



Machine Learning Model: House Price Prediction

Team 8: Huiying Ba, Danting Huang, Ziqin Ma, Jiao Sun, Senbo Zhang

Data Preliminary

- ▶ The dataset describes every aspect of residential homes in Ames, Iowa.
- ▶ 1460 observations with 81 columns, including character variables and numeric variables.
- ▶ Without Id's, the dataframe consists of 79 predictors and the response variable `SalePrice`

Numeric

Scale
(Quality and
condition)

Type
(roof type,
alley)

Data Cleaning

Missing Value

- ▶ Numeric variables => column means
- ▶ Type variables => string "N/A"
- ▶ Scale variables => as a level of 0

Years

- ▶ Year Sold - Year Built = Age of property when sold
- ▶ Year Sold - Year Remodeled = Remodeled age when sold
- ▶ Year Sold - Garage Built = Garage of Age when dols

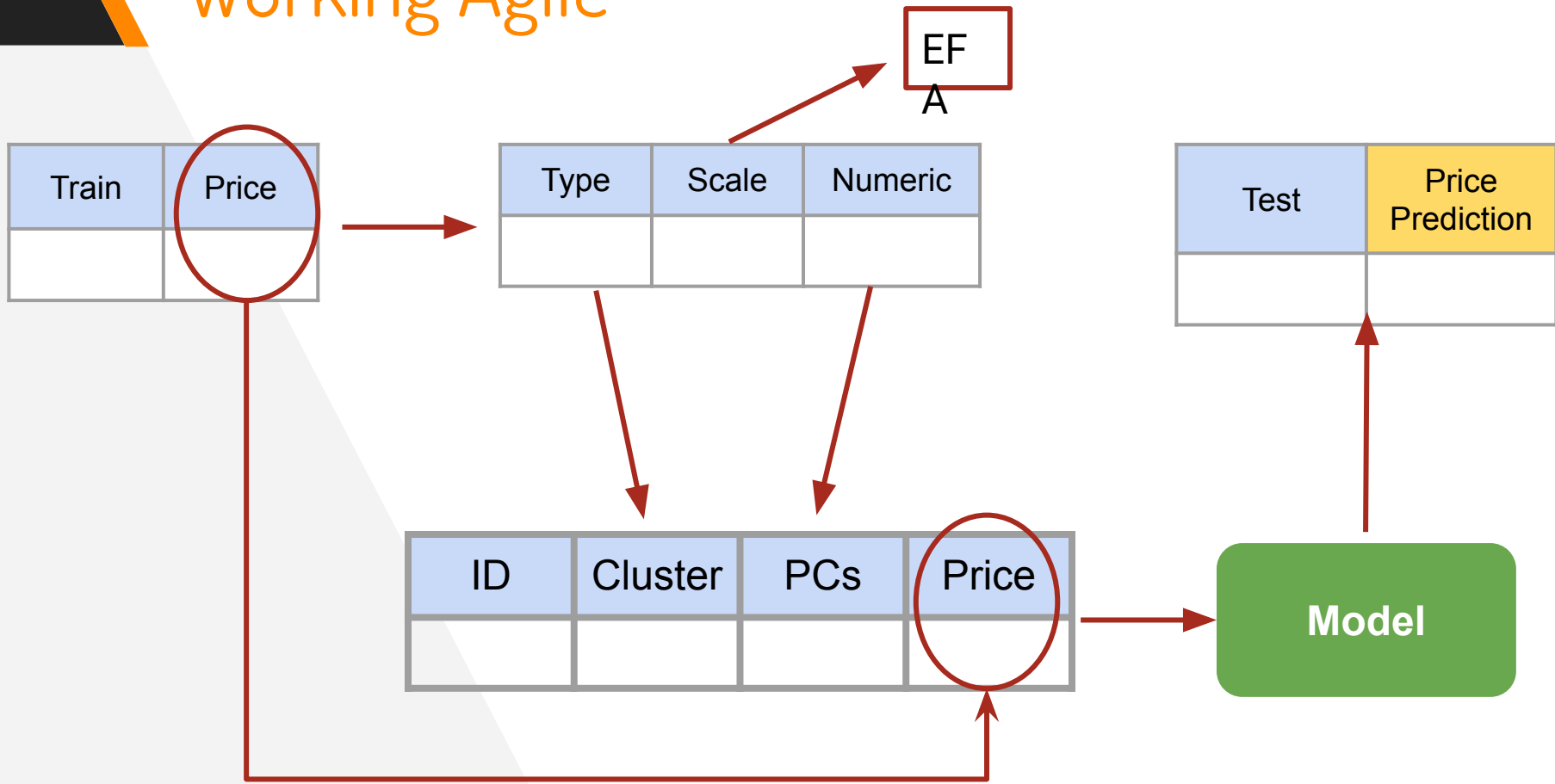
Months

- ▶ Converted to quarters (I, II, III, IV) as a categorical variable

Goal

- ▶ Combine unsupervised/supervised techniques on past dataset to build a model that can **estimate property prices** with certain features
- ▶ Provide information on **factors/aspects** that closely relates to the price in Ames, Iowa

