Become a Kaggle Master-HW1

Team ZYMAA

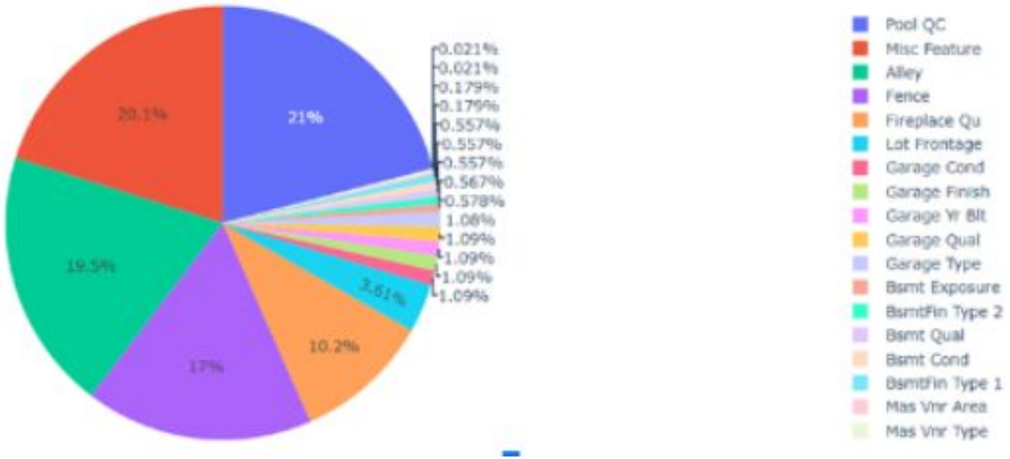
AYARI Mohamed Aziz
YBEGGAZENE Zakaria

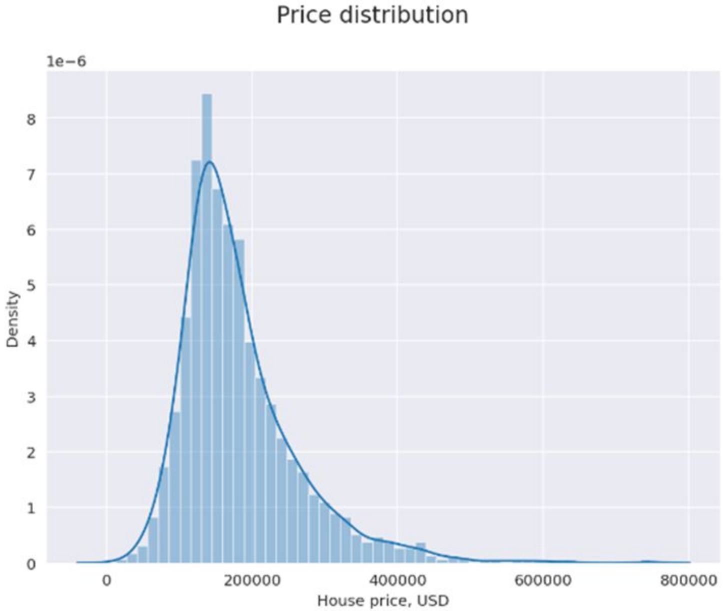
PLAN

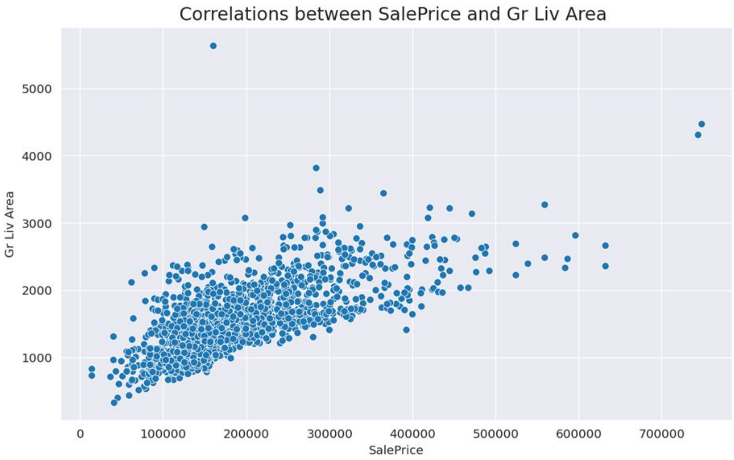
1. Introduction
2. Dataviz
3. Pre-processing and Feature engineering
4. Tested Models
5. Results
6. Conclusion

