Prediction of Ames House Price and Analysis of Impact of Proximity to Amenities

Pei-Chieh (Paget) Hsiao

Paget.Hsiao@gmail.com

Project Overview

- Business Questions
  - Does the proximity to amenities (Parklands, Iowa State University and airport) have an impact on the predicted sale price of Ames properties in 2006~2010?

- Stakeholders
  - House buyers & house sellers

Flow Chart of Action

Data wrangling

- Load data
- Identify outliers
- EDA
- Correlation

Prediction

- Encode categorical features
- Forward feature selection
- Build regressor models
- Cross Validation

Summary

- Predicted sale price
- Proximity to amenities
- Findings

Identify Outliers

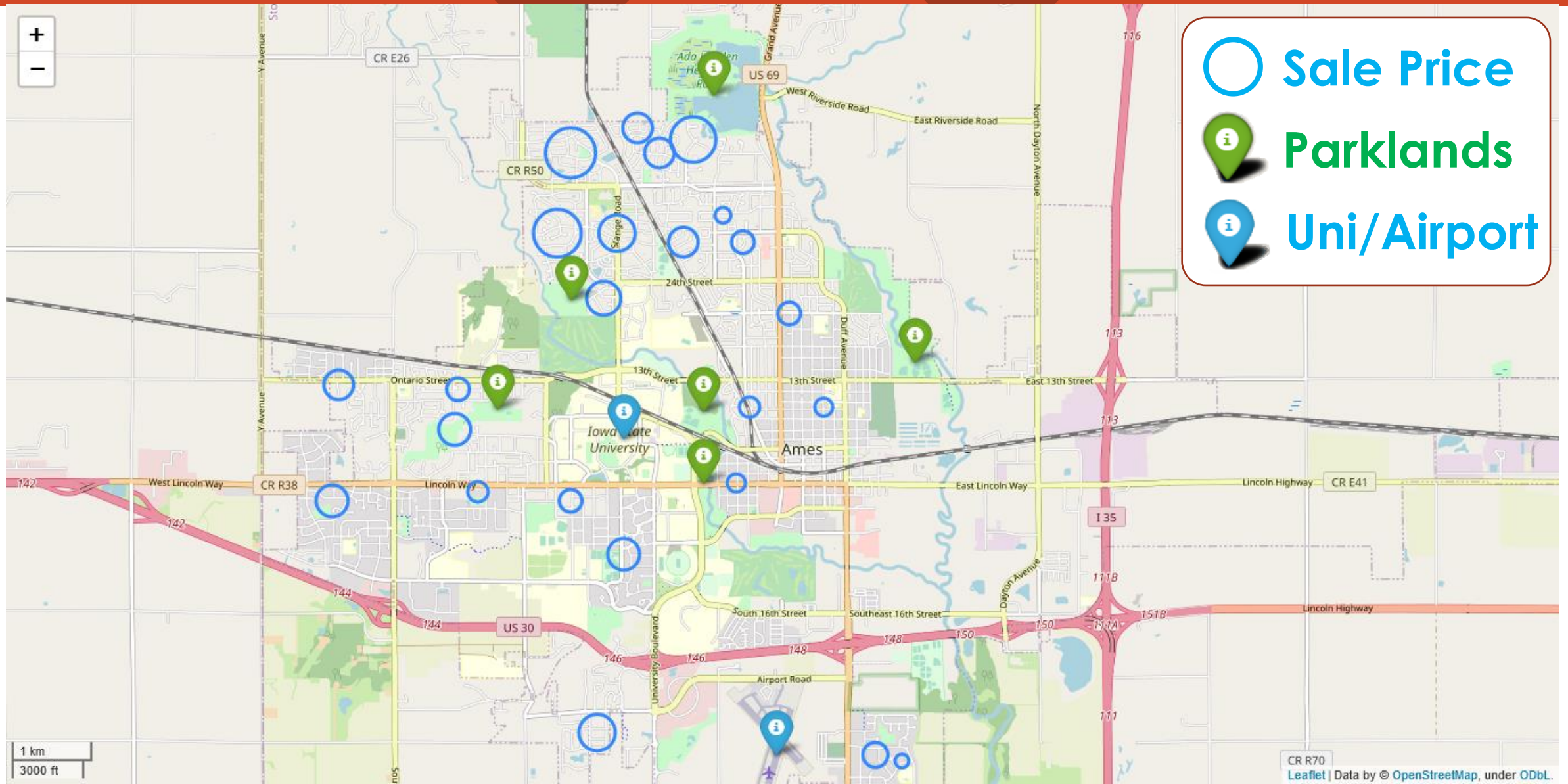
- Ames property dataset
  - 1460 rows
  - 80 features (ignore column ‘Id’)
- Remove commercial properties
  - ‘C’ in column ‘MSZoning’ – 10 rows
- Remove insignificant neighborhood
  - ‘Blueste’ in column ‘Neighborhood’ – 2 rows

