PROJECT 2 –

AMES HOUSING PRICE PREDICTION

Alicia Tay

AGENDA

- Background
- Business Problem
- Data, Methods, Models
- Assumptions
- Findings
- Future Improvements
- Kaggle Score

BACKGROUND

We are a team of property consultants engaged by the District Council of Iowa to help valuate the residents' properties.

BUSINESS PROBLEM

- Build a model to predict prices for house listings in Ames, Iowa USA
- Identify 30 most influential features, identifying strategies for potential home sellers to maximize profit

DATA, METHODS, MODELS

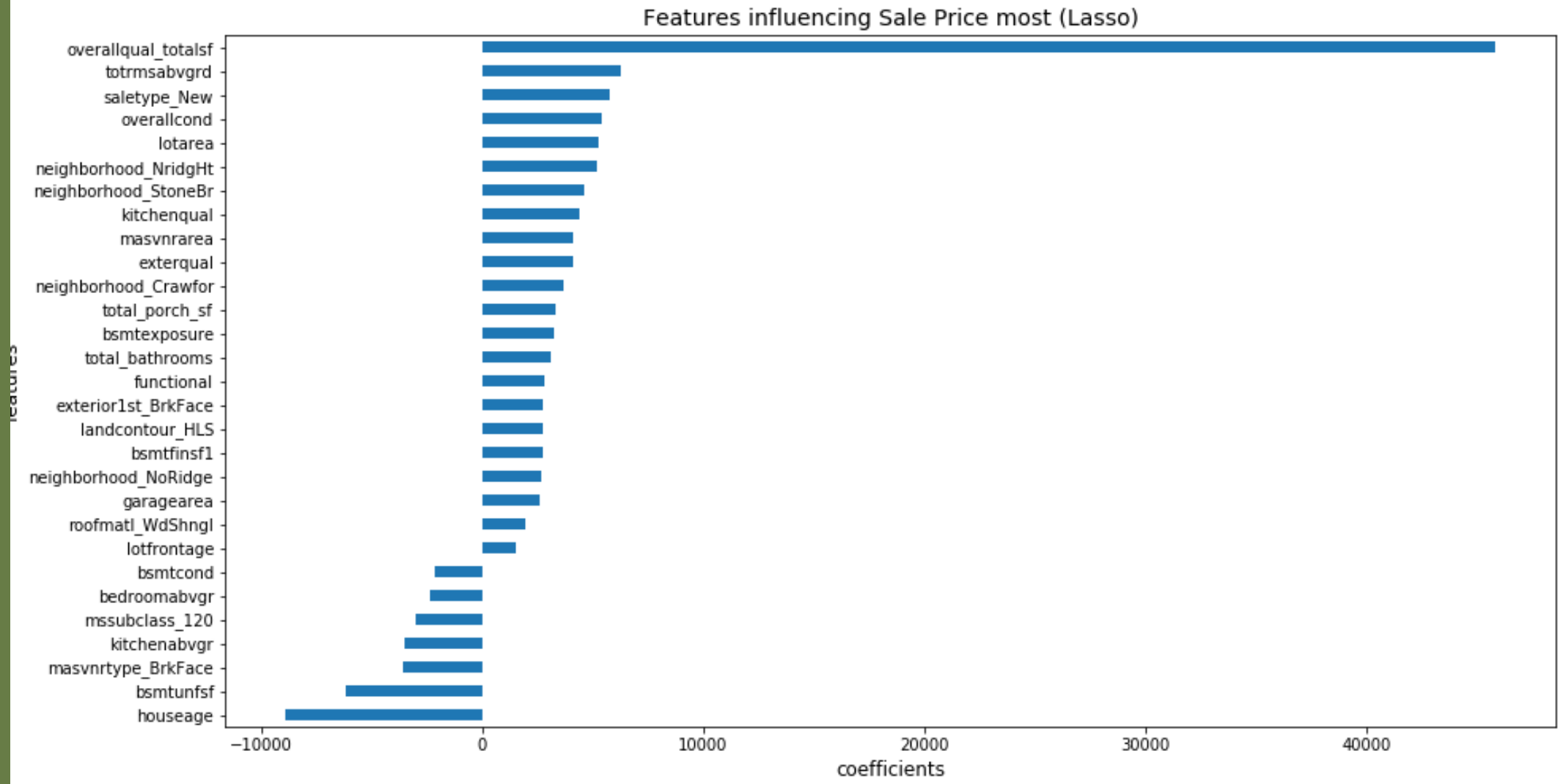
- Data: Ames housing prices between 2006-2010
- Model: Linear Lasso Regression Model
- Methods:
  - Feature Elimination: Lasso & Ridge, Multi-Collinearity
  - Feature Engineering: New variables such as Season, TotalSF
  - Polynomial Features: Interaction elements created to help with the accuracy of model

