Project 2

Prediction of Ames Housing Data

Fnu Parshant

CONTENTS:

Audience: Real Estate Agents

Problem Statement

- The **aim** of this project is to predict property sales prices for each house using data processing.
- To predict the sale price of the property, I used Linear Regression Model.

- Linear Regression: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.

- https://medium.com/@srishtisawla/linear-regression-data-science-algorithm-every-data-scientist-should-know-34d5fcb51c03

WORKFLOW:

1) Data Gathering.
2) Fixed all null values.
3) Dropped low correlated features.
4) Saved cleaned datasets into new .csv files.
5) Joined both cleaned train-test csv files.
6) Used dummy variables and feature engineering.
7) Scaled the data.
8) Split the data to run different models like Linear Regression, Ridge, Lasso.
9) Compared RMSE of all the models and found out the LR has the lowest RMSE.
10) At the end predicted the SalesPrice using LR model.

Brief Summary

According to Mateus Realty:

Realtors determine house values by looking at various features. Some of them are location of the property, condition of the property, year built, square feet area, heating-AC, garage type, total number of bedrooms and bathrooms, and many more. After this, they will look for similar houses in the neighborhood that were sold recently to get a better estimate of the property.

Data Cleaning and EDA

1. Replace null values

a. Understanding the Data Dictionary provided on Kaggle

b. Most null values correspond to ‘NA’

c. Cannot drop all rows .

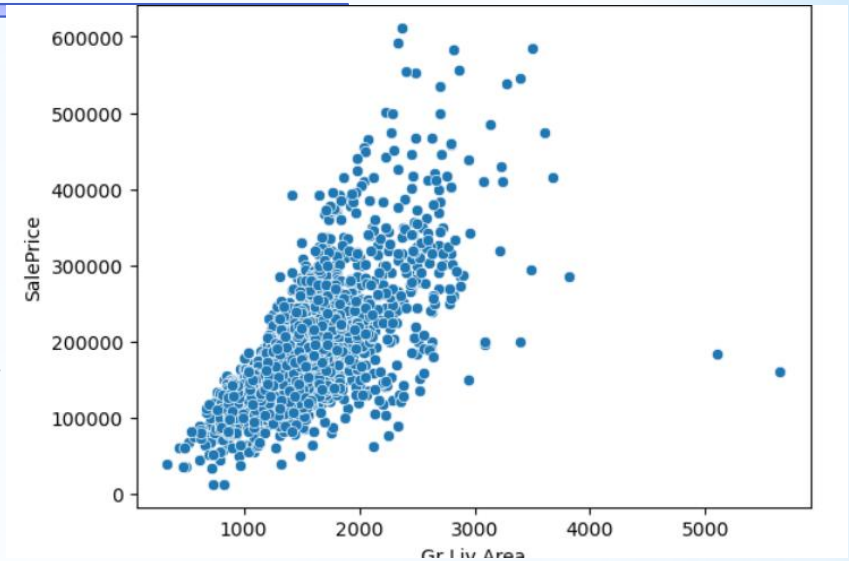
2. Removing outliers

3. Scale numerical values

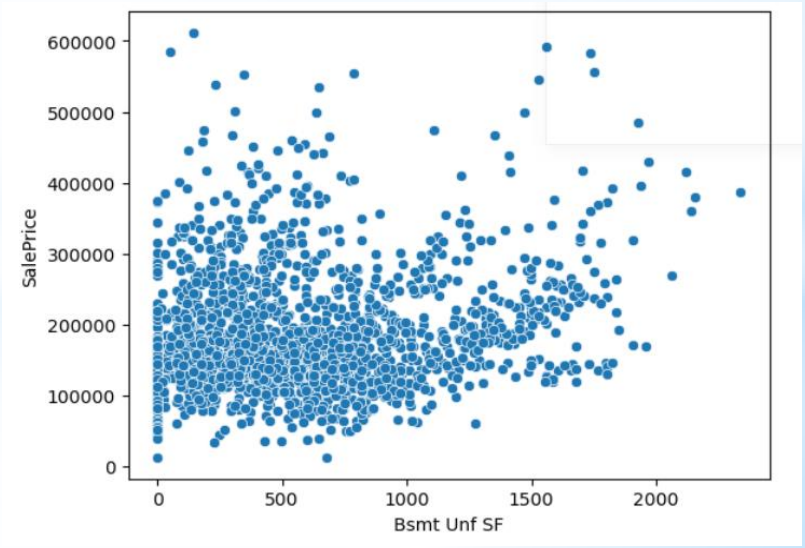
	Id	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	...	Screen Porch	Pool Area	Pool QC	Fence	Misc Feature	M
0	109	533352170	60	RL	NaN	13517	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	NaN	
1	544	531379050	60	RL	43.0	11492	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	NaN	
2	153	535304180	20	RL	68.0	7922	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	
3	318	916386060	60	RL	73.0	9802	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	
4	255	906425045	50	RL	82.0	14235	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	NaN	

