House Price Prediction Using Machine Learning: A Case in Iowa

Experiment Findings · February 2022

DOI: 10.13140/RG.2.2.19846.86086

Ozancan Özdemir
Middle East Technical University

House Price Prediction Using Machine Learning: A Case in Iowa

1stOzancan Ozdemir
Department of Statistics
Middle East Technical University
Ankara, TURKEY
ozancan@metu.edu.tr

***Abstract*—House price prediction is an important concept in the real estate industry. Thus, many researchers from different fields are interested in developing a regression model for the house price to obtain an accurate prediction and explore the factors affecting the house price. In this study, we aim to develop an accurate regression model using tree-based algorithms and explain the type of information which has an impact on the house price. For this purpose, we use the Ames House Price dataset being available on Kaggle.**

***Index Terms*—House Price Prediction, Lasso, Random Forest, XGBoost, LightGBM, Gradient Boosting, CatBoost.**

I. Introduction

House price prediction has been a popular problem in research for years since the traditional house price prediction depending on cost and sale price comparison does not satisfy the accepted standards and certification process. [1] An accurate house price prediction has importance for the stakeholders of the real estate industry, which is a constantly rising one in many countries, like homeowners, customers, and estate agents. In addition to getting accurate predictions, it is important to know the factors that have a significant impact on the house price because it is known that house price is affected by many factors like location, house type, build year, etc.

A variety range of regression models involving tree-based algorithms, Support Vector Regression, and Least Square Based Linear Regression have been applied to this problem to develop an accurate house price prediction model and explore the most effective factor on the house price. In 2017, Lu et al. developed a hybrid Lasso and Gradient Boosting algorithm for individual house price prediction using the Iowa state data available in Kaggle for competition. They use the log of the sale prices as the target variable and their final prediction consists of the combination of 65% Lasso and 35% Gradient Boosting. This suggested approach was ranked top 1% out of all submissions. [2] Phan proposed the combination of Stepwise and Support Vector Machine as the most accurate predictive algorithm for house price prediction by using a dataset taken from Melbourne City being available in Kaggle. He also revealed that the number of bedrooms, distance to CBD, latitude, longitude, and type of houses are the most important factors on the house prices in Melbourne. [3]Macpherson and Silman showed that neighborhood diversity is a factor in house price in certain counties in Tampa and Orlando, Florida. They proved that the level of the Hispanic population has a positive effect on the house price. [4]

The goal of this study is to develop a predictive model for house prices using tree-based methods which are Random Forest, Gradient Boosting, XGBoosting, Light GBM, and CatBoost. We will also find out the important factors on the house prices and how they can affect it. Besides, some factors are investigated particularly by using Lasso Regression and Random Forest.

In this study, we will use the Ames Housing Price data set which is available on Kaggle for the use of regression competition data set. The details of this data set will be explained in the next part of the paper. Then, we will talk about data preprocessing and quality checks with some descriptive statistics. After this, sub-research questions will be represented. This is followed by the modeling part, and it will end with the conclusion and discussion part. This regression project is conducted using Jupyter Notebook and related Python libraries including Pandas, sklearn, xgboost, CatBoost, LightGBM, shap.

II. Data Description

The data set contains the home sales that occurred in Ames, Iowa between 2006 and 2010. It is arranged and released by De Cock as an alternative for Boston House Price Index data. [5] The original data set has 80 variables and 3970 observations. However, we used the Kaggle competition version of this data. This version consists of both train and test sets. We only use the train set in this project instead of concatenating these two sets since the test set does not have the sale price column, which is the target. Thus, we study with a data set including 80 variables and 1490 observations. Out of 80 variables, 23 of them are continuous, 14 of them are discrete, 46 of them categorical (23 Ordinal 23 Nominal) variables. Most of these variables indicate the type of information a typical house buyer would want to know before buying a house such as various area dimensions, number of rooms, kitchens, etc. [5]

III. Data Pre-processing and Exploratory Data Analysis

In this part, we will represent the steps of making the data ready for the analysis, and then give some statistical information which helps to understand the data better.

