Predicting House Prices in Ames, Iowa

Dan Zylkowski

DSC 680

Fall 2021

Abstract/Executive Summary

Purchasing a house is a big decision, and it can be stressful. In fact, 40% of Americans say that buying a new home is the most stressful event in their lives[2]. For some people owning a home means simply having a place to live. But for most families a home represents a safe place to raise their children, provide them with a good education, and help secure their financial future. While real estate values move in cycles, home values have consistently appreciated [2]. For these reasons, owning a home is the American dream, but it isn't without risk. While the financial benefit of owning a home is clear, the risk of overpaying for a home should not be overlooked.

To properly determine the value of a home, many factors need to be considered. Homes are assessed by evaluating dozens of internal features (for example living area, lot size, and number of bedrooms and bathrooms), and external features (neighborhood and location). With so many features to account for, it can be very hard to manually calculate how much each feature adds or subtracts from a home's value.

Using the low-code machine learning library PyCaret, thirteen predictive modeling techniques were compared to determine the best performing model. The best model was the Cat Boost regressor model, and it was able to predict home prices with an R^2 of 0.9278. R^2, also known as the coefficient of determination, is a metric used in regression to measure goodness of fit, and how much variation in price can be explained by the independent variables. In addition, RMSE, or root mean squared error, can also be used as an evaluation metric where the lower the value, the better. The Cat Boost regressor had an RMSE of $19,534, which was an improvement of $54,268 over choosing the naïve model that predicted the average for each sale price. These results show that the model is skilled at predicting the sale price of a home. While this model is based on home prices in Iowa, I believe that it could be modified to add features that are more common in warmer climates (for example a pool or water features).

Problem Statement

Purchasing a home is one of the most important decisions a person can make. By extension, selling a home is also a life-changing decision. Overpaying for a home or undervaluing your home are two critical mistakes that can be financially devastating. No one wants to find out later that they could have asked for a higher price or end up with buyer's remorse after purchasing a home. Therefore, homebuyers, sellers, and real estate agents need a tool to help them understand how to value a house accurately. This project will use predictive modeling to create a robust data model that will accurately price a home based on key features.

Methods

Data Exploration

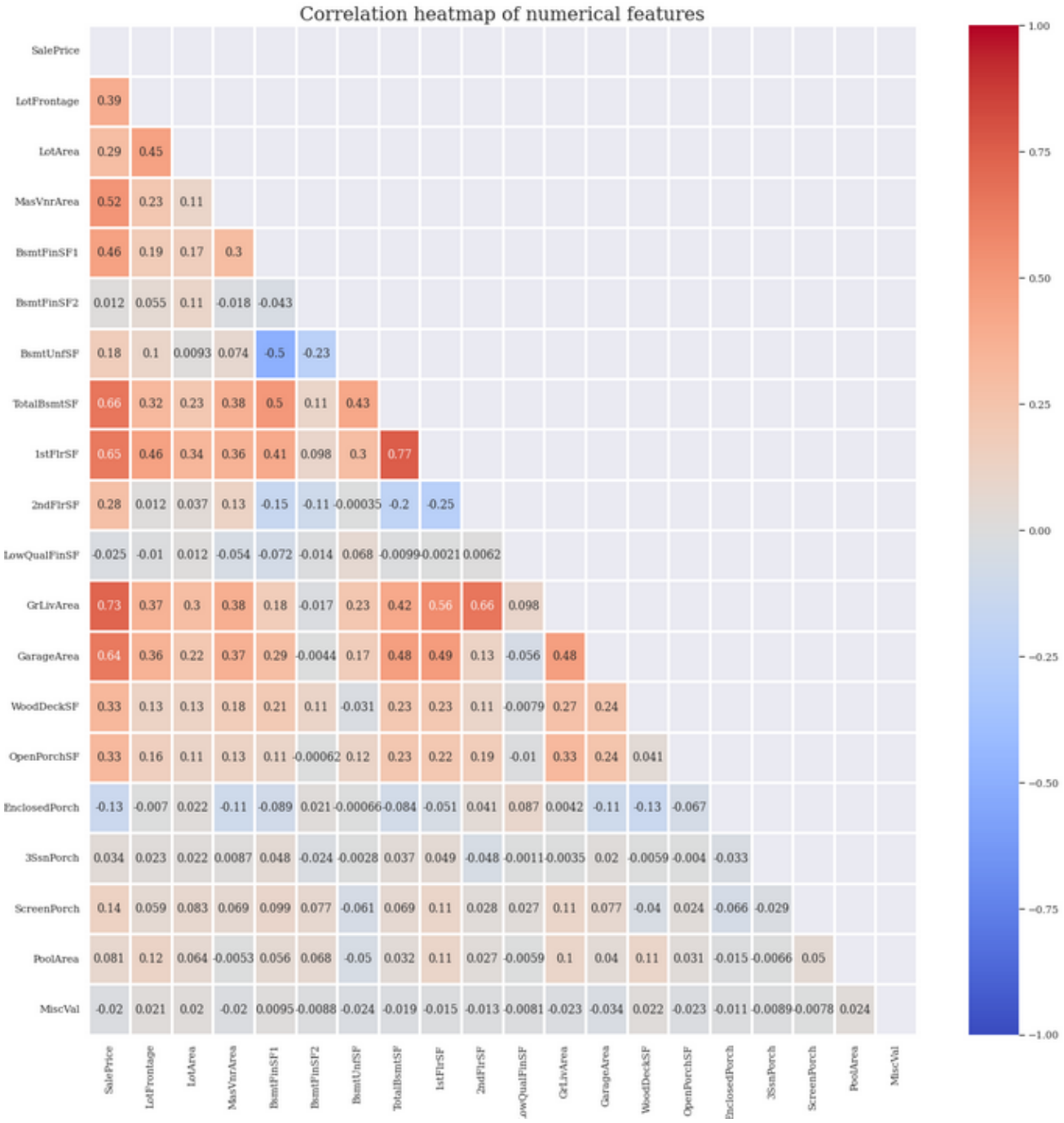
The data that was used for this study is from the "Iowa House Prices" dataset provided by Kaggle[9]. The dataset describes the sale of individual residential property in Ames, Iowa, from 2006 to 2010. The dataset contains 2919 observations. The data consists of 80 variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) that are involved in assessing home values. Detailed information about the dataset, including the full data dictionary, can be found here.

