Expanding Student Housing in Ames, Iowa

A Machine Learning Predictive Model

TABLE OF CONTENTS

01

WHO ARE WE

Cykit Learners

Darish Sakeesing

Master's Student in Databases

Chitra Sharathchandra

Data Engineer

02

Question to answer

Question we are answering…

Ames, Iowa is home to Iowa State University(ISU) which accounts for half of the city's population

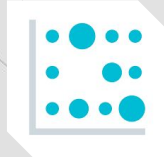
Housing is always a challenge for universities and a huge cost for students and their families.

As members of ISU Department of Residence Data Analysis team, we are studying the Ames Housing data set to determine whether there are opportunities to expand Housing options for students

03

DATA

DATA CHARACTERISTICS

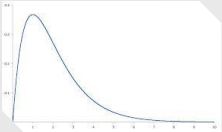


Large number of features

"Curse of Dimensionality"

Variable independence in question

Highly collinear



Skewed Distribution

Target variable skewed

Imputation mechanism required

Missingness

DATA ANALYSIS

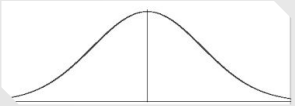
Imputation

Numerical variables

If related variables provide information, use that to derive grouping and calculate mean. Otherwise set to 0

Categorical variables

If related variables provide information, use it to derive grouping and calculate mode. Otherwise set to appropriate category and dummify if needed



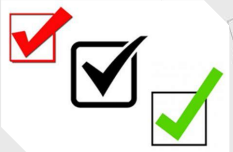
Handle Skewness

Target variable

Log the SalePrice variable

Scaled feature variables

Standardized all variables for linear regression



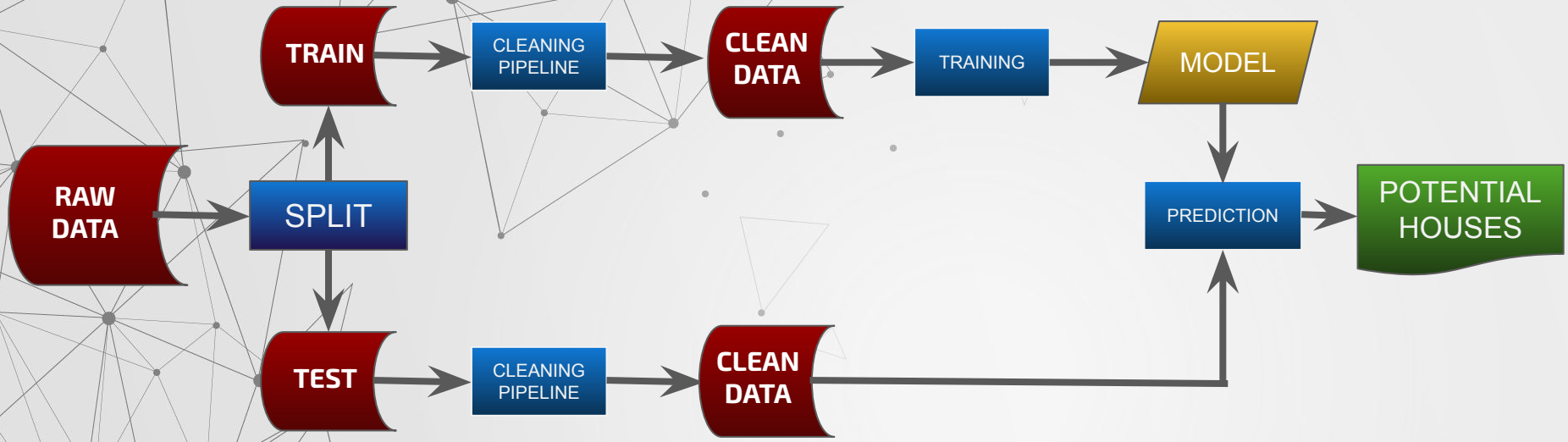
Feature Selection

Use models to perform feature selection

04

MODEL WORKFLOW

WORKFLOW


RAW DATA
SPLIT
TRAIN
TEST
CLEANING PIPELINE
CLEANING PIPELINE
CLEAN DATA
CLEAN DATA
TRAINING
MODEL
PREDICTION
POTENTIAL HOUSES

Some Key Steps

1. SPLIT BEFORE CLEANING

Since the split is random, we have to make sure that the test set is not *contaminated* with the mean of neighborhoods of houses in the training set. Therefore, the train and test sets have to undergo cleaning separately so that there is no *information leakage.*

2. BUILDING A CUSTOM PIPELINE

Since the cleaning will happen more than once and potentially many more times, it made sense to build a pipeline so that with one function call, the entire dataset is cleaned and formatted.

```
import CustomPipeline as cp
data = cp.clean('data/train.csv')
```