A. Data Pre-processing

In the beginning, we will check the existence and distribution of the missing observations in the data. It is seen that 19 variables out of 80 have missing observations. Among 19 variables, PoolQC(Pool Quality) has the highest missing percentage (99.52%), and it is followed by MiscFeature (96.3%). The variable with the lowest missing percentage is Electrical(Electrical System).(0.07%)

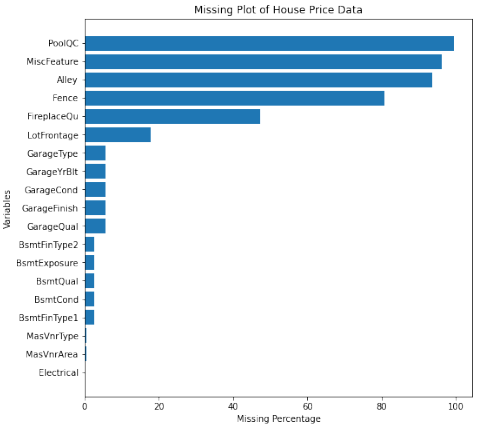

Fig. 1. Proportion of the missing observations in 19 Variables.

However, most of these missing observations indicate "None" rather than "Not Available" as stated in the data description. [5] For example, NA's in PoolQC indicate that the house does not have any pool. Thus, we will input NA values in categorical variables with "None". There is only one numeric variable with NA, which is Lot Frontage(Linear feet of street-connected to property). Here, we use median imputation since NA's in this variable do not depend on any other variables. After imputation, we see that there is no in the shape of the distribution.

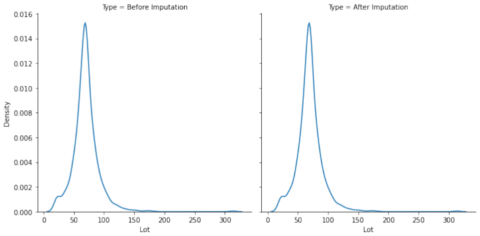

Fig. 2. The Distribution of Lot Frontage Before and After Imputation

Then, we will check the outlier in the data. As stated in the data description, the data has five observations which can be considered the outlier. It is said that these extreme observations can be observed in the scatter plot of GrLivArea (Ground Living Area) and Sales Price, and we can get rid of these observations by removing the houses whose Living area exceeds 4000 square feet in line with Cock's suggestion. [5]

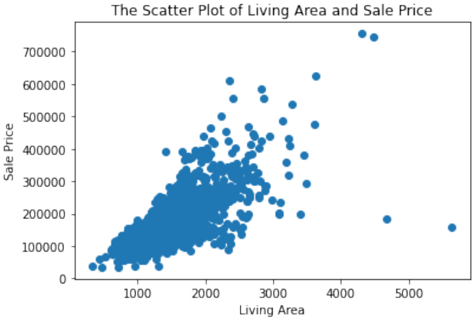

Fig. 3. The Association Between Ground Living Area and Sale Price

Thanks to the description file, we can handle with outlier and missing observations in the data. Then, we use **ydata** data quality check tool in Python to control the quality of the data in terms of bias and fairness, expectations, relations, drift, duplicates, labeling and erroneous data. For example, it uses unit tests for data expectation and evaluates it based on great expectation. On the other hand, it will fit the causality test and calculate variance inflation factor(VIF) or fit chi-square test to assess the association among the variables. After this quality checking, it is stated that the data has strongly associated variables in both numerical and categorical ones. This result will affect the way of analyzing the research questions which will be presented in the following sections.

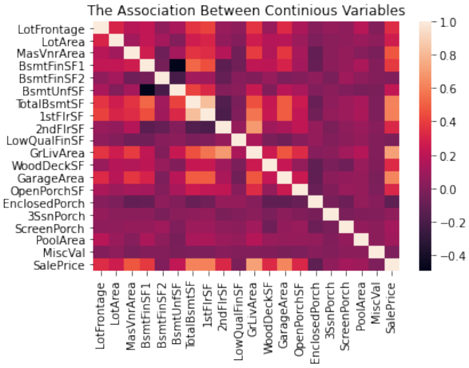

Fig. 4. The Association Between Continuous Variables

B. Exploratory Data Analysis

It is known that Sale Price is the variable of interest. The average house sale price in Ames between 2006 and 2010 is 180921$ while its minimum and maximum are 34900$ and 755000$, respectively. The median sale price is 163000$ which indicates that the sale price has the right-skewed distributions. Most of the numerical variable including sale price do not have normal distribution. However, we see that the distribution of sale prices becomes approximately normal after log transformation but it does not consider in the further analysis since they do not assume normality or other assumptions are not satisfied although the normality is assessed.

