



UNIVERSITY OF
MARYLAND

ROBERT H. SMITH
SCHOOL OF BUSINESS

BUDT 758J

Final Project Report

Spring 2024

Section 1: Team member names and contributions

Team name on the Kaggle.com: BUDT758J_spring24_Group_03

Chien-Jui Huang: Data Cleaning, Linear Model Predictions

Wendy Guan: Data Cleaning, Lasso Model Predictions

Kelan Quan: EDA, Data Cleaning, Organizing Report

Pravah Malunjar: Data Cleaning, XGBoost Model Predictions

Jooyoung Park: Data Cleaning, Neural Network Model

Section 2: Business Understanding and Project Objectives

The real estate market is inherently complex, influenced by various factors such as location, size, amenities, and price conditions. Accurately predicting house prices is crucial for real estate agents, homebuyers, sellers, and investors to make informed decisions. In this project, we leveraged advanced regression techniques to develop models for predicting house prices.

The predictive models present values to stakeholders in the real estate industry. Firstly, its accuracy is highlighted through the utilization of advanced regression techniques, ensuring precise price predictions that mitigate the risks associated with overpricing or underpricing properties. Moreover, its efficiency stands out as it automates the price estimation process, allowing stakeholders to save time and resources compared to traditional manual appraisal methods. Finally, the model facilitates informed decision-making by providing users with data-driven insights and reliable price estimates, ultimately leading to better outcomes in real estate transactions.

The predictive model for house prices provides a valuable tool for stakeholders in the real estate industry, enabling them to make informed decisions, optimize pricing strategies, and maximize returns on investment. Our model contributes to a more efficient and transparent real estate market by leveraging advanced regression techniques.

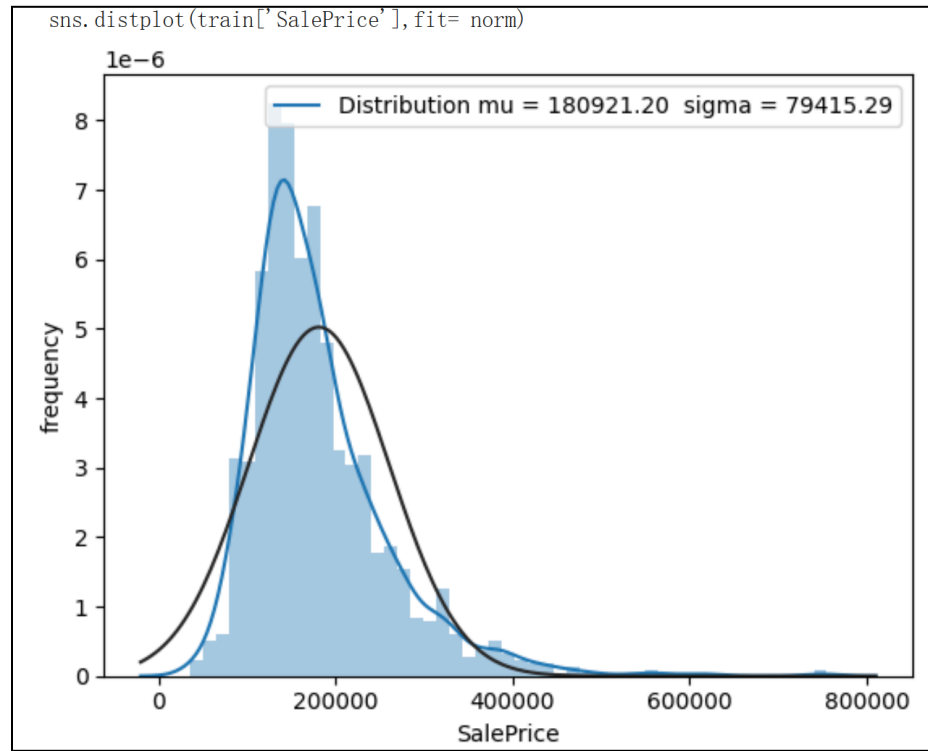
Section 3: Data Understanding and Exploratory Data Analysis