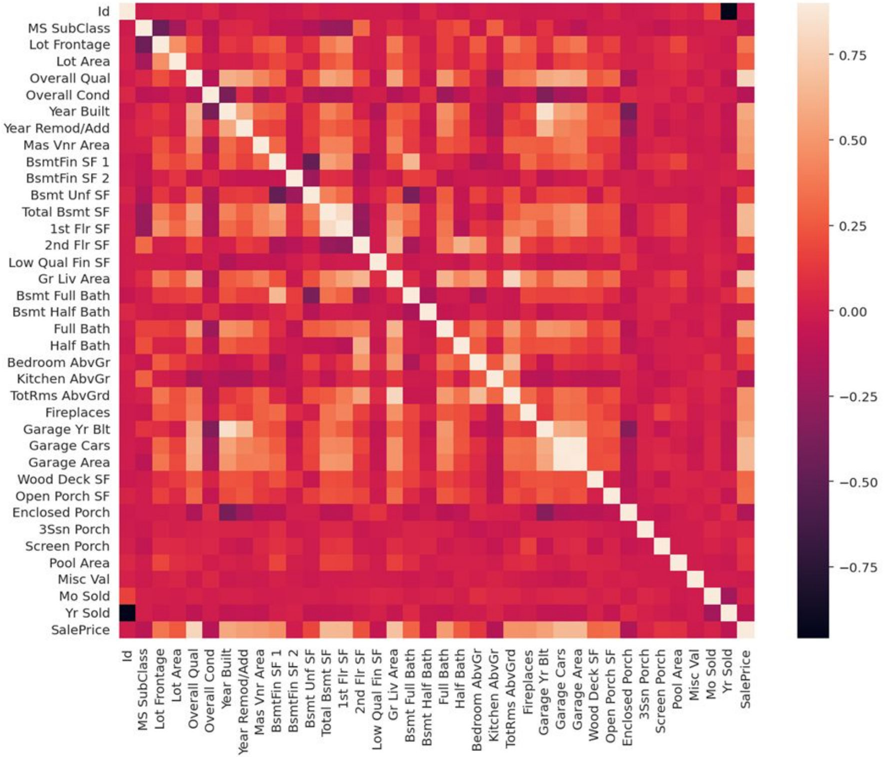
Dataviz

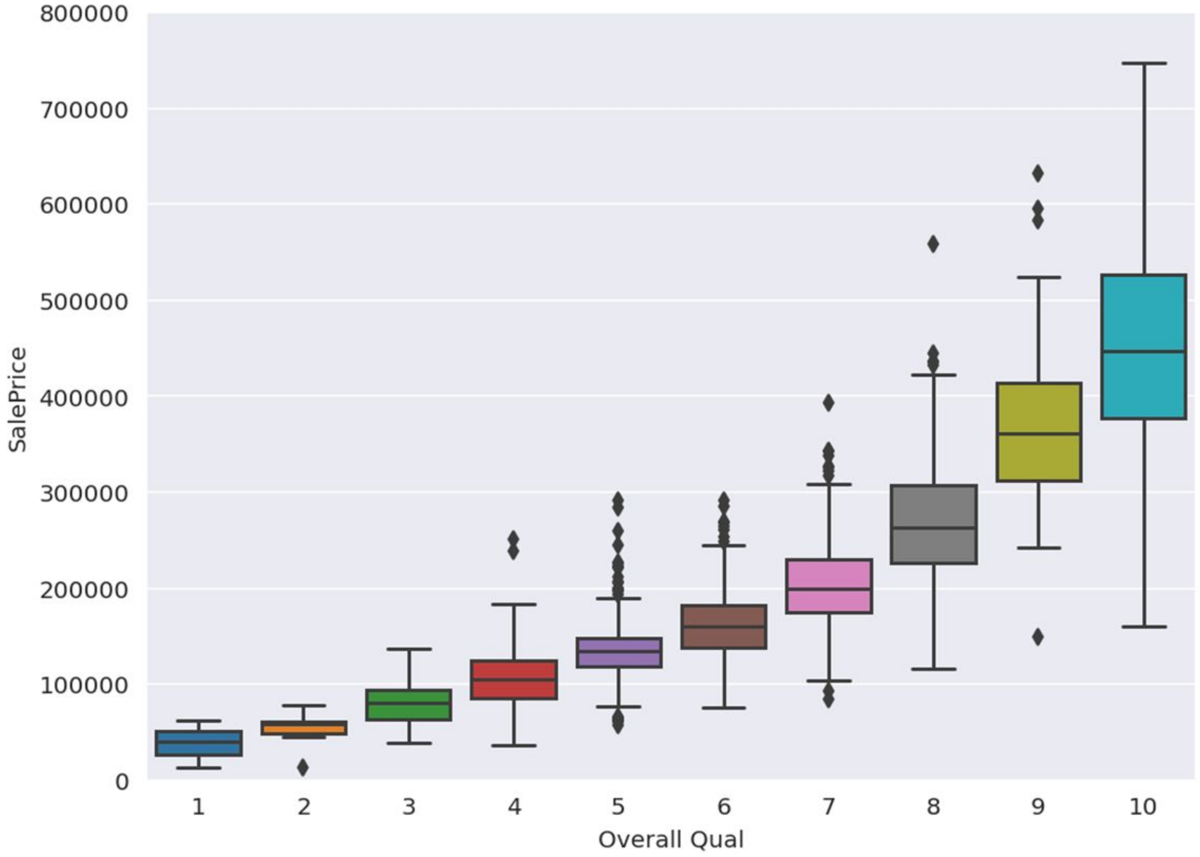
20 features with highest percentage of missing data






Correlations between SalePrice and Gr Liv Area
Gr Liv Area
5000
4000
3000
2000
1000
0
100000
200000
300000
400000
500000
600000
700000
SalePrice


Id
MS SubClass
Lot Frontage
Lot Area
Overall Qual
Overall Cond
Year Built
Year Remod/Add
Mas Vnr Area
BsmtFin SF 1
BsmtFin SF 2
Bsmt Unf SF
Total Bsmt SF
1st Flr SF
2nd Flr SF
Low Qual Fin SF
Gr Liv Area
Bsmt Full Bath
Bsmt Half Bath
Full Bath
Half Bath
Bedroom AbvGr
Kitchen AbvGr
TotRms AbvGrd
Fireplaces
Garage Yr Blt
Garage Cars
Garage Area
Wood Deck SF
Open Porch SF
Enclosed Porch
3Ssn Porch
Screen Porch
Pool Area
Misc Val
Mo Sold
Yr Sold
SalePrice
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


800000
700000
600000
500000
400000
300000
200000
100000
0
SalePrice
1
2
3
4
5
6
7
8
9
10
Overall Qual

Data Preprocessing

log transform to the target feature

Skewness measure to test dataset disparity.

