COMS-W4995 Applied Machine Learning Project Final Report

Team 7: Ananya Gandhi, YeongWoo (Janie) Kim, Austin Schaefer, Heather Song

1. Background and Context

In the realm of real estate, the accurate pricing of residential properties holds paramount importance for buyers, sellers, and industry professionals. This project centers around a dataset with 79 variables detailing residential properties in Ames, Iowa. Our primary objectives are twofold:

1. to accurately forecast the final selling price of houses and
2. to identify inherent trends within the dataset.

2. Exploratory Data Analysis (EDA)

2.1. Summary Statistics and Distribution of Sale Prices

The dataset exhibits variability in sale prices, ranging from below $100,000 to above $300,000. The mean sale price is approximately $180,000, while the median is around $163,000. The distribution is slightly right-skewed, suggesting the presence of higher-priced properties contributing to the higher mean.

2.2. Numerical and Categorical Features

The dataset contains 79 features, with 37 numerical and 42 categorical variables.

Scatter plots and box plots were generated to visualize the relationship between numerical and categorical features and the logarithmic sale prices, which aided in identifying potential patterns and outliers.

We decided to remove the features that exhibited no correlation or had highly imbalanced labels in order to lower the complexity of the model training.

2.3. Missing Value Imputation

The dataset contains missing values, requiring careful handling before model implementation. Notable features with missing values include "PoolQC," "MiscFeature," "Fence," "Alley," and "FireplaceQu." It's important to note that missing values in these features indicate the absence of the corresponding property attributes, such as a pool or fireplace.

Other features like “Electrical” and “LotFrontage” were either dropped if there were few missing values or median imputed.

2.4. Splitting and Encoding

The dataset was split into training, validation, and test sets. The training set comprises 60% of the data, while the validation and test sets each consist of 20%. Standard scaling was applied to the features to ensure uniformity in scale across the dataset.

Ordinal encoding was performed on selected categorical features with an inherent order, while one-hot encoding was applied to others. Target encoding was used on high-cardinality categorical columns.

2.5. Heatmaps

Correlation matrices and heatmaps were generated to explore relationships between features. Highly correlated features, identified by a threshold of 0.85, were removed to enhance model interpretability and performance.

3. Applying ML Techniques

3.1 Model Implementation and Analysis

To reduce complexity and improve performance of the models, we performed feature selection with Lasso regression. The initial 79 features were reduced to 23. With a simpler dataset, our models were able to focus on more important features and have higher accuracy.

Without lasso feature selection

	Root Mean Squared Error	R Squared
Linear Regression	2.495765e+15	-9.172415e+20
Ridge Regression	4.892156e+04	6.480000e-01
Lasso Regression	4.310900e+04	7.260000e-01
ElasticNet Regression	4.444866e+04	7.090000e-01
XGBoost	2.635572e+04	8.980000e-01
CatBoost	2.687239e+04	8.940000e-01
Support Vector Regression	4.493121e+04	7.030000e-01
Random Forest Regressor	3.199826e+04	8.490000e-01

With lasso feature selection

	Root Mean Squared Error	R Squared
Linear Regression	46244.556	0.685
Ridge Regression	46235.556	0.685
Lasso Regression	43108.733	0.726
ElasticNet Regression	44422.875	0.709
XGBoost	26232.224	0.899
CatBoost	27912.563	0.885
Support Vector Regression	43670.423	0.719
Random Forest Regressor	32590.369	0.844

As we can see above, implementing Lasso Feature Selection significantly improves the Root Mean Squared Error and R Squared values, especially for Linear Regression. In order to find the best model, we performed hyperparameter tuning with RandomizedSearchCV and GridSearchCV.

3.2. Model Performance and Comparison

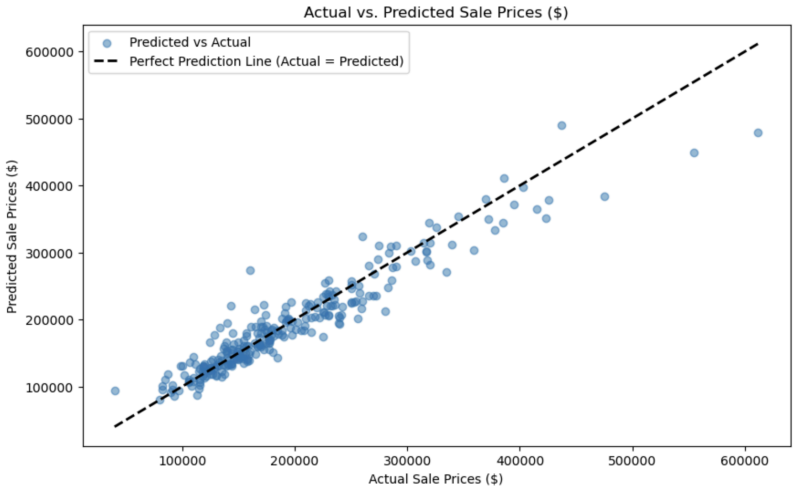
After implementing all 8 models, we used Root Mean Squared Error (RMSE) and R-squared to evaluate the regression models. RMSE can supplement the limitations of R-squared by giving us the average error in the units of the measured variable, and R-squared can show us the proportion of the outcome's variance that is explained by the model.

