Predicting Housing Prices

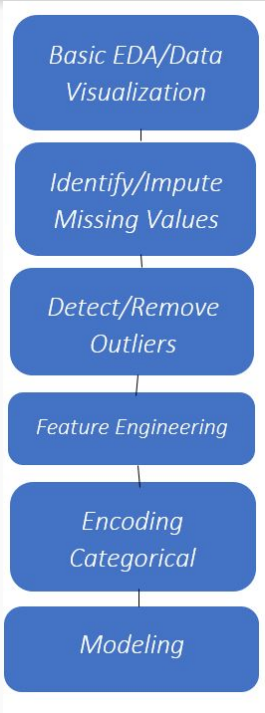
Lingdi, Luke, Stan, Radhika
August 26, 2019

Data Overview

- Data contains details of houses sold between 2006 and 2010.
- Details include specifics with anything from the garage quality to the masonry veneer type.
- There are 79 variables to consider, of which most are introduced as factors.
- Goal is to predict a sale price of an output of sample houses.

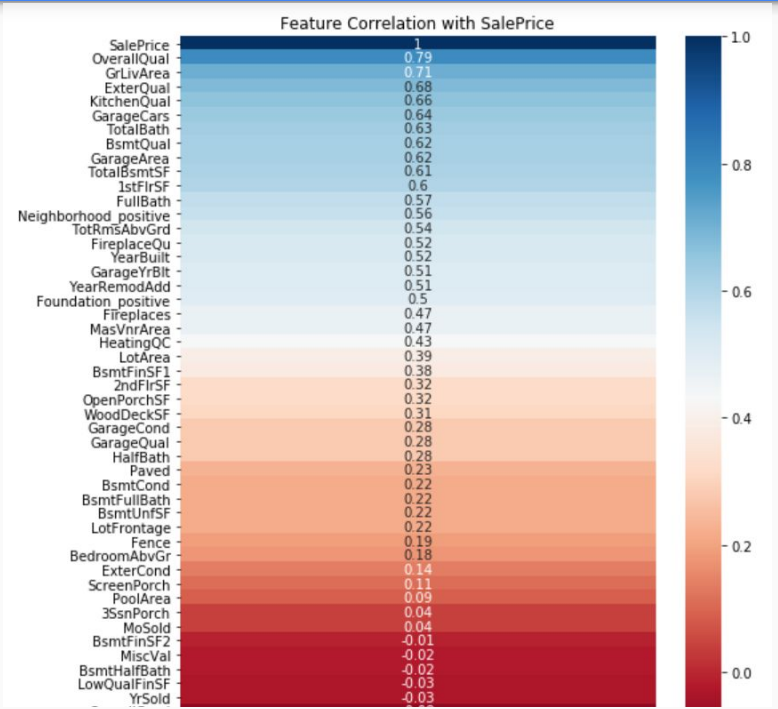
Project Motivation and our approach

- Expand upon programming and data analysis skills
- Skillfully apply a machine learning model to a dataset
- Predict other datasets with a high degree of accuracy
- In order to accomplish this we took the following steps:



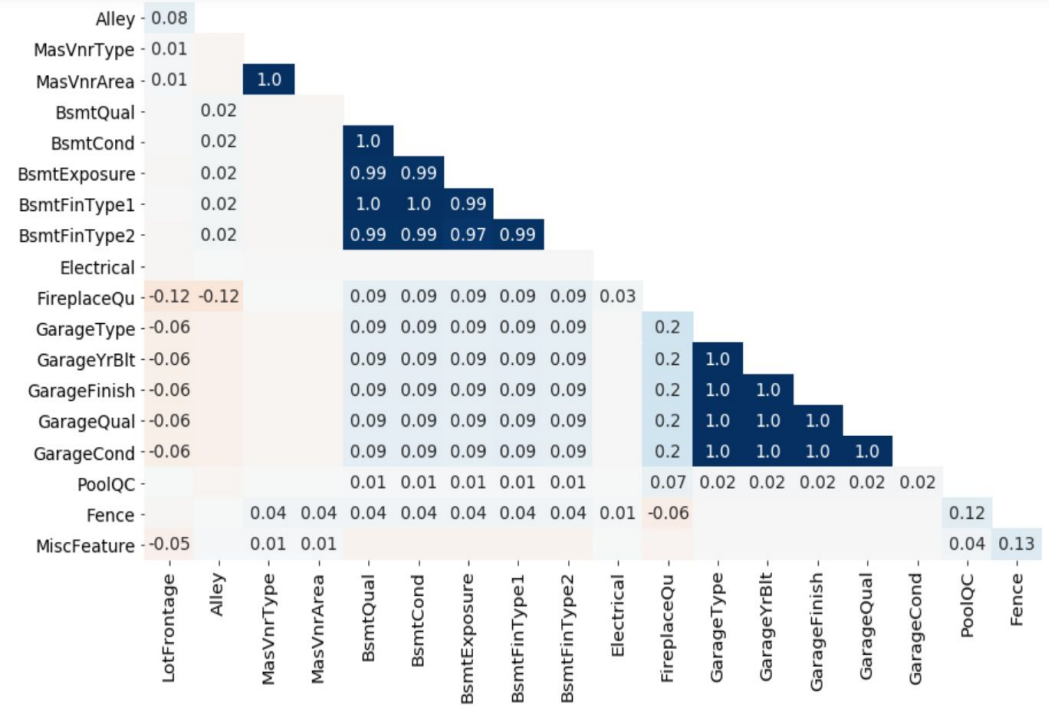
Exploring Correlated Features

- By looking at each feature's correlation with Sale Price, the most correlated features is Overall Quality.
- Understanding each feature's correlation with Sale Price will assist us in feature engineering and feature selection
- All correlated features are positively correlated



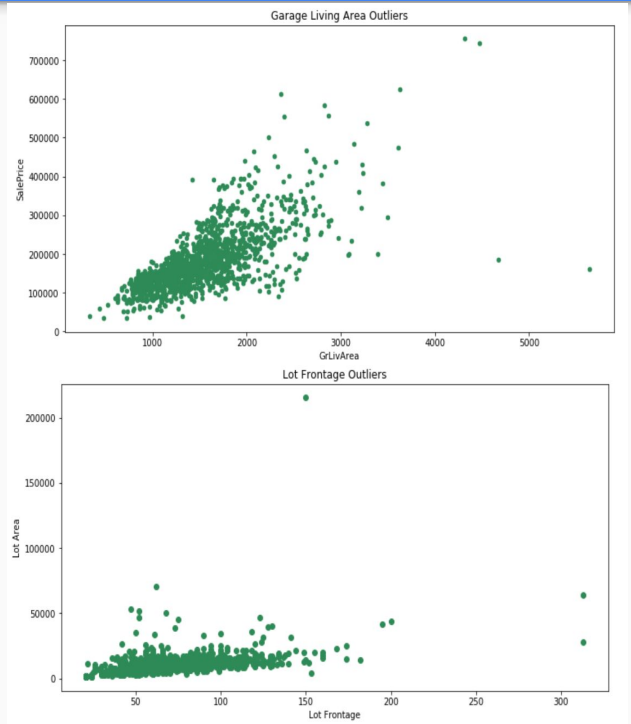
Missingness

- 35 columns with missing values.
- High correlation among most missing values.
  - Similar features can mostly be imputed as a group.
    - I.e. Garage features and Basement features.
- Most data missing at random.



Outliers - Train dataset

- In order to best express each features' correlation with sale price, we removed outliers.
- The most pertinent outliers were identified in 'Above Ground Living Area' and 'Lot Frontage'.



