House Prices: Advanced Regression Techniques

Final Report (Final Project)

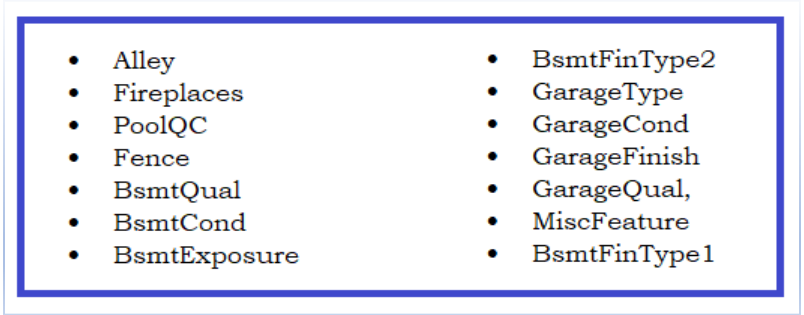
Preethi Prasobh(U1194871), Sukanksha Totade(U1192422) and Vivek Mishal(U1141856)

Introduction:

The Kaggle competition is to create generalizable model with low variance which predicts sales price of residential homes in Ames, Iowa. The data is partitioned into 50-50 set, naming train and test data set. The train data sets having total 81 variables with Sale Price as target variable and rest 79 are predictors (excluding ID since its unique to house# and is not good measurement for prediction) while test is having 80 variables (not having Sale Price).

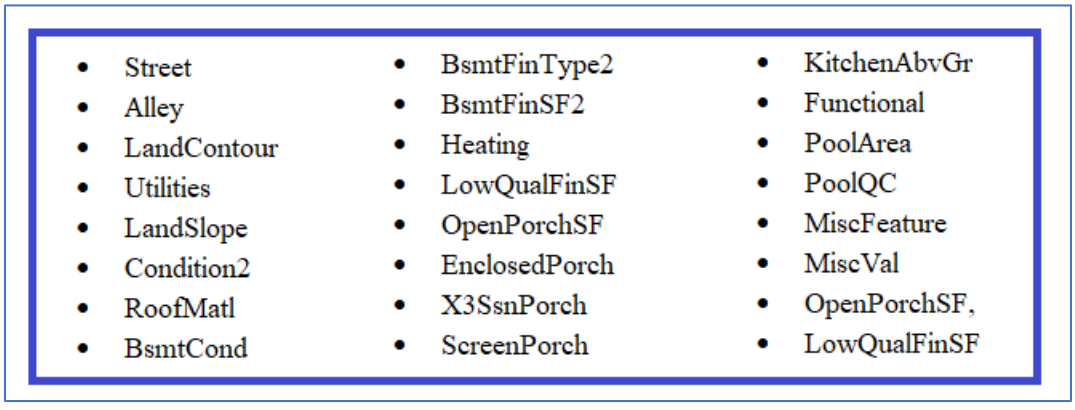
Data modeling and cleaning:

- In our train and test dataset, we transformed 'NA' values with meaningful values for following variables:



- Alley
- Fireplaces
- PoolQC
- Fence
- BsmtQual
- BsmtCond
- BsmtExposure
- BsmtFinType2
- GarageType
- GarageCond
- GarageFinish
- GarageQual,
- MiscFeature
- BsmtFinType1

- For more accurate imputation of remaining NA's, we combined both test and train data and ran miss-Forest imputation function.
- After imputing NA's, we divided data back into original train and test datasets.
- As we had 80 independent variables, we did a test to check their variance and removed following 22 variables which were found to have near zero variance.



- Street
- Alley
- LandContour
- Utilities
- LandSlope
- Condition2
- RoofMatl
- BsmtCond
- BsmtFinType2
- BsmtFinSF2
- Heating
- LowQualFinSF
- OpenPorchSF
- EnclosedPorch
- X3SsnPorch
- ScreenPorch
- KitchenAbvGr
- Functional
- PoolArea
- PoolQC
- MiscFeature
- MiscVal
- OpenPorchSF,
- LowQualFinSF

Final Report (Final Project)

Preethi Prasobh(U1194871), Sukanksha Totade(U1192422) and Vivek Mishal(U1141856)

- To check the distribution of dependent variable - Sale Price we decided to plot histogram.