Data Wrangling

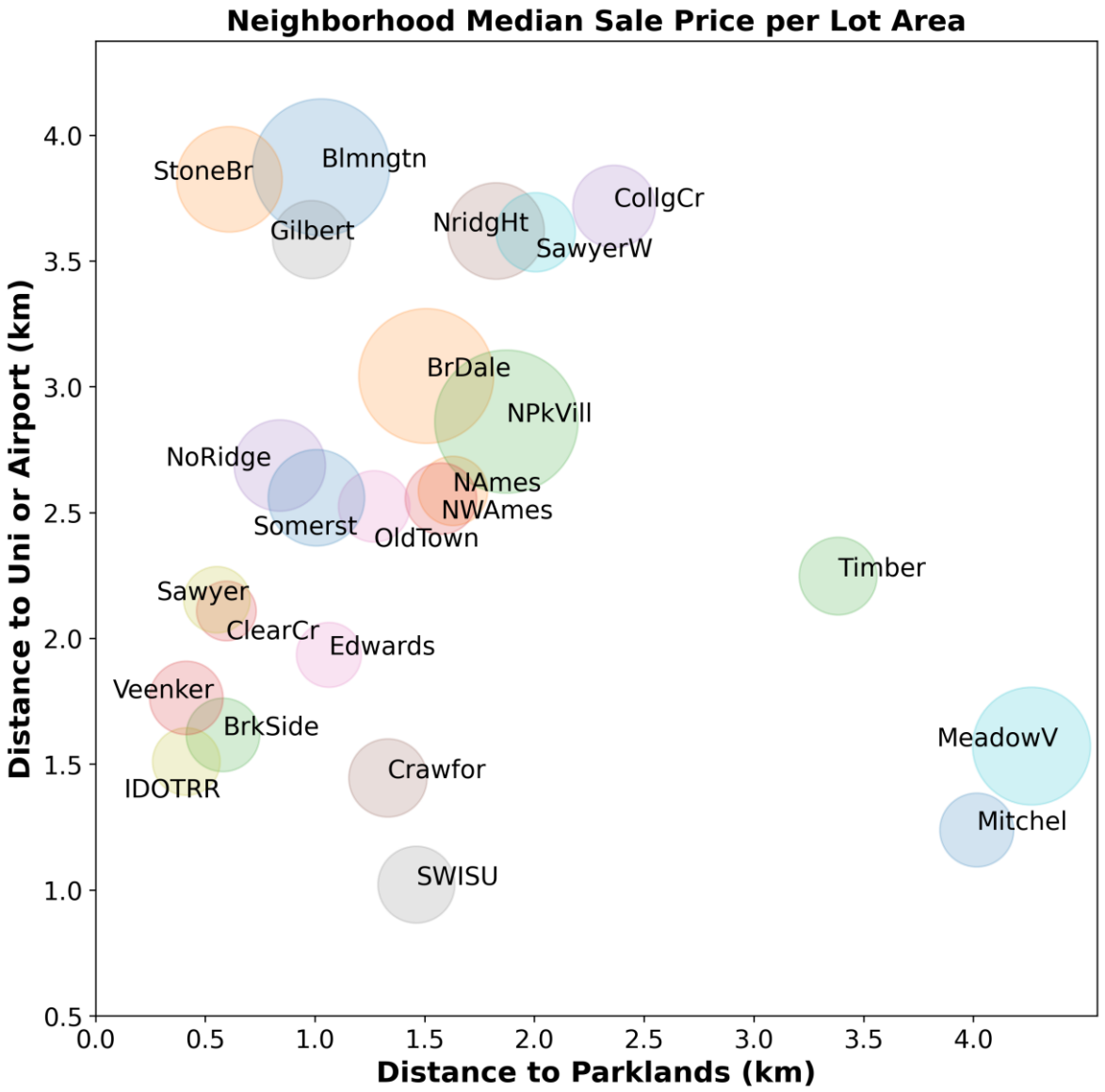
- Combine numerical features
  - Bathroom = ‘FullBath’ + ‘HalfBath’/2
  - BsmtBath = 'BsmtFullBath’ + 'BsmtHalfBath’/2
  - AgeBuilt = 'YrSold' - 'YearBuilt’
  - AgeRemod = ‘YrSold' - 'YearRemodAdd'

