Data-Driven House Price prediction

Yuqi Zeng

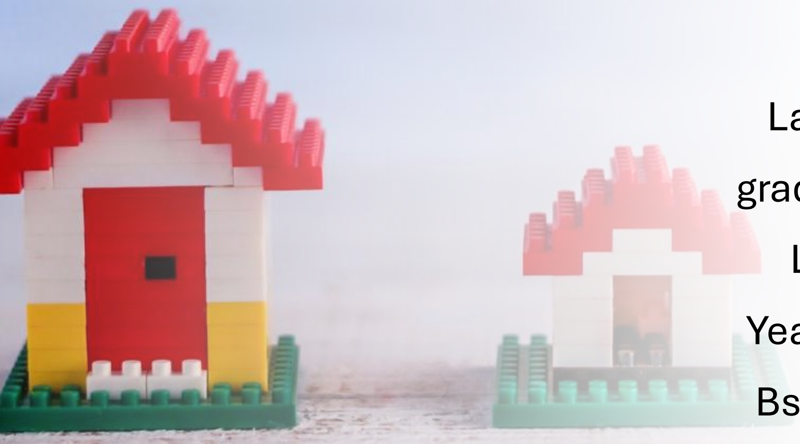
August 13, 2025

Pennsylvania state university

CONTENTS

1. Data Cleaning and EDA
2. Variable Selection
3. Model Selection
4. Model Interpretation

Data Cleaning



- Data size: 279 samples
- Delete unavailable variables

variables	**CollgCr**	**Edwards**	**OldTown**	Training model
BsmtUnfSF	✓	nan	✓	Not include
Street	nan	nan	nan	Not include
LandContour	nan	nan	nan	Not include
grade to building	nan	nan	nan	Not include
LandSlope	nan	nan	nan	Not include
YearRemodAdd	nan	nan	nan	Not include
BsmtExposure	nan	nan	nan	Not include
KitchenAbvGr	nan	✓	✓	Not include

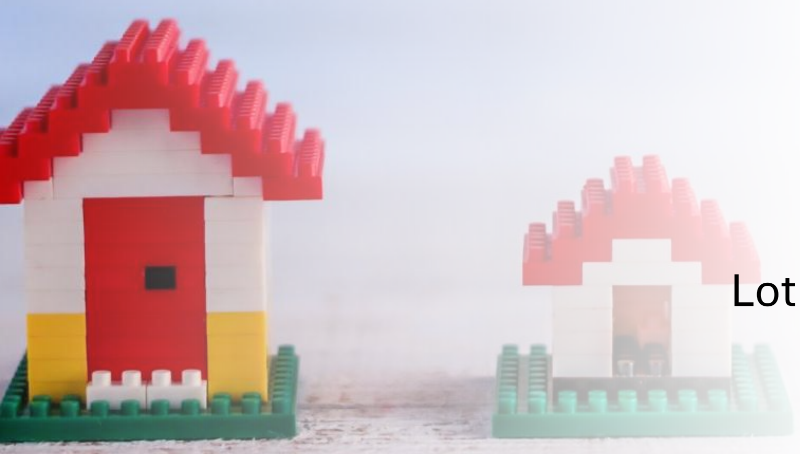
Data Cleaning

- Split concatenated variables

Exterior ⇨ Exterior1st, ExterQual, ExterCond

LotInfo ⇨ LotConfig, LotShape, LotFrontage, LotArea

P	Q	R	S	T	U	V
RoofStyle	Exterior	Utilities	BsmtFinSF	Heating	LotInfo	KitchenQua
Gable	VinylSd;TA;Gd	AllPub	706	GasA	Inside;Reg;8450;65	Gd
Gable	VinylSd;TA;Gd	AllPub	486	GasA	Inside;IR1;11250;68	Gd
notGable	VinylSd;TA;Gd	AllPub	0	GasA	Inside;Reg;9742;75	Gd
Gable	VinylSd;TA;Gd	AllPub	0	GasA	Corner;Reg;11049;85	Gd
notGable	VinylSd;TA;TA	AllPub	280	GasA	CulDSac;IR1;9200;NA	TA
Gable	VinylSd;TA;Gd	AllPub	0	GasA	Corner;IR1;11645;89	Gd
Gable	OtherSd;TA;TA	AllPub	632	GasA	Inside;Reg;7200;60	TA
Gable	VinylSd;TA;TA	AllPub	739	GasA	Inside;Reg;9375;NA	Gd
Gable	VinylSd;TA;Gd	AllPub	1013	GasA	Inside;IR1;10665;72	Gd
Gable	VinylSd;TA;TA	AllPub	588	GasA	Inside;Reg;8070;60	TA



