House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

OBJECTIVE

The objective of this project is to practice advanced regression techniques.

DESIGN CONSIDERATIONS

- **Programming Language:** Python 3.5.
- **IDE:** Spyder, Jupyter.
- **Packages:** Numpy, Pandas, Sklearn, Json, Xgboost, Matplotlib, Scipy.

DATA STRUCTURE

List, Dictionary, Array, Dataframe.

ABOUT DATA

- This data set was constructed for the purpose of an end of semester project for an undergraduate regression course. The original data (obtained directly from the Ames Assessor's Office) is used for tax assessment purposes but lends itself directly to the prediction of home selling prices.
- Data Set is describing the sale of individual property in Ames, Iowa from year 2006 to 2010.
- The data set contains 2930 observations & 82 variables.
- A large number of explanatory variables (23 nominal, 23 ordinals, 14 discrete and 20 continuous) are involved in assessing home values.
- 20 continuous variables relate to various area dimensions for each observation. Like: Area measurements on the basement, main living area etc.
- The 14 discrete variables typically quantify the number of items occurring within the house. Like: number of kitchens, bedrooms, and bathrooms etc.

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

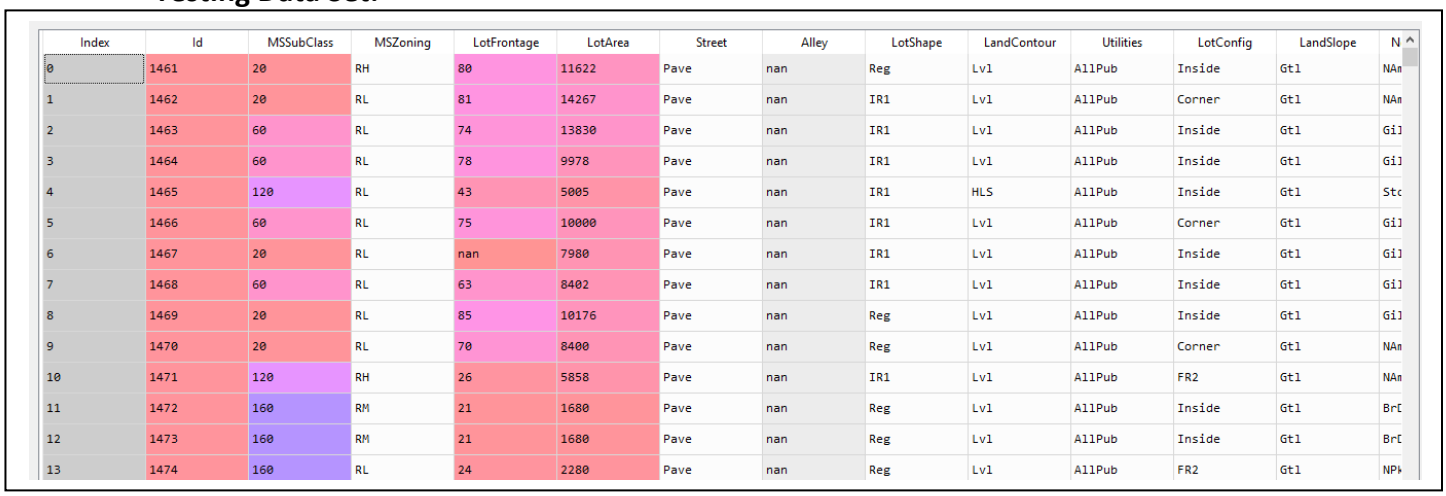
- There are a large number of categorical variables (23 nominal, 23 ordinal) associated with this data set. Like: types of dwellings, garages etc.

Training Data Set:

train_df - DataFrame

Index	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotCo
0	1	60	RL	65	8450	Pave	nan	Reg	Lvl	AllPub	Inside
1	2	20	RL	80	9600	Pave	nan	Reg	Lvl	AllPub	FR2
2	3	60	RL	68	11250	Pave	nan	IR1	Lvl	AllPub	Inside
3	4	70	RL	60	9550	Pave	nan	IR1	Lvl	AllPub	Corner
4	5	60	RL	84	14260	Pave	nan	IR1	Lvl	AllPub	FR2
5	6	50	RL	85	14115	Pave	nan	IR1	Lvl	AllPub	Inside
6	7	20	RL	75	10084	Pave	nan	Reg	Lvl	AllPub	Inside
7	8	60	RL	nan	10382	Pave	nan	IR1	Lvl	AllPub	Corner
8	9	50	RM	51	6120	Pave	nan	Reg	Lvl	AllPub	Inside
9	10	190	RL	50	7420	Pave	nan	Reg	Lvl	AllPub	Corner
10	11	20	RL	70	11200	Pave	nan	Reg	Lvl	AllPub	Inside
11	12	60	RL	85	11924	Pave	nan	IR1	Lvl	AllPub	Inside
12	13	20	RL	nan	12968	Pave	nan	IR2	Lvl	AllPub	Inside

Testing Data Set:



Index	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	N
0	1461	20	RH	80	11622	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	NAm
1	1462	20	RL	81	14267	Pave	nan	IR1	Lvl	AllPub	Corner	Gtl	NAm
2	1463	60	RL	74	13830	Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	Gil
3	1464	60	RL	78	9978	Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	Gil
4	1465	120	RL	43	5005	Pave	nan	IR1	HLS	AllPub	Inside	Gtl	Sto
5	1466	60	RL	75	10000	Pave	nan	IR1	Lvl	AllPub	Corner	Gtl	Gil
6	1467	20	RL	nan	7980	Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	Gil
7	1468	60	RL	63	8402	Pave	nan	IR1	Lvl	AllPub	Inside	Gtl	Gil
8	1469	20	RL	85	10176	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Gil
9	1470	20	RL	70	8400	Pave	nan	Reg	Lvl	AllPub	Corner	Gtl	NAm
10	1471	120	RH	26	5858	Pave	nan	IR1	Lvl	AllPub	FR2	Gtl	NAm
11	1472	160	RM	21	1680	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Brl
12	1473	160	RM	21	1680	Pave	nan	Reg	Lvl	AllPub	Inside	Gtl	Brl
13	1474	160	RL	24	2280	Pave	nan	Reg	Lvl	AllPub	FR2	Gtl	NPk

ABOUT PARAMETERS

- **Order (Discrete):** Observation number.

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

- **PID (Nominal):** Parcel identification number - can be used with city web site for parcel review.
- **MS SubClass (Nominal):** Identifies the type of dwelling involved in the sale.
- **MS Zoning (Nominal):** Identifies the general zoning classification of the sale.
- **Lot Frontage (Continuous):** Linear feet of street connected to property.
- **Lot Area (Continuous):** Lot size in square feet.
- **Street (Nominal):** Type of road access to property.
- **Alley (Nominal):** Type of alley access to property.
- **Lot Shape (Ordinal):** General shape of property.
- **Land Contour (Nominal):** Flatness of the property.
- **Utilities (Ordinal):** Type of utilities available.
- **Lot Config (Nominal):** Lot configuration.
- **Land Slope (Ordinal):** Slope of property.
- **Neighborhood (Nominal):** Physical locations within Ames city limits (map available).
- **Condition 1 (Nominal):** Proximity to various conditions.
- **Condition 2 (Nominal):** Proximity to various conditions (if more than one is present).
- **Bldg Type (Nominal):** Type of dwelling.
- **House Style (Nominal):** Style of dwelling.
- **Overall Qual (Ordinal):** Rates the overall material and finish of the house.
- **Overall Cond (Ordinal):** Rates the overall condition of the house.

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

- **Year Built (Discrete):** Original construction date.
- **Year Remod/Add (Discrete):** Remodel date (same as construction date if no remodelling or additions).
- **Roof Style (Nominal):** Type of roof.
- **Roof Matl (Nominal):** Roof material.
- **Exterior 1 (Nominal):** Exterior covering on house.
- **Exterior 2 (Nominal):** Exterior covering on house (if more than one material).
- **Mas Vnr Type (Nominal):** Masonry veneer type.
- **Mas Vnr Area (Continuous):** Masonry veneer area in square feet.
- **Exter Qual (Ordinal):** Evaluates the quality of the material on the exterior.
- **Exter Cond (Ordinal):** Evaluates the present condition of the material on the exterior.
- **Foundation (Nominal):** Type of foundation.
- **Bsmt Qual (Ordinal):** Evaluates the height of the basement.
- **Bsmt Cond (Ordinal):** Evaluates the general condition of the basement.
- **Bsmt Exposure (Ordinal):** Refers to walkout or garden level walls.
- **BsmtFin Type 1 (Ordinal):** Rating of basement finished area.
- **BsmtFin SF 1 (Continuous):** Type 1 finished square feet.
- **BsmtFinType 2 (Ordinal):** Rating of basement finished area (if multiple types).
- **BsmtFin SF 2 (Continuous):** Type 2 finished square feet.
- **Bsmt Unf SF (Continuous):** Unfinished square feet of basement area.