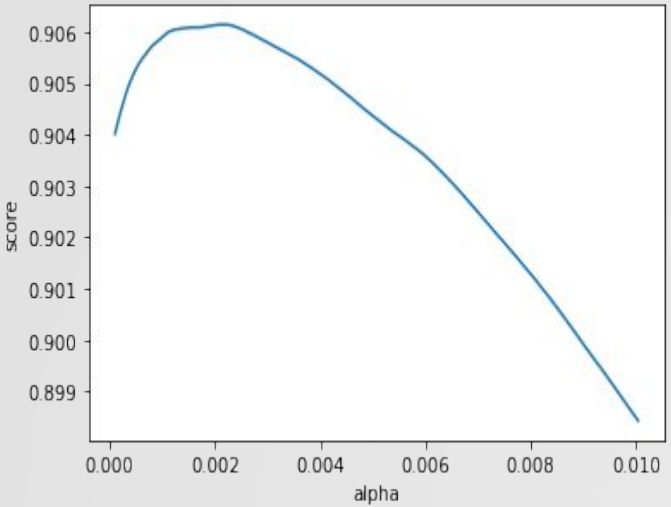
cp.clean() returns a dataframe

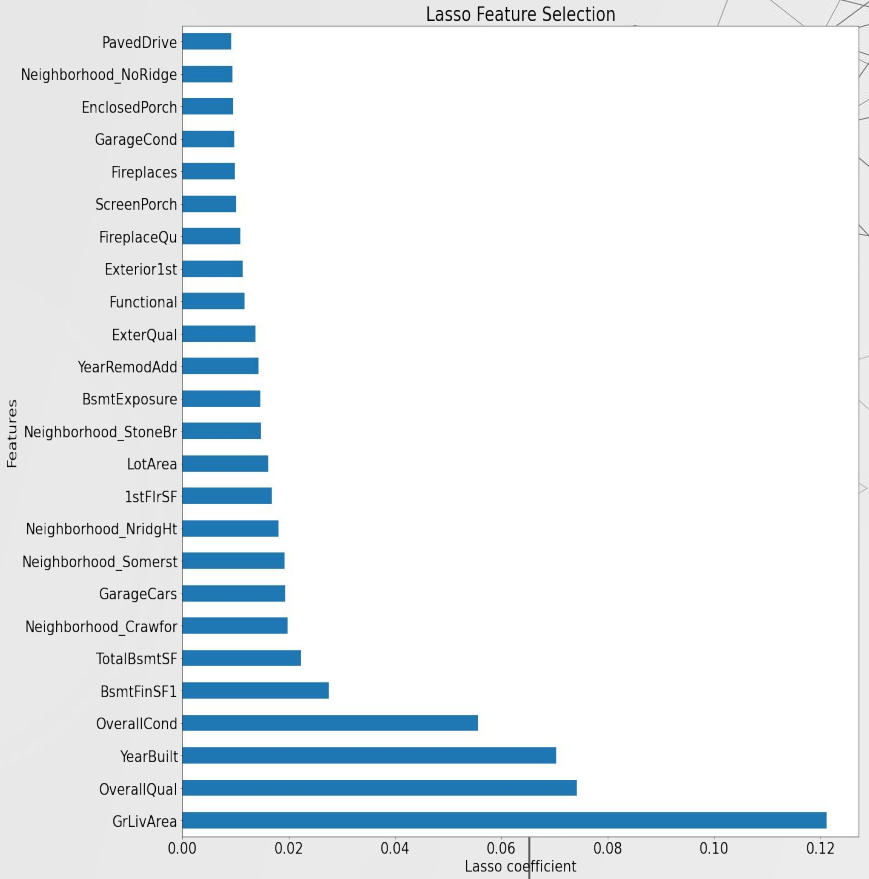
CustomPipeline.py

05

FEATURE SELECTION

Lasso Snapshot



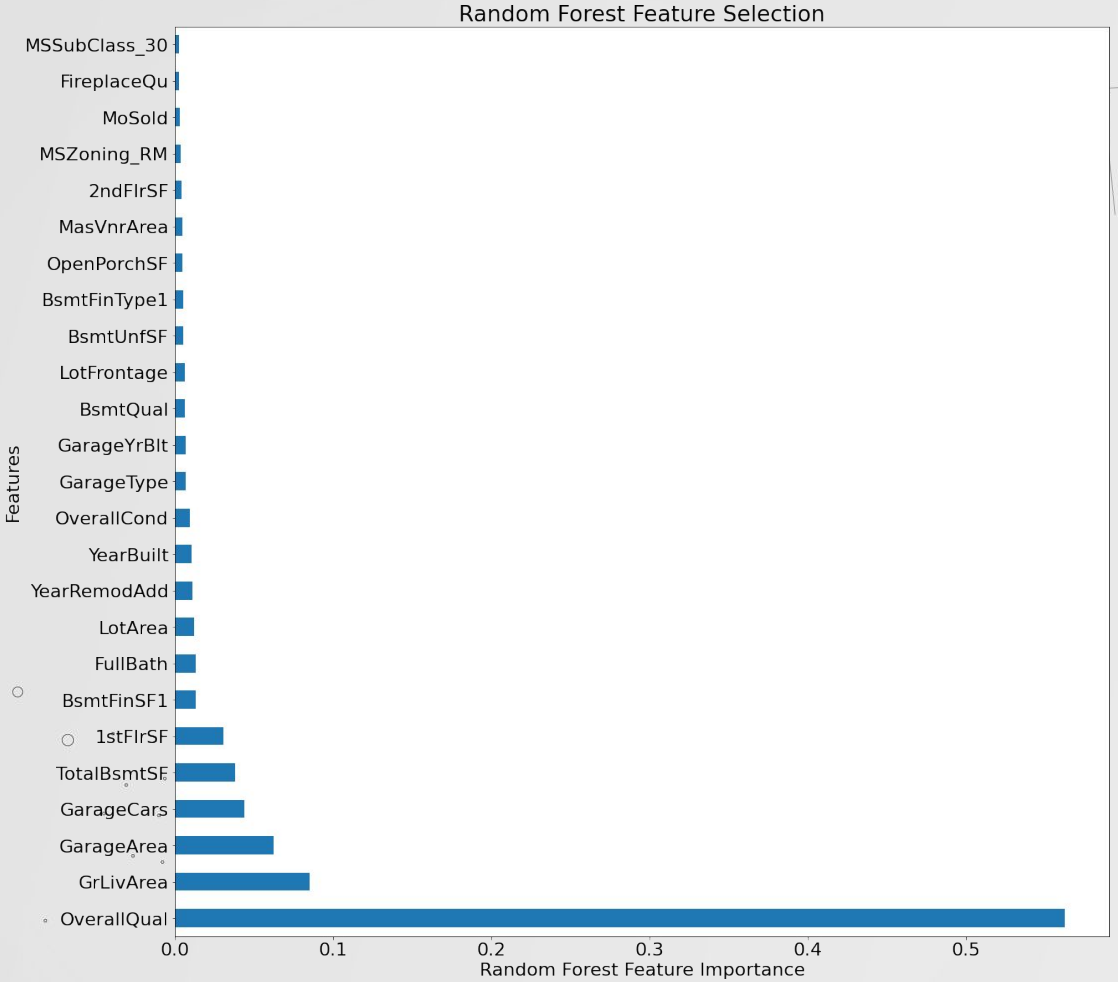


3 - Fold Cross Validation:

Optimal Alpha = 0.00209

Adjusted R2 = 0.921

Random Forest Snapshot



Adjusted R2 = 0.977

GridSearch Best Params:

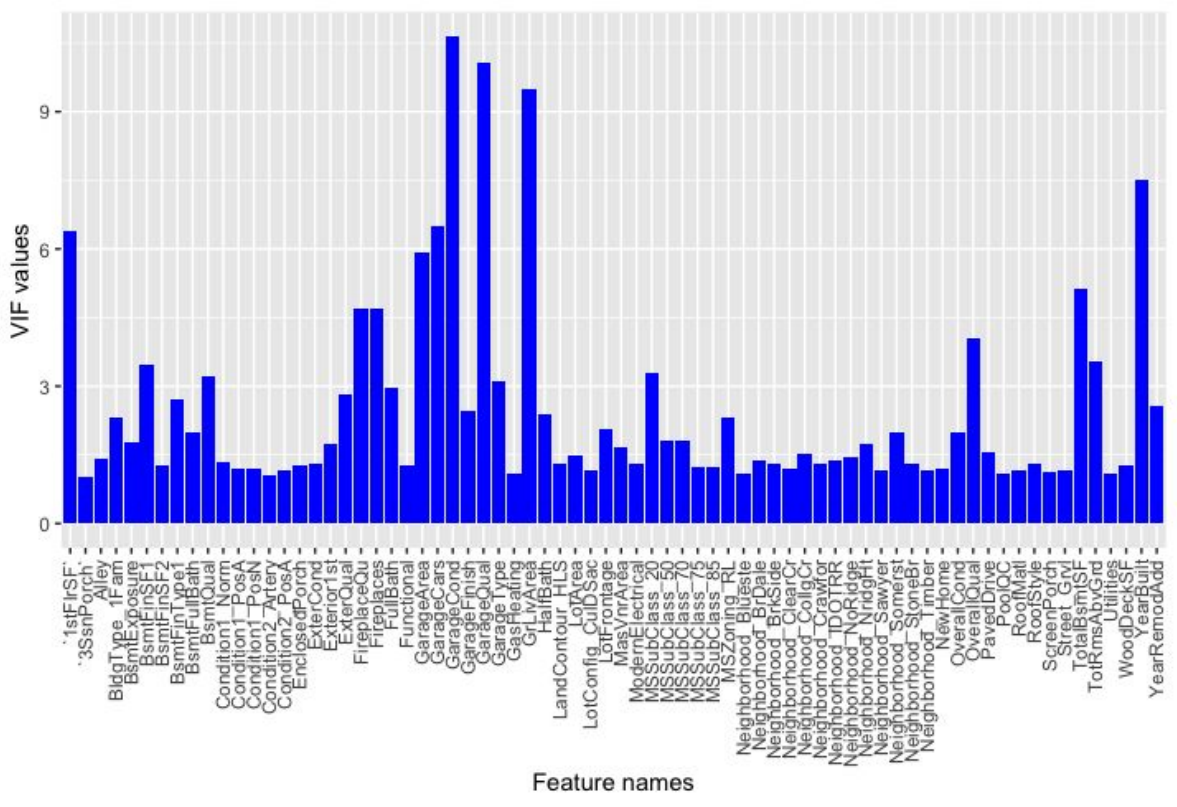
ccp_alpha:	0,
min_samples_leaf:	2,
n_estimators:	100,
Bootstrap:	False