Working Agile

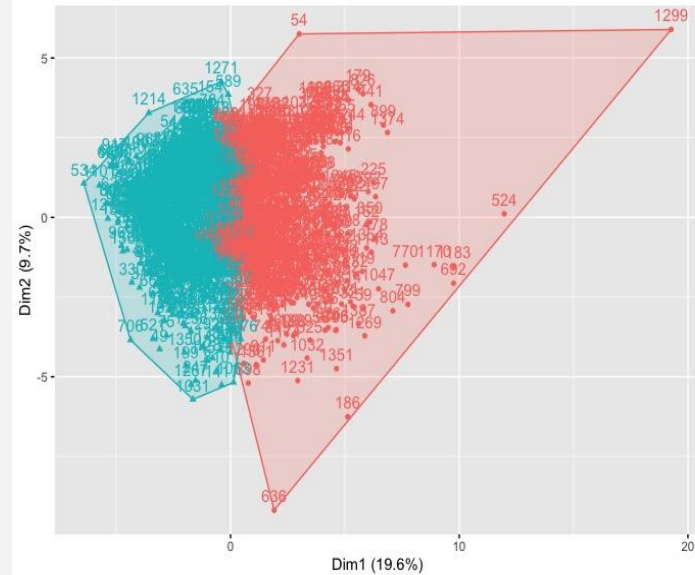


Modeling



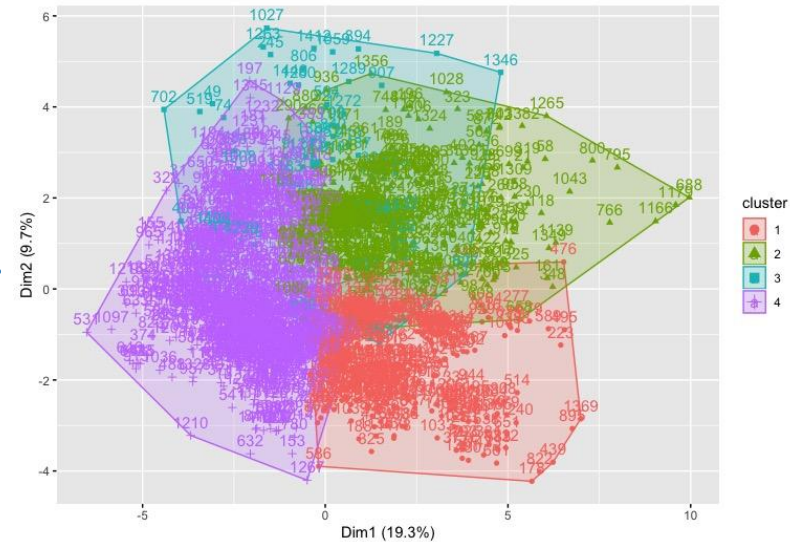
KMeans: identify outliers

Cluster plot



Removing
Outliers

Cluster plot



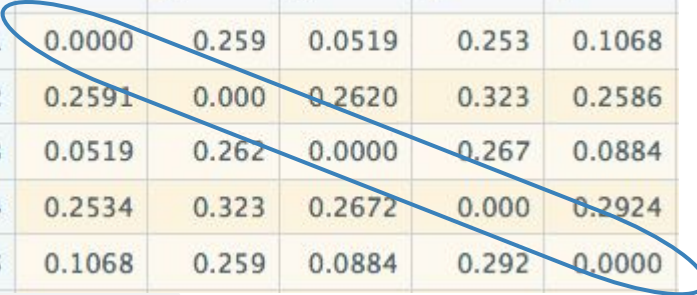
Will not use KMeans clusters to reduce dimensions

Partitioning using Gower Distance

Goal: want to reduce number of categorical columns

Gower Distance:

calculating distance in a mix of categorical and numerical variables

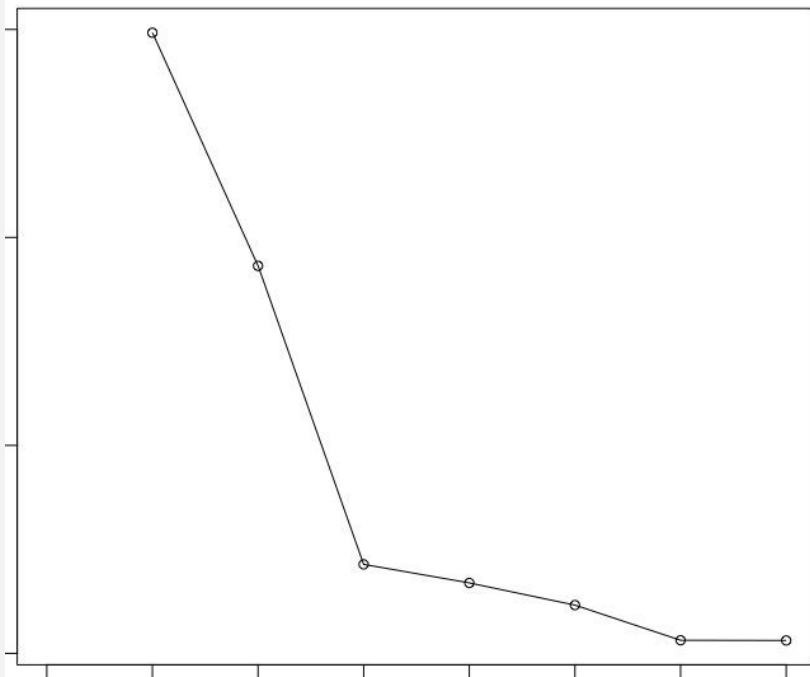


	1	2	3	4	5
1	0.0000	0.259	0.0519	0.253	0.1068
2	0.2591	0.000	0.2620	0.323	0.2586
3	0.0519	0.262	0.0000	0.267	0.0884
4	0.2534	0.323	0.2672	0.000	0.2924
5	0.1068	0.259	0.0884	0.292	0.0000

Dissimilarity matrix:

- Pairwise dissimilarities among observations using Gower distance
- 0's in diagonal

