House Price Prediction

Introduction

- Homebuyers have requirements when looking for a home to buy
  - Number of bedrooms
  - Square footage
  - Location
  - Age of house

- Goal: Create a machine learning model to confidently predict the price of homes in Ames, Iowa
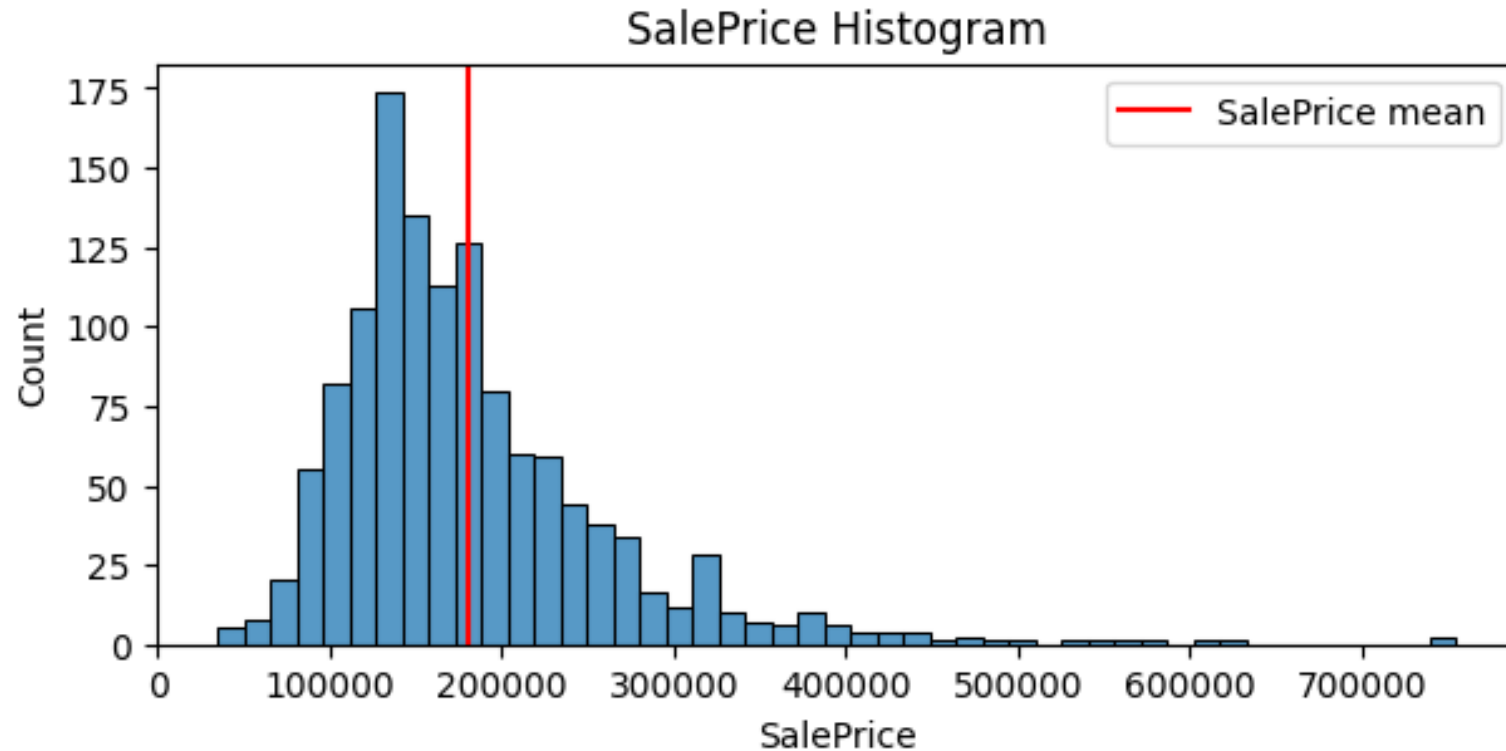
Data