Data Cleaning

- Variables Cleaning Process

Column	Treatment
GarageType	NA or blank indicates no garage, should be labeled as “noGarage”
LotFrontage	Convert to numerical
LotArea	Convert to numerical; Fill NA with mean
BsmtCond	NA should be labeled as “noBasement” (indicates no basement)
BsmtQual	Fill NA with mode
BsmtFinType1	Fill NA with mode
houseAge	Create : houseAge = YrSold - YearBuilt; And Row 31: age < 0 → likely swapped years; take absolute value, keep row
Neighborhood	Create : Neighbor, indicating the sample belongs to which neighborhood
Sample_id	Create : sample_id, as the unique ID of each sample

Data Cleaning

- Transform categorical variables to numerical matrix variables
- Method: One-hot encoding

```
"CentralAir"   "BsmtQual"     "HouseStyle"
"PavedDrive"   "SaleType"     "RoofStyle"
"Utilities"    "Heating"      "LotConfig"
"BsmtCond"     "Foundation"   "Electrical"

"HeatingQC"    "GarageType"   "RoofMatl"
"Exterior1st"  "ExterQual"    "ExterCond"
"LotShape"     "KitchenQual"  "BsmtFinType1"
"BldgType"     "Neighbor"
```

EDA

- Numerical Variables exploration data analysis
  - Average house price is $158, 035.38

Table: Numerical Variables Summary Statistics

variable	count	mean	median	sd	min	max	q1	q3
OverallQual	279	5.810	6	1.369	1	10	5.0	7.0
BedroomAbvGr	279	2.842	3	0.761	1	5	2.0	3.0
Fireplaces	279	0.362	0	0.570	0	3	0.0	1.0
FullBath	279	1.516	2	0.535	0	3	1.0	2.0
OpenPorchSF	279	45.753	24	68.781	0	547	0.0	64.0
BsmtFinSF1	279	345.140	203	493.965	0	5644	0.0	644.0
LotFrontage	279	9363.401	9100	4428.480	2522	63887	7495.5	10800.0
LotArea	279	67.683	67	22.195	24	313	60.0	72.0
HalfBath	279	0.272	0	0.454	0	2	0.0	1.0
WoodDeckSF	279	79.459	0	104.093	0	576	0.0	144.0
TotRmsAbvGrd	279	6.405	6	1.613	3	12	5.0	7.0
SalePrice	279	158035.380	143000	60810.233	37900	475000	114752.0	195750.0
OverallCond	279	5.656	5	1.262	1	9	5.0	7.0
GrLivArea	279	1457.477	1431	540.851	605	5642	1097.5	1722.0
houseAge	279	45.925	52	38.801	0	136	6.0	81.5

EDA

- Categorical Variables exploration data analysis

$CentralAir

	CentralAir	count	variable
1	Y	243	CentralAir
2	N	36	CentralAir

$BsmtQual

	BsmtQual	count	variable
1	TA	129	BsmtQual
2	Gd	117	BsmtQual
3	Fa	23	BsmtQual
4	Ex	10	BsmtQual

$HouseStyle

	HouseStyle	count	variable
1	1Story	143	HouseStyle
2	2Story	76	HouseStyle
3	1.5Story	51	HouseStyle
4	2.5Story	9	HouseStyle

$HeatingQC

	HeatingQC	count	variable
1	Ex	165	HeatingQC
2	TA	59	HeatingQC
3	Gd	42	HeatingQC
4	Fa	13	HeatingQC

$GarageType

	GarageType	count	variable
1	Attchd	146	GarageType
2	Detchd	105	GarageType
3	noGarage	28	GarageType

$RoofMatl

	RoofMatl	count	variable
1	CompShg	276	RoofMatl
2	notCompShg	3	RoofMatl

$PavedDrive

	PavedDrive	count	variable
1	Y	232	PavedDrive
2	N	35	PavedDrive
3	P	12	PavedDrive

$SaleType

	SaleType	count	variable
1	WD	250	SaleType
2	notWD	29	SaleType

$RoofStyle

	RoofStyle	count	variable
1	Gable	233	RoofStyle
2	notGable	46	RoofStyle

$Exterior1st

	Exterior1st	count	variable
1	VinylSd	136	Exterior1st
2	OtherSd	99	Exterior1st
3	MetalSd	44	Exterior1st