- Adjust features
  - Combine ‘Condition1’ & ‘Condition2’ as ‘Condition’
  - ‘Pos’⇐ [‘PosA’,’PosN’]
  - ‘Neg’⇐ ['RRNn','RRAn','Feedr','Artery','RRNe','RRAe']

Proximity to Amenities



Sale Price & Proximity to Amenities



- **Sale price per lot area tends to decrease with 'distance to Parklands'**
- **Sale price per lot area tends to increase with 'distance to Uni or airport'**

Numerical Feature Selection

- Drop column ‘Id’

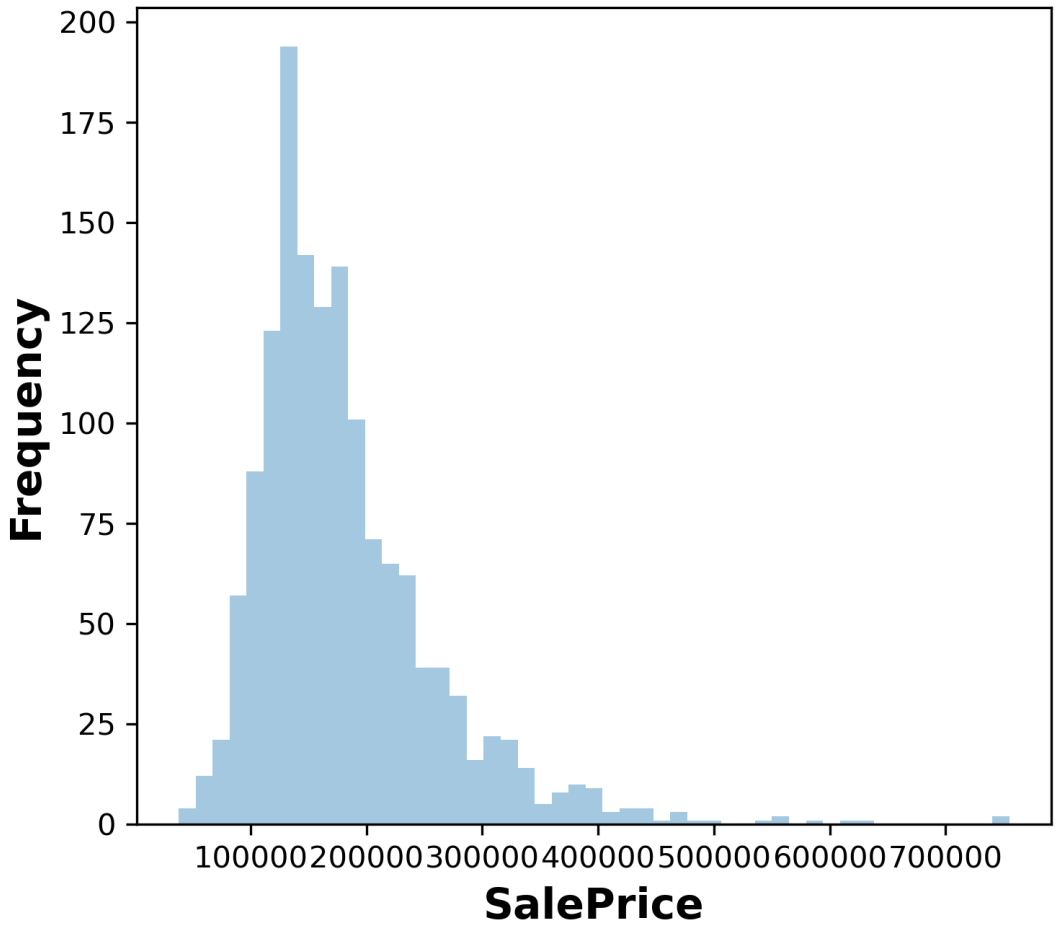
- Drop features containing NaN
  - Three features are excluded:
    'LotFrontage','GarageYrBlt','MasVnrArea'