- CSV file
  - 1259 entries with 81 descriptive columns for homes in Ames, Iowa
- Text file
  - Description of each column of the .csv file

Data Wrangling

- Missing data
  - Four columns with over 80% missing data were removed (PoolQC, MiscFeature, Alley, Fence)
  - FireplaceQu had 47.6% missing data
  - These columns were dropped
- Heavily imbalanced columns were dropped
- Redundant columns were dropped
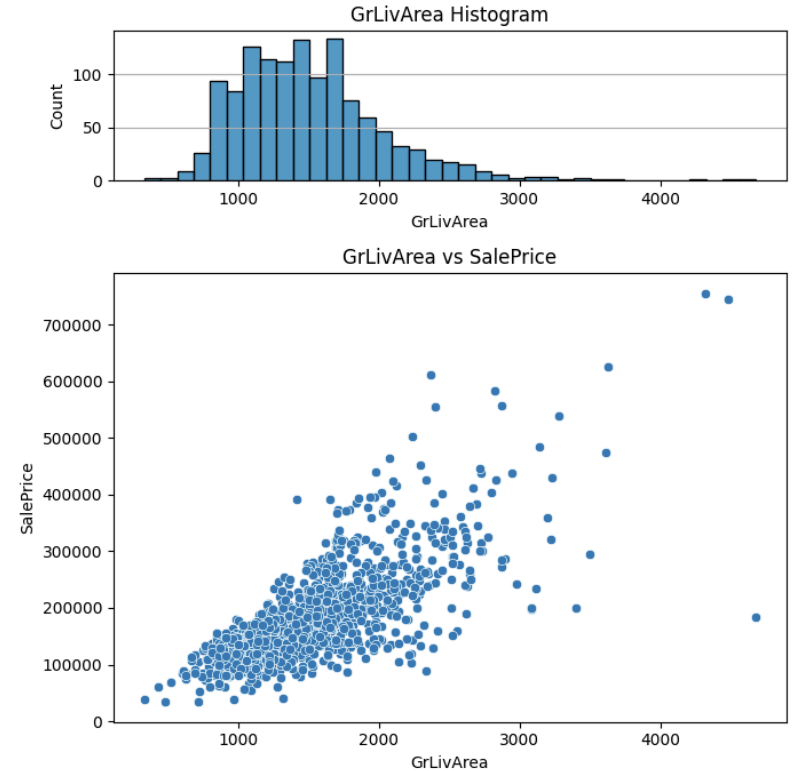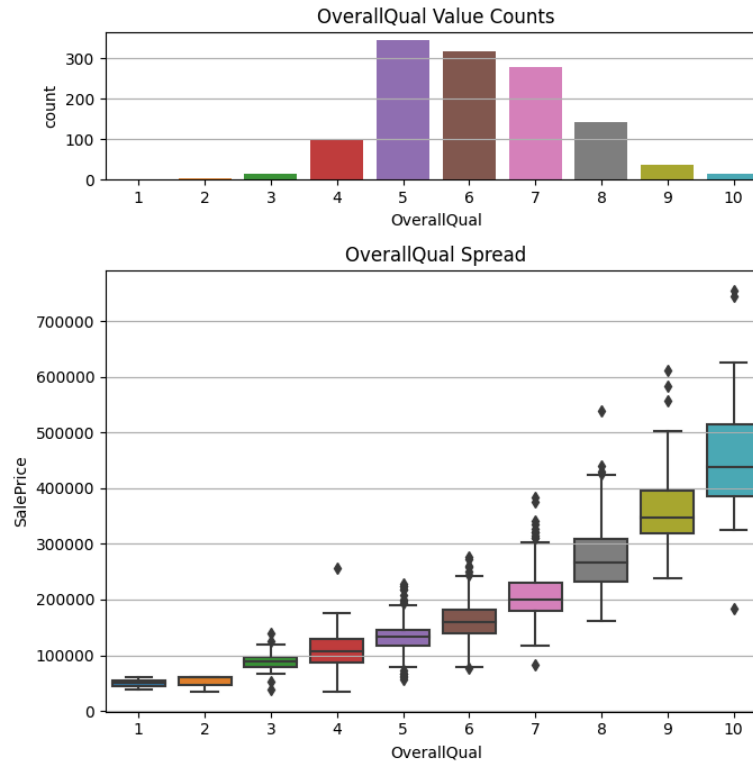 - GarageCars and GarageArea