Feature Selection

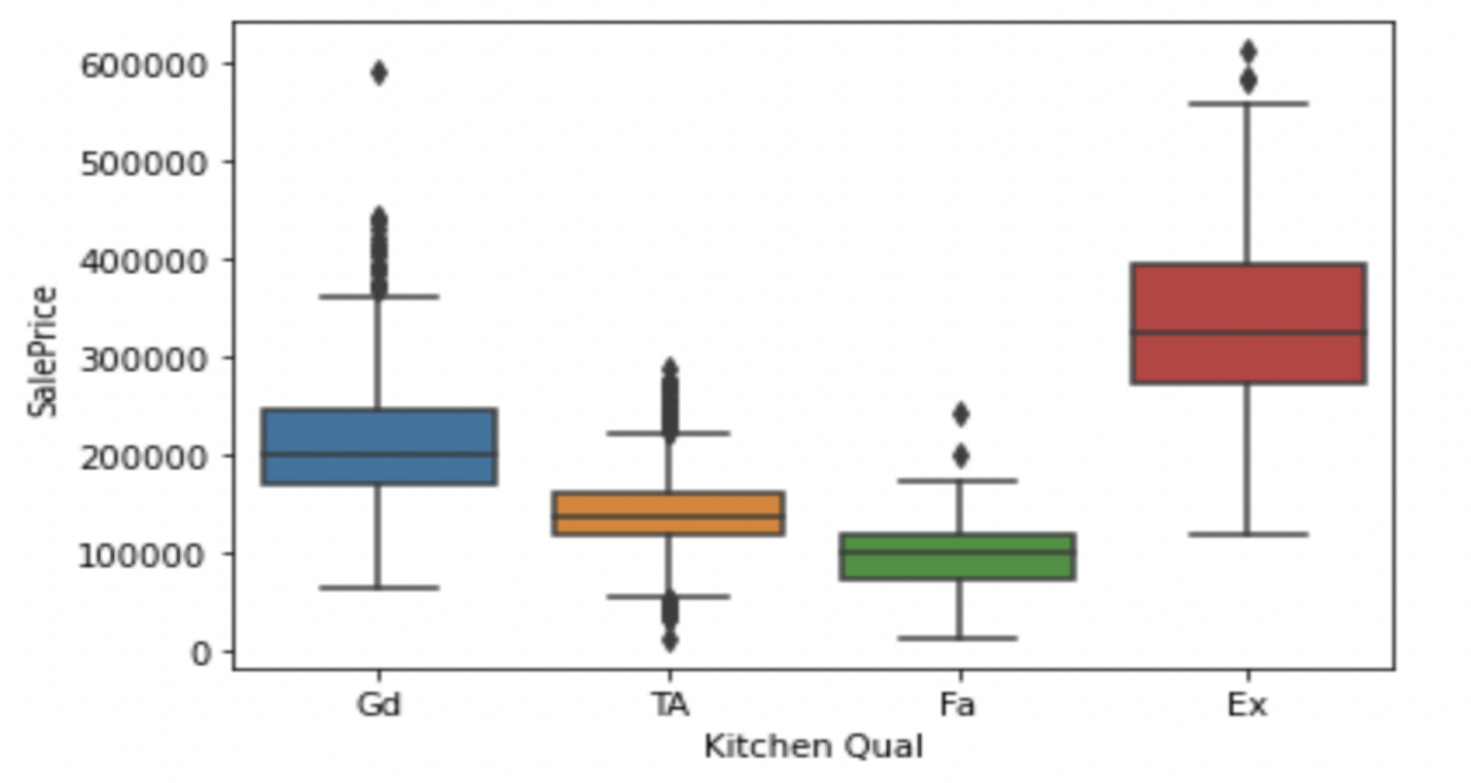
1. Features with higher correlations to saleprice were chosen.
For ex: Ground Level Area.



2. Features with weak correlations to saleprice were dropped.
For ex: Unfinished square Feet in basement.



Feature Selection



- I used boxplots (from Python) to correlate all other features with the Sales Price. BoxPlot draws a box plot to show distributions with respect to categories.