Partitioning using Gower Distance

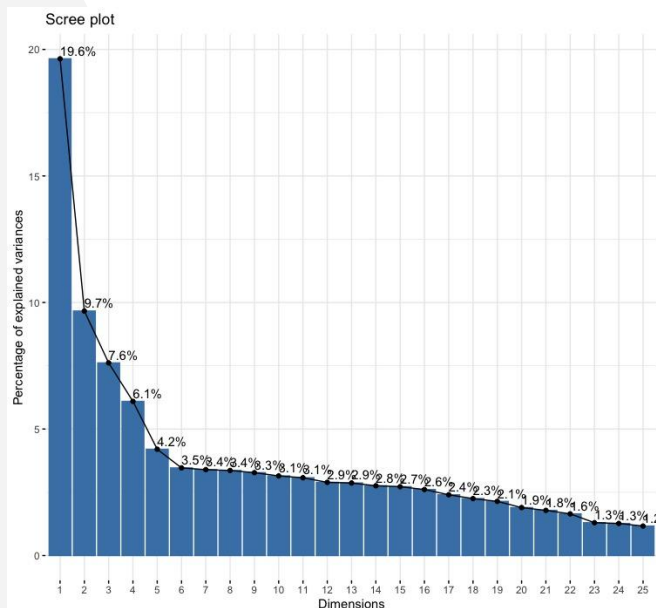


Ran a function to find the optimal k with the highest average silhouette score

Choose $k = 2$

Principal Component Analysis

Scree plot



Eigenvalues

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	6.47778897227242	19.6296635523407	19.6296635523407
Dim.2	3.18860384262738	9.66243588674963	29.2920994390903
Dim.3	2.51252161374115	7.61370185982166	36.905801298912
Dim.4	2.01060268904077	6.09273542133566	42.9985367202476
Dim.5	1.38667031585783	4.20203126017525	47.2005679804229
Dim.6	1.14298435699357	3.46358896058657	50.6641569410094
Dim.7	1.12057557244568	3.3956835528657	54.0598404938751
Dim.8	1.10945759702702	3.36199271826369	57.4218332121388
Dim.9	1.0811461485866	3.27620045026241	60.6980336624012
Dim.10	1.03864033451401	3.14739495307276	63.845428615474
Dim.11	1.01475591981658	3.07501793883812	66.9204465543121
Dim.12	0.95483768793424	2.89344753919467	69.8138940935068
Dim.13	0.94646348262937	2.86807115948294	72.6819652529897
Dim.14	0.910015318665588	2.75762217777451	75.4395874307642
Dim.15	0.898122791841415	2.72158421770126	78.1611716484655
Dim.16	0.861241090885593	2.6098214875321	80.7709931359976
Dim.17	0.792595877764342	2.40180569019497	83.1727988261926
Dim.18	0.742823760275179	2.25098109174297	85.4237799179355
Dim.19	0.706623734443835	2.1412840437692	87.5650639617047
Dim.20	0.625711021023537	1.89609400310163	89.4611579648064
Dim.21	0.589282934701082	1.78570586273055	91.2468638275369
Dim.22	0.543907540488141	1.64820466814588	92.8950684956828
Dim.23	0.428551119142284	1.29863975497662	94.1937082506594

92.9%

ID	PCs	Cluster	Price

Model Input

Validation

New dataframe for supervised models

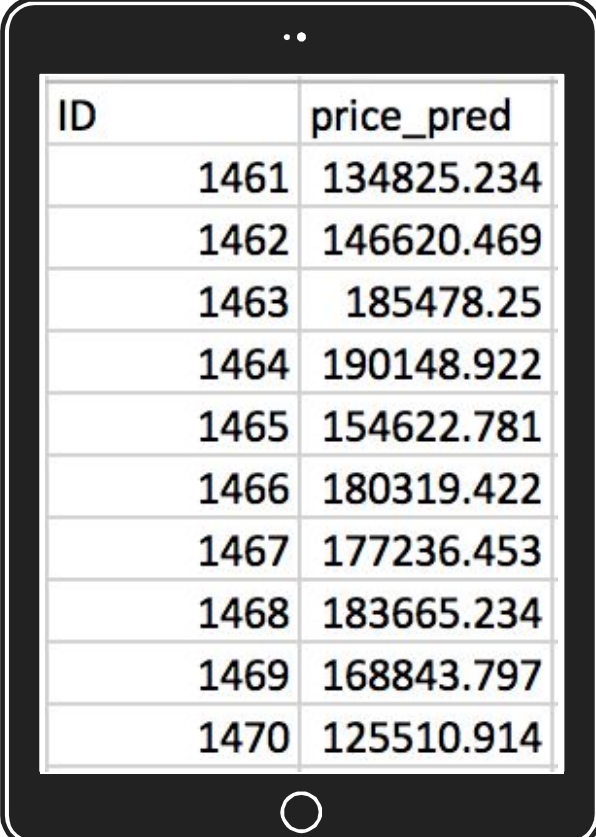
Supervised Techniques

- ▶ LASSO/Random Forest/XGBoost
- ▶ Compared MSE in validation sets
- ▶ XGboost: accuracy of 87.59%