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

- **Total Bsmt SF (Continuous):** Total square feet of basement area.
- **Heating (Nominal):** Type of heating.
- **HeatingQC (Ordinal):** Heating quality and condition.
- **Central Air (Nominal):** Central air conditioning.
- **Electrical (Ordinal):** Electrical system.
- **1st Flr SF (Continuous):** First Floor square feet.
- **2nd Flr SF (Continuous):** Second floor square feet.
- **Low Qual Fin SF (Continuous):** Low quality finished square feet (all floors).
- **Gr Liv Area (Continuous):** Above grade (ground) living area square feet.
- **Bsmt Full Bath (Discrete):** Basement full bathrooms.
- **Bsmt Half Bath (Discrete):** Basement half bathrooms.
- **Full Bath (Discrete):** Full bathrooms above grade.
- **Half Bath (Discrete):** Half baths above grade.
- **Bedroom (Discrete):** Bedrooms above grade (does NOT include basement bedrooms).
- **Kitchen (Discrete):** Kitchens above grade.
- **KitchenQual (Ordinal):** Kitchen quality.
- **TotRmsAbvGrd (Discrete):** Total rooms above grade (does not include bathrooms).
- **Functional (Ordinal):** Home functionality (Assume typical unless deductions are warranted).
- **Fireplaces (Discrete):** Number of fireplaces.

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

- **FireplaceQu (Ordinal):** Fireplace quality.
- **Garage Type (Nominal):** Garage location.
- **Garage Yr Blt (Discrete):** Year garage was built.
- **Garage Finish (Ordinal):** Interior finish of the garage.
- **Garage Cars (Discrete):** Size of garage in car capacity.
- **Garage Area (Continuous):** Size of garage in square feet.
- **Garage Qual (Ordinal):** Garage quality.
- **Garage Cond (Ordinal):** Garage condition.
- **Paved Drive (Ordinal):** Paved driveway.
- **Wood Deck SF (Continuous):** Wood deck area in square feet.
- **Open Porch SF (Continuous):** Open porch area in square feet.
- **Enclosed Porch (Continuous):** Enclosed porch area in square feet.
- **3-Ssn Porch (Continuous):** Three season porch area in square feet.
- **Screen Porch (Continuous):** Screen porch area in square feet.
- **Pool Area (Continuous):** Pool area in square feet.
- **Pool QC (Ordinal):** Pool quality.
- **Fence (Ordinal):** Fence quality.
- **Misc Feature (Nominal):** Miscellaneous feature not covered in other categories.
- **Misc Val (Continuous):** $Value of miscellaneous feature.
- **Mo Sold (Discrete):** Month Sold (MM).