Exploratory Data Analysis
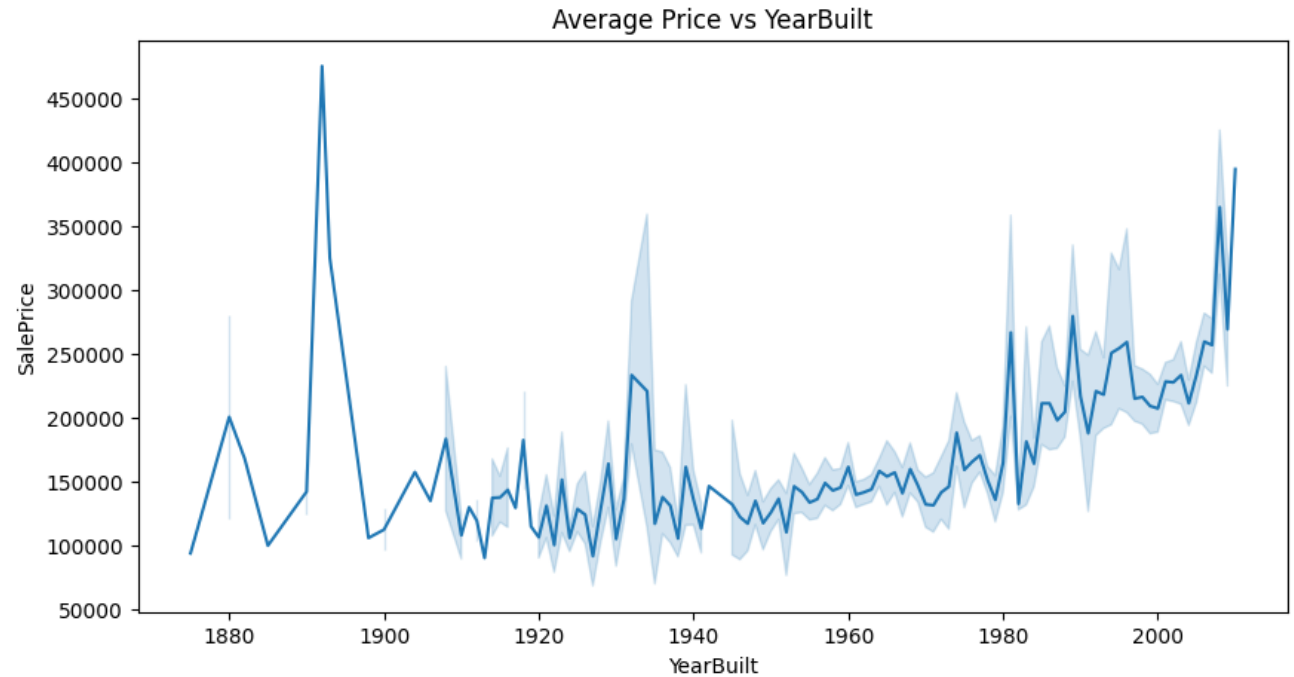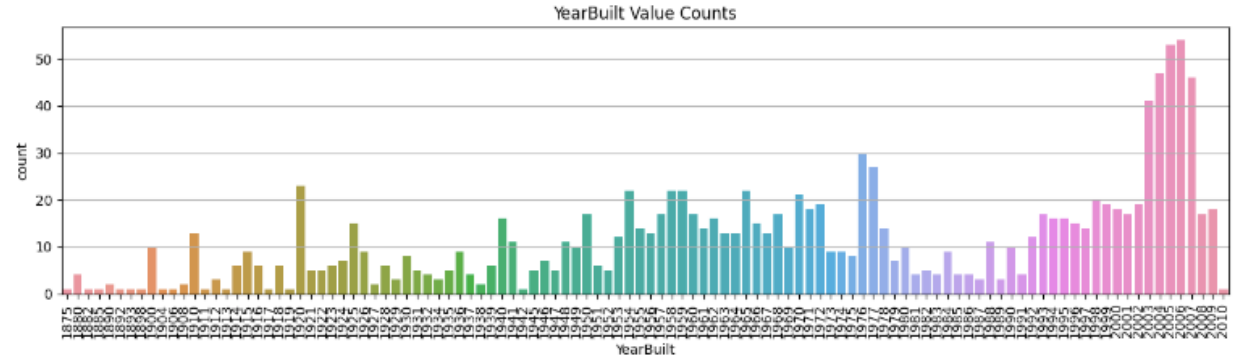
- Target feature (SalePrice) is right-skewed