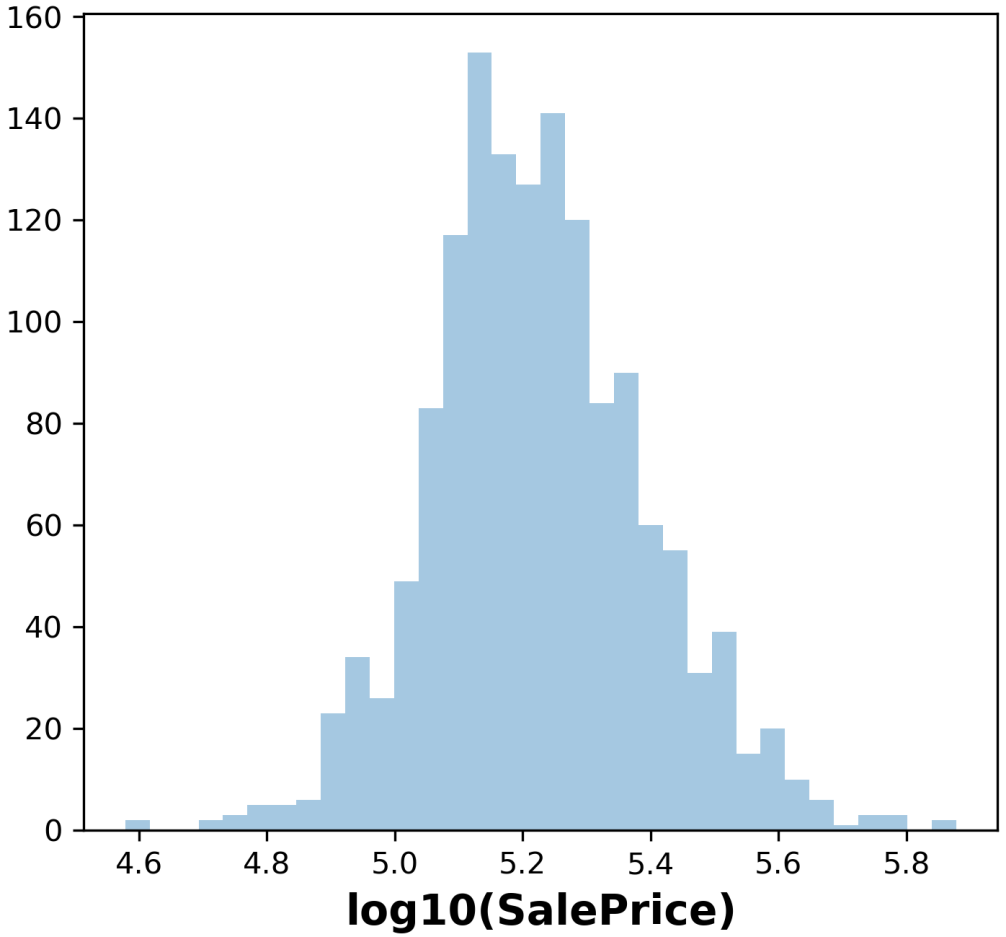
Categorical Feature Selection

- Ordinal Encoding – **categories are in ascending order of median SalePrice**

Feature Name	Categories
?Qual, ?Cond	['NA', 'Po', 'Fa', 'TA', 'Gd', 'Ex']
MSZoning	['RM','RH','RL','FV']
BldgType	['2fmCon','Duplex','Twnhs','1Fam','TwnhsE']
HouseStyle	['1.5Unf','1.5Fin','2.5Unf','SFoyer','1Story','SLvl','2Story','2.5Fin']
MasVnrType	['NA','BrkCmn','None','BrkFace','Stone']
Foundation	['Slab','BrkTil','Stone','CBlock','Wood','PConc']
BsmtExposure	['NA','No','Mn','Av','Gd']
BsmtFinType1	['NA','LwQ','BLQ','Rec','ALQ','Unf','GLQ']
GarageType	['NA','CarPort','Detchd','Basment','2Types','Attchd','BuiltIn']
GarageFinish	['NA','Unf','RFn','Fin']
CentralAir	['N','Y']
Condition	['NegNeg','NegNorm','NormNorm','NegPos','PosNorm','PosPos']
SaleCondition	['AdjLand','Abnorml','Family','Alloca','Normal','Partial']
Neighborhood	['MeadowV','IDOTRR','BrDale','OldTown',……,,'Timber','StoneBr','NoRidge','NridgHt']