Outliers - Test dataset

- **No Basement in the below indices: Impute with "Zero" for all the columns below**

Id	BsmtQual	BsmtFullBath	BsmtHalfBath	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF
2121	None	NaN	NaN	NaN	NaN	NaN	NaN
2218	None	0.0	0.0	0.0	0.0	173.0	173.0
2219	None	0.0	0.0	0.0	0.0	356.0	356.0

Outliers - Test dataset

- **Impute with "YearBuilt" for the ‘GarageYrBlt’ column in the following index:**

Id	YearBuilt	GarageYrBlt
2593	2006	2207.0

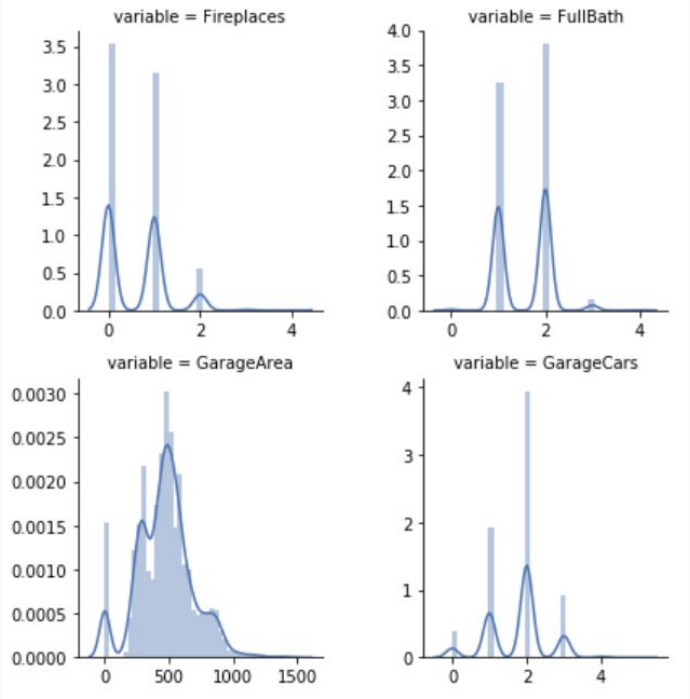
Imputing Missing Data - Train and Test

- **Impute with "None" when NaN means something:**
  - Alley: no alley access
  - BsmtQual/ BsmtCond/ BsmtExposure/ BsmtFinType1/ BsmtFinType2: no basement
  - FireplaceQu: no fireplace
  - GarageType/ GarageFinish/ GarageQual/ GarageCond: no garage
  - PoolQC: no pool
  - Fence: no fence
  - MasVnrType: no masonry veneer type
  - MiscFeature: no miscellaneous feature
- **Impute with “Mode” for Categorical:**
  - 'Electrical', 'KitchenQual', 'Functional', 'Exterior1st', 'Exterior2nd', 'SaleType', “MSZoning', 'Utilities', 'GarageCars'
- **Impute with “Mean” for Numeric:**
  - 'GarageYrBlt', 'GarageArea', 'MasVnrArea'
- **Impute with “Zeros”:**
  - 'BsmtQual', 'BsmtFullBath', 'BsmtHalfBath', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF'
- **Groupby ‘neighbourhood’ and ‘Mean’:**
  - ‘LotFrontage'

	Total	Percent
PoolQC	2909	0.996574
MiscFeature	2814	0.964029
Alley	2721	0.932169
Fence	2348	0.804385
SalePrice	1459	0.499829
FireplaceQu	1420	0.486468
LotFrontage	486	0.166495
GarageFinish	159	0.054471
GarageCond	159	0.054471
GarageQual	159	0.054471
GarageYrBlt	159	0.054471
GarageType	157	0.053786
BsmtCond	82	0.028092
BsmtExposure	82	0.028092
BsmtQual	81	0.027749
BsmtFinType2	80	0.027407
BsmtFinType1	79	0.027064
MasVnrType	24	0.008222
MasVnrArea	23	0.007879
MSZoning	4	0.001370
BsmtFullBath	2	0.000685
BsmtHalfBath	2	0.000685
Utilities	2	0.000685
Functional	2	0.000685
Electrical	1	0.000343
Exterior2nd	1	0.000343
KitchenQual	1	0.000343
Exterior1st	1	0.000343
GarageCars	1	0.000343
TotalBsmtSF	1	0.000343
GarageArea	1	0.000343
BsmtUnfSF	1	0.000343
BsmtFinSF2	1	0.000343
BsmtFinSF1	1	0.000343
SaleType	1	0.000343

