HOUSING PRICE PREDICTIONS

Deepshika Sharma

PROBLEM STATEMENT

Identifying features helpful in predicting housing prices and using a model to give us the best r2 score

DATASET

- Housing prices from Ames
- The training set had more than 2000 observations and 80 features
- Some observations: Fireplaces, Year Built, Garage Cars, Kitchen Quality, etc.
- The validation set had less than 900 observations
- The model has to be evaluated on the training data and then tested on the testing data
- Predictions are made and then it is tested against the dataset from Kaggle which calculates the MSE

WHERE DO WE START? STEP 1

DATA CLEANING

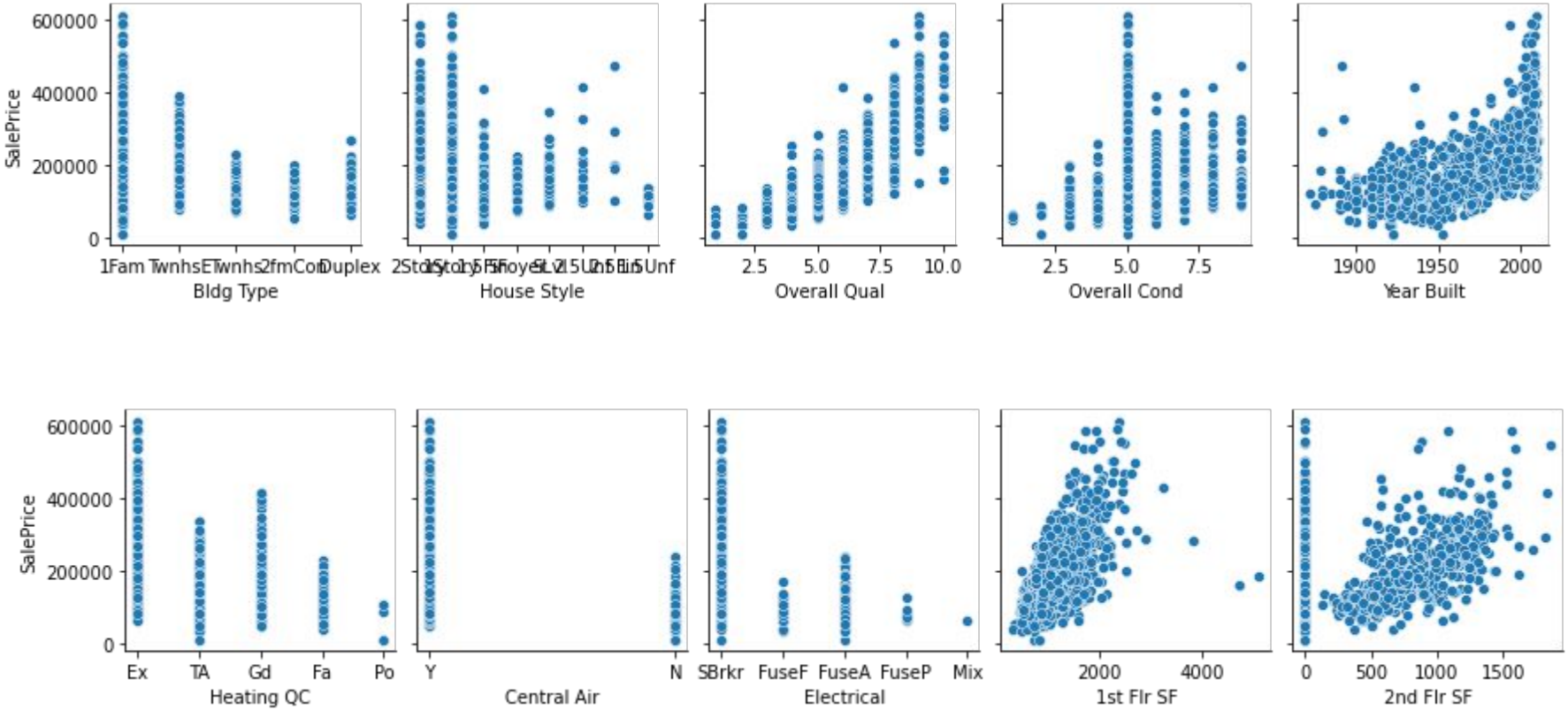
- Cleaning the training and validation sets.
- Making sure they had the same number of features
- Deciding what to do with null values, replace, remove or just leave it.