Lasso vs Random Forest

Top 25 Features

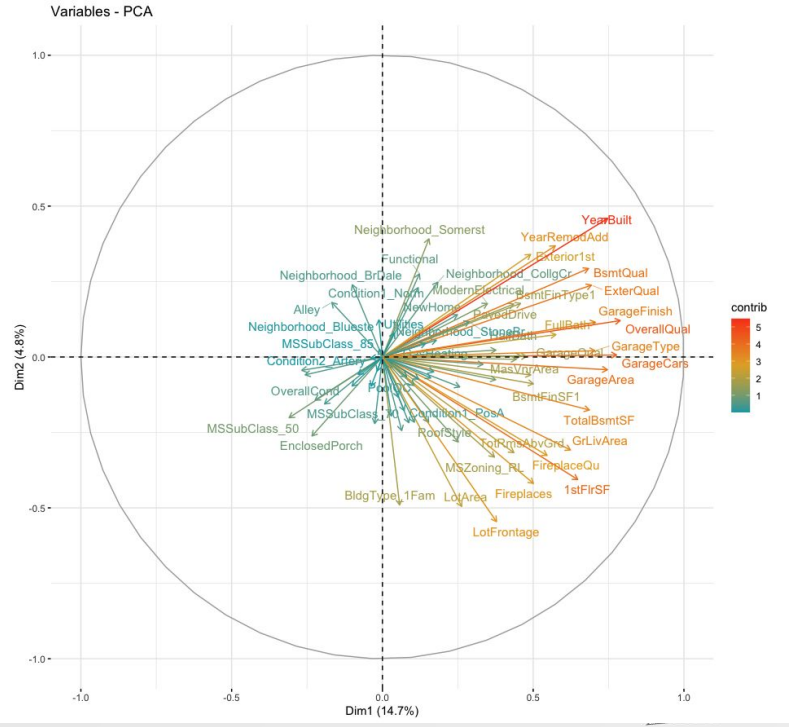
Both	Lasso only	Random Forest only	Mutually Exclusive
GrLivArea	Neighborhood_Somerst	BsmtQual	Appear in the total feature list
OverallQual	Neighborhood_NoRidge	GarageYrBlt	
YearBuilt	Exterior1st	GarageArea	
OverallCond	Neighborhood_Crawfor	2ndFlrSF	
BsmtFinSF1	ScreenPorch	BsmtUnfSF	
TotalBsmtSF	Fireplaces	MSZoning RM	
GarageCars	GarageCond	FullBath	
1stFlrSF	BsmtExposure	LotFrontage	
LotArea	PavedDrive	GarageType	
YearRemodAdd	ExterQual	BsmtFinType1	
FireplaceQu	EnclosedPorch	MasVnrArea	
	Neighborhood_NridgeHt	MoSold	
	Neighborhood_StoneBr	MSSubClass_30	
	Functional	OpenPorchSF	