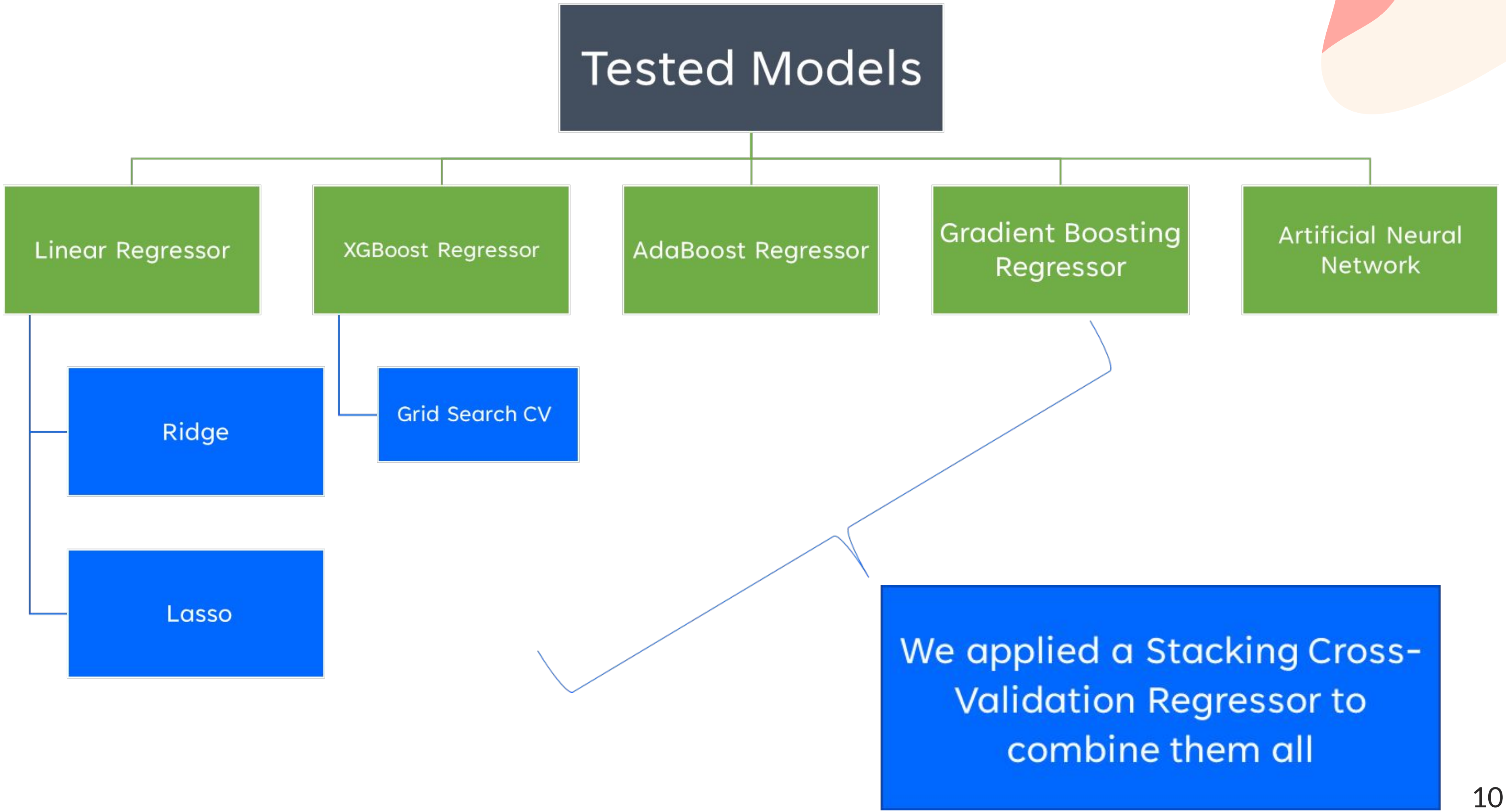
Detect outliers

one-hot encoder get_dummies()

