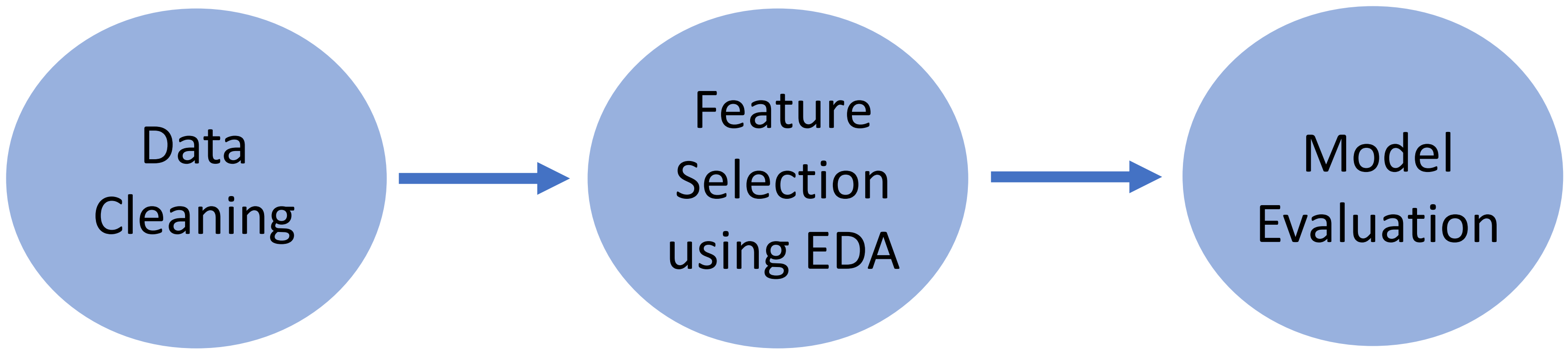
Build a Regression Model for Prediction of House Price in Ames

Jasper

Problem Statement

- Identify features that are influential in predicting the Sale Price in Ames
- Determine the best regression model for predicting Sale Price with the influential features
- With a set of features, we are able to predict the price of a house

Workflow


Data
Cleaning
Feature
Selection
using EDA
Model
Evaluation

Data Cleaning

```python
# all can change to 0
train[cont_vars_na] = train[cont_vars_na].fillna(0)
```

```python
# all can change to NA
train[ordinal_vars_na] = train[ordinal_vars_na].fillna('NA')
```

```python
# all change to NA except Mas Vnr Type to None, 4 vars with null values
for var in nominal_vars_na:
    if var == 'Mas Vnr Type':
        train[var] = train[var].fillna('None')
    train[var] = train[var].fillna('NA')
```

Fig.1. Code snippet that impute missing values

- Impute the missing values for the continuous, nominal and ordinal features

```python
# Overall Qual, Overall Cond already in int datatype
# so only need to convert for the rest of ordinal variables
# NA is assigned to 0
def convert_ordinal_features(df, features):
    for feature in features:
        if feature == 'Lot Shape':
            df[feature] = df[feature].map({'IR3':1, 'IR2':2, 'IR1':3,'Reg':4})
        elif feature == 'Exter Qual' or feature == 'Heating QC' or feature == 'Kitchen Qual':
            df[feature] = df[feature].map({'Po':1, 'Fa':2, 'TA':3,'Gd':4, 'Ex':5})
        elif feature == 'Bsmt Qual' or feature == 'Fireplace Qu':
            df[feature] = df[feature].map({'NA':0, 'Po':1, 'Fa':2, 'TA':3,'Gd':4, 'Ex':5})
        elif feature == 'Bsmt Exposure':
            df[feature] = df[feature].map({'NA':0, 'No':1, 'Mn':2, 'Av':3,'Gd':4})
        elif feature == 'BsmtFin Type 1':
            df[feature] = df[feature].map({'NA':0, 'Unf':1, 'LwQ':2, 'Rec':3,'BLQ':4, 'ALQ':5, 'GLQ':6})
        elif feature == 'Garage Finish':
            df[feature] = df[feature].map({'NA':0, 'Unf':1, 'RFn':2, 'Fin':3})
```