Step 1 - VIF Analysis

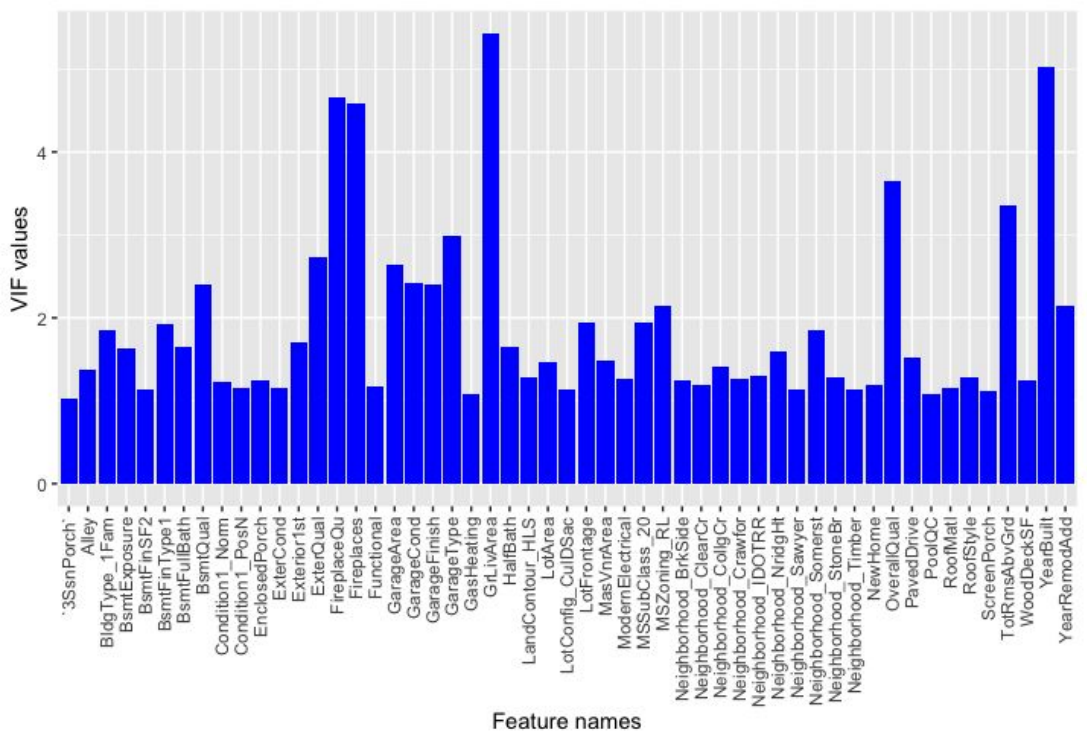


Step 2 Regress across features

- Eliminated with backward selection using adjusted R2 as metric - GarageQual, 1stFloorSF, TotalBsmtSF, GarageCars with minimal impact on adjusted R2
- To study the high VIF for GrLivArea, ran a Linear Regression between GrLivArea and other features
- Attempted forward and backward stepwise regression with no success
- Used PCA to find high impact features and eliminated those