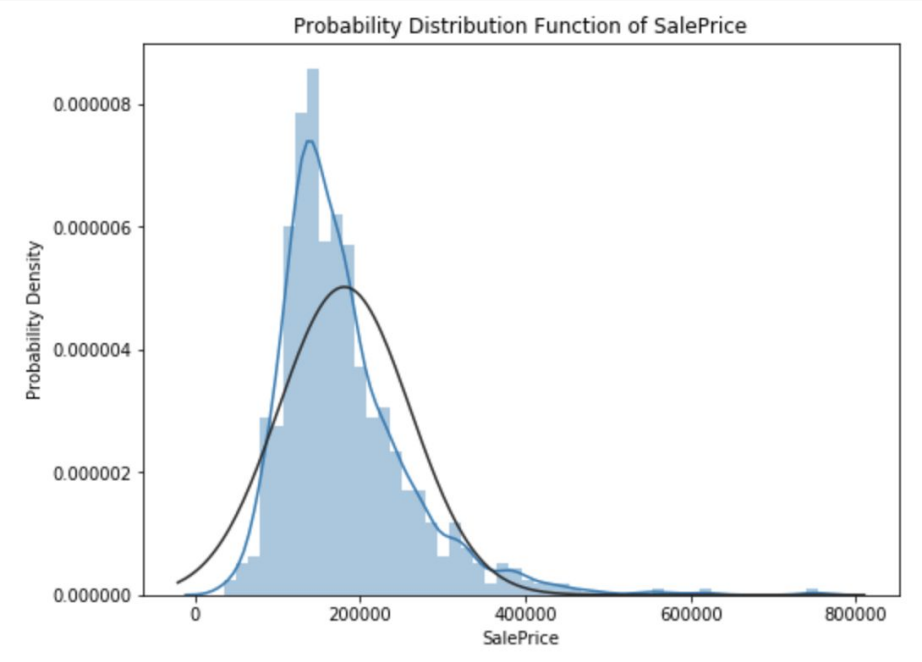
Dealing with Skew

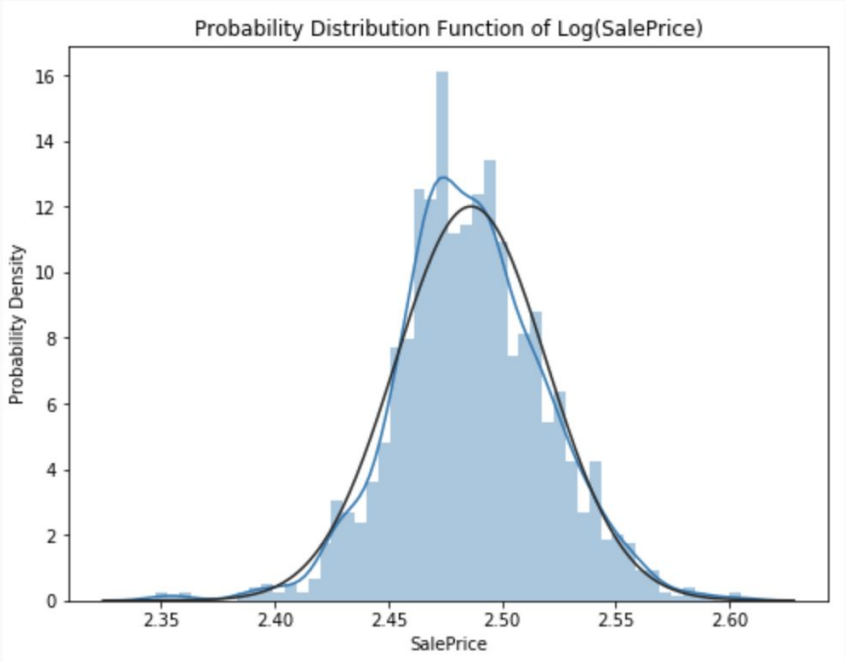
- To correct skewed distributions, we took the log of these numerical predictors so the features don't violate the assumption of normality.
- Based on the data, **'Misc Val'**, **'Pool Area'**, and **'Lot Area'** were the most skewed.