$ExterQual

	ExterQual	count	variable
1	TA	242	ExterQual
2	Gd	37	ExterQual

$ExterCond

	ExterCond	count	variable
1	TA	180	ExterCond
2	Gd	99	ExterCond

$Utilities

	Utilities	count	variable
1	AllPub	279	Utilities

$Heating

	Heating	count	variable
1	GasA	265	Heating
2	GasW	8	Heating
3	Grav	4	Heating
4	OthW	1	Heating
5	Wall	1	Heating

$LotConfig

	LotConfig	count	variable
1	Inside	203	LotConfig
2	Corner	58	LotConfig
3	CulDSac	12	LotConfig
4	FR2	6	LotConfig

$LotShape

	LotShape	count	variable
1	Reg	211	LotShape
2	IR1	57	LotShape
3	IR2	9	LotShape
4	IR3	2	LotShape

$KitchenQual

	KitchenQual	count	variable
1	TA	135	KitchenQual
2	Gd	122	KitchenQual
3	Ex	13	KitchenQual
4	Fa	9	KitchenQual

$BsmtFinType1

	BsmtFinType1	count	variable
1	Unf	120	BsmtFinType1
2	GLQ	72	BsmtFinType1
3	ALQ	49	BsmtFinType1
4	BLQ	38	BsmtFinType1

$BsmtCond

	BsmtCond	count	variable
1	TA	244	BsmtCond
2	noBasement	35	BsmtCond

$Foundation

	Foundation	count	variable
1	PConc	134	Foundation
2	CBlock	83	Foundation
3	BrkTil	62	Foundation

$Electrical

	Electrical	count	variable
1	SBrkr	239	Electrical
2	Fuse	40	Electrical

$BldgType

	BldgType	count	variable
1	1Fam	245	BldgType
2	2fmCon	16	BldgType
3	Twnhs	12	BldgType
4	Duplex	6	BldgType

$Neighbor

	Neighbor	count	variable
1	CollegeCr	116	Neighbor
2	OldTown	89	Neighbor
3	Edwards	74	Neighbor

Variable Selection

- The total number of predictors is 38

- Delete Utilities
- Because the variable *Utilities_* is perfectly linearly dependent (completely predictable by other variables), causing multicollinearity in the model.

- Pay attention to high VIF variables (VIF>10), higher VIF indicates multicollinearity problem, not good for prediction model.
  - houseAge, BsmtQual_Gd, BsmtQual_TA, KitchenQual_Gd, KitchenQual_TA

Model Selection

Model	Description	Root mean square error (RMSE)
MLR	Multiple linear regression model	Not an appropriate model Due to non-constant variance, abnormality, high VIF
Ridge	Regularized Regression with all predictors and alpha = 0 and lambda = 10	44364.41
Lasso	Regularized Regression with all predictors and alpha = 1 and lambda = 10	52453.66
Elastic	combines **Lasso (L1)** and **Ridge (L2)** with all predictors and alpha = 0.1 and lambda = 39896.7	37954.48
Ridge with feature selection	Regularized Regression and Delete high VIF variables	55953.55
Lasso with feature selection	Regularized Regression and Delete high VIF variables	35596.42
Elastic with feature selection	combines **Lasso (L1)** and **Ridge (L2)** and Delete high VIF variables	34904.15
XGBoost	**Boosted tree model with all predictors**	**24604.36**
XGBoost with feature selection	Boosted tree model and Delete high VIF variables	26650.51

- Based on 10-fold cross validation

Model Selection

Final Model Information

- Algorithm: eXtreme Gradient Boosting (XGBoost)
- Predictors: all variables
- Validation: 10-fold cross validation
- Tunning grid:
  - learning rate (eta): 0.05, 0,10, 0.30
  - Max_depth: 3, 5, 7
  - Reiterate round (nrounds): 100, 200
  - Fixed parameters: Gamma=0, colsample_bytree=0.8, min_child_weight=1, subsample=0.8
- Best model:
  - parameters:

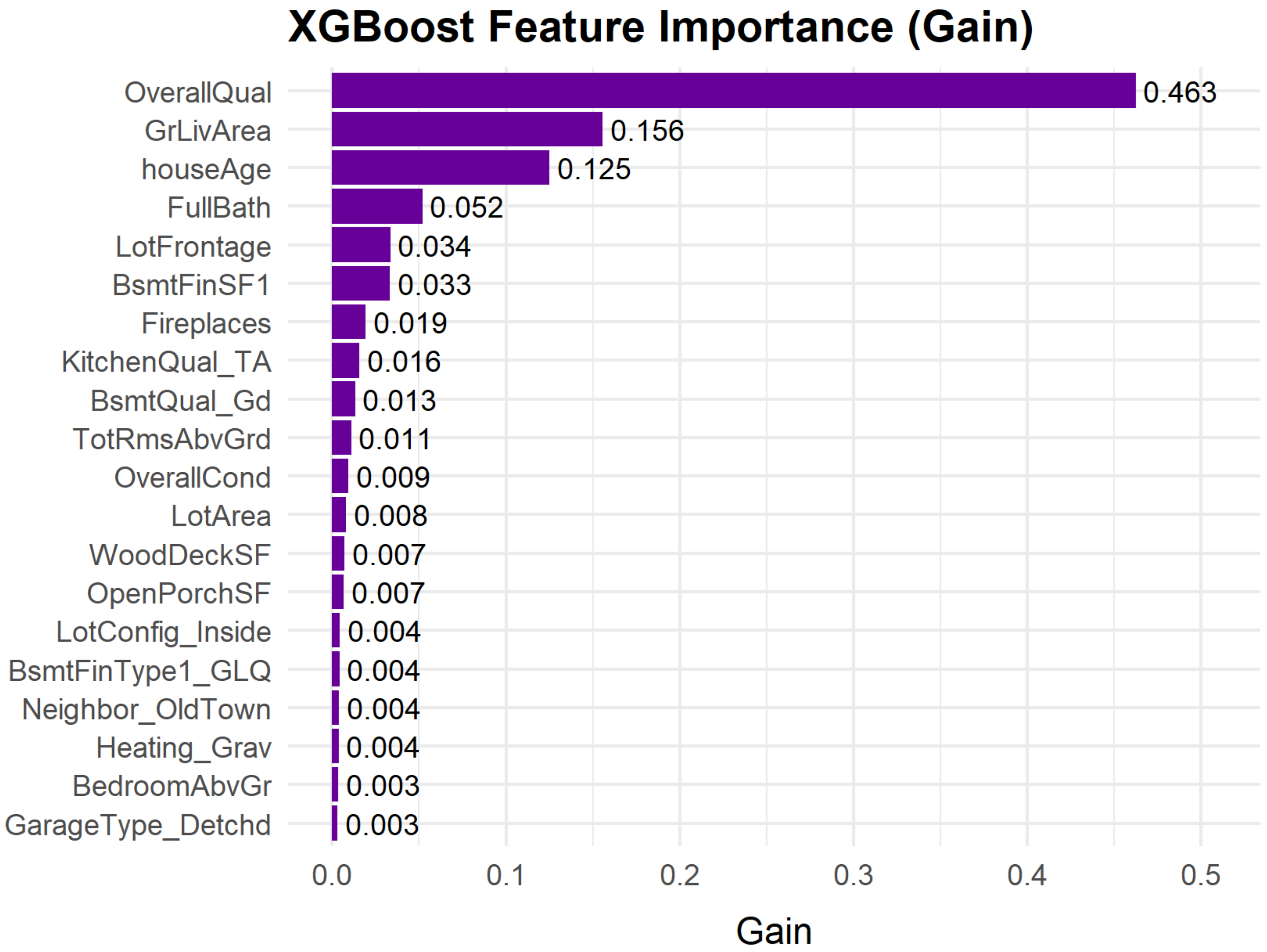
nrounds = 100
max_depth = 5
eta = 0.05
gamma = 0
colsample_bytree= 0.8
min_child_weight= 1
subsample = 0.8