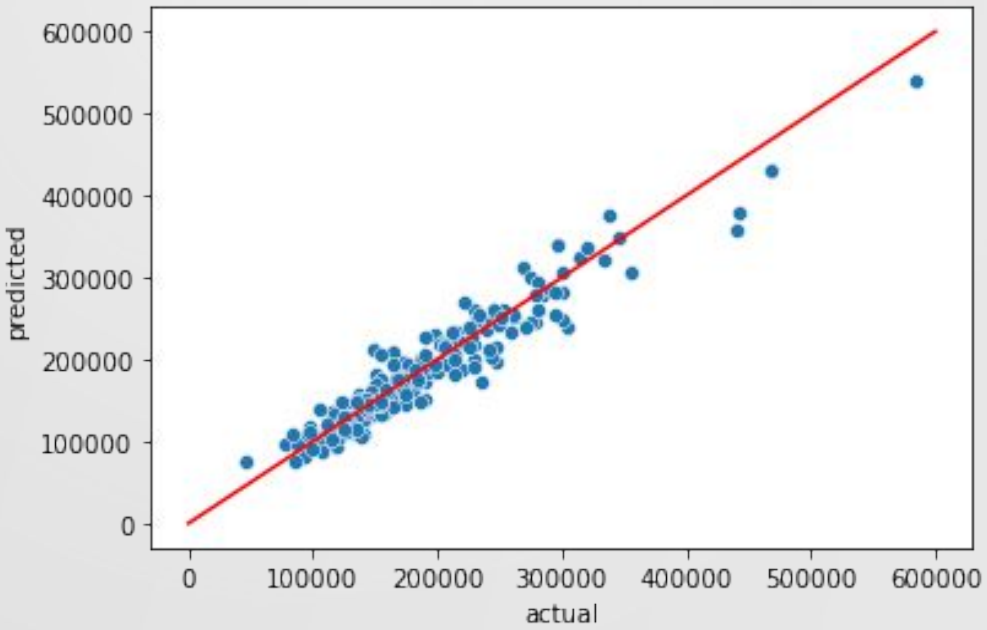
Step 3 - post elimination VIF analysis



06

FINAL MODEL

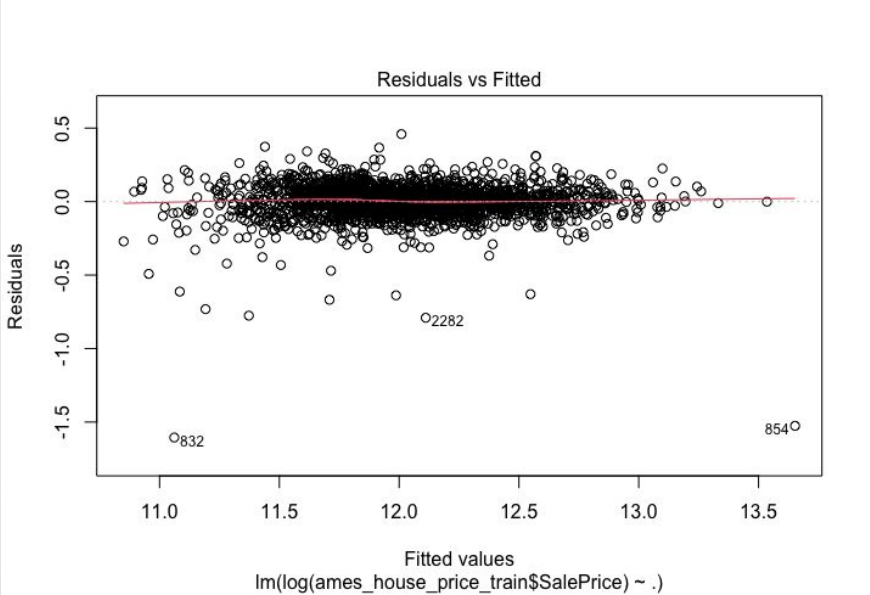
LINEAR REGRESSION PERFORMANCE

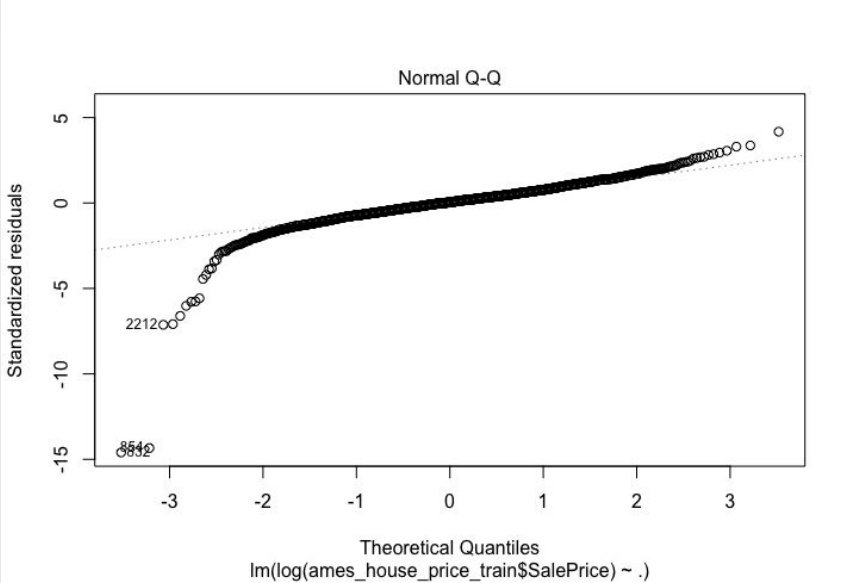


Adjusted R2 = 0.901
RMSE = 0.108

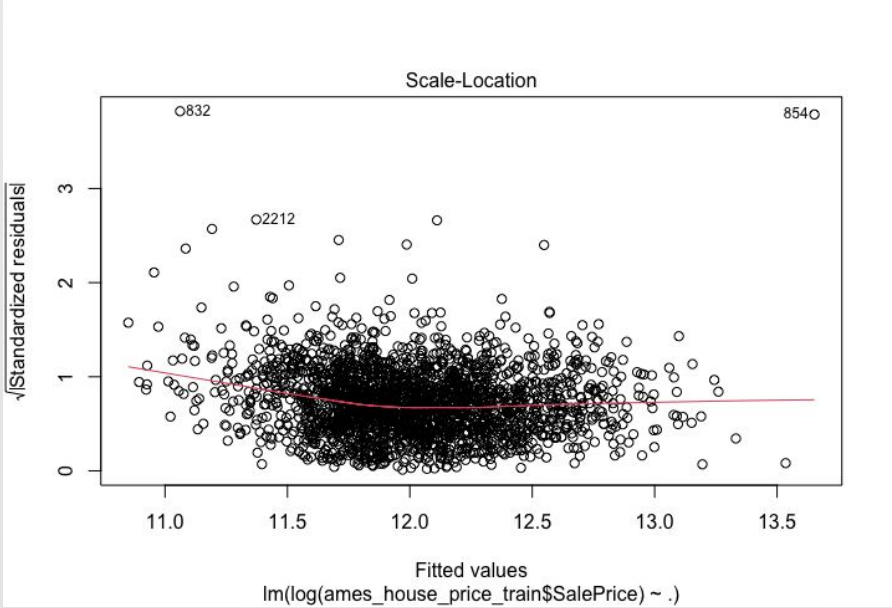
<u>In Plain Terms:</u>
Minimum Error: **$60**
Median Error: **$10k**

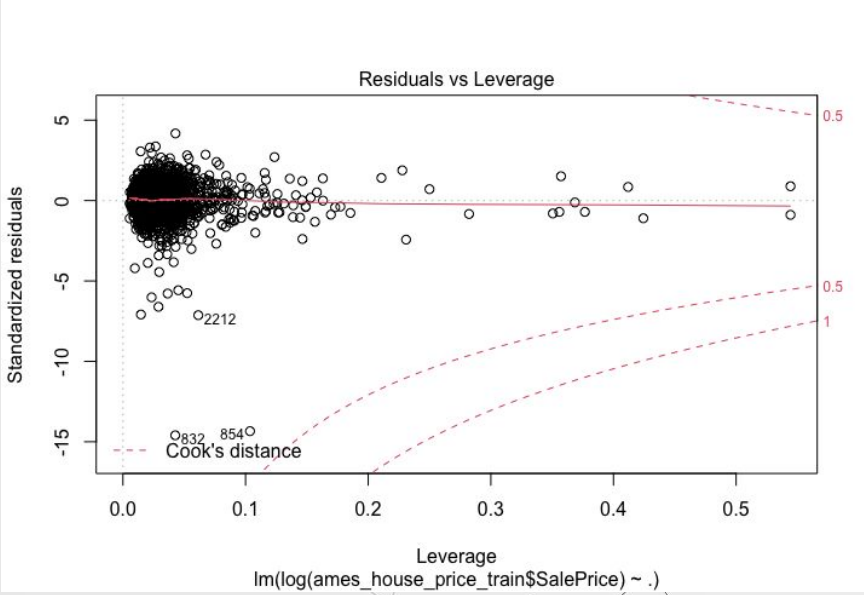
Step 4 - Post elimination model analysis