STEP 2: FEATURE ENGINEERING

- Lots of features that were numerical but could be classified as categorical
- Dummify columns

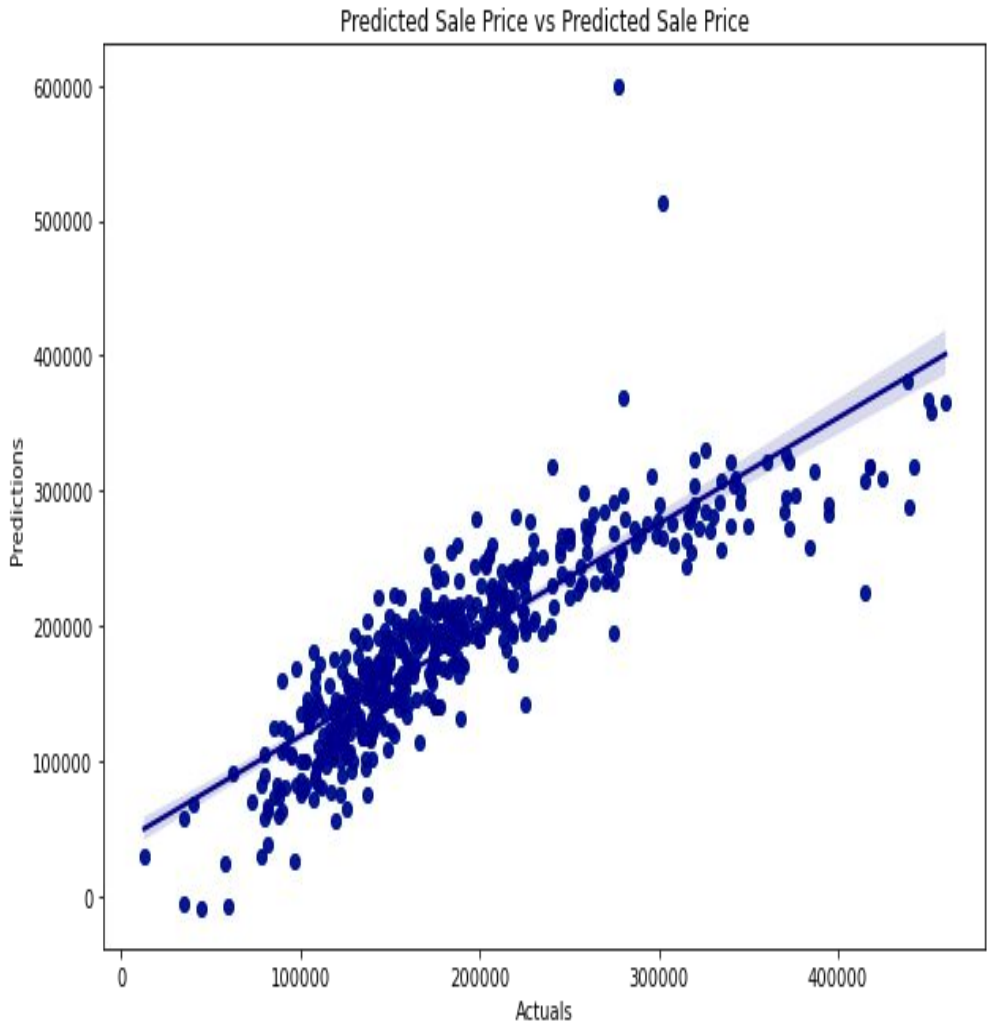
STEP 3: FEATURE SELECTION

- Correlation
- On going process
- Using judgement to select features

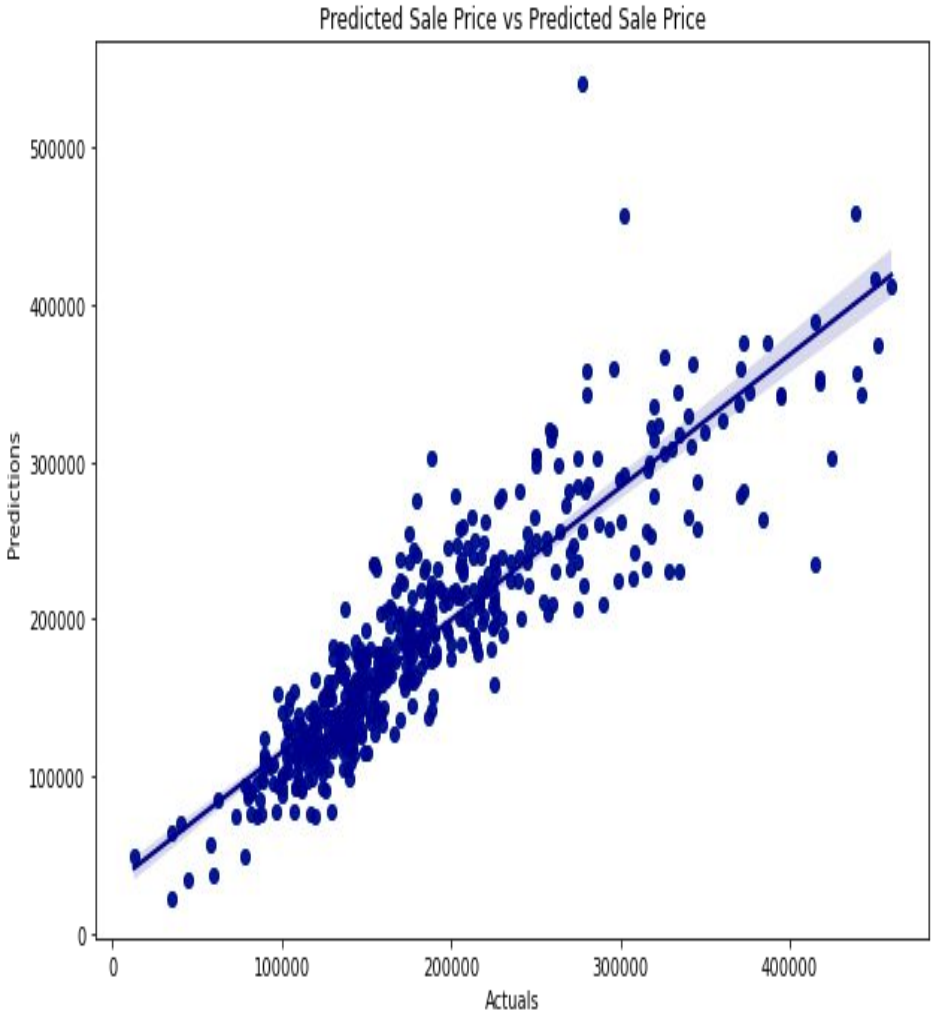

SalePrice
600000
400000
200000
0
1Fam
TwnhsE
Twnhs
2fmCon
Duplex
Bldg Type
2Stor
1Story
1.5Fin
SFoyer
SLvl
1.5Unf
2.5Fin
2.5Unf
House Style
2.5
5.0
7.5
10.0
Overall Qual
Overall Cond
1900
1950
2000
Year Built
Ex
TA
Gd
Fa
Po
Heating QC
Y
N
Central Air
SBrkr
FuseF
FuseA
FuseP
Mix
Electrical
2000
4000
1st Flr SF
0
500
1000
1500
2nd Flr SF