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

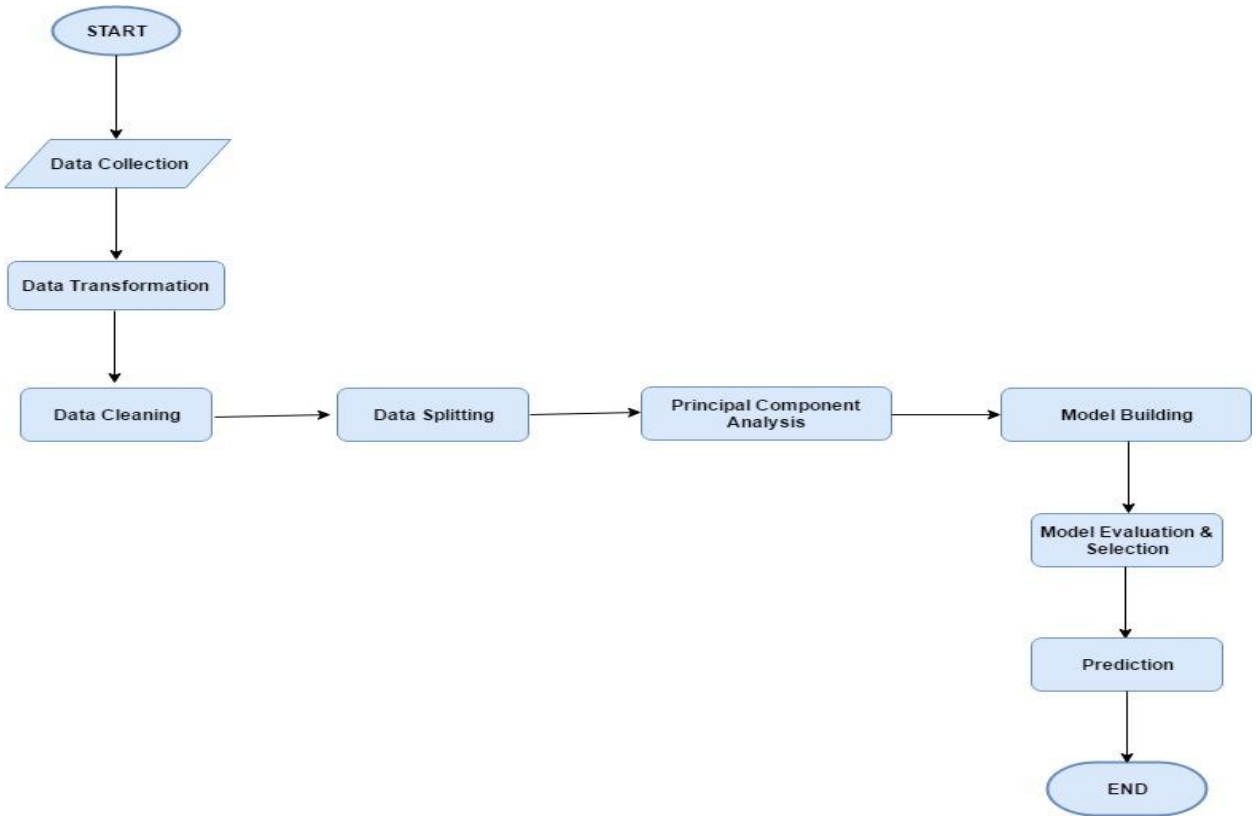
By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

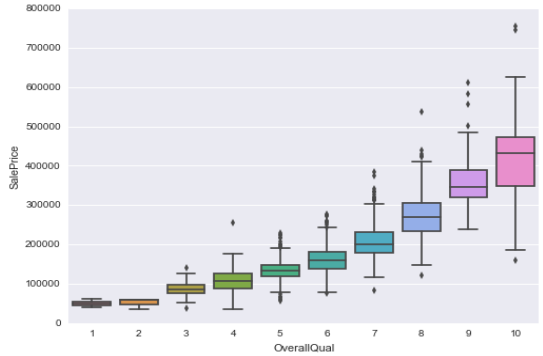
- **Yr Sold (Discrete):** Year Sold (YYYY).
- **Sale Type (Nominal):** Type of sale.
- **Sale Condition (Nominal):** Condition of sale.
- **SalePrice (Continuous):** Sale price $$.

ALGORITHM

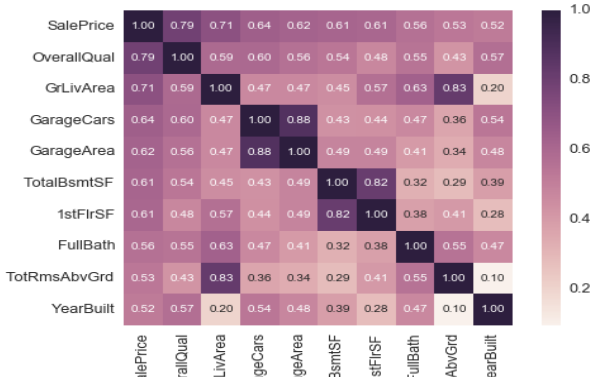


DATA EXPLORATION

Sales Price & Overall Quality:



Correlation Matrix:



House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

- **Sales Price:**

df_train['SalePrice'].describe()