Step 4 - Post elimination model analysis





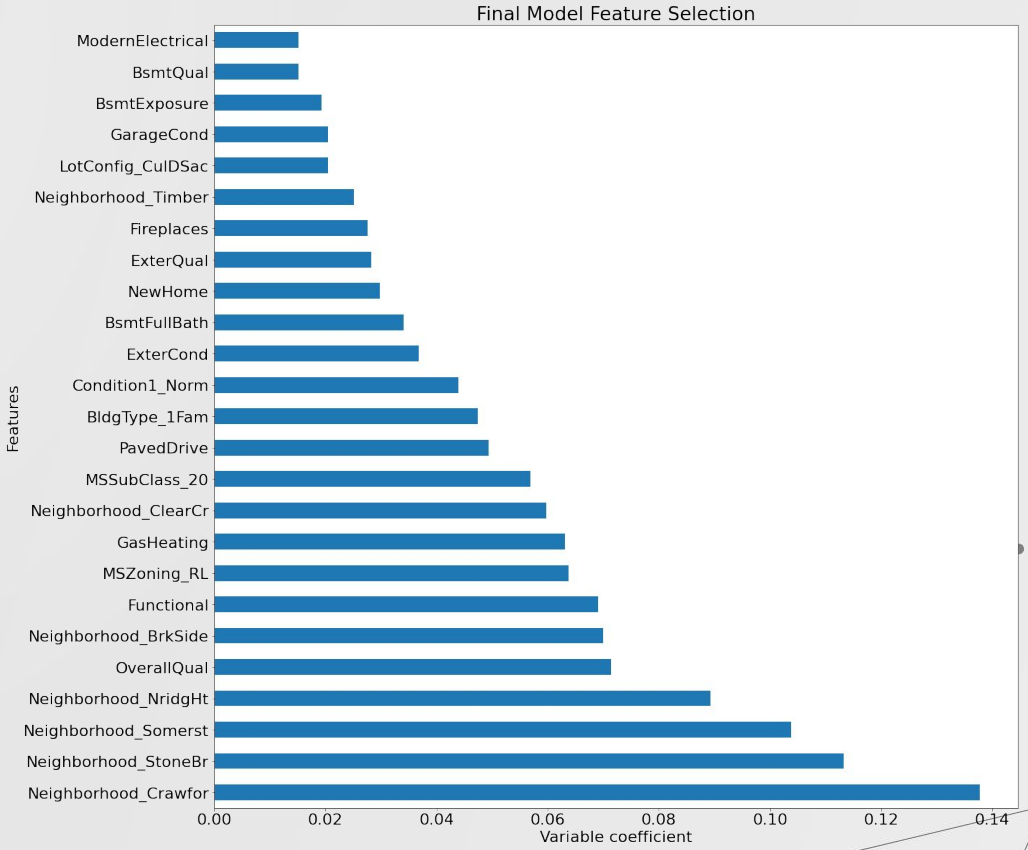
And the top 25 features are...

Not surprising finds

- Neighborhoods are important in determining house price
- Overall quality of the house is an important indicator for house prices
- Many features contribute to the price of the house

Surprising finds

- Only OverallQual shows up from original common list of 25
- Having a basement full bath and basement quality matters so homeowners should finish their basements
- Single story new homes are more valuable