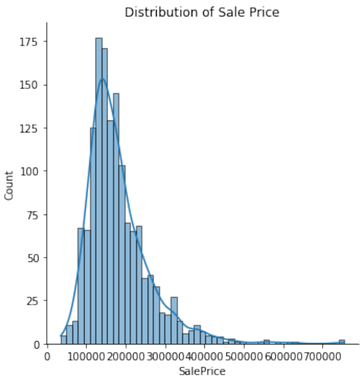

Fig. 5. The Distribution of Sale Price

If we check the distribution of the categorical variables, we also see that some of the categorical variables such as Street, Utilities, Condition1, Condition 2 have almost only one class which may makes these variables ”useless” in the modelling part due to the low variability in terms of sale prices.

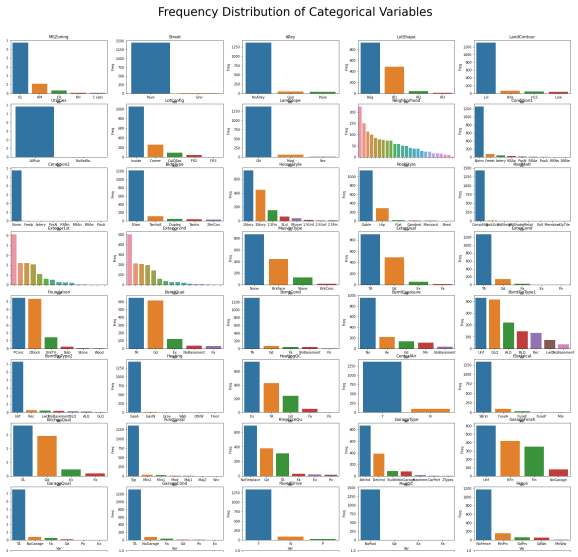

Fig. 6. The Distribution of Categorical Variables

After these descriptive ones, we can go further by checking the association among the variables. In previous part, we have already stated that there is an association among the variables, and represented the pearson correlation heat map for the continuous variables, and thereby we can measure the association between continuous features and sale price. Then, we also calculate the spearman correlation coefficient to assess the association between the discrete features and sale prices. Thus, we see that the GrLivArea, GarageCars,YearBuilt and Fullbath have the highest positive association with sale price while KitchenAbvGrd and EnclosedPorch have the highest negative association.

We also check the association of the sale price with categorical variables by drawing a matrix of box-plot. Then,

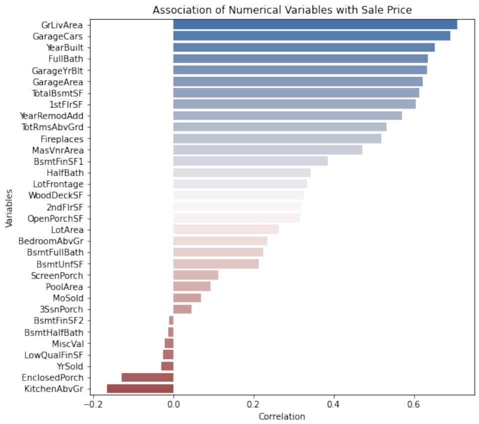

Fig. 7. The Strength of the Association of Numerical Variables with Sale Price

it is observed that the sale price may change in terms of neighborhood, basement quality, etc.