To better understand the variables and how they impact the prediction of price, was split into training and testing sets (50% training) prior to performing exploratory data analysis (EDA) to avoid data snooping. EDA was then performed on the numerical and categorical variables. As is typical of datasets involving predictions of expensive items, the data is right-skewed due to having a handful of homes with extremely large values (See Figure 1 below). Skewed target variables are good candidates for log transformation, as predictive models often perform better after a skewed variable has been transformed. There were 8 homes with values greater than the mean home price plus four standard deviations, and these were treated as outliers.

Figure 1: EDA of Price Variable

A correlation matrix for the combined dataset showed low to moderate correlations between the variables, so there is no multicollinearity observed. In the correlation heatmap (see Figure 2 below), the above grade living area, first floor living area, garage area, and basement area were shown to have strong positive correlations with price. This indicates that buyers value larger houses, with larger garages and basements. have higher prices, which makes sense.

Figure 2: Correlation matrix of continuous features

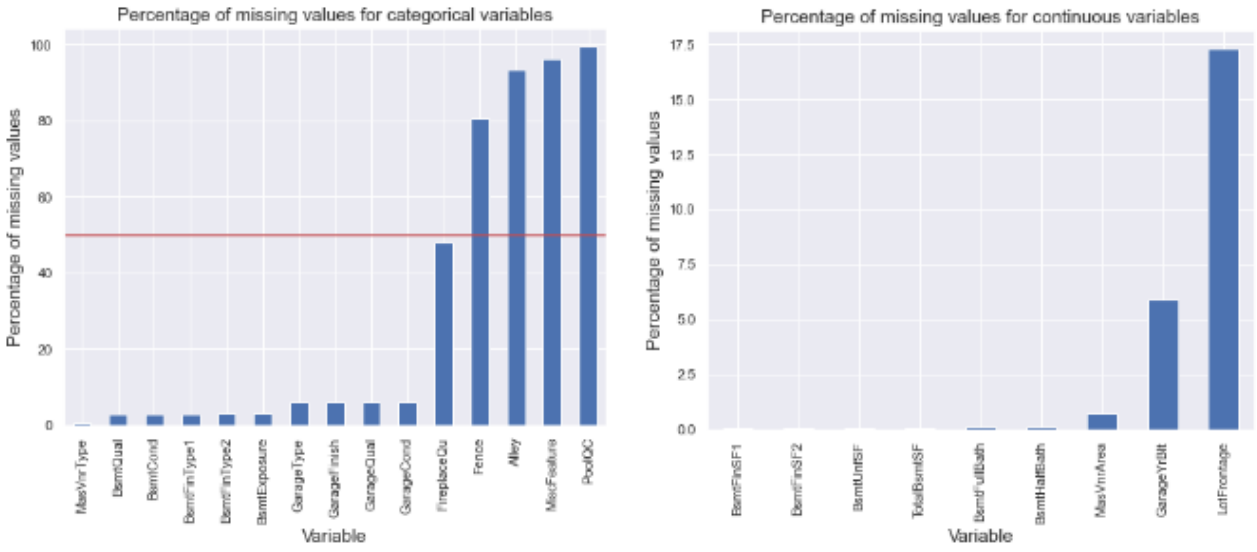


Handling missing values

The dataset contained missing values across both categorical and numerical variables (See Figure 3 below). Four of the categorical variables were missing more than 80% of their values, and so these were removed. The other missing values in the categorical variables were