Preprocessing

- Convert Categorial features into dummies

	Year Built	Year Remod/Add	Mas Vnr Area	BsmtFin SF 1	Total Bsmt SF	Gr Liv Area	Full Bath	Half Bath	TotRms AbvGrd	Fireplaces	...	Paved Drive_Y	Sale Type_CWD	Sale Type_Con	Sale Type_ConLD	Sale Type_ConLI	Type_C
0	1976	2005	289.0	533.0	725.0	1479	2	1	6	0	...	1	0	0	0	0	
1	1996	1997	132.0	637.0	913.0	2122	2	1	8	1	...	1	0	0	0	0	
2	1953	2007	0.0	731.0	1057.0	1057	1	0	5	0	...	1	0	0	0	0	
3	2006	2007	0.0	0.0	384.0	1444	2	1	7	0	...	1	0	0	0	0	
4	1900	1993	0.0	0.0	676.0	1445	2	0	6	0	...	0	0	0	0	0	

- Scaling

	Year Built	Year Remod/Add	Mas Vnr Area	BsmtFin SF 1	Total Bsmt SF	Gr Liv Area	Full Bath	Half Bath	TotRms AbvGrd	Fireplaces	...	Paved Drive_Y	Sale Type_CWD	Typ
0	0.153973	0.994528	1.055637	0.200601	-0.741084	-0.039049	0.785682	1.235207	-0.281050	-0.924835	...	0.323830	-0.064139	-0.0
1	0.815388	0.610939	0.174708	0.429501	-0.313110	1.241600	0.785682	1.235207	0.992261	0.618664	...	0.323830	-0.064139	-0.0
2	-0.606655	1.090425	-0.565946	0.636391	0.014699	-0.879537	-1.024910	-0.754849	-0.917705	-0.924835	...	0.323830	-0.064139	-0.0
3	1.146096	1.090425	-0.565946	-0.972512	-1.517355	-0.108758	0.785682	1.235207	0.355605	-0.924835	...	0.323830	-0.064139	-0.0
4	-2.359406	0.419145	-0.565946	-0.972512	-0.852630	-0.106766	0.785682	-0.754849	-0.281050	-0.924835	...	-3.088037	-0.064139	-0.0

- Tain test split

Modeling

RMSE

- Linear Regression: 26262.833
- LASSO : 26461.007
- Ridge : 26286.375

-Note: Ridge and LASSO are just modifications of the LR model.

* The linear regression model has the smallest **RMSE**.
* **RMSE** is the square root of MSE (Mean squared error).
* **MSE** is the average of the squared residuals.
* Residuals = y_test – y_preds
* **Low RMSE** means better fit.

Interpretation of Results

	coefficient
Kitchen Qual_Po	4.625108e+15
Sale Type_VWD	1.082276e+13
Gr Liv Area	2.212819e+04
Overall Qual_9	1.722132e+04
Overall Qual_8	1.666925e+04
...	...
MS SubClass_160.0	-8.556375e+03
Exter Qual_Gd	-8.814951e+03
Kitchen Qual_Gd	-1.022183e+04
Exter Qual_TA	-1.169633e+04
Kitchen Qual_TA	-1.256738e+04

Out of the given 80 features, following **32 features** are highly correlated with the Sales Price of a property :
['MS SubClass', 'MS Zoning', 'Street', 'Land Contour',
'Neighborhood', 'Condition 1', 'Overall Qual', 'Year Built',
'Year Remod/Add', 'Mas Vnr Area', 'Exter Qual', 'Exter Cond',
'Foundation', 'BsmtFin SF 1', 'Total Bsmt SF', 'Heating QC',
'Central Air', 'Electrical', 'Gr Liv Area', 'Full Bath', 'Half Bath',
'Kitchen Qual', 'TotRms AbvGrd',
'Functional', 'Fireplaces',
'Garage Type', 'Garage Finish', 'Garage Cars', 'Garage Area',
'Paved Drive', 'Sale Type', 'SalePrice']

Conclusions and Recommendations

* Area of the grade (ground) living area appear to add the most value to a home price.
* **Kitchen Qual_Po** has the largest negative coefficient. This tells us poor Kitchen quality hurt the value of a home the most.
* **Main Recommendation:** To correctly predict the SalesPrice of a property you all (real estate agents) should focus on the above mentioned 32 features (in the Findings). In order to increase the saleprice of a house try to improve **Gr Liv Area, overall quality of home, increase car capacity of garage, increase Type 1 finished area and Kitchen Qual_Po quality,** as you all know that with the increase in sale price your profit will increase too.
* **Note:** In New York State, the standard real estate agent commission is 5- 6% of the property price.

Thank You!