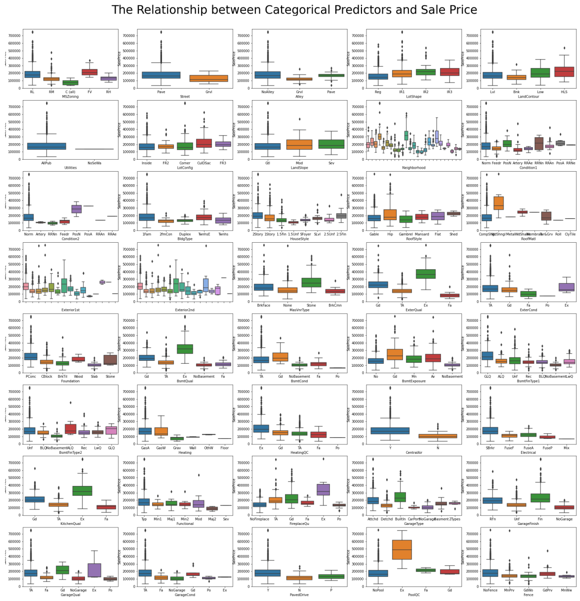

Fig. 8. The Association Between Sale Price and Categorical Variables

However, all of these results are some preliminary results, and they are needed to be confirmed by applying more advanced techniques. As stated at the beginning, our aim is not only to develop a predictive model for sale price, but also detect the most important factor affecting the sale price in Iowa.

C. Research Questions

1) How does the 28 neighborhoods in Ames, Iowa affect the Sale Price? Which neighbourhoods are important?: Firstly, we attempt to use one-way ANOVA to answer this question, but the homogeneity of the variance is not satisfied. Since we have 28 neighborhoods, we do not prefer a non-parametric approach, and then we consider Lasso Regression to understand which neighborhood is important for sale price. Thus, we transform Neighborhood variable using one-hot encoding. Since these new binary features are correlated with each other, Lasso regression is an ideal choice. The box plot above shows that the sale price in Noridge, NridgHt, StoneBr, and Somerst are higher than others on average, while it is lower in IDOTRR and Meadov compared to others. We also see that NridgHt, Noridge, Somerst, and StoneBr have a positive correlation while OldTown, Edwars, and IDOTRR have a negative association with sale price according to the spearman coefficients.

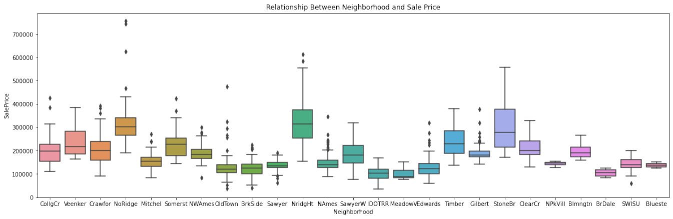

Fig. 9. The Distribution of Sale Price over Neighborhoods

In lasso regression, the log of the sale price is our target variable since it follows the approximately normal distribution, and 28 neighborhoods are the features. Since the model is used for exploration, it is not divided into test and train sets. The model has only one parameter, which is α, regularization parameter and it is determined by 10-fold cross-validation which is $3.14x10^{-}5$. In the end, it is shown that IDOTRR, Meadow, Noridge, and NridgHT are the most important neighborhoods on the sale price in Ames. By looking at the correlation and coefficients, we can see that if the house is in IDOTRR or Meadow, the sale price degrades significantly. If it is in Noridge or NridgHt, the sale price increases significantly. On the other hand, SawyerW neighborhood does not have any effect on the sale price. We calculate the R^2 score of this model via 10-fold cross-validation. Thus, 55.8% ± 0.025% variability of the sale price is explained by neighborhoods.

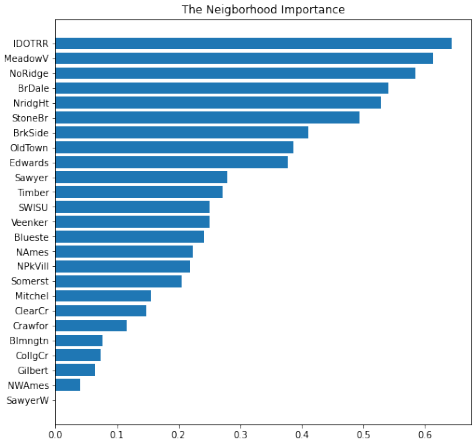

Fig. 10. Neighborhood Importance by Lasso Regression

2) How residential above grade properties affect the sale price between 2006 and 2010 in Ames, Iowa? What is the most efficient home properties on the sale price?: It is known that the physical properties of the house like the number of rooms, number of bathrooms are efficient on the sale price. In this question, we will try to explore how above-ground properties of homes in Ames, Iowa affect the sale price and what is the most important one by using a shap plot obtained from the Random Forest model. To do so, a subset including only sale price and variables related to above grade properties, which are GrLivArea(Living area),FullBath(Full baths above grade), HalfBath (Half baths above grade),BedroomAbvGr (Number of bedroom), KitchenAbvGr (Number of kitchen), TotRmsAbvGrd (Total number of rooms above grade). As in the previous research question, the data is not divided into a test set and a train set since we also aim to explain how much variability of the sale price is expressed by those features.

