELINE MODEL

Linear Regression,
Lasso, Ridge
& ElasticNet

BASELINE MODEL PERFORMANCE

	REGRESSION MODEL	TRAIN SCORE	TEST SCORE	CROSS-VAL (R2 SCORE)	RMSE SCORE	REMARKS
BASELINE MODEL	LINEAR REGRESSION	0.936	-9.811E+23	-7.681E+22	3.62E+15	Fail
BASELINE MODEL	LASSO	0.929	0.903	0.908	24167	Selected for Model Tuning
BASELINE MODEL	RIDGE	0.931	0.898	0.906	24718	Selected for Model Tuning
BASELINE MODEL	ELASTICNET	0.929	0.903	0.908	24163	Similar to Lasso (l1 ratio = 1)

FEATURE ENGINEERING



FEATURE ENGINEERING

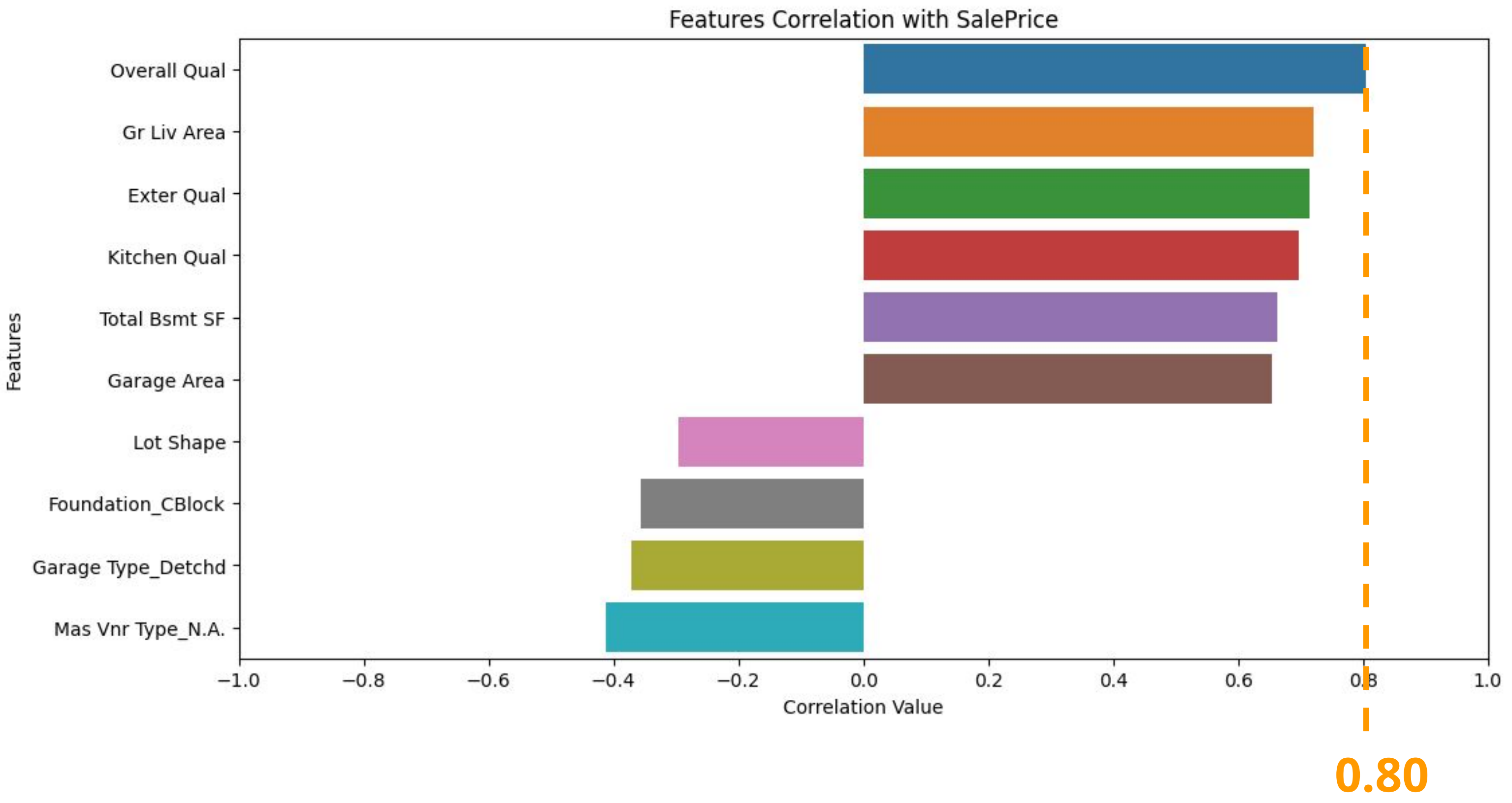
HOUSE AGE
Current Year - Year Built

BATHROOMS
Full Bath + 0.5*Half Bath +
Basement Full Bath + 0.5*Basement Half Bath

