GB 656: Machine Learning for Business Analytics
Group 16
Final Project Analysis
December 9, 2024

Housing Prices – Advanced Regression Techniques

Introduction:

Source:

This project comes courtesy of a competition that we found on Kaggle. The author recommended that it was targeted towards students who had recently taken a course in machine learning, and we decided this was perfect!

It is also important to note that this is known as a *Getting Started* competition: It is intended for those who do not have the most exposure to machine learning methodology. We chose this because it allowed us to devote more time to the analytical aspect of machine learning; we can use the data to make more well-informed decisions and recommendations. We hope you enjoy reading up on our findings!

Problem Statement:

In an increasingly competitive real estate market, accurate property valuation is critical for both buyers and sellers. The volatility of real estate prices can arise from many geopolitical factors. Some of which include location, structural characteristics and market condition. These yield challenges for realtors and appraisers. Traditional valuation methodology is both time-consuming and prone to human error; this can result in suboptimal pricing strategies and missed opportunities. Clearly, there is potential to apply machine learning techniques to the real estate industry.

Value Add:

In this competition, we were tasked with predicting the selling price of a given house. This is crucial to advancing the real estate market because firms can dedicate more technological resources to the intricacies of valuation. In turn, human capital can be spent on the seller-to-buyer relationship.

The starting dataset contained 79 features that describe residential homes in Ames, Iowa. chosen to address this challenge by utilizing advanced regression techniques to predict home sales prices. The goal is to build a robust predictive model using a dataset containing comprehensive property characteristics and historical sales prices.

Operational Framework:

Use Cases:

Sellers:

Real estate firms need an efficient and accurate system to estimate property values. This not only provides for a timelier process, but it also enhances their competitive strategy. The most well-equipped firms are often the most innovative, so this has the potential to yield increased profitability.

Buyers:

First-time home buyers can often be deceived by illusory sales techniques. The final sales price can often seem arbitrary at best, so using prior data can help make better assumptions about the future. A well-documented model can alleviate some of the confusion that exists.

Value Proposition:

- We aim to improve accuracy by incorporating machine learning techniques that could identify linear (or nonlinear) relationships and interactions between property features.
- Automation of property valuation will reduce the manual labor required, and this will allow firms to put more focus on the strategic elements involved with the business.
- The deployment of predictive models will allow firms to make data-driven decisions, and these processes will adapt well to a rapidly evolving market.

Business Impact:

- Optimizing the pricing process will ensure fair rates for buyers and sellers
- Realtors benefit from transparent pricing because it builds brand recognition
- Customer loyalty will grow through reputation and lead to future sales

Data Preparation:

We cleaned and preprocessed the dataset to address any discrepancies that existed. For example, “LotFrontage” and “GarageYrBlt” contained many missing values. To account for this, they were filled with the necessary null values.

Categorical values, those that were of type ‘object’ were corrected by encoding them via a label encoder. Following this, we could move on with the model creation phase of our project.

Model Selection:

Feature Engineering:

We created meaningful features such as the total square footage of the house, an interaction term between mass and square footage, and time-based features such as the year since the last remodel. These engineered features enhanced the predictive power of the model.

Potential Models:

1. **Linear Regression:** To establish a baseline for performance benchmarking
2. **Ridge/LASSO Regression:** To address multicollinearity and perform feature selection
3. **Random Forest Regression:** To capture nonlinear relationships and interactions
4. **Gradient Boosting (XGBoost):** To handle complex data structures that other cannot

Evaluation Metrics:

Per the competition instructions, our model was measured by its Root Mean Square Error (RMSE). The RMSE was be calculated on the validation set to generalize the results.

To employ this, our models were compared between one another based on their RMSE. We utilized feature selection methodology to determine those with the greatest prediction power. Grid Searches were then determined to find the optimal model, and it was the XGBoost.

Results:

Benchmark Comparison:

The linear regression model yielded an RMSE of 0.35. In comparison, the final XGBoost model was abled to reduce its RMSE to roughly 0.14. This shows a significant improvement in overall prediction accuracy. This improvement highlights the effectiveness of feature engineering and the model's ability to capture complex data patterns.

Overall Performance:

We can interpret the XGBoost model's low RMSE score as a high ability to closely match actual sales prices. Secondly, with respect to robustness, cross-validation confirms the model's reliability across various subsets of data. The final model requires minimal human intervention and can, therefore, be scaled up for real-world applications.