Random forest parameters are tuned via the random search method. We tuned **n estimators, min samples split,min samples leaf,max leaf nodes,max depth** to prevent the model from over-fitting problem. As a result of this model, we see that GrLivArea(Living area), which is the variable being positively correlated with sale price, is the most important above-ground properties of the houses being efficient on the sale price. It can be said that the house sale price increases as the living area increases or vice-versa. We can also see that a lower number of bedrooms above grade results in a higher sale price, while a higher number of kitchens above grade results in a lower sale price. On the other hand, the number of bathrooms and the total number of rooms does not affect the sale price significantly. By 10-fold cross-validation, we also see that 56.5% ± 0.049% variability of sale price is explained by these features.

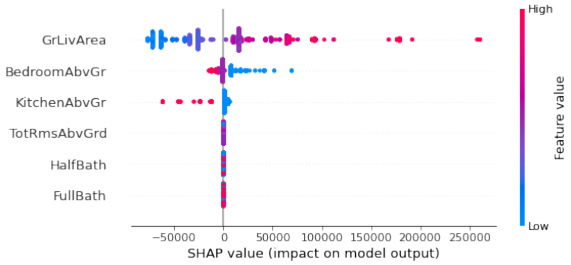

Fig. 11. Beeswarm Plot by Random Forest

3) What is the most effective basement residential property in predicting sale price in Iowa?: The previous question focused on finding the most efficient above-ground property on the sale price. Here, we will consider the basement properties, and random forest will be used again. Before constructing the model, the categorical basement variables are converted into dummy variables using one-hot encoding transformation.

Random forest parameters are tuned via the random search method as in the previous question. To prevent the model from a possible over-fitting problem, **n estimators, min samples split,min samples leaf,max leaf nodes,max depth** parameters are tuned. In the end, it is seen TotalBsmtSF(Basement area) is the most effective basement property on the sale price. If it increases, the house sale price also increases. Moreover, we can say that the excellent quality grade for the basement is an important factor for high sales prices although it is not the most important basement feature in predicting sale price. However, it can be also said that most of the basement properties do not affect the sale price. By 10-fold cross-validation, we also see that 54.7% ± 0.048% variability of sale price is explained by basement features.

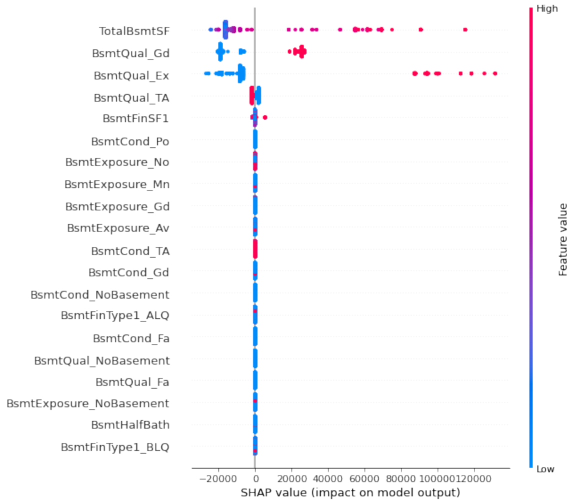

Fig. 12. Beeswarm Plot by Random Forest

IV. Modelling

In the previous part, we try to examine the factors that have affected the sale price in Iowa in some subsets of the features. Here, we aim to answer the following questions, and thereby we can develop a predictive model producing accurate house sale prices prediction and clarify the most efficient factors that have affected house sale price in Iowa over years.

- Which machine learning algorithm performs better and has the most accurate result in house price prediction? And why?
- What are the factors that have affected house prices in Ames, Iowa over the years?