Out[4]:

count 1460.000000

mean 180921.195890

std 79442.502883

min 34900.000000

25% 129975.000000

50% 163000.000000

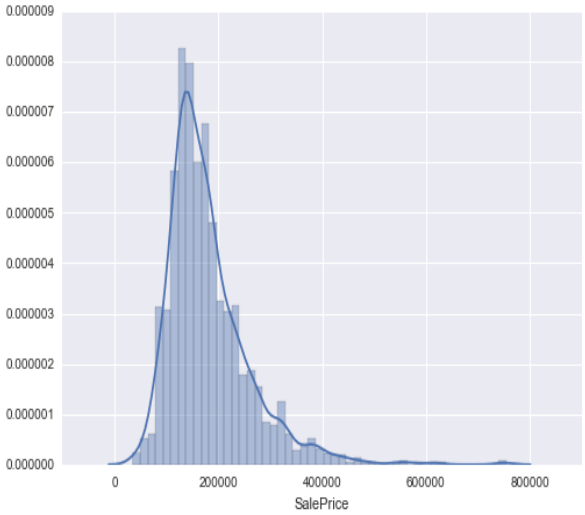
75% 214000.000000

max 755000.000000

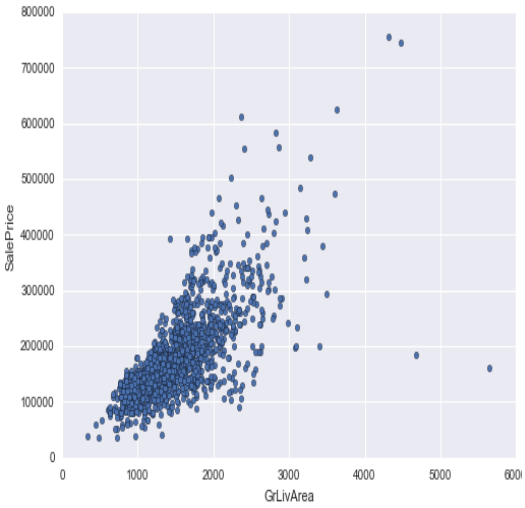
Name: SalePrice, dtype: float64

#skewness and kurtosis

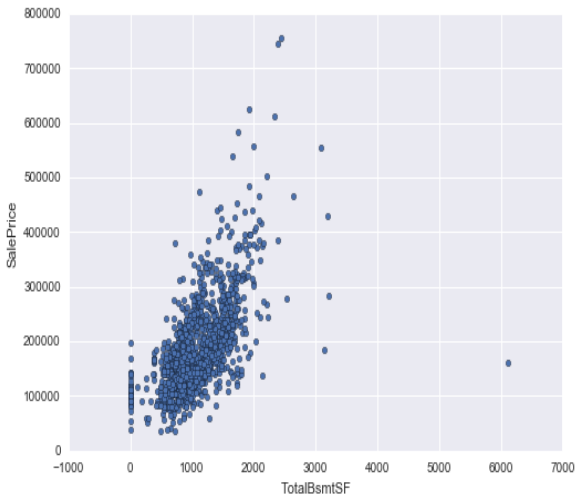
Skewness: 1.882876 **Kurtosis:** 6.536282



Sales Price & Living Area Square Feet:



Sales Price & Total Basement Area:



MODEL EVALUATION

Accuracy Score of Models:

```
XGBoost score is 0.788796
Ridge score is 0.766273
Lasso score is 0.764264
Linear Regression score is 0.449918
```

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering

By

Aakash Parwani

DS-680: Marketing Analytics, Saint Peter's University, Jersey City, NJ

Root Mean Square Error of Models:

[' Root mean squared error of lasso is: 0.14593883217',

'Root mean square error of rigde: 0.1392427785',

'Root mean square error of randomforest: 0.157068919051',

'Root mean square error of xboost: 0.128156202379']

MODEL SELECTION:

From model evaluation section it is clear that **Xboost & Ridge** models are performing well**.**

Finally, **Xboost** model has been selected for the prediction of sales price on test data set.

Id	SalePrice
1461	116983.1295
1462	153552.7201
1463	185779.9498
1464	202041.9957
1465	193216.7502
1466	170346.912
1467	175506.1006
1468	161380.1524
1469	187963.524
1470	119402.2029
1471	188773.8889
1472	94940.38675
1473	93876.59373
1474	144017.4029
1475	113980.3212
1476	362268.0075
1477	241285.3421
1478	285767.1012