Fig.2. Code snippet that convert values of ordinal features to numerical

- Convert value of ordinal features to numerical values

Feature Selection – Continuous/Discrete

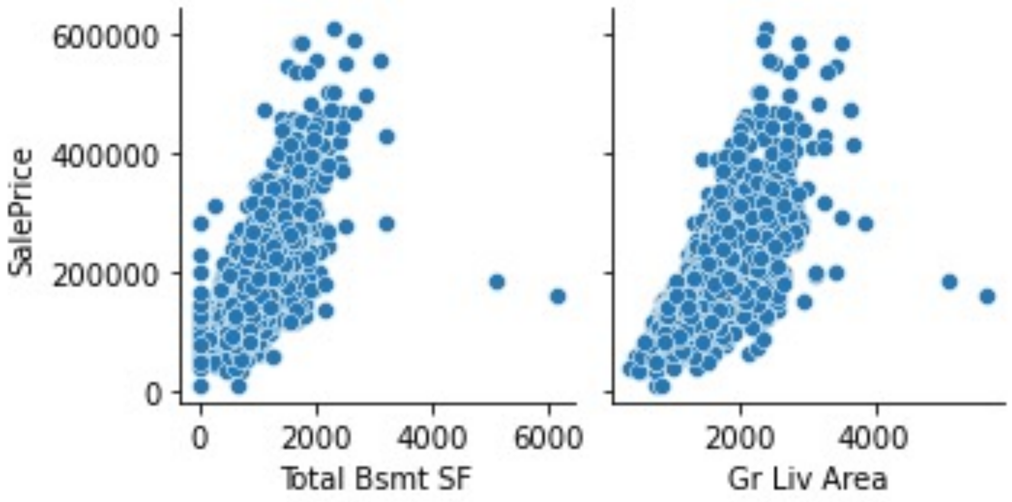


Fig.3. Continuous features that are correlated with Sale Price

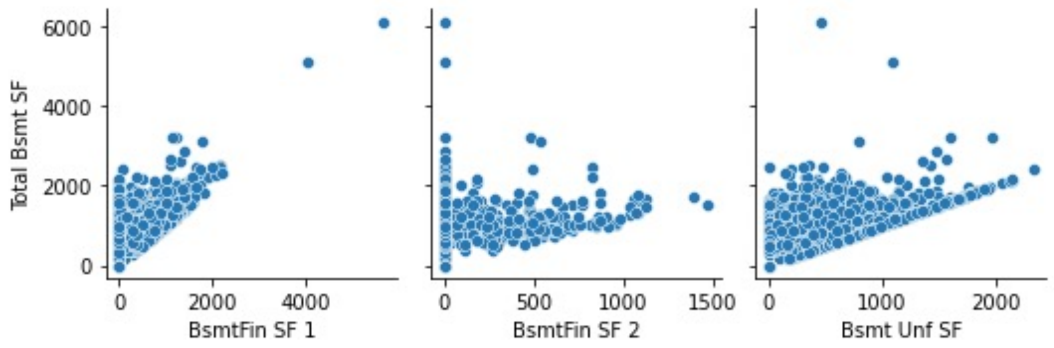


Fig.4. Collinearity between Continuous features

- Pair plots are used to find the correlation between continuous/discrete relationship
- Collinearity between continuous features can be removed
- Reduced down to 2 continuous features: Total Bsmt and Gr Liv Area