Exploratory Data Analysis

- Categorical and discrete features were visualized with count plots and box plots and ontinuous features were visualized with histograms and scatter plots

- OveralQual and GrLivArea had strong relationship with SalePrice

Exploratory Data Analysis

- Notable increase in average price with YearBuilt

- High variation before 1910 is due to low number of samples

YearBuilt Value Counts

Average Price vs YearBuilt
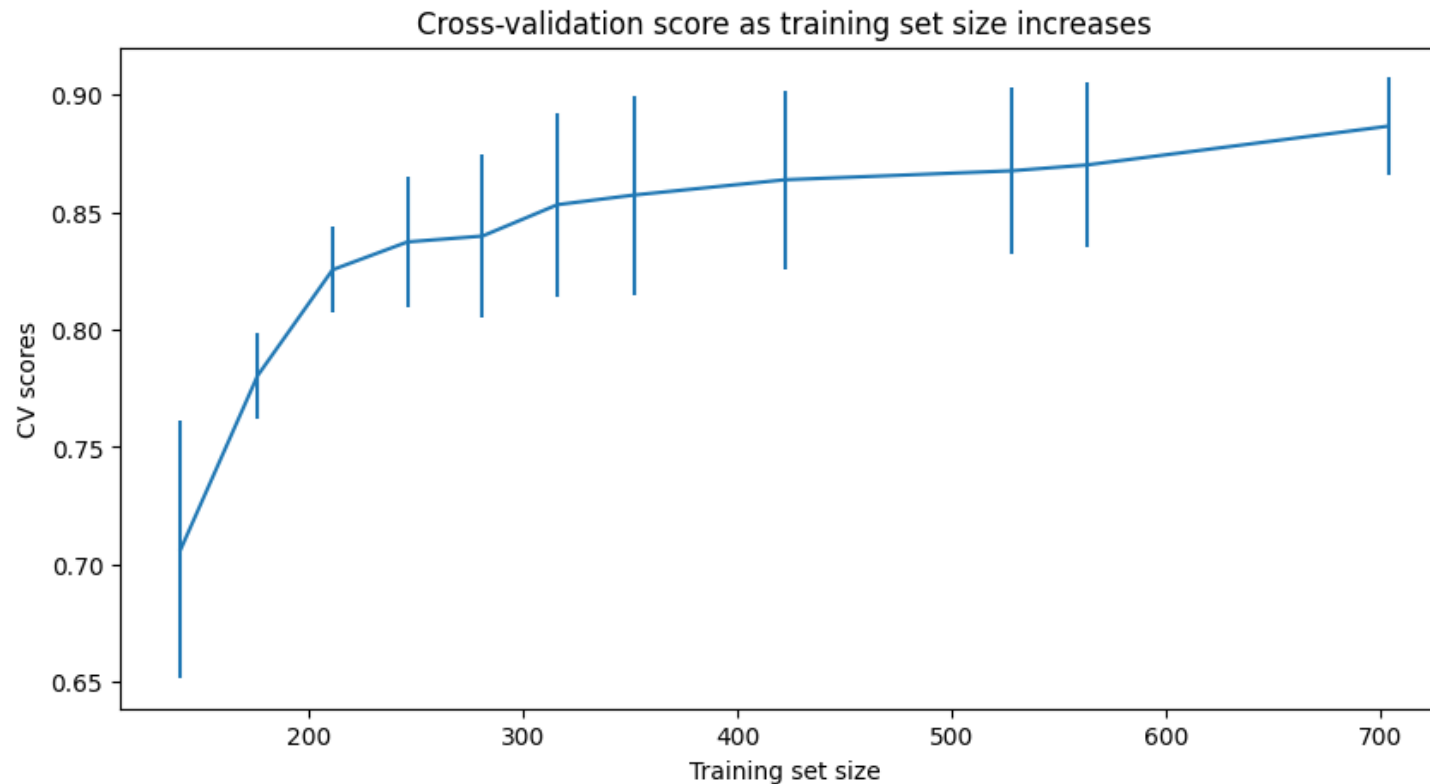
Preprocessing and Training