Table 1: Data Understanding

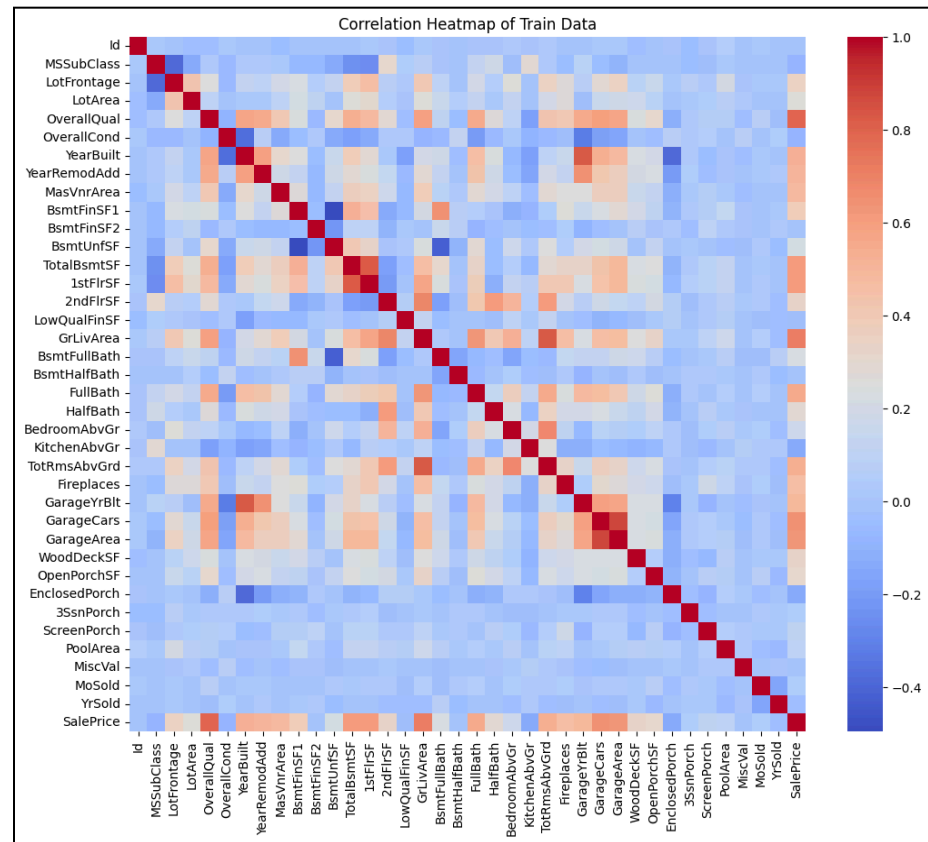
Property Characteristics	<ul style="list-style-type: none">● Lot size (LotArea)● Building size (1stFlrSF, 2ndFlrSF, GrLivArea)● Number of bedrooms, bathrooms (Bedroom, FullBath, HalfBath)● Overall quality and condition(OverallQual, OverallCond)● Year built and remodeled (YearBuilt, YearRemodAdd)● Exterior and interior material quality (ExterQual, ExterCond, KitchenQual)
Location	<ul style="list-style-type: none">● Zoning classification (MSZoning)● Neighborhood (Neighborhood)● Proximity to amenities, main roads, railroads (LotConfig, Condition1, Condition2)● Type of road access (Street)● Lot configuration and slope (LotShape, LandSlope)
Property Amenities	<ul style="list-style-type: none">● Type of foundation, roof, and exterior covering (Foundation, RoofStyle, RoofMatl, Exterior1st, Exterior2nd)● Type and quality of basement (BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2)● Basement square footage and number of bathrooms (BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath, BsmtHalfBath)● Heating and cooling systems (Heating, HeatingQC, CentralAir)● Garage features (GarageType, GarageYrBlt, GarageFinish, GarageCars, GarageArea, GarageQual, GarageCond)
Additional Features	<ul style="list-style-type: none">● Lot frontage, alley access, utilities available (LotFrontage, Alley, Utilities)● Home functionality rating (Functional)● Fireplace presence and quality (Fireplaces, FireplaceQu)● Presence of additional features like pools, fences, and miscellaneous features (PoolArea, PoolQC, Fence, MiscFeature, MiscVal)