Target Variable



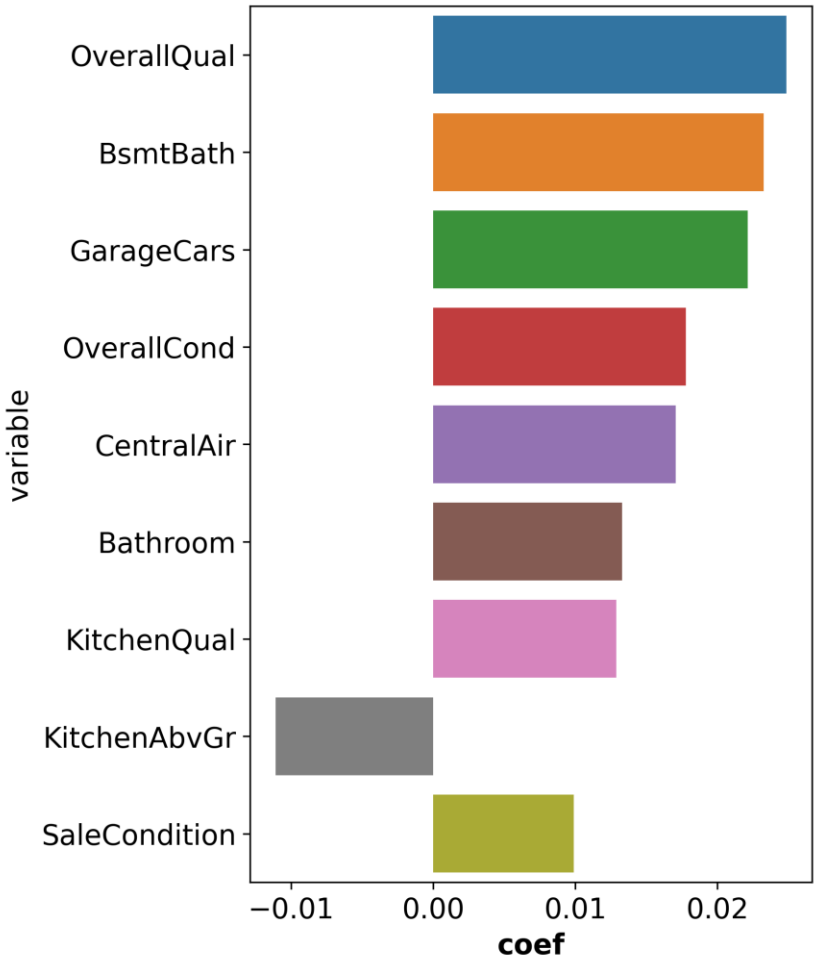


- Log10(SalePrice) is used (approximate to normal distribution)