Ordinal Encoder for categorical data

Delete Features with highly missing data

Models

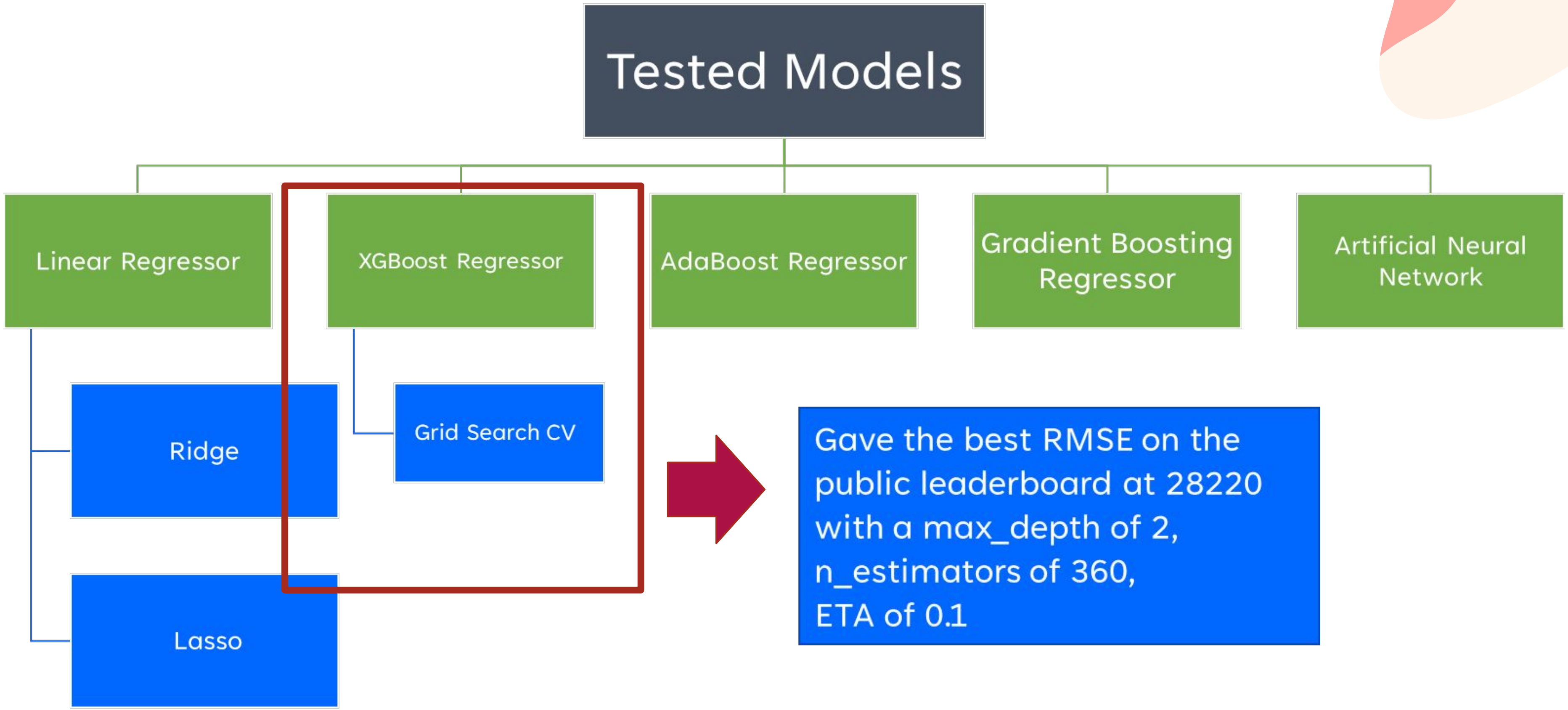


Tested Models

- Linear Regressor
  - Ridge
  - Lasso
- XGBoost Regressor
  - Grid Search CV
- AdaBoost Regressor
- Gradient Boosting Regressor
- Artificial Neural Network

We applied a Stacking Cross-Validation Regressor to combine them all

Results



Tested Models

- Linear Regressor
  - Ridge
  - Lasso
- XGBoost Regressor
  - Grid Search CV
- AdaBoost Regressor
- Gradient Boosting Regressor
- Artificial Neural Network

Gave the best RMSE on the public leaderboard at 28220 with a max_depth of 2,
n_estimators of 360,
ETA of 0.1

Conclusion

- Exploratory Data Analysis should be given the most time to understand each feature
- Data preprocessing plays a substantial role in the performance of any model
- Start with simple models and add complexity
- Learn to control the bias and variance tradeoff (overfitting vs underfitting)

Merci pour votre attention

Avez-vous des questions ?