- Categorical features were one-hot encoded

- 70/30 train/test split

- Missing values were imputed with the median

- Scaled around 0 with a standard scaler algorithm

Modeling and Evaluation

- Random forest regressor model
- GridSearch method with five-fold cross-validation to find optimal parameters
  - Max depth: 15
  - N estimators: 1000
- Average r-squared score from 5-fold cross-validation: 0.886
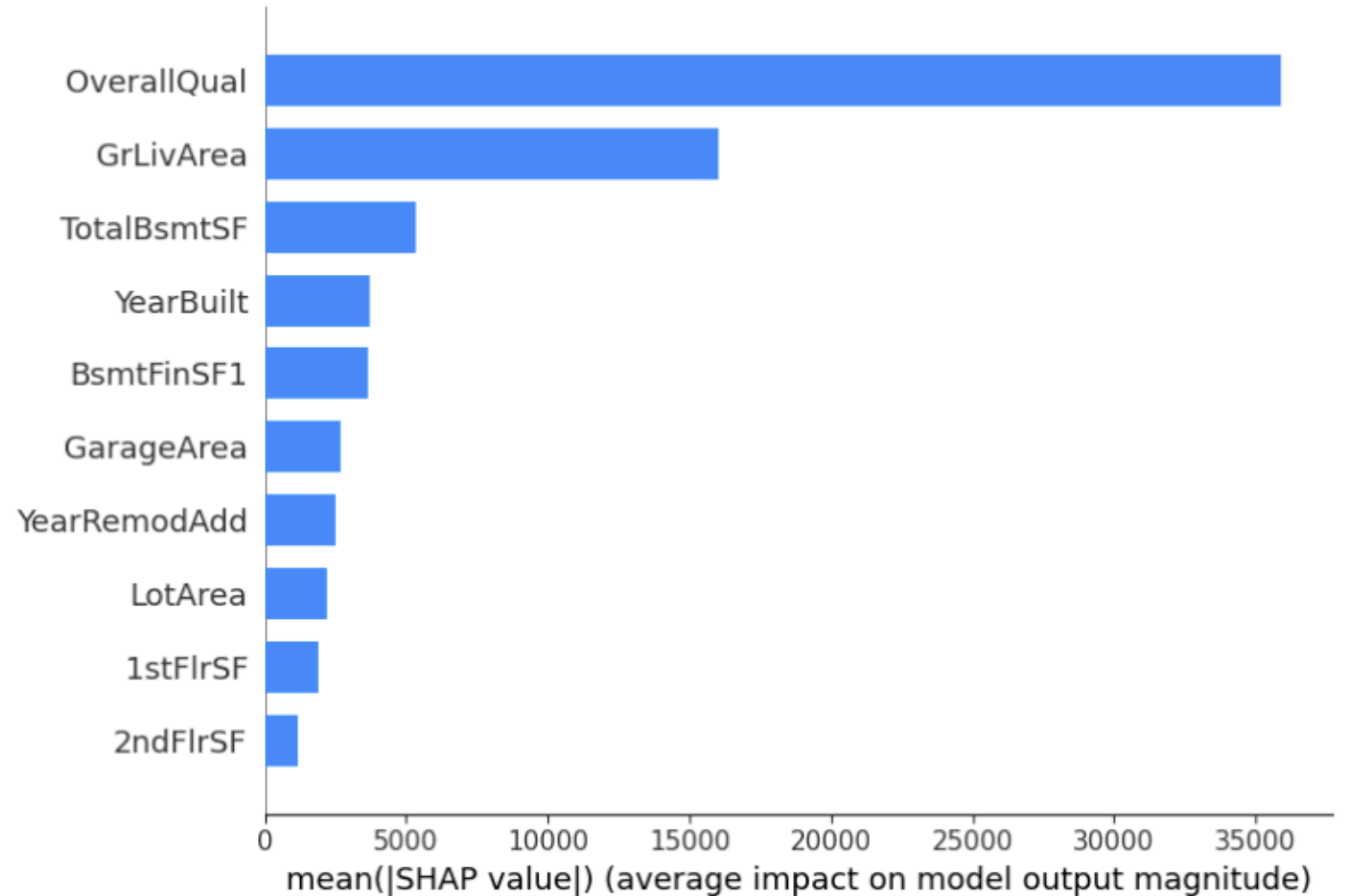- Average MAE: 17,302.68
- MAE std: 746.15

