Technical Report: Ames Housing Price Prediction

John Zeiders

Section 1: Technical Details

1. **Feature Selection:** Dropped several columns based on predefined configurations:

- Forced drop columns: 'Latitude', 'Longitude'
- Highly correlated features: 'TotRms_AbvGrd', 'Garage_Yr_Blt', 'Garage_Area', 'Latitude'
- Potential non-linear features: 'Lot_Frontage'
- Sparse categories: 'Street'

2. **Feature Type Handling:**
    - Identified numeric features as those with types float or int in Pandas
    - Identified categorice features as those with type object in Pandas
    - The following numeric feature were then manually coereced as categorical:
    - Treated certain numeric features as categorical: 'MS_SubClass', 'Overall_Qual', 'Overall_Cond', 'Mo_Sold', 'Bsmt_Full_Bath', 'Bsmt_Half_Bath', 'Full_Bath', 'Half_Bath', 'Bedroom_AbvGr', 'Kitchen_AbvGr', 'Garage_Cars'. This is despite being interpreted as numeric by pandas.

3. **Preprocessing Pipeline:** A pre-processing pipeline was implemented using sklearn.
    - Numeric features:
        - Imputation: Median strategy for missing values
        - Scaling: StandardScaler applied
        - Outlier handling: Custom OutlierCapper transformer (capping at 5th and 95th percentiles)
    - Categorical features:
        - Imputation: Constant strategy, filling with 'Missing'
        - Encoding: OneHotEncoder with use_cat_names=True

Model Implementation

Two models were selected & trained using the implementations from SKLearn.

1. **Elastic Net Regression:**
    - Used ElasticNetCV with the following parameters:
        - Cross-validation folds: 5
        - L1 ratio range: [0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.8, 0.9, 0.95, 0.97, 0.99, 1]
        - Random state: 42

    - Utilized all available CPU cores (n_jobs=-1)
2. **XGBoost Regressor:**
    - Parameters:
        - Number of estimators: 5000
        - Learning rate: 0.05
        - Max depth: 6
        - Subsample: 0.8
        - Column sample by tree: 0.8
        - Random state: 42
        - Utilized all available CPU cores (n_jobs=-1)

Section 2 Performance Metrics

MacBook Pro, Apple M2 Max, 32GB.

model	RMSE	Train Time (seconds)	fold_num
XGBoost	0.12062	8.97608	1
ElasticNetCV	0.12012	1.80893	1
XGBoost	0.12271	7.76506	2
ElasticNetCV	0.11578	1.44031	2
ElasticNetCV	0.11275	1.83849	3
XGBoost	0.115	7.25294	3
XGBoost	0.11781	7.03987	4
ElasticNetCV	0.11922	1.74771	4
ElasticNetCV	0.11135	1.61322	5
XGBoost	0.1127	6.26483	5
XGBoost	0.13068	8.685	6
ElasticNetCV	0.13436	1.50211	6
ElasticNetCV	0.13293	1.66389	7
XGBoost	0.13067	9.04411	7
XGBoost	0.1245	7.57617	8
ElasticNetCV	0.12754	1.46828	8
XGBoost	0.13171	7.38593	9
ElasticNetCV	0.13241	1.52945	9
ElasticNetCV	0.12568	1.61462	10
XGBoost	0.11988	6.33288	10