due to mislabeling of 'NA' values as nan in the data. Missing values for the LotFrontage continuous variables was imputed using the mean value of LotFrontage for the associated LotShape. The GarageYrBlt variable also had missing values due to mislabeling of 'NA' values as nan in the data. Other missing values were less than 1%, and they were imputed using the mean or mode of the variable.

Figure 3: Missing values in categorical and numerical variables

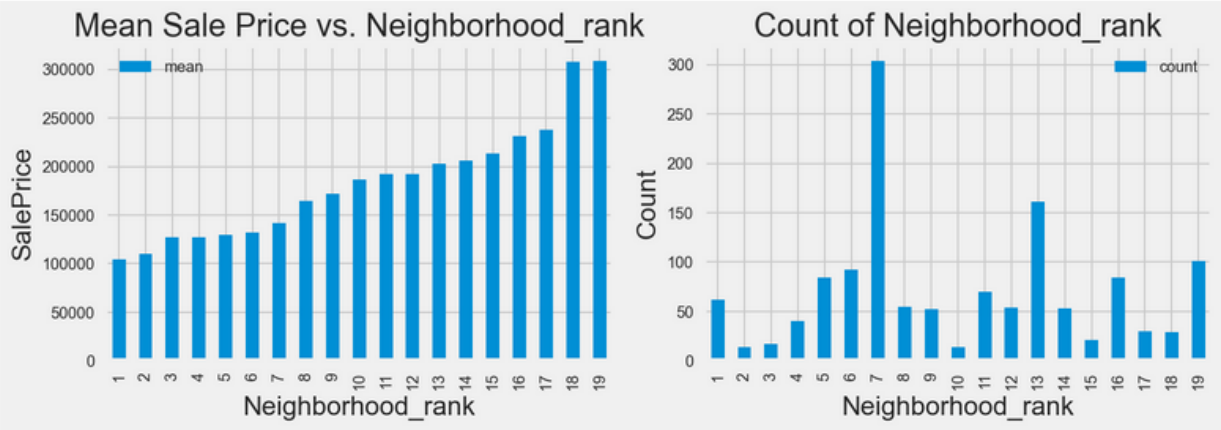


Encoding Categorical variables

The dataset contained 43 categorical variables that had non-numerical values. To deal with so many categorical features, I first combined rare levels of categorical features (< =1% occurrence). Then I used the training data to create an encoding scheme that ranked the levels of each categorical variable to maximize the correlation with Sale Price. By encoding the variables this way, the number of levels in the categorical variables was additionally reduced,

and a natural ordering strategy was implemented. For example, some neighborhoods have higher prices, and the strategy was able to determine the best way to group the neighborhoods together to maximize the correlation between neighborhoods an SalePrice (See Figure 4). Other variables were encoded similarly, and I think this was an informed strategy in the absence of domain knowledge about specific home features.

Figure 4: Countplot and barplot of optimal Neighborhood ranking.




```
Correlation for 19 bins=0.7235
```

Modeling

After handling all missing values and encoding numerical features, I used the PyCaret

library to build and compare 13 different models. Because PyCaret makes it easy to process and transform datasets, I decided to see how the models would perform with and without scaling the data and transforming the target variable. Some models are more performant on scaled and transformed data, and I wanted to see the difference.