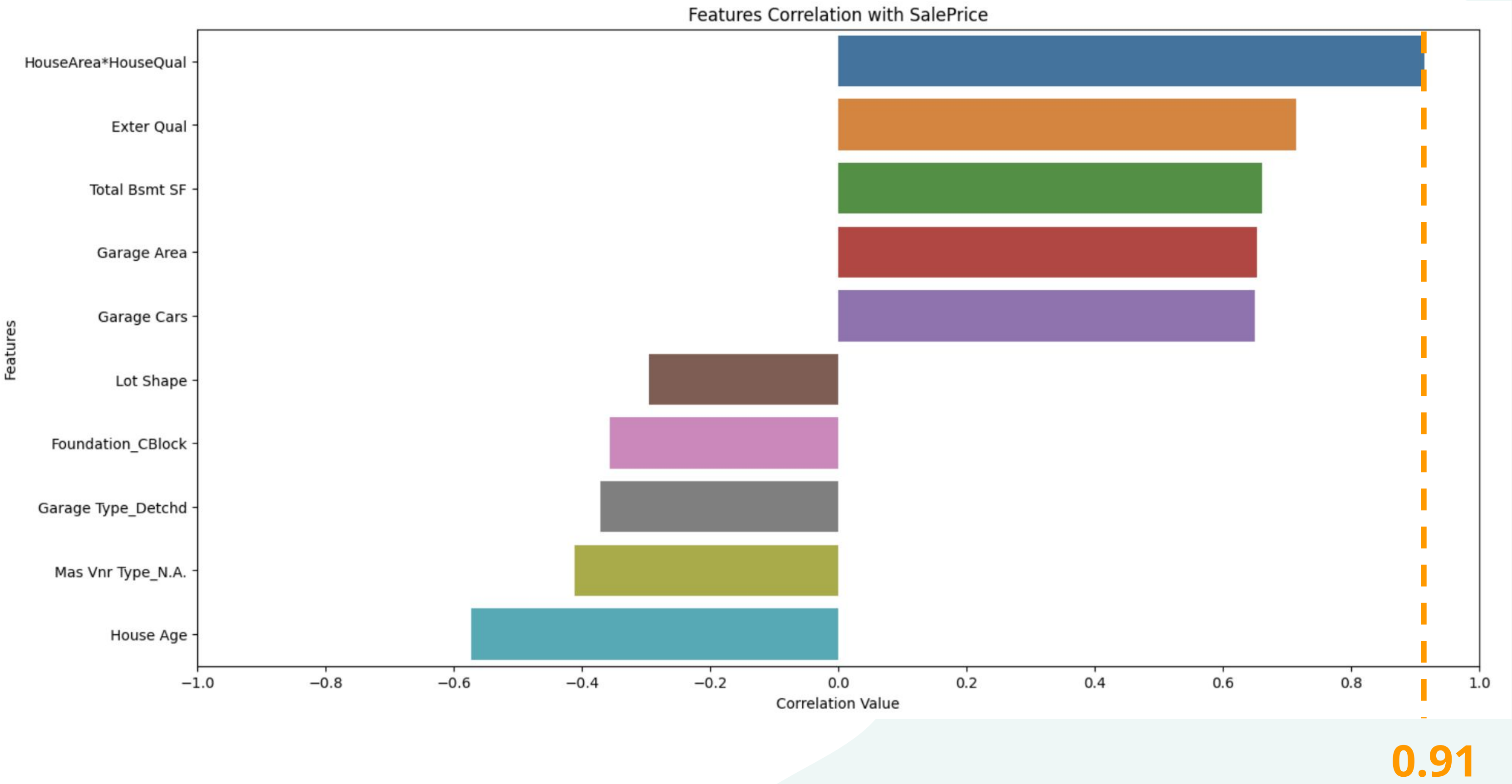
HOUSE QUALITY
Overall Quality * Basement Quality * Kitchen Quality

HOUSE AREA*HOUSE QUALITY
Ground Living Area * House Quality

BEFORE FEATURE ENGINEERING



AFTER FEATURE ENGINEERING



FINAL MODEL PERFORMANCE

	REGRESSION MODEL	TRAIN SCORE	TEST SCORE	CROSS-VAL (R2 SCORE)	RMSE SCORE	REMARKS
BASELINE MODEL	LASSO	0.929	0.903	0.908	24167	-
BASELINE MODEL	RIDGE	0.931	0.898	0.906	24718	-
FINAL MODEL	LASSO	0.942	0.940	0.925	21990	Selected for Kaggle Submission
FINAL MODEL	RIDGE	0.946	0.913	0.920	22853	-