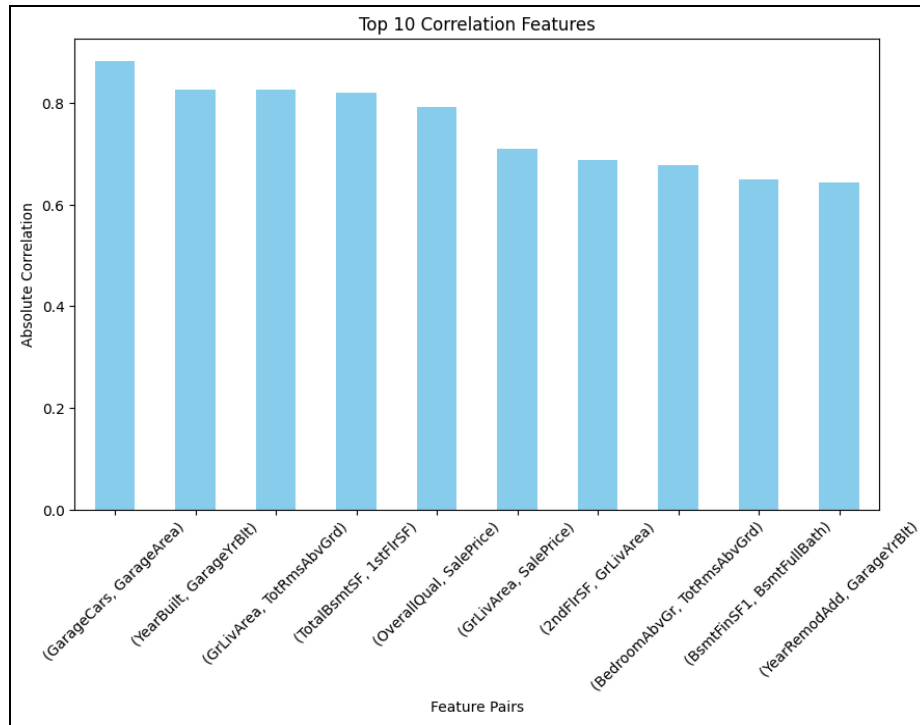
Exploratory Data Analysis

The distribution of the SalePrice (target variable) reveals numerous outliers, causing a right skew that could complicate model training. To address this challenge, one effective strategy involves applying a logarithmic transformation to the target variable. This adjustment assists machine learning models in achieving improved generalization performance by mitigating the influence of extreme values. Alternatively, excluding outlier instances altogether presents another viable approach to enhance model robustness. Each method offers distinct advantages, providing flexibility in addressing outliers and optimizing model accuracy.