Data pre-processing and exploratory data analysis show that the data suffer from the multicollinearity problem, and the response variable, sale price, does not satisfy the normality assumption. Thus, we apply tree-based algorithms since they do not require distributional assumption, and they are not affected by the multicollinearity problem seriously.

Unlike the previous research questions, we consider all features rather than selecting some subsets since we are looking for the most efficient factor among 80 features in the data. As opposed to most of the house price prediction studies in the literature, even the ones using Iowa data, we do not consider one-hot encoding in encoding the categorical variables in this study since it results in high cardinality in the data, almost 400 features, which may reduce the success of the models. Instead of this, we use the CatBoost encoder since it overcomes the problem of target leakage which may occur in the usage of the Target encoder. In this encoder, categorical feature values are encoded using the following formula: $\frac{TargetSum+prior}{FeatureCount+1}$ where target sum is the sum of the target value for that particular categorical feature, feature count is the total number of categorical variables, and prior is a constant value determined by the ratio between the sum of target values and the total number of observations. The performance of the models is compared by using two metrics which are R^2 score and Root Mean Square ($RMSE$) value whose formulas are given below.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{y}\right)^2} R^2 = 1 - \frac{\sum\left(y_i-\hat{y}\right)^2}{\sum\left(y_i-\bar{y}\right)^2}$$

where, $\hat{y}$ is the predicted value of target, $\bar{y}$ is the average of target, N is the length of the target.

Also, we create a baseline which is the average of sale prices by neighborhoods to evaluate whether the developed models are significantly better or not. Before the modeling, we randomly split the data into two sets; train set (80%) and test set (20%). Then, we randomly divide the train set into train set (80%) and validation set (20%) used for parameter tuning. The idea of random search is used to tune the hyperparameters of all models. In the hyperparameter tuning process, we aim to consider the ones which improve accuracies such as n estimator and learning rate, prevent the over-fitting such as depth, max leaf, and regularize the features such as eta or l2 leaf reg. The performance of the models with the tuned parameter for the test set is given below. It is seen that all algorithm outperforms baseline in terms of both R^2 and $RMSE$. On the other hand, CatBoost algorithm has the most accurate prediction among all models. It is followed by Gradient Boosting and Light GBM. After this, we also create a hybrid model to check whether CatBoost model can predict even rare cases or not. In this hybrid approach, first of all, we get the prediction from CatBoost model and calculate the residuals by substracting those predictions from the actual observations. Then, these residuals are used as target variable by the Gradient Boosting model which is the second-ranked model in terms of prediction performance. After this, the prediction of CatBoost and Gradient Boosting for the given test data is summed up as the final result. In this approach, the performance metrics are $R^2 : 0.918$ and $RMSE : 22349.079$ which are not significantly better than CatBoost performance. Thus, it can be said that CatBoost can also predict rare cases along with its greatest performance.

After determining the predictive model, we can explore the most efficient factor in predicting sale price in Iowa between 2006 and 2010 by looking at the importance plot of CatBoost. The most important residential property in predicting sale price is GrLivArea(Living Area) which we have already seen in the second sub-research question. We can say that the sale price of the house increases as the living area of the house increases.

Model	Parameters	R^2	RMSE
Random Forest	'n_estimators': 50, 'min_samples_split': 5,' min_samples_leaf': 5, 'max_leaf_nodes': 7, 'max_features': 'auto', 'max_depth': 10, 'bootstrap': True	0.809	34150.546
Gradient Boosting	'learning_rate': 0.114, 'max_depth': 5, 'max_leaf_nodes': 4,' min_samples_split': 70, 'n_estimators': 90, 'subsample': 0.823	0.907	23880.368
XGBoost	'colsample_bylevel': 0.545, 'eta': 0.066, 'gamma': 6.655, 'learning_rate': 0.144, 'max_depth': 5, 'min_child_weight': 3.306, 'n_estimators': 90, 'reg_alpha': 0.787, 'subsample': 0.760	0.887	26252.205
Light GBM	'learning_rate': 0.275, 'max_depth': 7, 'min_child_weight': 0.372, 'n_estimators': 50, 'num_leaves': 100, 'reg_alpha': 0.261, 'subsample': 0.874	0.894	25528.921
CatBoost	'depth': 4,'iterations': 2000, 'l2 leaf reg': 2.414, 'learning rate': 0.145	0.918	22348.181
Baseline	None	0.648	46392.926