Conclusion:

The employment of advanced regression algorithms showed that real estate firms could benefit from machine learning. Our proposed model, the XGBoost, yielded an RMSE of 0.14. It overcomes may of the drawbacks associated with traditional valuation such as the cost of time and potential for human error. Our model can capture complex data that exists in the housing market and make predictions that accurately reflect this.

Business Impact:

Seller confidence, buyer trust, and brand reputation are all enhanced by trustworthy pricing tactics. Time on the market is shortened, transaction volume is increased, and accurate valuations will eventually fuel revenue growth.

Scalability:

This methodology can be implemented across several geographies, property kinds, and market situations with little manual involvement. It has the potential to move beyond Ames, Iowa.

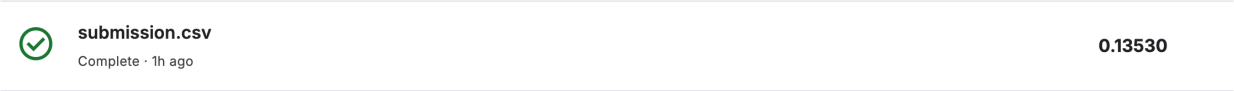
Future Directions:

To maintain forecasting integrity, the model should be retrained with updated data. Such data should consider new patterns or demographic changes that may emerge.

Our research showed that real estate pricing strategies could be revolutionized by fusing data-driving insights with machine learning approaches. This will increase operational efficiencies and raise competitive advantages.

Visuals:

Proof of Submission:

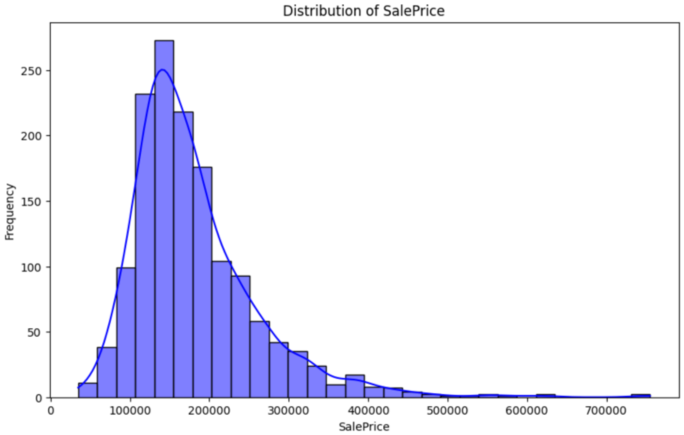


submission.csv

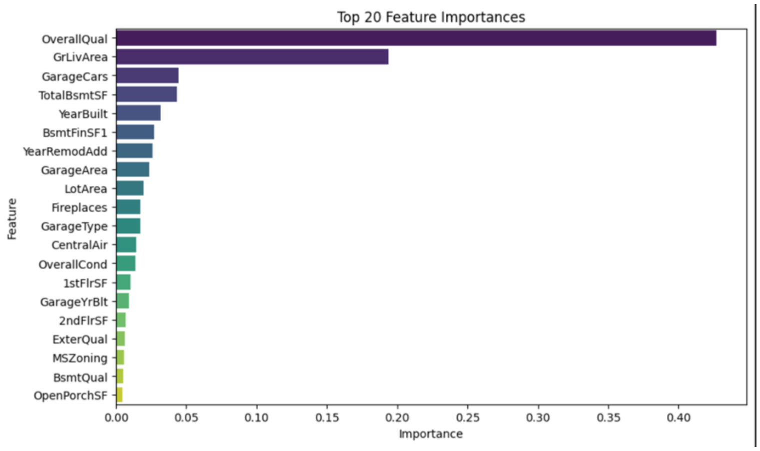
Complete · 1h ago

0.13530

Sale Price Distribution:



Feature Selection:


Top 20 Feature Importances
Feature
OverallQual
GrLivArea
GarageCars
TotalBsmtSF
YearBuilt
BsmtFinSF1
YearRemodAdd
GarageArea
LotArea
Fireplaces
GarageType
CentralAir
OverallCond
1stFlrSF
GarageYrBlt
2ndFlrSF
ExterQual
MSZoning
BsmtQual
OpenPorchSF
Importance