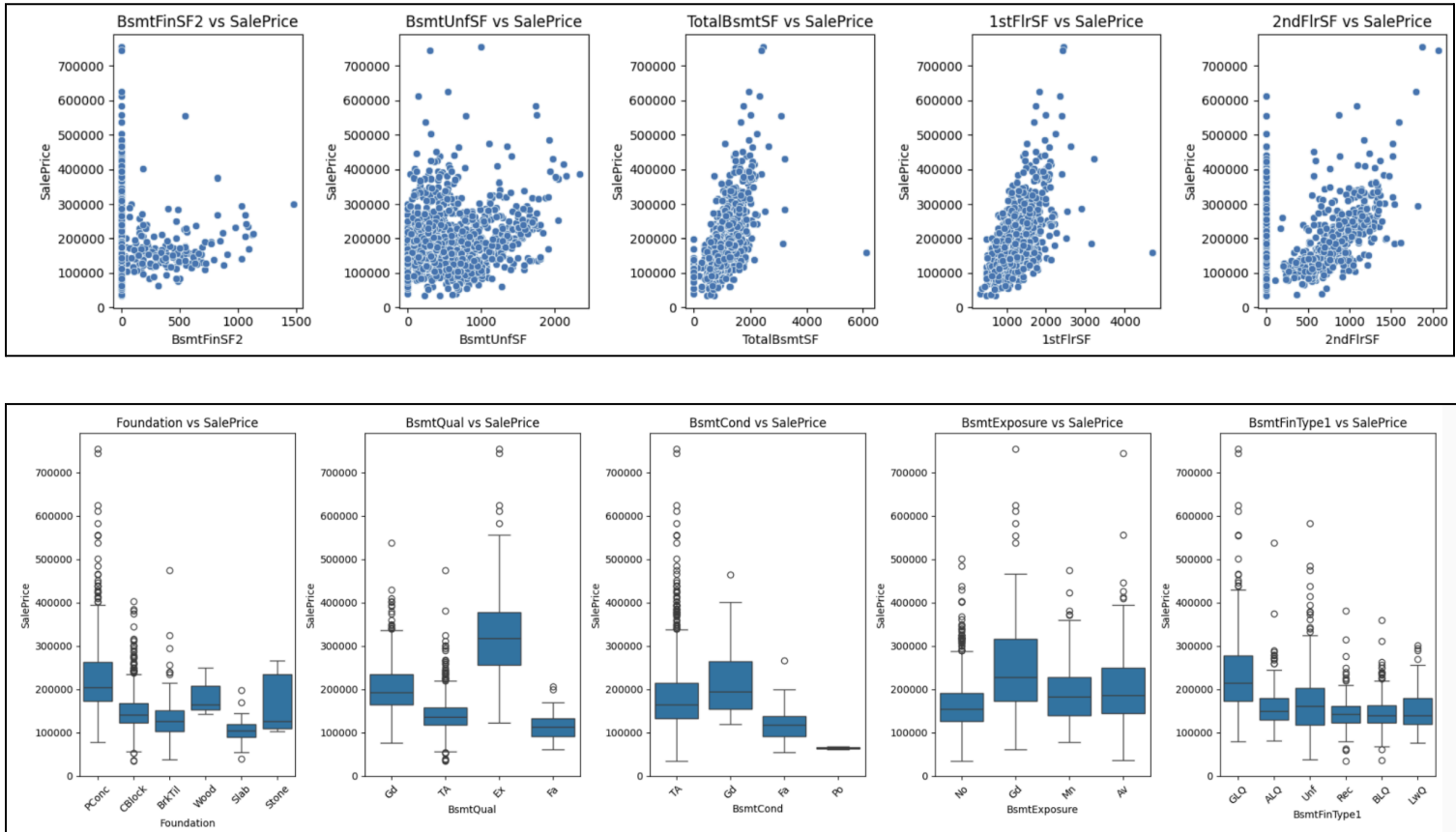


Correlation Heatmap of Train Data for Numerical Variables. We delve into correlation heatmap to help identify patterns between multiple variables. From this graph, we can see that most of the variables correlate 0.2 to 0.6, and for our target variable SalesPrice, OverallQual has the highest correlation except itself. The below graph shows the top ten highest correlation feature groups.





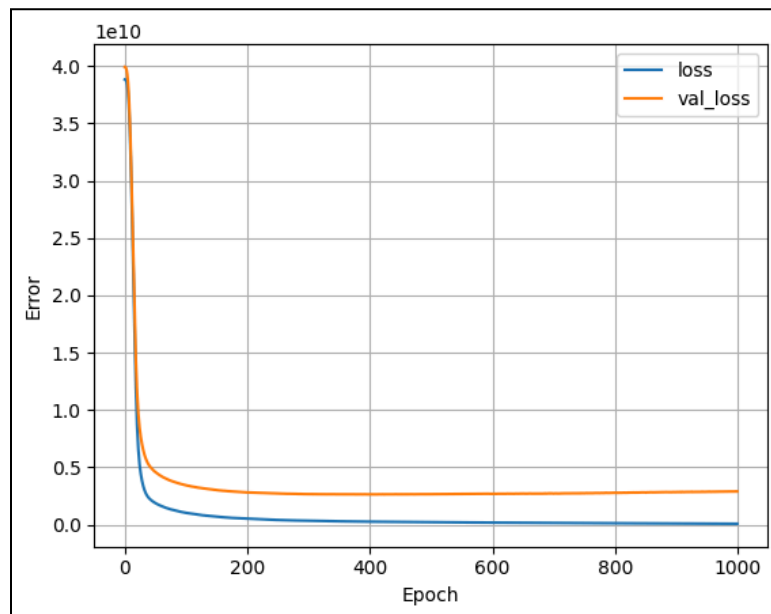
We also create scatter plots and boxplots between SalePrice and numeric features and categorical features to visualize the correlation which better helps us determine which features might have a more meaningful impact on the model training.



Section 4: Evaluation and Modeling

1. Neural Network Model

Loss Curve for Neural Network



After around 100 epochs, the root squared loss for the validation set starts to become stable at around 0.3. Metrics for neural networks were expected to be worse than other models since there was an insufficient amount of data for the data to be independent and thus unable to provide accurate predictions. The score on Kaggle was also low, therefore the neural network model was depreciated.

2. Lasso Regression Model

The evaluation performance of the Lasso model is shown below:

Validation RMSE: 32007.35

Best alpha: 10.0

Training R^2 : 0.90

Validation R^2 : 0.87

The lasso regression has a regularization parameter alpha that optimizes the balance between model complexity and performance. Through 10-fold cross-validation, the optimal alpha was selected to minimize RMSE. For this model, numerical features were standardized, and categorical features were processed through on-hot encoding. The model achieves a training R^2 of 0.90 and validation R^2 of 0.87, indicating strong predictive performance. The validation RMSE of approximately 32007.35, demonstrating the model's average prediction error. The Lasso model was used as a feature selection tool as well for the Xgboost model.

3. XGBoost Model

The evaluation performance of the Lasso model is shown below:

R^2 : 0.9107554005566659

MSE: 622757182.5730572

RMSE: 24955.10333725463

Best Learning Rate: 0.079

Best Lambda: 18

Significance: An exceptional R-Square value of 0.9107 indicates that more than 91% of the variance in the target variable from the predictors can be explained by the model. This suggests a well-fitting model

and excellent predictive power. The model's predictions are now more accurate than in previous iterations, as seen by the declines in MSE and RMSE, which indicate a drop in average prediction error. The model appears to converge more quickly to the ideal solution, as indicated by the higher learning rate of 0.079. This aggressive learning rate seems to be well-balanced by strong regularization, demonstrating XGBoost's adaptability to varying hyperparameters. The efficacy of regularization controlling model complexity and avoiding overfitting—a frequent problem in many predictive modeling tasks—is demonstrated by the sustained usage of a high lambda value.