- From the histogram, it was evident that dependent variable Sale Price is right skewed as shown below in figure1:

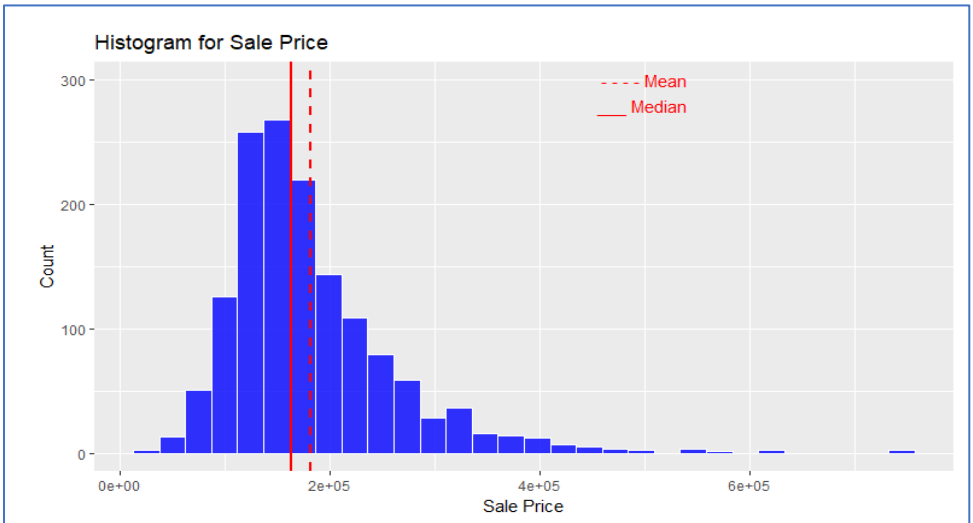
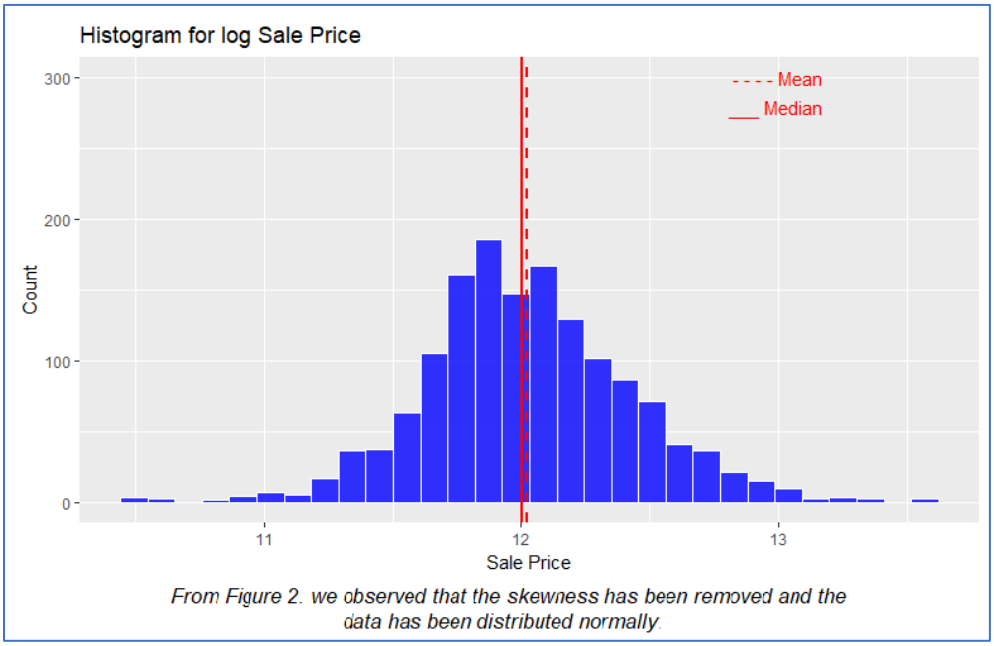


Figure 1. Sale Price data distribution is right skewed

- Model is more efficient for predictions if dependent variable is normally distributed, and therefore we log transformed Sale Price to make it normally distributed as shown below in figure 2:



From Figure 2. we observed that the skewness has been removed and the data has been distributed normally.