Feature Selection – Nominal Features

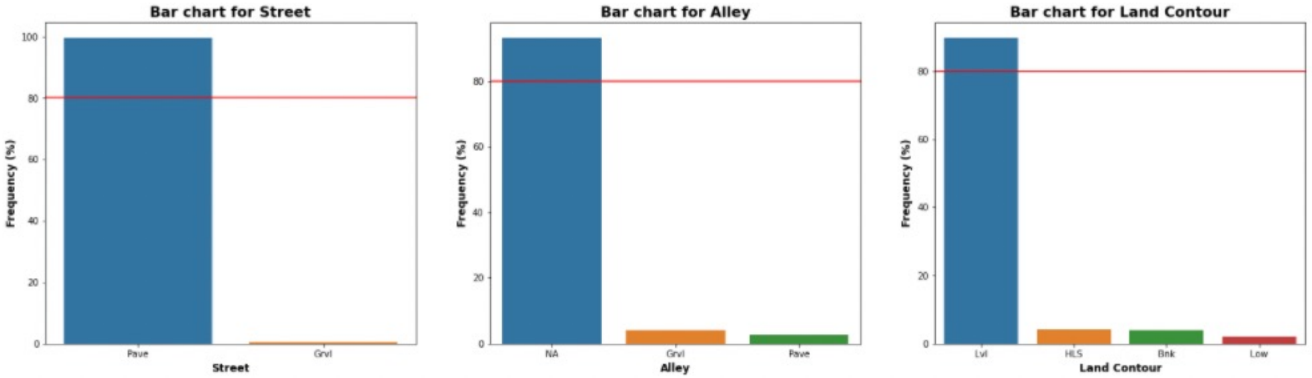


Fig.5. Nominal features with overwhelming occurrence of a single category

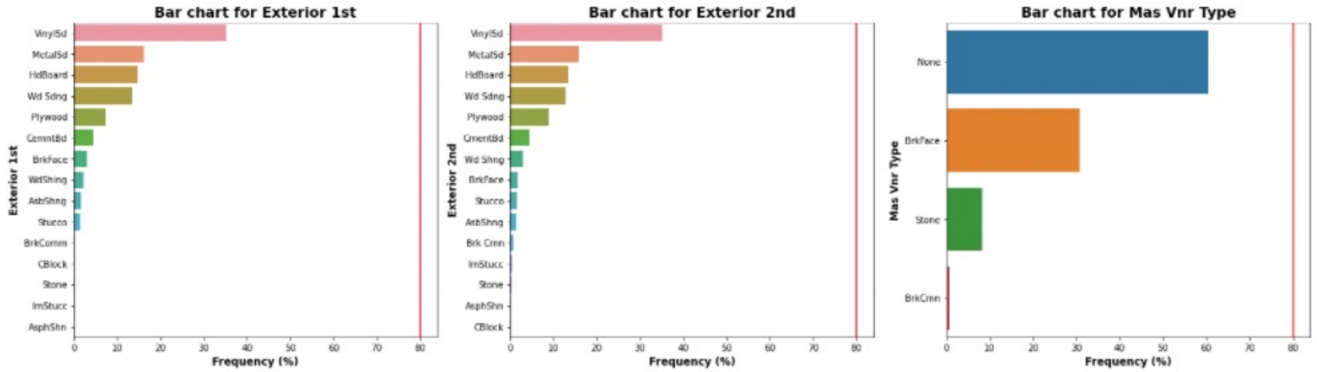


Fig.6. Nominal features without overwhelming occurrence of a single category

- Bar charts used to visualize nominal features with a single category > 80% occurrence
- Reduced to 11 nominal features

Feature Selection – Ordinal Features

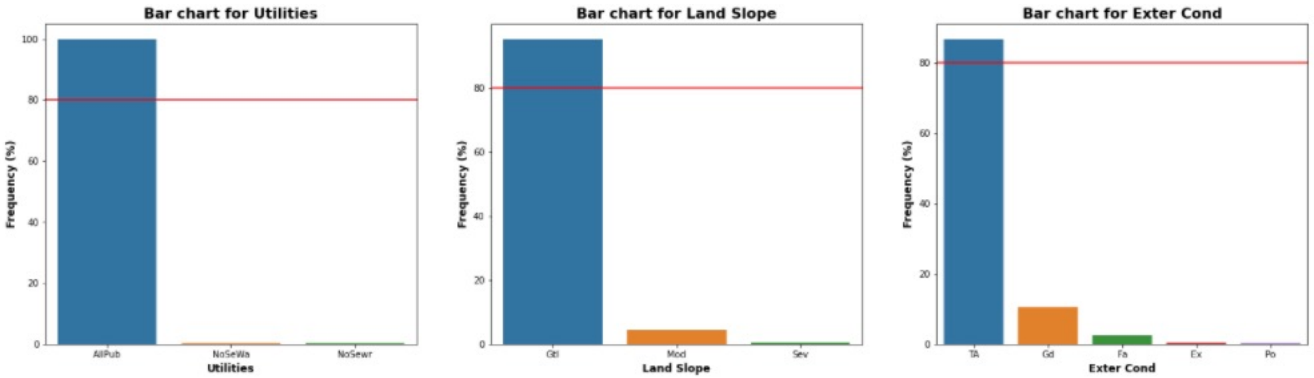


Fig.7. Ordinal features with overwhelming occurrence of a single category

- Bar charts used to visualize ordinal features with a single category > 80% occurrence
- Reduced to 11 ordinal features