Benefits of XGBoost: XGBoost is well-known for its versatility in handling various data kinds, connections, and distributions, which makes it an excellent tool for extracting intricate patterns from data. It makes use of a gradient-boosting framework, which improves prediction performance over iterations by gradually correcting errors generated by prior trees in order to maximize accuracy. The L1 (lasso regression) and L2 (ridge regression) regularization features of XGBoost help lessen overfitting and enhance model performance on omitted data. Strong regularization is provided by the lambda value of 18, which helps to control the model's complexity and improves its ability to generalize. As a result, error measurements like MSE and RMSE have less volatility and bias.

4. Linear Regression Model

The evaluation performance of the linear regression model scores at RMSE 31418.61 which is slightly better than the Lasso but worse than xgboost. Linear regression is a simple and easy-to-understand model. Its simplicity makes it computationally efficient and less prone to overfitting, especially when dealing with small to medium-sized datasets. This characteristic is proved by the performance of this model on our leaderboard. Besides, linear regression can be robust to outliers, which is the case for our target variable. Our linear regression provides information about the importance of all features, not just the ones selected by Lasso. This can be valuable for gaining insights into the predictive power of different features.

Section 4: Result

Score from Kaggle.com

After the submission, the leaderboard demonstrates that the XGBoost model is the best one with the highest prediction ranking.

Overview

Data

Code

Models

Discussion

Leaderboard

Rules

Team

Submissions

828

BUDT758J_spring24_Group_03







0.12849

5

30m

Your Best Entry!

Your submission scored 0.18910, which is not an improvement of your previous score. Keep trying!

Submission and Description		Public Score ①
	linear_house_price_predictions.csv Complete · Chien-Jui Huang · 31m ago	0.18910
	neural_net_predictions.csv Complete · Chien-Jui Huang · 1h ago	0.62783
	neural_net_predictions.csv Error · Chien-Jui Huang · 1h ago	
	Lasso_prediction_filled.csv Complete · Chien-Jui Huang · 1h ago	0.21511
	Lasso_prediction_filled.csv Complete · Chien-Jui Huang · 2h ago	0.19259
	XGBoost_house_price_predictions.csv Complete · Chien-Jui Huang · 2h ago	0.12849

Section 5: Reflection/takeaways:

Things our group learned from the project

During the project, our group demonstrated excellence in a number of critical areas. Our success was largely due to our ability to work together as a team, successfully combining our skills to take on a variety of responsibilities like feature engineering, data preparation, model selection, and evaluation. Every time a problem arose, we tackled it methodically, drawing on the individual abilities of each team member to come up with original and workable answers. Our dedication to effective communication made sure that everyone was on the same page regarding project goals and deadlines, which promoted a harmonious and effective work atmosphere. Furthermore, to find the best model for precisely projecting home prices, we experimented with a variety of regression techniques, ranging from conventional linear regression to more sophisticated approaches like random forests and gradient boosting. All in all, our team's cooperative

The main challenges our group faced in the project

The project faced several significant difficulties. Because the dataset had outliers, category variables, and missing values, data preprocessing became a major obstacle. Tackling these problems necessitated thorough preprocessing and cleaning to guarantee the accuracy and consistency of the data for further analysis. Another difficult task was feature engineering, which required careful thought and domain knowledge to identify relevant features from the dataset that could improve predictive performance. Ultimately, it took a lot of effort to optimize the hyperparameters for each regression model, requiring a great deal of trial and error to reach the right performance level. Despite these obstacles, resolving them with careful feature engineering, rigorous data pretreatment, and model tweaking eventually cleared the path for the project's

Reflection on the project

After giving our project some thought, we decided that if we could do it over again, we would take more time to prepare and explore the dataset in depth. To get useful insights, we would need to go deeper into the subtleties and intricacies of the data. Furthermore, to capture intricate relationships between features and the target variable and improve the predictive power of our models, we would place a higher priority on using more advanced feature engineering techniques. We would also investigate the possibility of using ensemble methods

and neural networks to improve performance even more. We would carry out a more thorough feature selection process with a longer timetable to determine the most important factors for precisely projecting home prices. In addition, we would explore more sophisticated methods like neural networks, utilizing their regression capabilities.