Final Report (Final Project)

Preethi Prasobh(U1194871), Sukanksha Totade(U1192422) and Vivek Mishal(U1141856)

Model Development:

1. Initially, we ran linear regression model on log of dependent variable using all the independent variables with 10-fold cross validation.

Linear Regression	log(RMSE)	R Squared
In-Sample	0.114383	0.9179476
Out-Sample	0.1471565	0.8651398

In-Sample R Squared of 0.9179476 is more than out-sample R Squared of 0.8651398, which indicated that model has high variance and thus was overfitting.

2. As there were many independent variables, we decided to implement LASSO regression model so that we could cut down on the insignificant variables. LASSO has the property of shrinking the coefficients of the insignificant variables close to zero.

LASSO	log(RMSE)	R Squared
In-Sample	0.1168565	0.9143604
Out-Sample	0.1538496	0.8537313

LASSO did not shrink any variables and thus performance of the model did not improve as compared to Linear Regression model.

3. Later, we used GLMNET, it uses combination of RIDGE and LASSO and finds the best models' performance

GLMNET	log(RMSE)	R Squared
In-Sample	0.125062	0.9019112
Out-Sample	0.1431038	0.8712661

Here, the best result uses α=0.55 so this result is somewhere between ridge and lasso, but closer to lasso with out of sample R squared value of 0.87 and logged RMSE 0.14.

4. To further improve the model and reduce the log RMSE, we log transformed independent variables which had skewed distribution and used interaction between two significant variables – Full Bath and Neighborhood.

Final Report (Final Project)

Preethi Prasobh(U1194871), Sukanksha Totade(U1192422) and Vivek Mishal(U1141856)

GLMNET with Log & Interaction	log(RMSE)	R Squared
In-Sample	0.1192373	0.9108353
Out-Sample	0.1249806	0.8870376

The log-log model with an interaction performed well with an out sample RMSE of 0.1249806 and R squared of 0.8870376.

Model performance:

In our best model, following variables were found to be significant in predicting sale price:

(Intercept)	12.02	NeighborhoodEdwards	-0.01	BsmtFinSF1	0.01
OverallQual	0.09	NeighborhoodIDOTRR	-0.01	HeatingQCFa	0.00
log(GrLivArea)	0.11	NeighborhoodMitchel	0.00	HeatingQCGd	0.00
FullBath	0.01	NeighborhoodNridgHt	0.01	HeatingQCTA	-0.01
OverallCond	0.05	NeighborhoodOldTown	-0.01	ElectricalMix	0.00
YearBuilt	0.05	NeighborhoodSomerst	0.01	log(X1stFlrSF)	0.03
BsmtFullBath	0.02	NeighborhoodStoneBr	0.01	X2ndFlrSF	0.01
TotRmsAbvGrd	0.00	NeighborhoodVeenker	0.00	Fireplaces	0.01
BsmtFinType1GLQ	0.00	Condition1Norm	0.02	GarageYrBlt	0.00
BsmtFinType1NoBasement	-0.01	Condition1RRAe	0.00	GarageFinishUnf	0.00
BsmtFinType1Unf	-0.02	BldgTypeDuplex	-0.01	GarageArea	0.00
CentralAirY	0.01	BldgTypeTwnhs	0.00	GarageQualGd	0.00
MSZoningFV	0.00	RoofStyleMansard	0.00	GarageCondFa	0.00
MSZoningRL	0.01	Exterior1stBrkComm	0.00	PavedDriveY	0.00
YearRemodAdd	0.02	Exterior1stBrkFace	0.01	WoodDeckSF	0.01
GarageCars	0.04	Exterior1stHdBoard	0.00	FenceGdWo	0.00
TotalBsmtSF	0.00	Exterior1stWd	0.00	YrSold	0.00
FireplaceQuGd	0.00	Exterior2ndImStucc	0.00	SaleTypeConLD	0.00
FireplaceQuNoFireplace	-0.01	Exterior2ndStucco	0.00	SaleTypeNew	0.02
log(LotArea)	0.04	Exterior2ndVinylSd	0.00	SaleConditionAdjLand	0.00
MSSubClass	-0.01	Exterior2ndWdShng	0.00	SaleConditionNormal	0.01
log(LotFrontage)	.	MasVnrArea	0.00	FullBath:NeighborhoodBrkSide	0.00
LotShapeIR2	0.00	ExterQualTA	0.00	FullBath:NeighborhoodEdwards	-0.01
LotShapeIR3	-0.01	ExterCondFa	0.00	FullBath:NeighborhoodMeadowV	0.00
LotConfigCulDSac	0.01	ExterCondTA	0.00	FullBath:NeighborhoodNAmes	0.00
LotConfigFR2	0.00	FoundationPConc	0.01	FullBath:NeighborhoodNoRidge	0.02
LotConfigFR3	0.00	FoundationStone	0.00	FullBath:NeighborhoodNridgHt	0.02
NeighborhoodClearCr	0.00	FoundationWood	0.00	FullBath:NeighborhoodSawyer	0.00
NeighborhoodCrawfor	0.02	BsmtQualNoBasement	0.00	FullBath:NeighborhoodSomerst	0.00