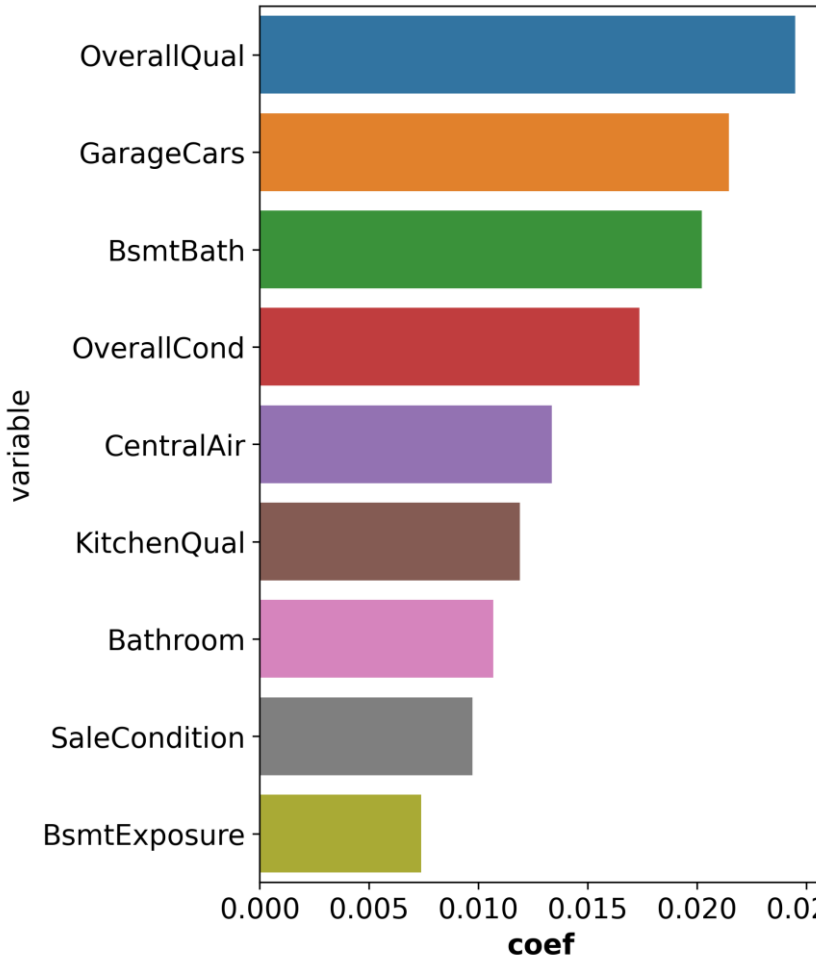
Feature Coefficient

Only 9 most significant features are shown

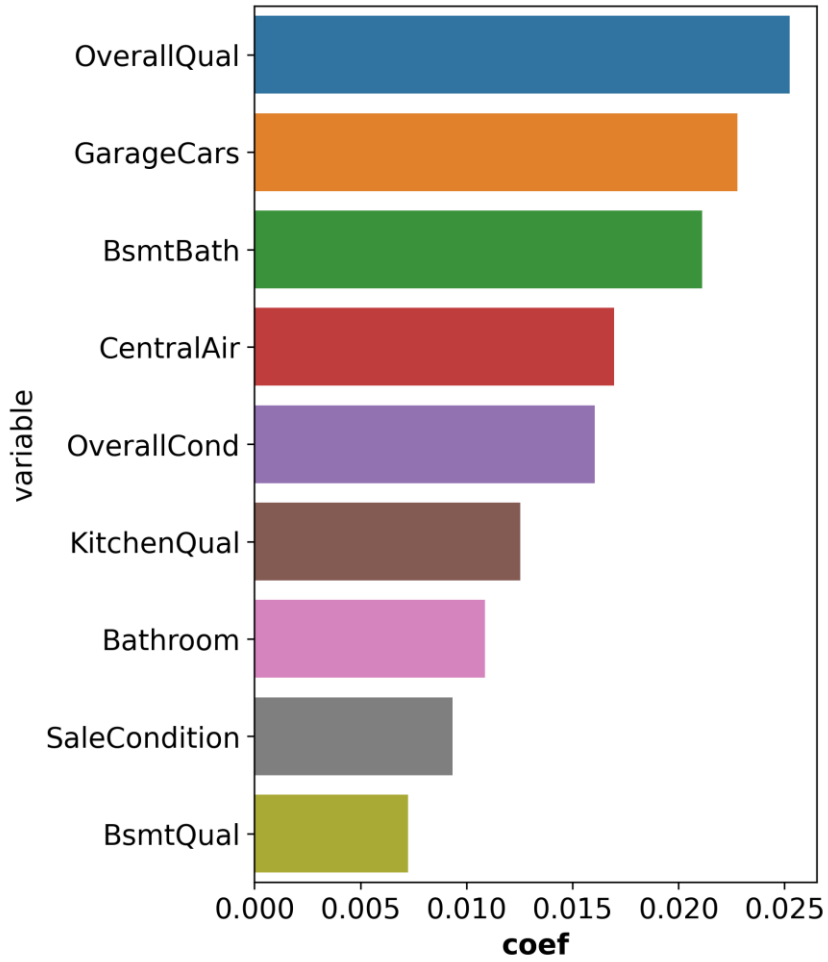
- **OLS - Forward feature selection**



- **RidgeCV**



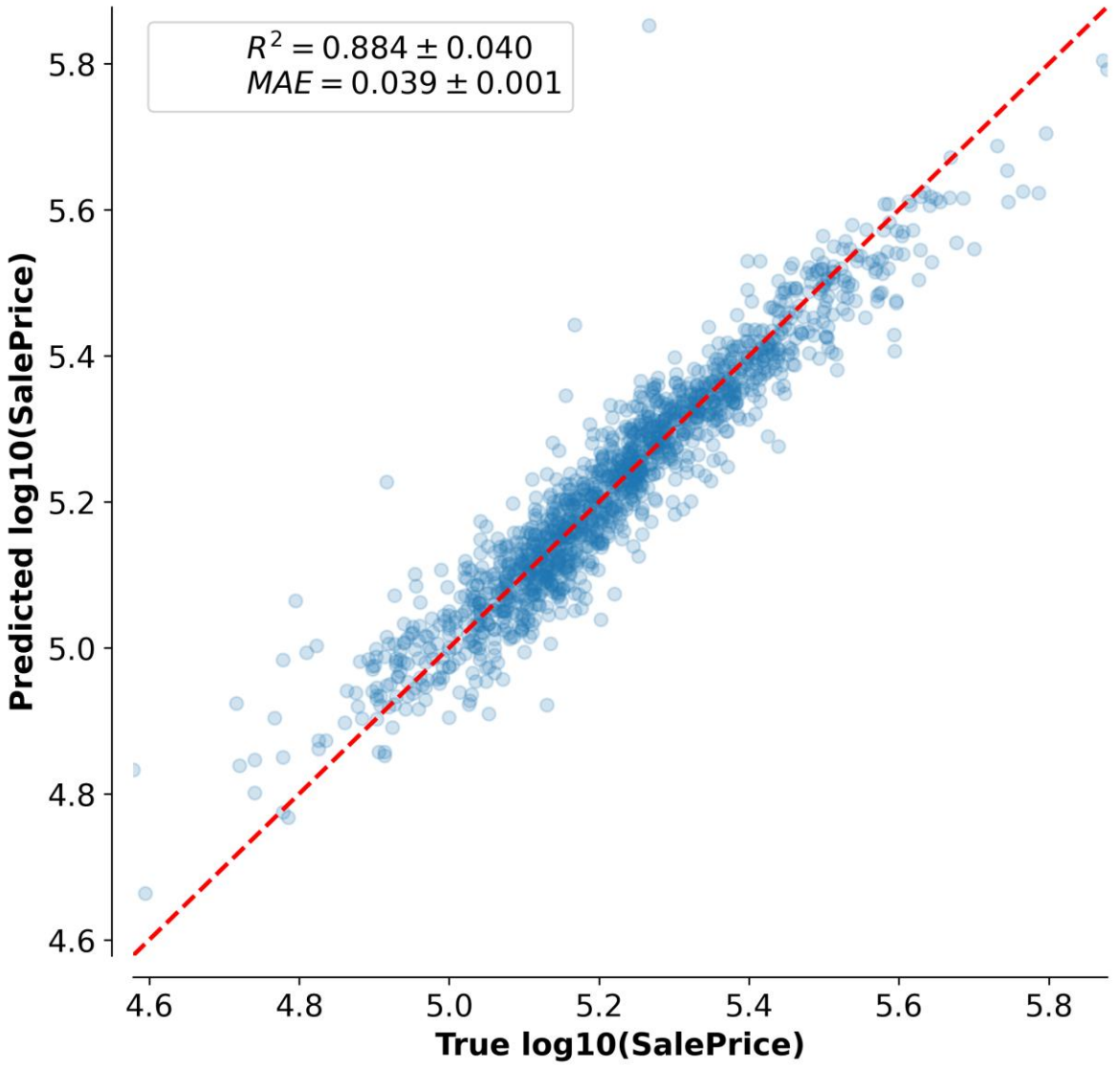
- **LassoCV**



R^2 Score and Cross Validate

Regression	R^2 Score	R^2 Cross validate
OLS	0.8991	0.8842 ± 0.040
Ridge	0.8993	0.8784 ± 0.050
Lasso	0.8994	0.8777 ± 0.046

- **Almost identical R^2 score between OLS, ridge and lasso regressors**
- **High cross validated R^2 score of 0.8842 ± 0.040**



OLS with One Hot Encoding

- Ordinal encoding transforms categorical features into ordinal integers in a single column of integers (0 to n_categories - 1) per feature
- It implies the feature category is linearly correlated to the target variable.
- Use one hot encoding in preprocessor in pipeline for seamless modelling.

Regressor	R^2 validate (ordinal encoding)	R^2 validate (one-hot encoding)
OLS	0.8842 ± 0.040	-
RidgeCV	0.8784 ± 0.050	0.8778 ± 0.047
LassoCV	0.8777 ± 0.046	0.8776 ± 0.051

- Applying one-hot encoding in OLS model results in overfitting. Regularization is necessary.

Summary

- Impact of proximity to amenity on the Neighborhood sale price of Ames residential properties is analysed.
- Sale price tends to decrease with 'distance to Parklands' and increase with 'distance to Uni or airport'.
- Only 12 properties identified as outliers are excluded in the prediction of sale price.
- Ordinal encoding and one hot encoding are used for categorical features. Both achieved identical cross-validated R^2 score up to 0.88. However, regularisation is required to prevent overfitting.