LASSO	Random Forest	XGBoost
1,244,322,758	1,136,864,724	854,295,790

Running model on future data

- Input:
Original test dataset with 79 variables, no price column
- Output:
ID + Price estimate



A tablet device is shown, displaying a table with two columns: 'ID' and 'price_pred'. The table contains 10 rows of data, representing predicted prices for specific IDs. The tablet has a black bezel and a white home button at the bottom.

ID	price_pred
1461	134825.234
1462	146620.469
1463	185478.25
1464	190148.922
1465	154622.781
1466	180319.422
1467	177236.453
1468	183665.234
1469	168843.797
1470	125510.914

Inference: PC1 from LASSO

Age

- ▶ House
- ▶ Garage

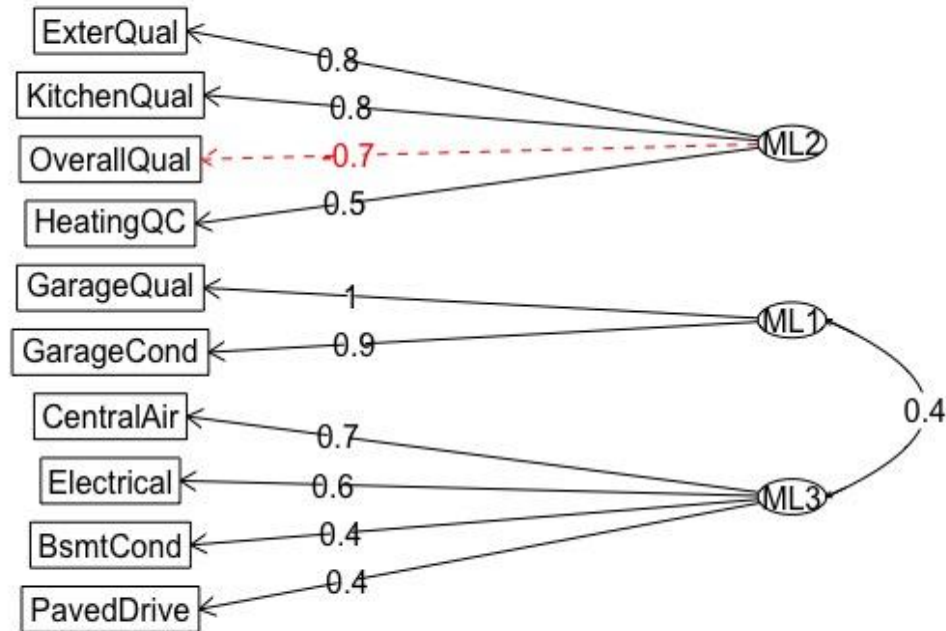
Square feet

- ▶ Living area
- ▶ 1st floor
- ▶ Garage
- ▶ Basement

- ### Number of Rooms/facilities
- ▶ Bedrooms
 - ▶ Fireplaces
 - ▶ Full bathrooms
 - ▶ Other rooms

Inference: Factor Analysis

Factor Analysis



Qualities

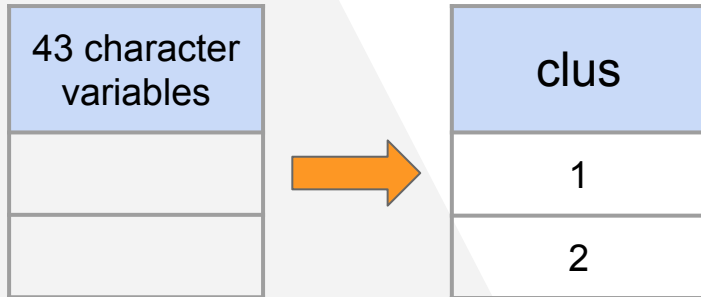
Garage

Amenities

Discussion

Limitations:

- Loss of interpretability

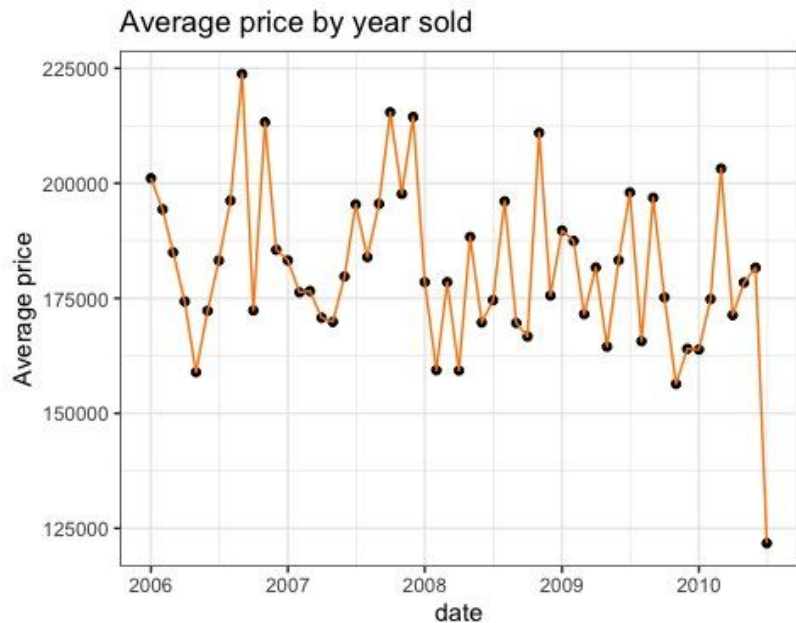
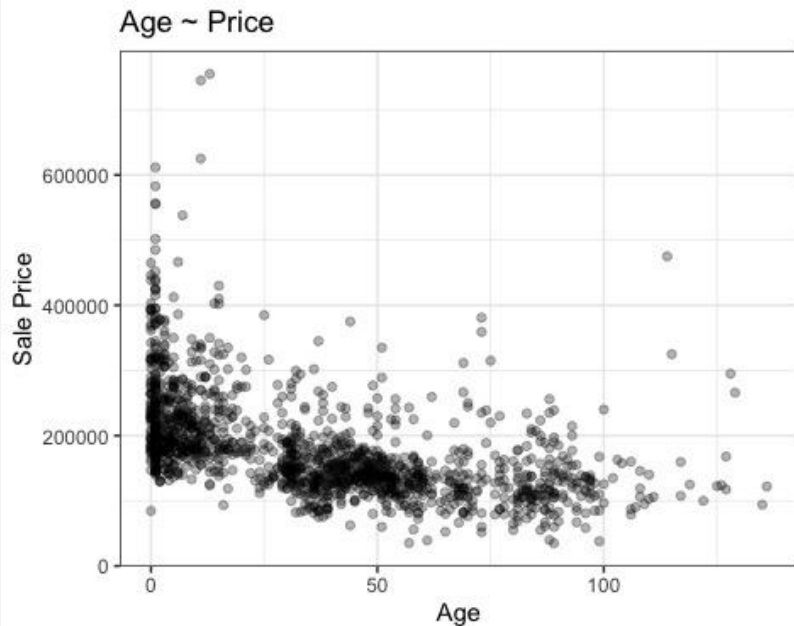


- Type variables aggregation
Alley access to property,
Paved driveway

★ Better approach with geographic data

Evaluating Years

We parsed Year Sold into Age of property when sold



Applications

- ▶ Sellers & Buyers

Price debrief of properties
with similar features

- ▶ Real estate agencies

Pricing constructions
and significant factors

THANKS!

Any questions?

Github Repository:

<https://github.com/ziginm/BA-820-Project.git>