ASSUMPTIONS

- Features are independent of each other, follows a Gaussian distribution, and has a linear relationship with Saleprice.
- Features with more than 80% null values are not considered significant to this study as we have sufficient data points

FINDINGS X TOP FEATURES

- The more the better!
  - Renovate to include more features
  - Maximise space by converting any unused space into another bedroom/bathroom.
- Substance is key to the pot of Gold
  - Overall quality of the material and finishing is very important to prospective buyers
- Not all are created equal
  - Neighborhoods Northridge Heights, Stone Brook, Northridge, Crawford seems to be premium attractive areas that have big impact on the sale price.


Features influencing Sale Price most (Lasso)
overallqual_totalsf
totrmsabvgrd
saletype_New
overallcond
lotarea
neighborhood_NridgHt
neighborhood_StoneBr
kitchenqual
masvnrarea
exterqual
neighborhood_Crawfor
total_porch_sf
bsmtexposure
total_bathrooms
functional
exterior1st_BrkFace
landcontour_HLS
bsmtfinsf1
neighborhood_NoRidge
garagearea
roofmatl_WdShngl
lotfrontage
bsmtcond
bedroomabvgr
mssubclass_120
kitchenabvgr
masvnrtype_BrkFace
bsmtunfsf
houseage
features
−10000
0
10000
20000
30000
40000
coefficients

FINDINGS X MODELS

- Linear Regression model was a close second, but Lasso Regression gave us the best RMSE score

	Model	R2 Square Train Set	R2 Square Test Set	Adjusted R2 Test Set	RMSE
0	Linear	0.911583	0.905041	0.899130	24501.902697
1	Ridge	0.911493	0.904746	0.898817	24539.911107
2	Lasso	0.911583	0.905041	0.899130	24501.902599
3	Elastic	0.911552	0.904885	0.898965	24521.948345
4	Linear_poly	0.923621	0.920983	0.916238	22350.749627
5	Ridge_poly	0.923616	0.920849	0.916096	22369.723389
6	Lasso_poly	0.923621	0.920983	0.916238	22350.749567
7	Elasticnet_poly	0.923617	0.920870	0.916119	22366.662993

KAGGLE

- I achieved a score of 25098.835 on Kaggle
- More could be done to improve the model's accuracy, such as:
  - Being more sensitive to outliers rather than removing them, and experimenting with more polynomial features.
  - This might include the need of with Neural Networks
  - More features (>30)could be added so that there are more predictors that the model can learn from to predict the nuances in price changes
  - Skewed variables could be treated with a log transformation
  - However, linear regression models relative to more complex machine-learning techniques is its simple, easy interpretability.

FUTURE IMPROVEMENTS

- Our dataset was collected between 2006-2010. The subprime mortgage crisis happened during the period 2007 - 2010. This could have influenced that data, and might not be accurate for a 2020 time frame
- Understanding the difference between the listed and sale price, along with the negotiations and discussions between homeowners and agents could help us understand more on the factors that influences price