07

MODEL APPLICATION

Find homes that meet our criteria

Iowa State University Department of Residence

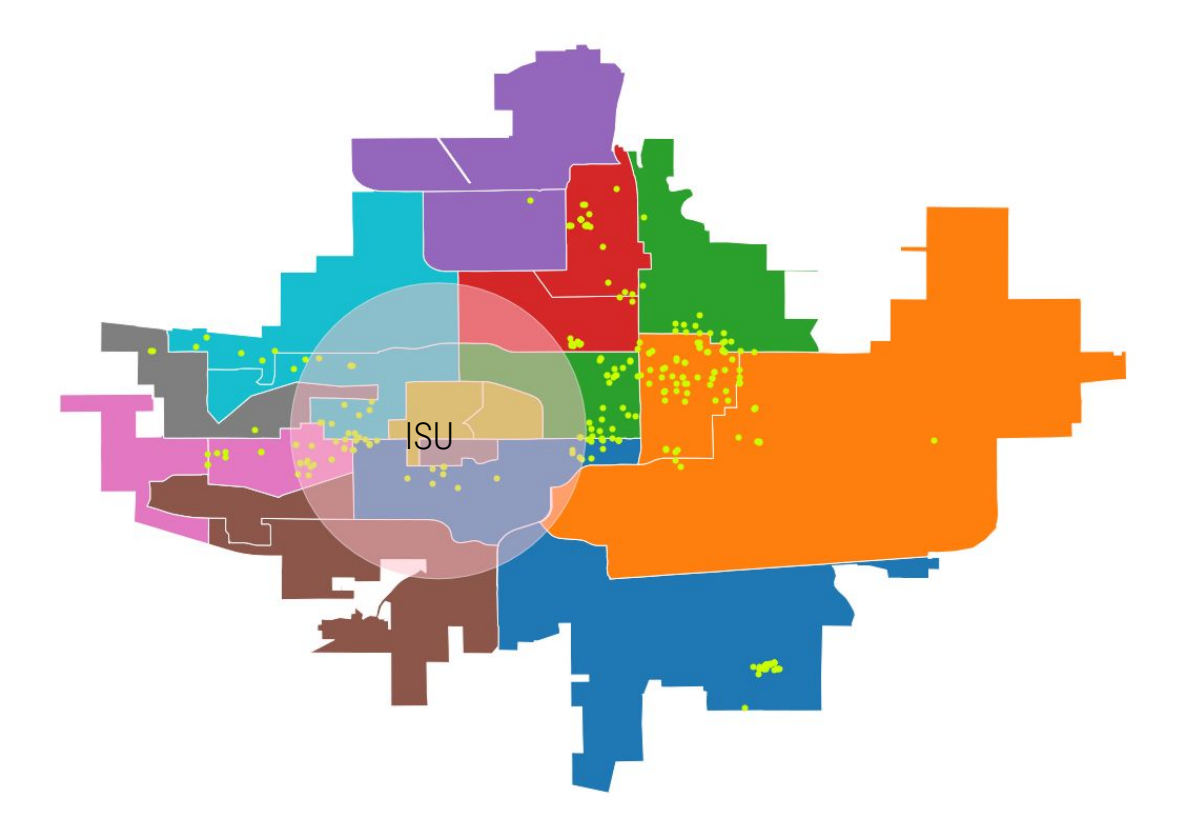
Home Selection Criteria

- Use the model to generate predicted Sales Price of newly provided data set
- Retrieve 10% of homes with the lowest Sales Price
- Filter further to meet University requirement of the following criteria:
  - 1 mile distance from campus
  - Bedroom to bathroom ratio of less than or equal to 2
  - Overall condition and quality is Fair and above

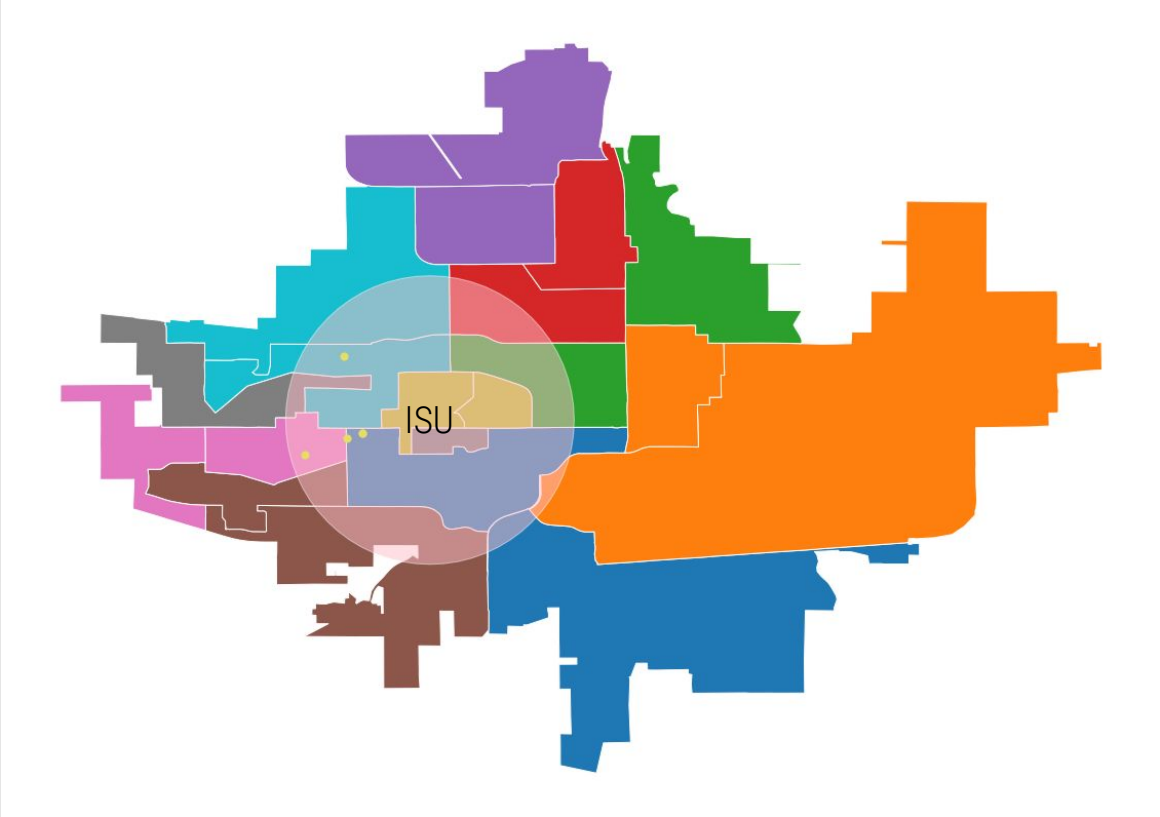
```
import PredictPipeline
PredictPipeline.getUniHouses('path/to/your/data')
```

Produces a .csv file with list of potential houses

Map of candidate homes



Map of selected homes



08

LOOKING FORWARD

What next?

- Build a user friendly app for the university to enter criteria to find potential homes to expand housing
- Study other models to improve performance

THANKS

Does anyone have any questions?