Modeling and Evaluation

- Learning curve indicates the presence of sufficient data

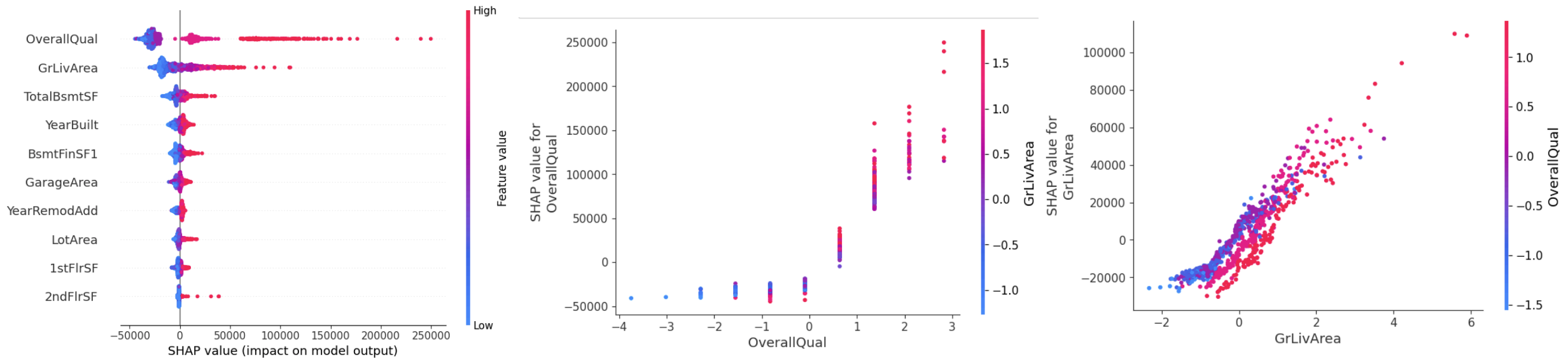Cross-validation score as training set size increases

Modeling and Evaluation

- OveralQual and GrLivArea were most influential features in accurately predicting the sale price

Modeling and Evaluation

- Dependency plots of 10 most important features indicate strong relationship with these features and SalePrice

Results

- Overall quality of house and above ground living area are most important features in accurately determining the price of a house