Submission and Description	Private Score	Public Score
kaggle_submission.csv a day ago Final Submission	20833.27812	22350.95249

KAGGLE LEADERBOARD (PUBLIC)

#	Team	Members	Score	Entries	Last	Code
1	CharlesRice		0.00000	3	2Y	
2	rhys		19309.24031	16	2Y	
3	Griffin		19333.48211	16	2Y	
4	weisja4		20286.14799	30	2Y	
5	Luke McKinley		20507.68177	11	2Y	
6	Stephanie Caress		20817.46675	11	2Y	
7	JulKel		21539.40770	3	2Y	
8	Marina Baker		21860.96015	39	2Y	
9	Jeong Huh		22182.01037	5	2Y	
9A	Group 2		22350.95249	10	1D	
10	Scott Armstrong		22764.22164	11	2Y	

PREDICTED VS ACTUAL PRICE

- Model are relatively accurate in predicting house prices under $500,000
- Only 12 houses (0.5%) are priced above $500,000 in the datasets
- Need more data for houses priced above $500,000 to improve model accuracy

TOP FEATURES AFFECTING SALE PRICE

POSITIVE CORRELATION

- Ground Living Area
- Overall Quality
- Kitchen Quality
- Basement Quality
- External Quality

HouseArea*HouseQual

NEGATIVE CORRELATION

- House Age
- Mas Veneer Type - N.A.
- Garage Type - Detached
- Foundation - CBlock
- Lot Shape

Conclusion

- Summary
- Limitation/ Opportunities
- What’s next?

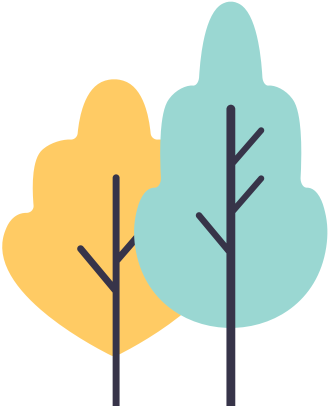
SUMMARY

BEST MODEL: LASSO REGRESSION MODEL

- RMSE Score: **21990**
- Kaggle Private Score: **20833**
- Performance increases for houses below $500,000 (that is 99.5% of houses in the dataset)

TOP FEATURES WITH BEST CORRELATION (> 0.6)

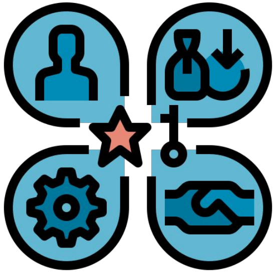
- House Area*House Quality (Combination of 4 features): **0.91**
- External Quality
- Total Bsmt SF
- Garage Area/ Garage Cars
- 1st Floor SF
- Number of Bathrooms



LIMITATION

Coverage

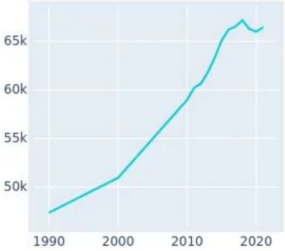
Time Frame



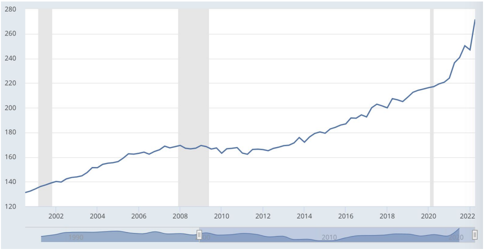
Other Factors

OPPORTUNITIES

Coverage

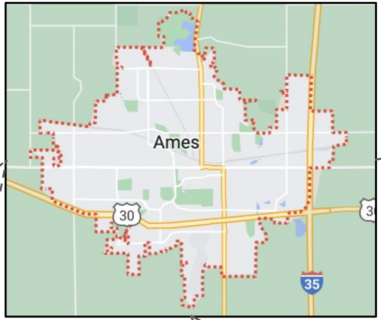


Growing Population



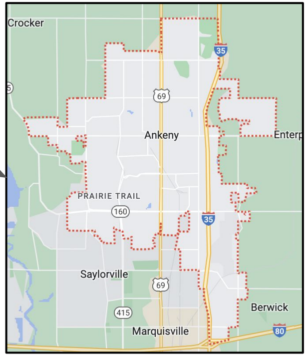
Growing Housing Prices

Ames



Iowa City

West Des Moines



Ankeny

OPPORTUNITIES

Time Frame

FORTUNE

FINANCE · ECONOMY

Prepare for a 'long and ugly' recession, says Dr. Doom, the economist who predicted the 2008 crash