	Skew
MiscVal	21.947195
PoolArea	16.898328
LotArea	12.822431
LowQualFinSF	12.088761
3SsnPorch	11.376065
KitchenAbvGr	4.302254
EnclosedPorch	4.003891
ScreenPorch	3.946694
MasVnrArea	2.613592
OpenPorchSF	2.535114
WoodDeckSF	1.842433
1stFlrSF	1.469604
MSSubClass	1.375457
GrLivArea	1.269358
2ndFlrSF	0.861675
TotRmsAbvGrd	0.758367
Fireplaces	0.733495
HalfBath	0.694566
OverallCond	0.570312

Log Transform Target Variable





Categorical Variables

Divide Categorical Variables into two groups:

- **Nominal** - One-Hot Encoding (pd.get_dummies)

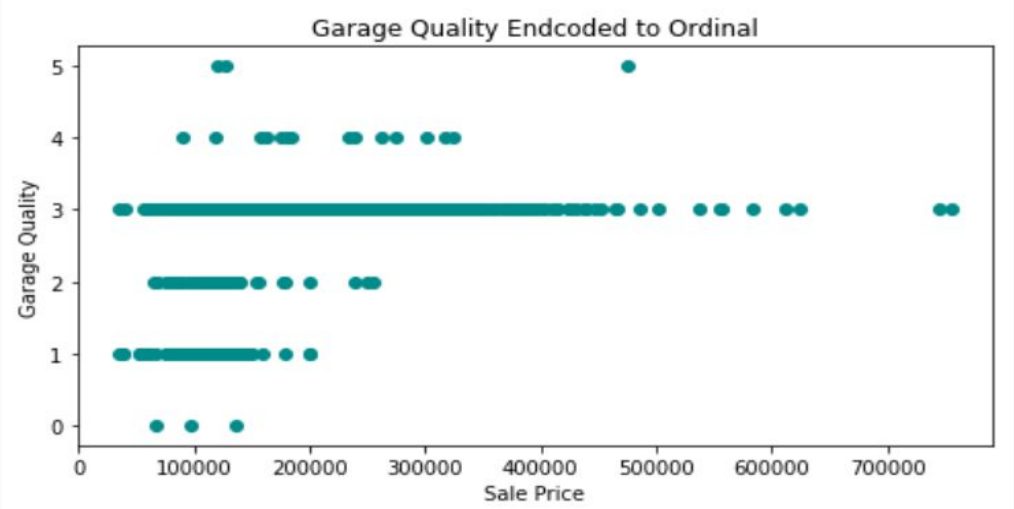
```
nominal_cols = ('MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities',
       'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType',
       'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType',
       'Foundation', 'Heating','CentralAir', 'Electrical', 'Functional', 'GarageType',
       'Fence', 'MiscFeature','SaleType', 'SaleCondition', 'MSSubClass', 'MoSold', 'YrSold')
```

- **Ordinal** - Manually Code with Integer (make dictionary with {key: integer value})

```
ordinal_cols = ('ExterQual', 'ExterCond', 'BsmtQual', 'BsmtCond', 'BsmtExposure',
            'BsmtFinType1', 'BsmtFinType2', 'HeatingQC', 'KitchenQual', 'FireplaceQu',
            'GarageFinish', 'GarageQual', 'GarageCond', 'PavedDrive', 'PoolQC')
```

Changing Factor Variables to Ordinal

- We iterated through factor variables that had descriptions that could be quantified.
- An example is ‘GarageQual’, which had the following existing values:
  - Na, Po, Fa, TA, Gd, Ex
  - These can be mapped to [0, 1, 2, 3, 4, 5]




```
enc_dict = {'Ex':10, 'Gd':8, 'TA':6, 'Fa':4, 'Po':2, 'None':0,
            'Av':6, 'Mn':4, 'No':2,
            'GLQ':10, 'ALQ':8, 'Rec':8, 'BLQ':6, 'LwQ':4, 'Unf':2,
            'Fin':10, 'RFn':6, 'Unf':2,
            'Y':10, 'P':6, 'N':2}
```

Feature Engineering

- Feature engineering based on:
  - Feature similarity -
    - I.e. ‘TotalArea’ , ‘TotalBath’.
  - Feature importance -
    - I.e. ‘Neighborhood’, ‘OverallQual’.

