Predicting Housing Sale Prices

Understanding features associated with higher sale price

Ben Poh, DSI24

Problem Statement

- You work for the local housing authority. Using the Ames housing dataset, your manager is keen to know how features of a property can determine its sale price. Your manager would also like to know if a better basement (size, quality, exposure etc.) will lead to a higher sale price.

Methodology

- **Part I:**
  - Data cleaning
  - Feature engineering
- **Part II**
  - Modelling
    - Linear Regression, Lasso, Ridge
  - Train and predict on test set for Kaggle competition

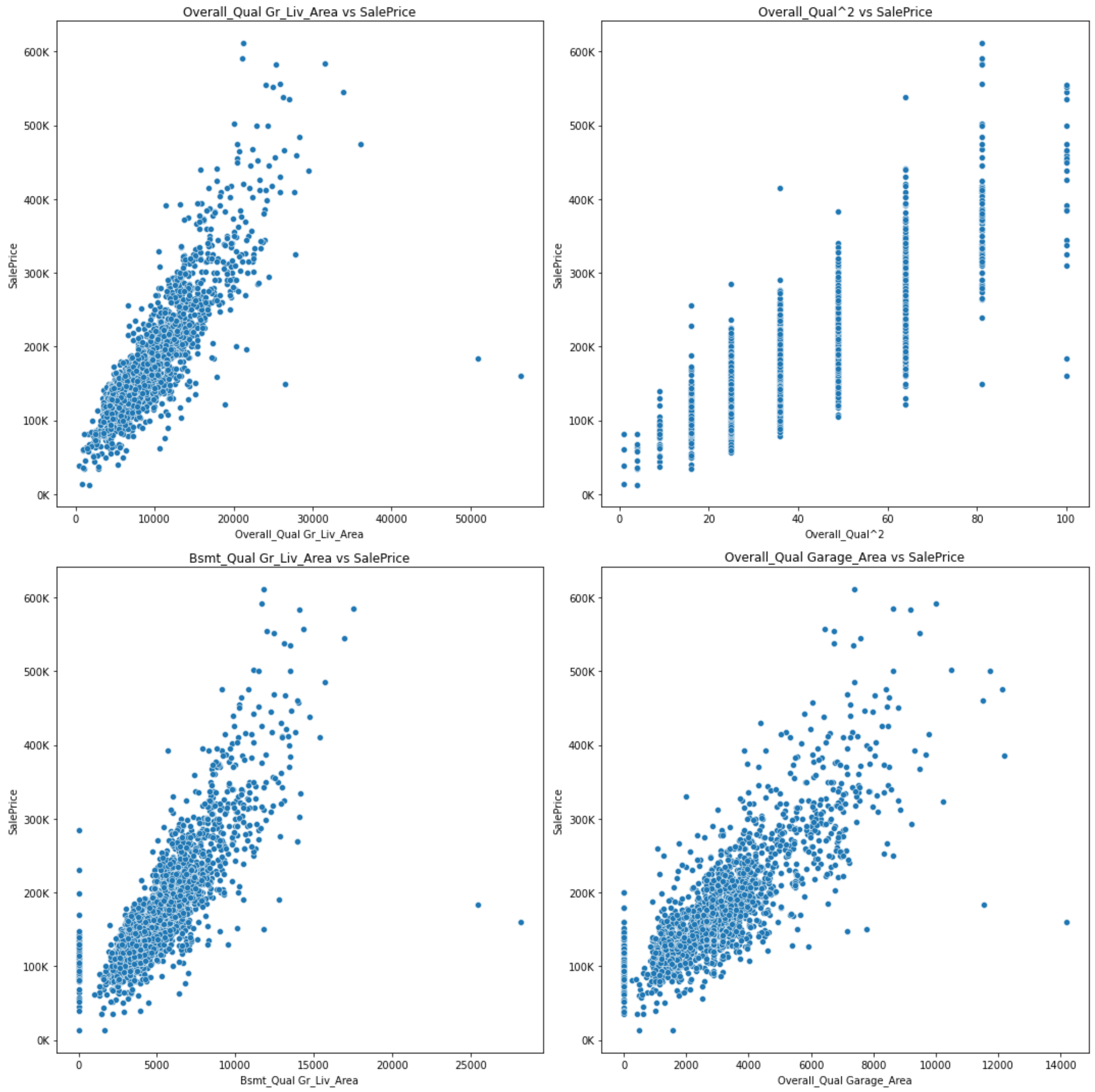
Data Cleaning

- 81 features – segregated into continuous, categorical, ordinal and discrete features for data cleaning/EDA
- Null values:
  - Imputed mean ‘Lot Frontage’ based on ‘Lot Shape’ and ‘Lot Config’ categorization
  - Replaced incorrect fields in basement and garage features
  - Update ‘None’ when the property does not have that feature
- Ordinal features – mapped to numeric formats
- Nominal features – binarise with OneHotEncoder