Figure 5: Comparing top 10 models with NO feature scaling or target transformation

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	13208.6122	407233844.0466	19908.0699	0.9252	0.1073	0.0770	3.1920
gbr	Gradient Boosting Regressor	14591.8653	476330232.5290	21553.9715	0.9126	0.1166	0.0849	0.1500
lightgbm	Light Gradient Boosting Machine	14280.0063	488902021.0982	21875.3669	0.9103	0.1160	0.0828	0.0800
et	Extra Trees Regressor	14743.6004	502812636.8402	22258.1490	0.9077	0.1212	0.0865	0.2820
rf	Random Forest Regressor	15485.6533	563190992.2501	23498.7099	0.8967	0.1274	0.0911	0.3980
br	Bayesian Ridge	17429.6295	578135397.3033	23968.8851	0.8932	0.1453	0.1064	0.0120
xgboost	Extreme Gradient Boosting	15908.4920	598275635.2000	24165.1992	0.8903	0.1237	0.0896	0.3720
ridge	Ridge Regression	17758.6258	600349696.0000	24423.6391	0.8891	0.1502	0.1091	0.4400
lr	Linear Regression	17809.8777	604234086.4000	24500.6664	0.8884	0.1508	0.1094	0.6600
ada	AdaBoost Regressor	20321.1789	765487524.3908	27603.7506	0.8588	0.1664	0.1306	0.0920

Figure 6: Comparing top 10 models after BOTH feature scaling and target transformation

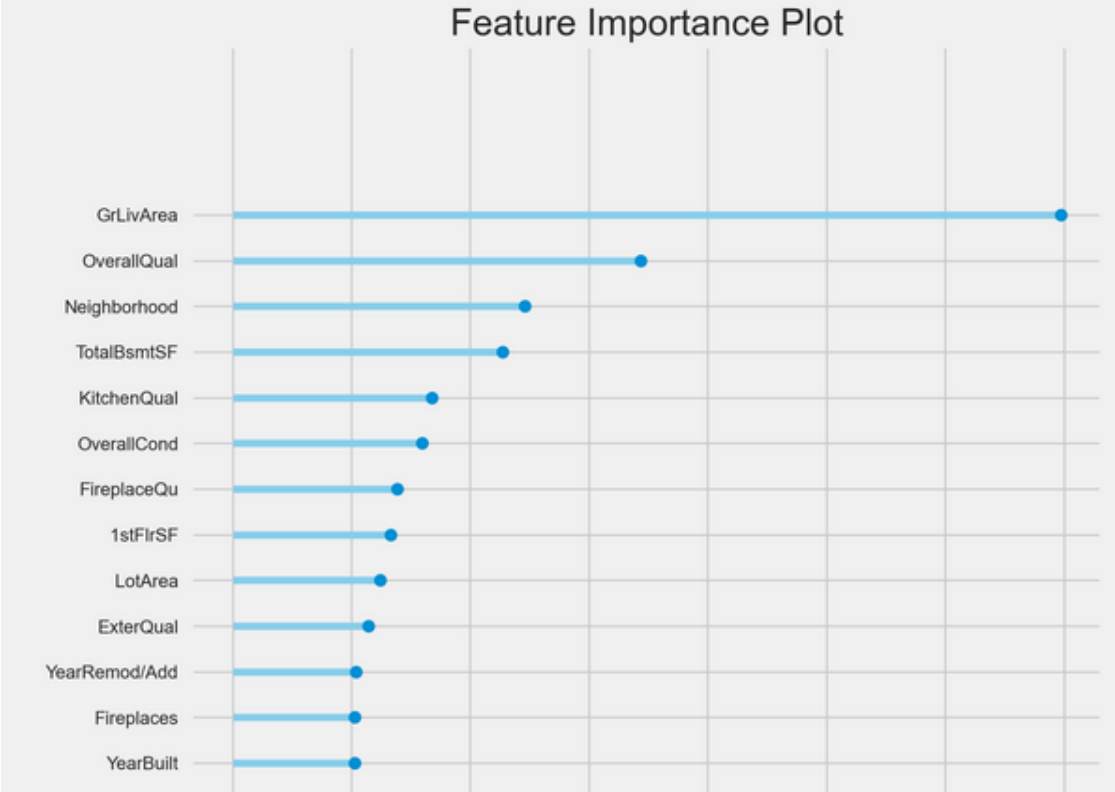
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	13197.5033	419367213.7831	20117.2234	0.9231	0.1072	0.0759	3.0660
br	Bayesian Ridge	14276.1883	430594280.8272	20716.7630	0.9205	0.1159	0.0842	0.0100
ridge	Ridge Regression	14511.4432	448494054.4000	21133.6301	0.9172	0.1175	0.0854	0.0100
lr	Linear Regression	14548.0119	451718342.4000	21206.0711	0.9166	0.1178	0.0856	0.0100
lightgbm	Light Gradient Boosting Machine	14385.2763	488756017.9795	21838.2298	0.9105	0.1171	0.0829	0.1000
gbr	Gradient Boosting Regressor	14573.0877	515108233.4828	22337.7221	0.9059	0.1170	0.0829	0.1280
et	Extra Trees Regressor	14869.8694	531407215.8674	22832.5041	0.9027	0.1202	0.0842	0.2520
xgboost	Extreme Gradient Boosting	15492.8205	550963942.4000	23300.4016	0.8988	0.1249	0.0890	0.2780
rf	Random Forest Regressor	15407.1111	594420440.7316	24096.7921	0.8913	0.1266	0.0882	0.3080
omp	Orthogonal Matching Pursuit	18121.3158	723489678.8187	26827.1491	0.8670	0.1423	0.1038	0.0080