Inference of 5 main Predictors:

- **OverallQual**: With one unit increase in Overall Quality of the house, Sale Price goes up by 9.4%
- **GrLivArea**:1% increase in ground living area is associated with 0.11% increase in Sale Price.
- **OverallCond**: With one Unit increase in Overall Condition of the house, Sale Price goes up by 5%
- **YearBuilt**: With one-year increase in year built for the house, Sale Price goes up by 5.1 percent.
- **GarageCars**: With one unit increase in parking for cars in garage, sale price of the house goes up by 4.08%

Final Report (Final Project)

Preethi Prasobh(U1194871), Sukanksha Totade(U1192422) and Vivek Mishal(U1141856)

Out of above independent variable, following variables were found to have skewed distribution and were log transformed.

- log(LotFrontage)
- log(X1stFlrSF)
- log(GrLivArea)
- log(LotArea)

In our best model,In-sample log RMSE was 0.1192373 and R squared was 0.9108353 and out of sample log RMSE as 0.12498 and R squared was 0.8870376, Kaggle rank was 1173.

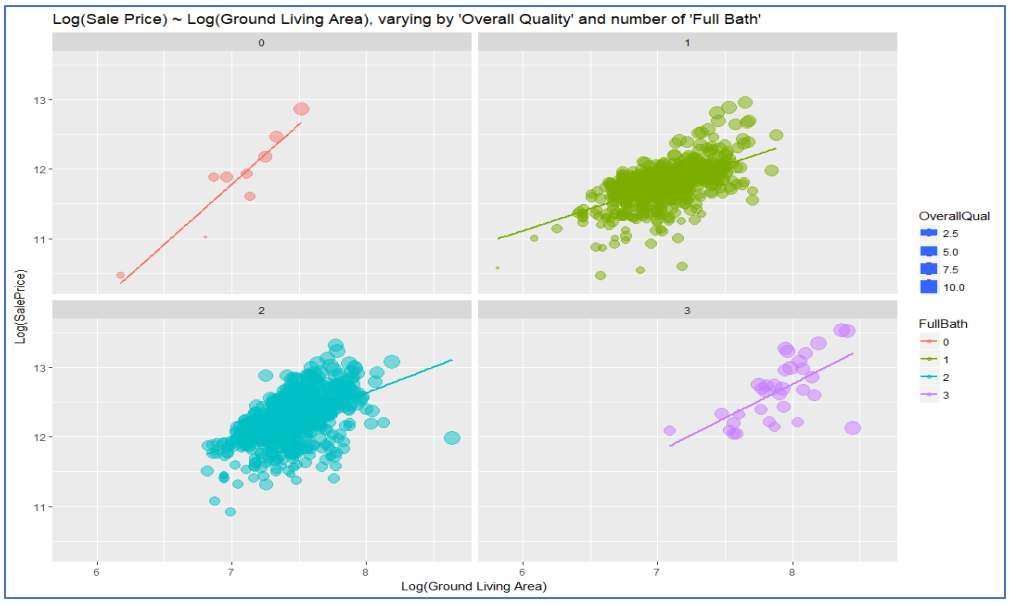


Figure 3

From the interaction plot, we could observe that, the sale price increases with increase in ground living area, overall quality and number of full bath.

Performance:

GLMNET with Log & Interaction	log(RMSE)	R Squared
In-Sample	0.1192373	0.9108353
Out-Sample	0.1249806	0.8870376

Kaggle	log(RMSE)	Rank
Score	0.12708	#1173