Continuous	Categorical	Ordinal	Discrete
Lot_Frontage	MS_Zoning	Overall_Cond	Bsmt_Full_Bath
BsmtFin_SF_1	Street	Overall_Qual	Full_Bath
BsmtFin_SF_2	Alley	Lot_Shape	Year_Remod/Add
Bsmt_Unf_SF	Land_Contour	Utilities	Kitchen_AbvGr
Total_Bsmt_SF	Lot_Config	Land_Slope	TotRms_AbvGrd
Garage_Area	Neighborhood	Exter_Qual	Half_Bath
Lot_Area	Condition_1	Exter_Cond	Bsmt_Half_Bath
Gr_Liv_Area	Condition_2	Bsmt_Qual	Bedroom_AbvGr
Low_Qual_Fin_SF	Bldg_Type	Bsmt_Cond	Garage_Yr_Blt
1st_Flr_SF	House_Style	Bsmt_Exposure	Fireplaces
2nd_Flr_SF	Roof_Style	BsmtFin_Type_1	Mo_Sold
Wood_Deck_SF	Roof_Matl	BsmtFin_Type_2	Yr_Sold
Open_Porch_SF	Exterior_1st	Heating_QC	Year_Built
Enclosed_Porch	Exterior_2nd	Electrical	Garage_Cars
3Ssn_Porch	Mas_Vnr_Type	Kitchen_Qual	
Screen_Porch	Foundation	Functional	
Pool_Area	Heating	Fireplace_Qu	
Mas_Vnr_Area	Central_Air	Garage_Finish	
Misc_Val	Garage_Type	Garage_Qual	
	Misc_Feature	Garage_Cond	
	Sale_Type	Paved_Drive	
	MS_SubClass	Pool_QC	
		Fence	

Feature Engineering

- Polynomial Features:
  - Added **4** interaction terms due to their high correlation with ‘SalePrice’
- Unified size representation for basement and gross liveable area



Modelling

- Baseline
  - X variables: Overall Quality and Gross Liveable Area (two highest +ve correlated with Sale Price)
  - Based on OLS, test RMSE = 39,513 / R2 = 0.75
- 3 sklearn models used:
  - Linear Regression
  - Ridge
  - Lasso

