Final Project

Brady France, Andrew Bingen, Nick Elias, Shida Ye

Business Framing:

Accurate property valuation is a cornerstone of the real estate industry, influencing decisions for homebuyers, investors, mortgage providers, and real estate agents. However, traditional valuation methods often rely on manual appraisals, which can be inconsistent, time-consuming, and subject to bias. As a result, there is a growing need for automated, data-driven solutions that provide precise and reliable property valuations.

Our team's machine learning model addresses this need by leveraging advanced regression techniques to predict house sale prices using a diverse set of property features. This model provides significant business value by enabling real estate agencies to optimize their pricing strategies, reducing over or underpricing risks, and helping them remain competitive in a dynamic market. Similarly, mortgage lenders can use these predictions to assess financial risks more accurately, ensuring loan terms align with a property's true market value.

For homebuyers, the model brings transparency, helping them understand fair market prices and make informed purchasing decisions. Moreover, real estate investors can use these predictions to identify profitable opportunities in undervalued or overvalued properties. By combining data science with real estate market dynamics, the project delivers a practical and scalable solution to improve decision-making across multiple stakeholders in the industry.

Accurate property valuations are crucial for businesses to make informed pricing decisions, optimize profitability, and minimize risks. Overestimating can lead to prolonged sales and missed revenue, while underestimating results in undervalued assets and lost profits. Inaccurate valuations can waste resources and damage brand credibility, so ensuring precision helps businesses stay competitive and efficient in a dynamic market. Even small improvements in accuracy can have large impacts on overall profitability.

Problem Statement:

The real estate market is complex, with house prices influenced by a variety of factors, including physical property attributes, neighborhood characteristics, and broader economic conditions. Predicting house prices is a critical task for multiple stakeholders,

including real estate agencies, homebuyers, and mortgage lenders, as accurate valuations reduce financial risks and optimize decision-making processes. However, traditional appraisal methods often fall short due to their subjective nature and inability to scale efficiently.

The Kaggle competition "*House Prices - Advanced Regression Techniques*" presents a supervised machine learning challenge to predict house sale prices based on a dataset of residential properties in Ames, Iowa. The dataset includes over 70 explanatory variables (79) covering property features such as lot size, building quality, year built, and neighborhood demographics. The primary objective is to build a regression model that leverages these features to accurately estimate a property's sale price.

Key challenges include handling missing data, identifying and engineering the most predictive features, and selecting appropriate modeling techniques to minimize prediction error. Success is measured by the model's performance on the Root Mean Squared Error (RMSE) metric, which prioritizes models capable of producing precise and consistent predictions.

By solving this problem, the project demonstrates the practical application of machine learning in addressing a real-world business challenge, providing actionable insights for accurate property valuations and bridging the gap between raw data and informed decision-making.

Approach we took:

Our modeling process began with a thorough examination of the data to understand its distribution and identify missing values. Since the dataset contained both numerical and categorical features, careful consideration was given to how best to handle each type. For numerical features, rather than relying on simplistic methods such as mean imputation, we employed median imputation to reduce the impact of outliers and achieve more robust estimates for missing values. For categorical variables, we replaced missing entries with meaningful placeholders, such as 'None', to preserve their predictive capacity without artificially skewing the distributions.

We then conducted feature encoding, opting for label encoding to transform categorical variables into numeric form, which our chosen model could interpret more effectively. Although one-hot encoding is a common technique, label encoding allows us to preserve information in a more compact form, especially when combined with a model that inherently manages categorical features well. We also applied MinMaxScaler to normalize numerical features, reducing the dominance of large-scale attributes and helping our model achieve more stable performance.

To further enhance model accuracy, we engineered features that captured interactions, and nonlinear relationships present in the data. For example, we combined basement areas, first-floor, and second-floor areas into a total square footage metric to better represent the property's overall usable space. We created features that reflected property age and remodeling age, introduced an indicator for whether the property had a pool, and considered the multiplication of overall quality by total square footage to capture properties that are not only large but also of high quality. Such engineered features aimed to provide the model with more meaningful inputs that correlate strongly with sale price.

For model training, we split our dataset into training and validation sets to ensure reliable performance estimates before making final predictions on the test data. We then tested the performance of many different types of models. We tested out a random forest model, a XGBoost model, a lightGBM model, and ensemble and stacked models incorporating multiple models into the predictions. Within each of these approaches we adjusted the hyperparameters to tune it for optimal results and tested out different ways of handling missing data.

In the end, we selected CatBoostRegressor as our primary model due to its ability to handle categorical features effectively, robust default hyperparameters, and strong performance on tabular data. CatBoost inherently reduces the need for extensive hyperparameter tuning and gracefully manages complex, nonlinear relationships, ultimately producing consistent and accurate predictions.

Upon evaluating the model on the validation set, we monitored the RMSE to ensure that our predictions were both precise and stable. After several iterations of refinement, including adjustments to model depth, learning rate, and the number of estimators, we settled on a configuration that balanced model complexity with generalization, minimizing validation errors and preventing overfitting.

Superiority of Our Approach:

For our preprocessing rather than utilizing simple techniques such as mean imputation and one-hot encoding, we decided to do things such as median imputation in the numerical columns. We also utilized methods such as feature scaling with MinMaxScaler and labeled encoding.

In terms of why our feature engineering is superior to others, the features we engineered are highly interpretable and help to improve the predictive power with the use of combining related information. In our feature importance analysis, many of the

engineered features that we created were rated near the top of importance and without them included, the model suffers significantly.

Rather than using a common benchmark model such as simple linear regression, we decided to utilize CatBoostRegressor. What stands out about this is that it is a type of gradient boosting algorithm that can optimize for categorical features as well as fast training; it also works to reduce overfitting.

Our CatBoostRegressor model had a root mean squared error of 0.123, which in comparison, the root mean squared error for a simple linear regression without our engineered features was 0.159. This may not seem like much but this moved us from rank 4460 all the way to 577 out of a total of 6416 competitors. This significant leap in the leaderboard highlights the impact of even marginal improvements in model performance within a highly competitive environment.

Conclusion:

In this project, we tackled the "*House Prices - Advanced Regression Techniques"* Kaggle competition by developing a machine learning model to predict house sale prices. Using advanced preprocessing methods like median imputation, label encoding, and feature scaling, alongside meaningful feature engineering, we enhanced the predictive power of our model. We selected CatBoostRegressor as our final model due to its ability to handle categorical variables effectively and minimize overfitting. This approach achieved an RMSE of 0.123, outperforming the benchmark linear regression (RMSE 0.159) and significantly improving our ranking from 4460th to 577th out of 6416 competitors.

Our results highlight the business value of accurate property valuation, offering real estate agencies, mortgage lenders, and homebuyers the tools to make data-driven decisions, optimize pricing strategies, and reduce financial risks. This project demonstrated our ability to combine technical expertise with practical problem-solving, resulting in a scalable and competitive solution. It also reinforced the importance of continuous iteration and refinement in machine learning workflows, preparing us for future challenges in data science and business analytics.