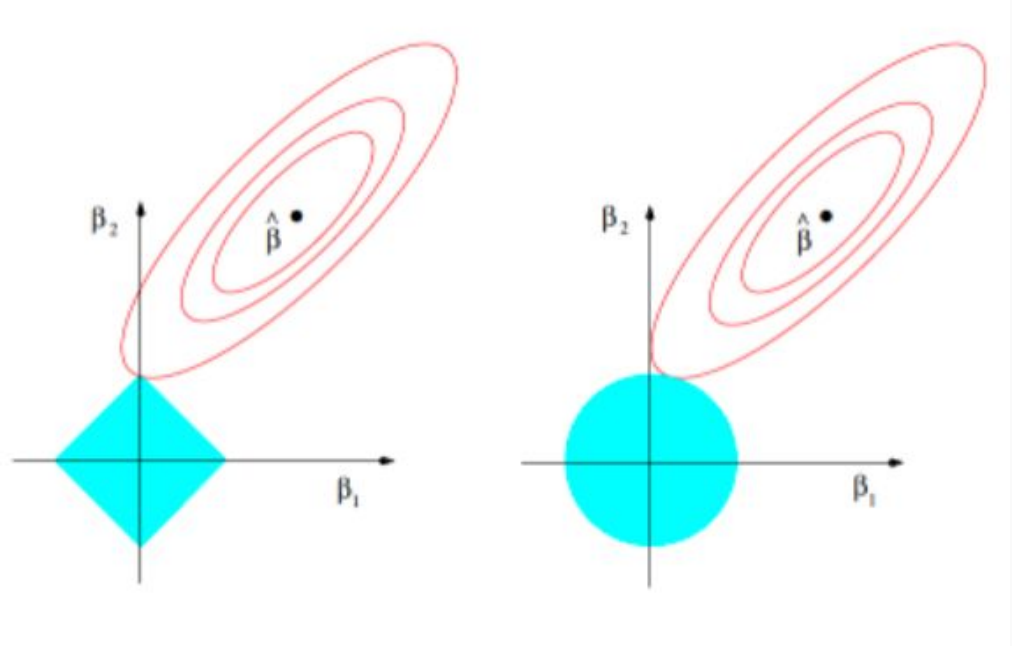
Feature Selection

- Feature Selection was conducted via Lasso Regularization
- Lasso parameters were found via GridSearchCV
- A sample of the important features were:
  - Neighborhood(Crawford, Stone Brook, Northridge Heights)
  - OverallQual
  - CentralAir
  - GarageCars
  - MSZoning
  - SaleType
  - SaleCondition

Crawfor	1.122394
StoneBr	1.092350
BrkFace	1.057598
NridgHt	1.054343
OverallQual	1.053084
BrkSide	1.051840
Norm	1.049203
OverallCond	1.040325
Somerst	1.036810
Functional	1.035041
New	1.031406
ClearCr	1.031122
PConc	1.026339
GarageQual	1.026335

Model Tryouts

- Models considered:
  - RandomForest (poor)
  - Lasso (0.1230)
  - Ridge (0.12579)
  - ElasticNet (0.12640)
  - Huber (0.18112)
  - Gradient Boost (0.125)
  - xgBoost (0.13)



Final Model Selection

- Lasso:
  - Score of 0.123
- Gradient Boost (keeping outliers):
  - Score of 0.125
- Lasso+Gradient Boost (50/50 split) (keeping outliers):
  - Score of 0.121
- Remarks:
  - Lasso by itself has high interpretability
  - The combination improves our score - depends on which we value more
- Lasso Interpretation example: LotArea has a coefficient of 0.000002, while mean LotArea is 9,820. For every unit increase in LotArea, log(price) increases by 0.000002. Therefore, a LotArea of 10,000 would imply that the corresponding increase in price is approximately $10,000.

Future Improvements

- Improve Feature Selection and Feature Engineering
  - Introduce a larger number of novel features
  - Combine and drop multicollinear features
- Hierarchical Linear Models
- Implement Robust Stacked Models
  - average performance of multiple models