From above, we can see that feature scaling and target transformation greatly increased model performance for Bayesian Ridge, Ridge, and Linear Regression. None of these three models were in the top 5 performing models when there was no feature scaling or target transformation, but all three jumped into the top 4 models when both feature scaling and target transformation occurred. The Cat Boost Regressor remained the top performing model in both cases.

Feature Importance

The feature importance plot shows that Above Grade Living Area, Overall Quality, and Neighborhood are the three most important features. Looking further down the list, feature Kitchen Quality, Year Built, and Year Remod/Add are also important. The model therefore informs buyers, sellers, and realtors that house size, quality, and location are very good predictors of house price. Additionally, if homeowners want to increase the value of their house, a good way to do it would be to remodel the kitchen.

Figure 6: Feature importance



Limitations

While the model was shown to be skillful at predicting home prices based on features, it does have obvious limitations. Not only is the data from the years 2006-2010, but it is also from the Midwestern state of Iowa. Some of the features in the model would not be present in houses in other states (for example Fireplaces and Basements are extremely rare in south Florida). Adding new features to the data would require additional time for data preparation.

Discussion/Conclusion

This project had multiple challenges involving the dataset. There were 80 variables, and 43 of those were non-numeric and needed to be encoded. Having little domain knowledge of

the dataset, I needed to come up with a strategy to encode the categorical variables. I didn't want to simple use dummy encoding, as the new variables would not contain as much information and would likely have multicollinearity. I it was a good strategy to discretize the variables into bins based on the correlation with the SalePrice. Cleaning and transforming the data took well over 80% of the total time spent on this project, but the results were worth it.

Looking at the results, I think the models performed quite well. The best model had an RMSE of 19,534 on the test set, which was a huge improvement over the baseline naive model RMSE of 73,802. Several models improved after the data was scaled and the target variable was transformed. Overall, I was pleased with the way the project turned out.

Next Steps

As discussed above, while many of the features in the dataset would generalize to other areas outside of Iowa, many of them would not. Therefore, to use this model in areas outside of Iowa, additional data would need to be collected. Once the data was collected, new features could be dropped easily, and new features could be added as necessary.

References:

1. 8 critical factors that influence a home's value. Opendoor. (2019, September 19). https://www.opendoor.com/w/blog/factors-that-influence-home-value. (This article explains which features of the data set are likely to influence the price of the house.)
2. Americans say buying a home is most stressful event in modern life. HousingWire. (2019, September 19). Retrieved September 19, 2021, from

https://www.housingwire.com/articles/46384-americans-say-buying-a-home-is-most-stressful-event-in-modern-life/