We can observe such an effect for OverallQual(Overal Quality), TotalBsmtSF(Total Basement Area), GarageArea(size of a garage in square feet), YearBuild(Built year), and LotArea which are the other important residential factors in predicting the sale price. The shap plot also represents that 2ndFlrSF (Second-floor area in square feet) may be considered as the important factor in the prediction of high prices although it is not in the top 10 list of the importance plot. The importance plot of CatBoost may not be sufficient in the determination of the most efficient factor because Gradient Boosting, LightGBM, and XGBoost also have satisfactory performance. When their importance plots are examined, it is seen that 1stFlrSF (First-floor area in square feet), BsmtTotal(Total basement area), GarageCars(Size of garage in cars), KitchenQuality(The quality of the kitchen) are also important residential factors in predicting house sale price in Ames, Iowa between 2006 and 2010. On the other hand, OverallQual is the most efficient residential factor in predicting sale price except CatBoost.

V. Conclusion and Discussion

This project mainly discusses developing accurate machine learning models for sale price by using tree-based algorithms and exploring the important residential factors which can be beneficial for the stakeholder of the real estate industry. It is shown that CatBoost is the model producing most accurate results. Among 80 factors, living area, which is also the most important above-ground residential property, overall quality, and neighborhood are the most efficient 3 residential properties in predicting the sale price. We see that area of the basement is also the most important basement property for sale price. Besides, overall quality, garage properties are also efficient. This study may be the first study using CatBoost encoder for house price prediction in place of one-hot encoding for categorical variables. In contrast to other house price sale studies, this study provides not a model for sale but also a factor affecting the sale price. That's why we cannot apply some feature engineering techniques which may affect the success but make the interpretation harder such as Principal Component Analysis, etc. Due to the same reason, we do not consider ensemble approaches for prediction. In the future, we may consider more predictive models such as neural network, Lasso and Ridge regressions whose success are proven before since we have only 5 models in this study. Also, we can give more importance to future engineering and consider the user-defined features obtained by the combination of some subset of the existing ones. Lastly, we can generate more sub-research questions like lot properties to explore the factors behind the sale price.

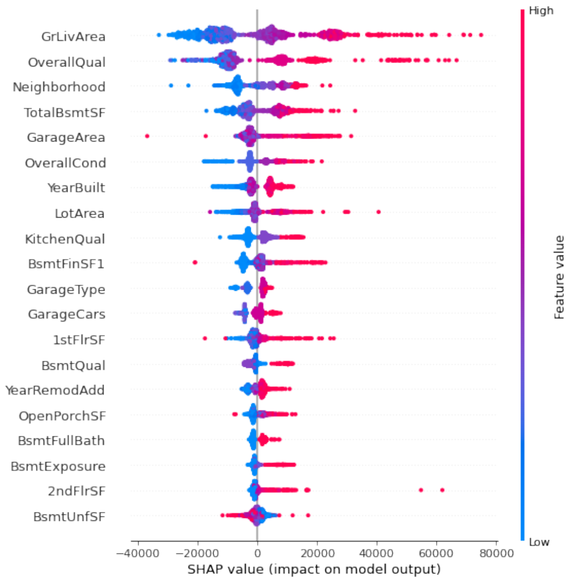

Fig. 13. Beeswarm Plot by CatBoost

References

[1] V. Limsombunchai, "House price prediction: hedonic price model vs. artificial neural network", New Zealand agricultural and resource economics society conference, 2004, pp. 25–26.

[2] S. Lu, Z. Li, Z. Qin, X. Yang and R. S. M. Goh, "A hybrid regression technique for house prices prediction," 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2017, pp. 319–323, doi: 10.1109/IEEM.2017.8289904.

[3] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 2018, pp. 35–42, doi: 10.1109/iCMLDE.2018.00017.

[4] D. Macpherson, and G. Sirmans, "Neighborhood Diversity and House-Price Appreciation," The Journal of Real Estate Finance and Economics, vol. 22, 2001, pp. 81–97, doi:10.1023/A:1007831410843.

[5] D. De Cock, "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project," Journal of Statistics Education, vol. 19, 2011, pp. 1—14.