	Root Mean Squared Error	R Squared
Linear Regression	46244.556	0.685
Ridge Regression	46235.556	0.685
Lasso Regression	43108.733	0.726
ElasticNet Regression	44422.875	0.709
XGBoost	25803.850	0.902
CatBoost	27912.563	0.885
Support Vector Regression	43670.423	0.719
Random Forest Regressor	32590.369	0.844

As shown in the table, XGBoost has the best performance. Our best-performance model can capture approximately 90% of the variance in the selling prices of houses in the dataset.

4. Analysis and Conclusion

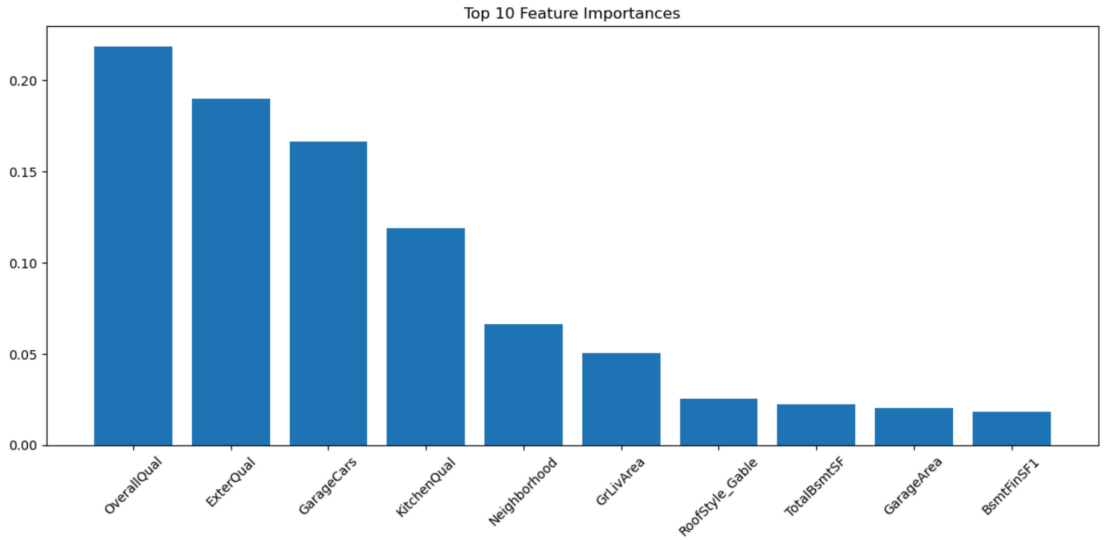
4.1 Model Predicting Power Analysis



As shown in the table above, the XGboost model that we implemented was able to accurately forecast the final selling price of houses since most of the predicted sale prices are aligned with the actual sale prices on the test dataset.

4.2 Iowa Housing Sales Trend Analysis

Since XGBoost has the best performance, we decided to apply XGBoost to draw further insights from the dataset.



Based on the feature importance output from the XGBoost model, we can draw several insights:

- Quality Over Quantity: While the size of the living areas is important, the overall quality of construction and finishes seems to have a more significant effect on sale prices.
- Functional Spaces: Functional spaces like kitchens, basements, and garages and the quality of these spaces is prioritized over other features deemed to be less significant, like roof style or pavement type.
- Modern and Updated: Newer homes with modern amenities (ex. updated kitchens) tend to fetch higher prices, which could be due to less maintenance required or a preference for contemporary designs.
- Function vs. Aesthetic: Overall quality and size of living spaces tended to dominate over other aesthetic features, such as the roof style or material used during construction.
- Investment Areas: For those looking to sell or improve their property value, focusing on overall quality, expanding or finishing living areas, and updating kitchens could potentially offer the best return on investment.

These findings are crucial for sellers, buyers, and real estate professionals as they provide a data-driven foundation for understanding what features are likely to drive home prices in the Ames, Iowa housing market.