3. Bay, H. (2019, May 23). 7 ways your Neighborhood impacts your home value. Real Estate Tips, Advice, Updates & More | Home Bay. https://www.homebay.com/tips/7-ways-your-neighborhood-impacts-your-homes-value/. (This article explains how the location of a home impacts the value.)
4. Best home improvements to Increase VALUE: ZILLOW. Home Sellers Guide. (2021, September 3). https://www.zillow.com/sellers-guide/best-home-improvements-to-increase-value/. (This article informs sellers of which features they can improve to obtain the most return on investment.)
5. DiClerico, D. (n.d.). 8 ways to boost your home value. Consumer Reports. https://inventory.consumerreports.org/home-improvement/8-ways-to-boost-your-home-value/. (This article explains the most important ways to boost home value and include some of the features in the dataset.)
6. Hargrave, M. (2021, September 4). Using hedonic pricing to determine the factors impacting home prices. Investopedia. https://www.investopedia.com/terms/h/hedonicpricing.asp. (This article explains how hedonic pricing identifies the internal and external factors and characteristics that affect an item's price in the market.)
7. Harrington, D. B. (2020, January 27). Here's what to look for when buying a house. Bob Vila. https://www.bobvila.com/articles/414-house-choosing-checklist/. (This article informs buyers what features in the home will be the most difficult to renovate.)
8. Hgtv. (n.d.). 30 tips for increasing your home's value. HGTV. https://www.hgtv.com/design/remodel/interior-remodel/30-tips-for-increasing-your-homes-value. (This article explains how homeowners can improve their homes to maximally increase the value.)

9. Taylor, N. (2018, February 21). *Iowa house prices*. Kaggle. Retrieved September 20, 2021, from https://www.kaggle.com/nickptaylor/iowa-house-prices.
10. Vogel, C. (2020, June 14). Home renovations that give you a return on your investment. This Old House. https://www.thisoldhouse.com/home-finances/21015466/renovations-that-give-you-a-return-on-your-investment. (This article identifies features that have a high ROI and thus are likely to be important in determining house prices.)
11. Voorhis, S. V. (2019, March 11). Six best and worst home improvements for your money. TheStreet. https://www.thestreet.com/personal-finance/real-estate/six-best-and-worst-home-improvements-for-your-money-14864740. (This article explains the best and worst features to improve based on ROI. This could indicate which features are most and least important in determining house prices.)
12. White, M. (2020, October 29). Why location is so important in real estate. Moving.com. https://www.moving.com/tips/why-location-is-so-important-in-real-estate/. (This article explains the reasons why location is important for home valuation.)

Appendix

Figure 7 : Prediction error plots from the Cat Boost model

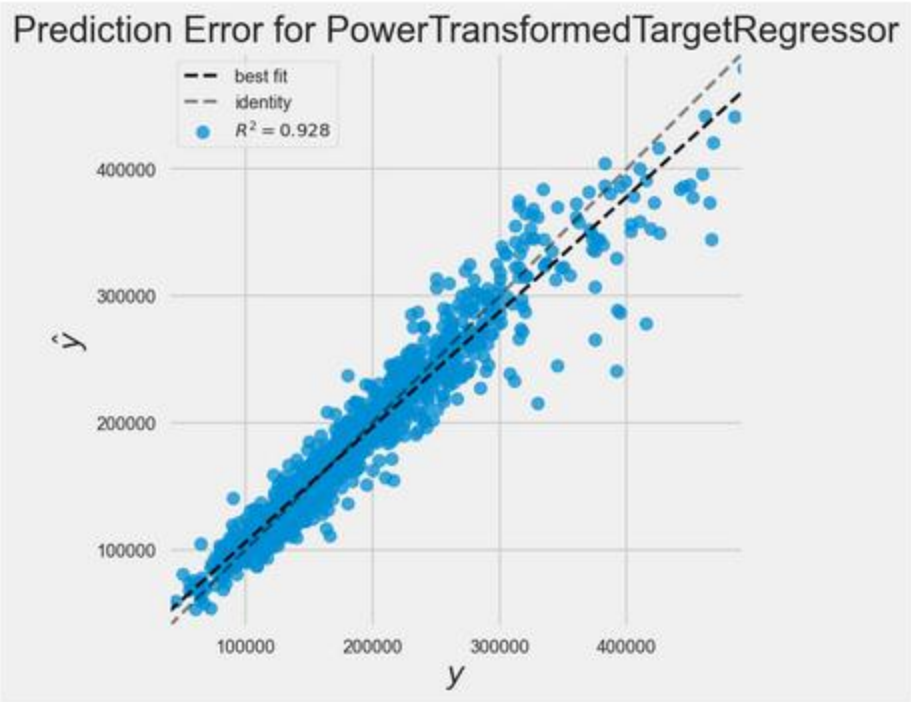


Figure 8: Residual plots from the Cat Boost model