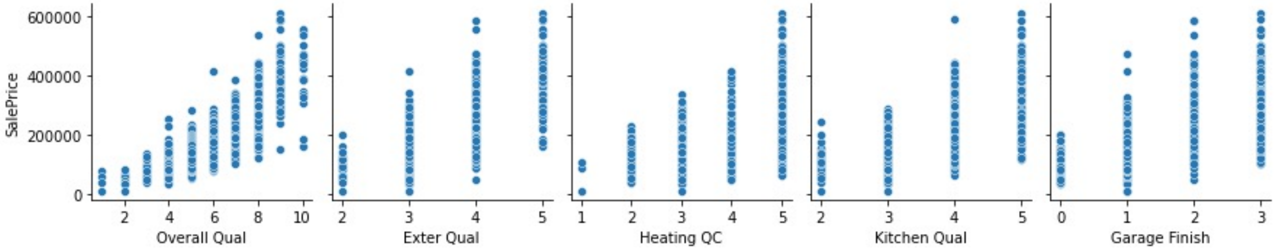


Fig.8. Ordinal features that are correlated with Sale Price

- Pair plots are used to find the correlation for ordinal features after converted to numerical values
- Further reduced to 5 ordinal features

Model Evaluation – Root Mean Square Error (RMSE)

Table 1. Table of comparison for different regression models

	Description	Hyperparameter	Number of Features	CV RMSE	Kaggle RMSE
Model 1	Linear Reg.	-	2	49996.5	46816.5
Model 2	Linear Reg.	-	85	34992.9	35610.1
Model 3	Ridge Reg.	alpha = 26	85	34541.8	34482.2
Model 4	Lasso Reg.	alpha = 97.7	38	34328.9	34264.7
Model 5	ElasticNet Reg.	alpha – 0.02, l1_ratio = 0.3	85	34546.9	34441.8

- Lasso Regression model has the best predictive performance in terms of RMSE
- Used lesser features than other models

Model Evaluation – Top 10 features

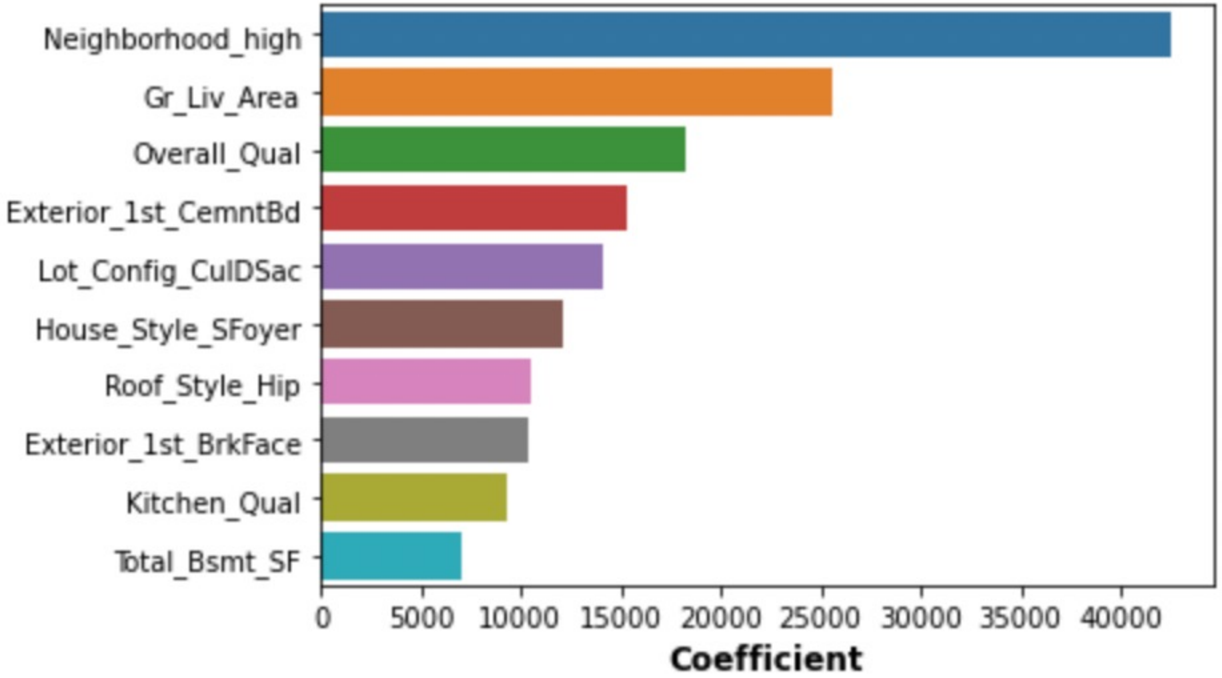


Fig.9. Top 10 features that affect the Sale Price

Conclusion

- Lasso regression has the best predictive performance among the models with 50% reduction in the number of features with RMSE of ~34K
- The Lasso regression model produce a set of coefficients for the respective features
- The top 3 features that will influence the price are location of the house, the total area of house and the overall quality of the house
- With a set of features, the model is able to predict the price of a house