BY TRISTAN BOVE

September 22, 2022 at 12:56 AM GMT+8

FINANCE · HOUSING

The U.S. housing market downturn will be worse in 2023, forecasts Goldman Sachs

BY LANCE LAMBERT

August 31, 2022 at 5:10 PM GMT+8

INSIDER

US home prices could plunge 20% by next summer as a housing recession kicks in, a top economist says

Theron Mohamed Sep 23, 2022, 5:51 PM

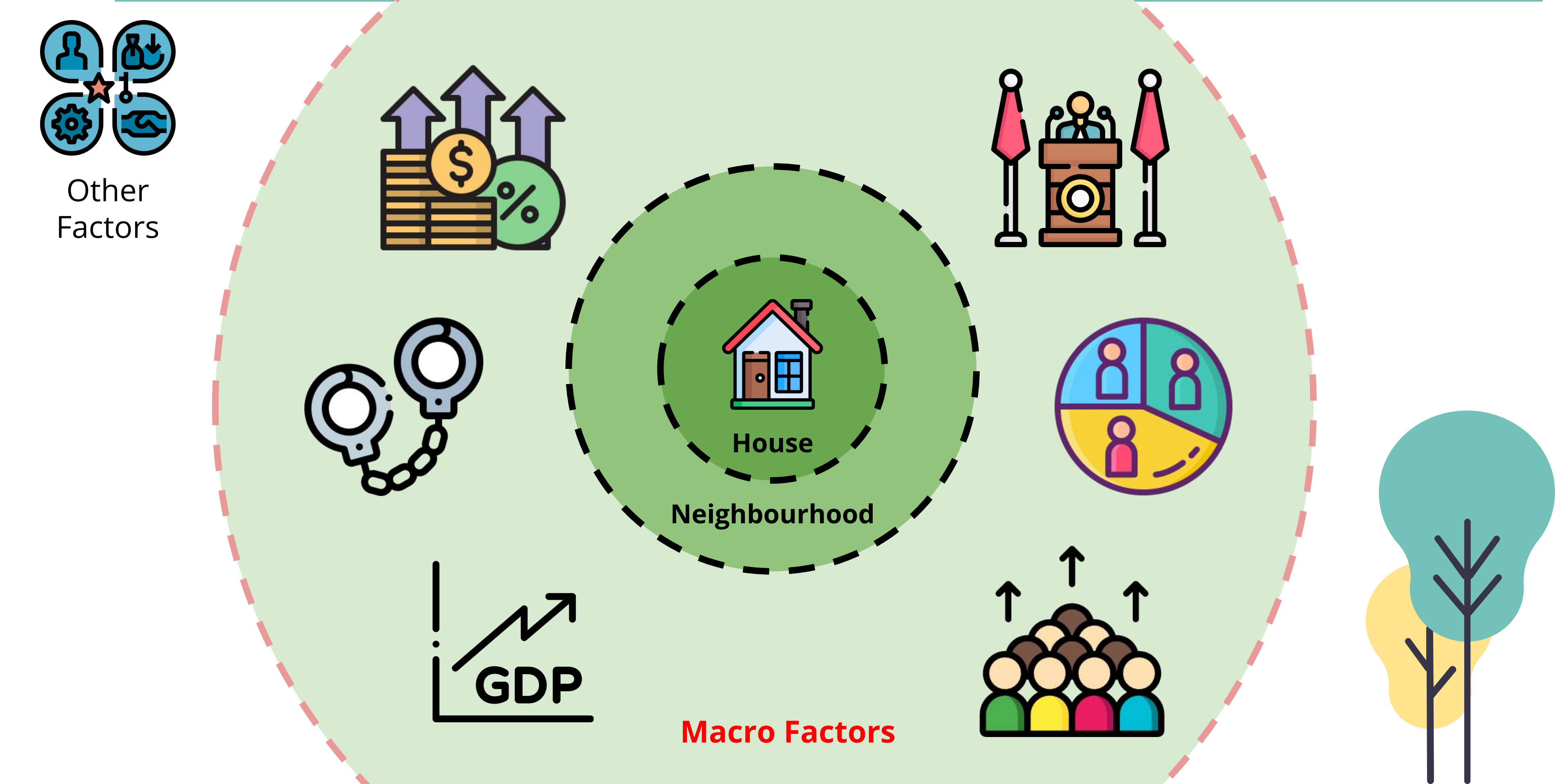
CNBC

ECONOMY

Danger ahead: The U.S. economy has yet to face its biggest recession challenge

PUBLISHED FRI, AUG 5 2022•3:41 PM EDT | UPDATED FRI, AUG 19 2022•8:58 PM EDT

OPPORTUNITIES



WHAT'S NEXT?

Improve Our Current Model

Introduce a working API for closed BETA

Instantiate a mobile application


THANK YOU!