⇨

- RMSE = 24604.36
- RMSE/Mean(salePrice)=15.57%
- R-Squared = 84.09%

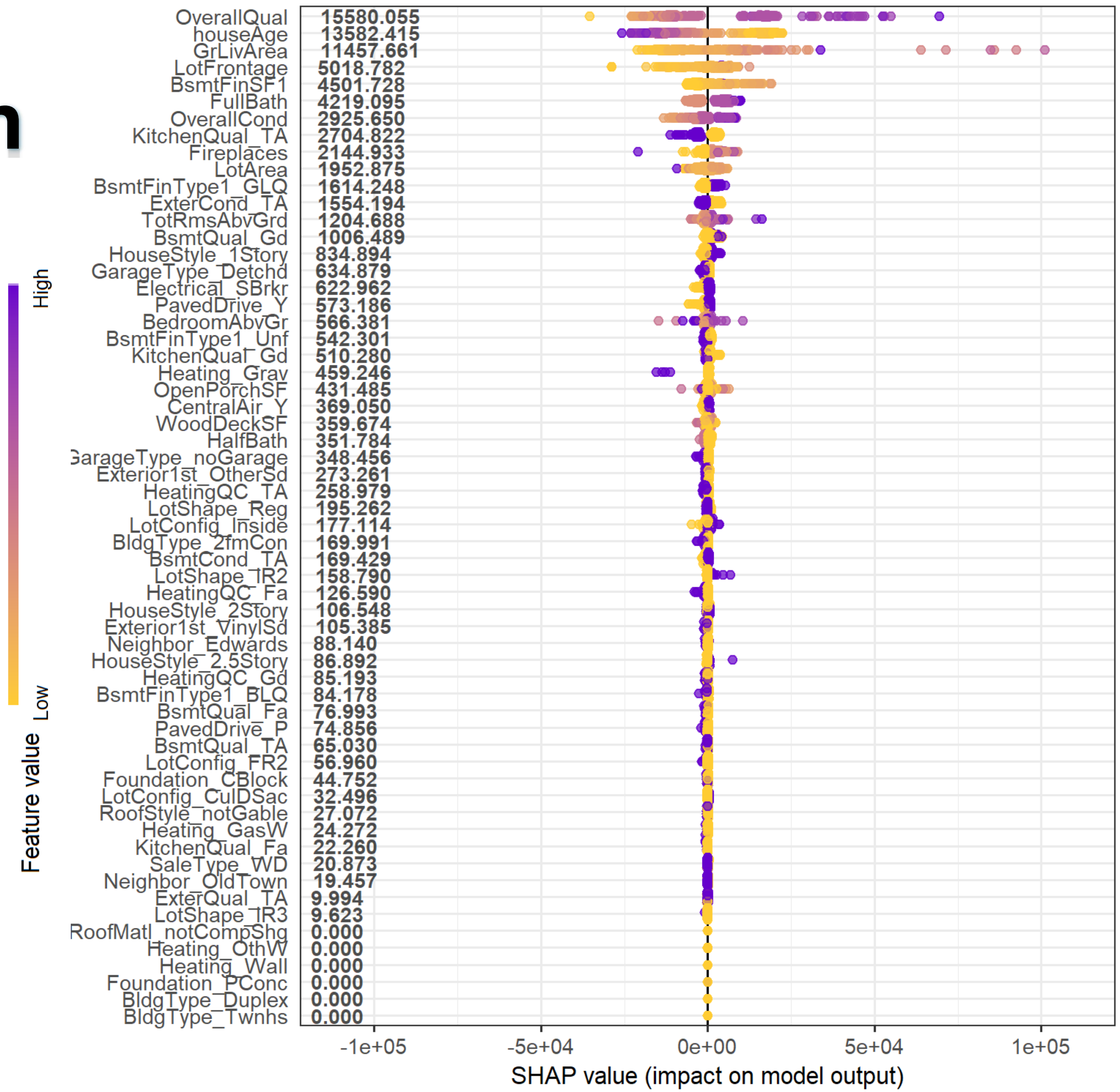
Model Interpretation

- The model identifies **Overall Quality (OverallQual)** as the most influential factor affecting house prices
- The second most important factor is **Above Ground Living Area (GrLivArea)**
- The third most important factor is **House age (houseAge)**
- Other variables, such as *full bathrooms above grade, linear feet of street frontage,and basement finish quality rating*, also contribute to price predictions, though to a lesser extent.



Model Interpretation

- **Overall Quality (OverallQual):** houses with higher quality ratings generally sell for higher prices.
- **Above Ground Living Area (GrLivArea):** larger living areas tend to be associated with higher prices.
- **House age (houseAge):** older houses generally having lower market values.



Model Prediction

Then we can use this model to predict the sale price of other houses

The prediction results as

uniqueID	SalePrice
House.1	281539.1
House.2	136949.4
House.3	128509.1
House.4	218124.2
House.5	208381.5
House.6	194810.9
House.7	121924.8
House.8	204269.4
House.9	265181.7
House.10	238017.8
House.11	223492.8
House.12	171497.8
House.13	282920.9
House.14	207321.7

… …

xgb_predictions.csv

House Price prediction

Pennsylvania state university

Yuqi Zeng

August 13, 2025

THANK YOU !