MODELING AND EVALUATION

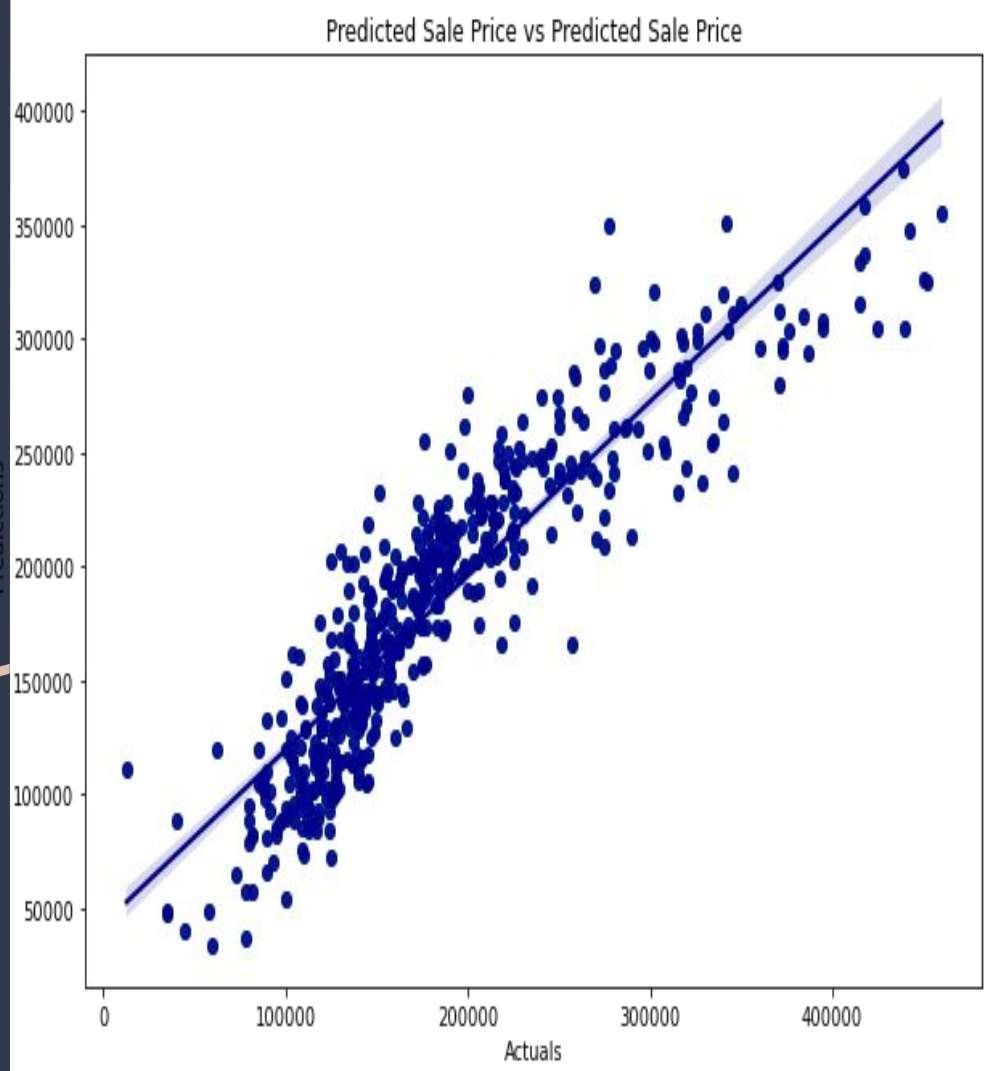
- Features -> Lot Area, Overall Qual, Overall Cond, Year Built, Year Remod/Add, Mas Vnr Area, Bsmt Full Bath, Bsmt Half Bath, Full Bath, Half Bath, Fireplaces, Bedroom AbvGr
- Kaggle RMSE - $38, 522
- Multicollinearity
- More numerical features



- Features -> Sale Type_New, Full Bath, Bsmt Exposure_Gd, Roof Style_Hip, House Style_1Story, Condition 1_Norm, Garage Cars, Year Built, Fireplaces,Neighborhood_StoneBr, Neighborhood_NridgHt, Neighborhood_NoRidge,Lot Area, Mas Vnr Area, Overall Qual, Overall Cond, Garage Cars, Screen Porch, MS Zoning_RL
- Tried to minimize multicollinearity
- Higher coefficients
- Kaggle MSE - $36, 693
- Combination of both numerical and categorical features



- Features -> Year Built, Fireplaces, Lot Area, Open Porch SF, Mas Vnr Area, Total Bsmt SF, Gr Liv Area, Year Remod/Add, Garage Cars, Wood Deck SF
- Accuracy score was low 0.72
- Only numerical features



CONCLUSION

From the coefficients, the following are a few of the top features I would recommend to look at when making predictions/buying houses

- Overall Qual
- Mas Vnr Area
- Neighborhood_NridgHt
- Kitchen Qual
- Neighborhood_StoneBr
- Bsmt Exposure

Features with negative coefficients

- Bldg_Type_Twnhse
- Neighborhood_Edwards
- Lnd_Contour_bnk

RECOMMENDATIONS

Some of the things homeowners/property investors could improve to increase the value of houses:

- Improve the kitchen quality -> positive coeff
- Improve the overall condition of the house
- Look in neighborhoods like Stonebrook and Northbridge