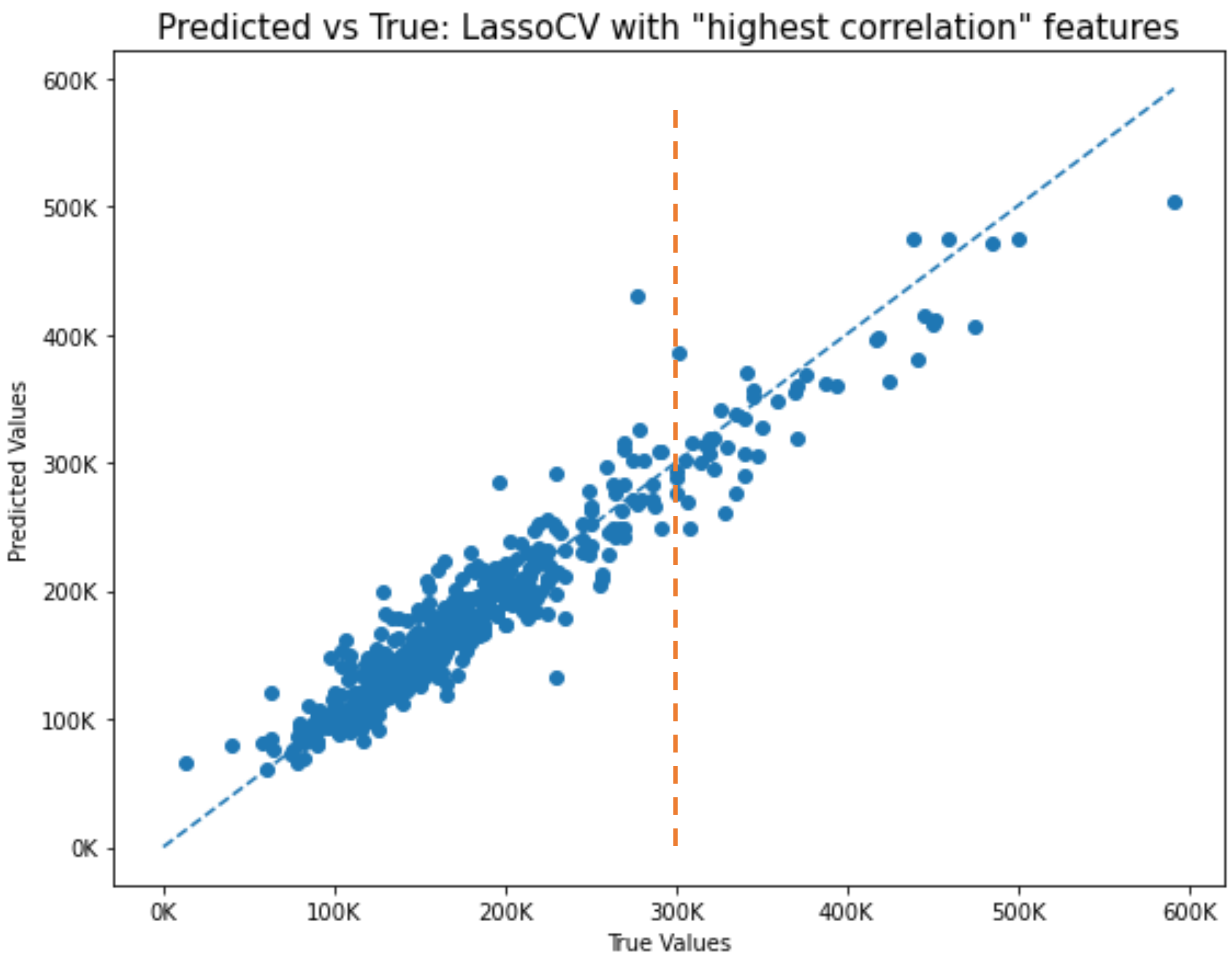
Modelling

Feature Selection	No. of Features	Model Description	Hyperparameters	Train RMSE	Test (Holdout) RMSE
Baseline	2	Linear Regression	-	39,779	39,513
Highest Correlation	49	Linear Regression	-	23,107	23,100
Highest Correlation	49	Ridge	α = 3.23	23,128	23,186
Highest Correlation	49	Lasso	α = 46.42	23,131	23,159
Highest Correlation + Reduced Collinearity	40	Linear Regression	-	23,269	23,225
Highest Correlation + Reduced Collinearity	40	Ridge	α = 3.24	23,288	23,297
Highest Correlation + Reduced Collinearity	40	Lasso	α = 36.784	23,280	23,238

- Model to deploy: **Lasso with high correlation features (49 features)**
- RMSE on test set = 23,238. Not the best, but close to results from Linear Regression – Lasso zeroise the ‘useless’ coefficients

Results: Predicted vs True

- Predicts pretty well for sale price < $300k
- Few properties with large sale prices in the train set



Results and Conclusion

- Overall quality – better quality always lead to higher prices
- Size (Lot Area, Gross Liveable Area) – bigger the better
- Age of property – younger the better
- Basement Features:
  - Quality, Size, Exposure, Type, # of Bathrooms all help increase sale price

Variable	Coefficient
Overall_Qual Gr_Liv_Area	48993.420682
Overall_Qual Garage_Area	22643.680871
Gr_Liv_Area	-17429.576591
Bsmt_Qual Gr_Liv_Area	16231.075533
Overall_Qual	-14835.665990
Garage_Area	-14014.872945
Bsmt_Qual	-8486.417885
Total_Bsmt_SF	7173.390682
Lot_Area	6800.263609
New	5151.427642
Bsmt_Exposure	4826.862267
Kitchen_Qual	4682.224378
Hip	4479.534916
Year_Built	4228.190032
BsmtFin_Type_1	4046.909321
Garage_Cond	3951.908237
Garage_Yr_Blt	-3906.650386
StoneBr	3849.534714
Exter_Qual	3707.697259
Bsmt_Full_Bath	3608.763063