House Price Prediction Advance Regression Techniques

Yukti Girdhar (18448940), Ashika Sethuraman (42367844), Amruta Kulkarni(17087028)

Abstract

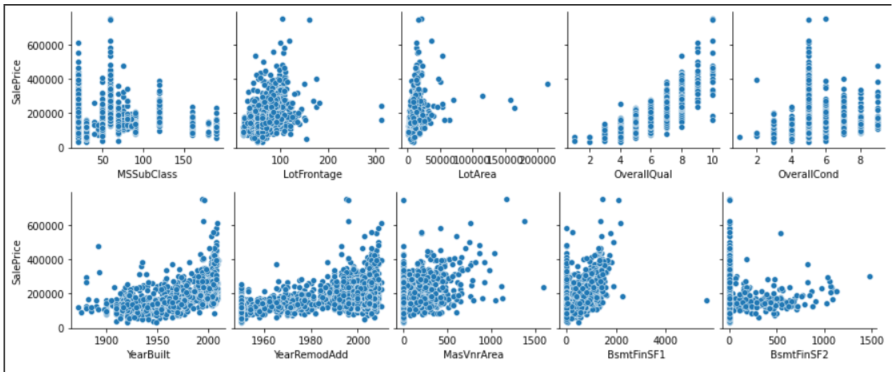
The aim of the report is to predict the prices of houses based on various features for example the type of dwelling, the zonal classification of sale, lot area etc. The initial dataset consists of 81 different features which are reduced using various feature engineering techniques like MI score, removing columns with missing values, low variance, correlated features. Once the data is preprocessed and ready to be used in the model, it is randomly divided into train and test data. There are around 3000 training examples which are divided into train and test in 4:1 ratio respectively. Various machine learning models like Linear Regression, Support Vector Machines, Gradient Boosting model i.e. XGBoost and Neural Networks are are trained and compared for the housing dataset. For evaluating the performance of the model various metric like RMSE, Mean Absolute Error, Variance Score and R2 curve are used. After testing of data is performed it is found that the Gradient boosting regressor outperforms the other models with least root mean square error.

1 Introduction

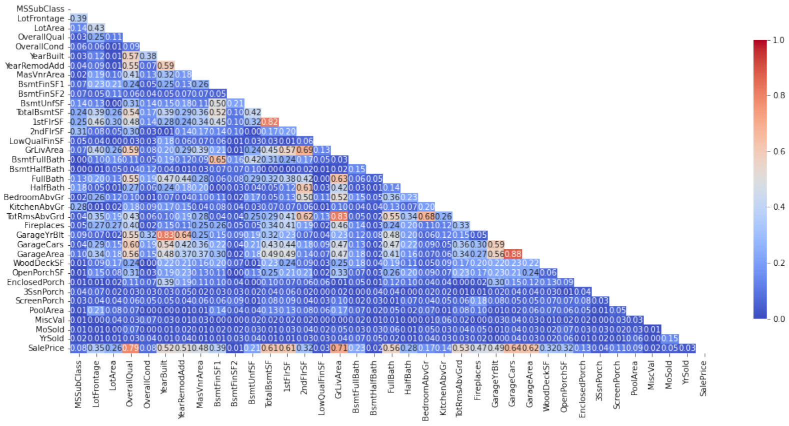
Since man came into being, he has been searching for houses. From austere settings to luxury homes, housing is every human being's need. There are different types of houses available depending on the location it is in. The coastal region have stilt houses while cities have apartments stacked one over other. Since housing is such an important market, the prediction of house prices becomes a very popular regression problem. In this project we are trying to play with various explanatory variables describing almost every aspect of residential homes. Someone who is looking for selling their house can also use the model to get a competitive pricing for their house. Our goal is also to explore and apply various principles of machine learning form data exploration, data pre-processing to training and performance evaluation techniques learnt in the CS273P course.

2 Data Exploration

The dataset contains 81 features (including ID and sales price columns) and 1490 records. After reading the data, we are comparing all the continuous features with respect to the target variable - sales price. Few visualizations are shown below.



Here the overallQual attribute seems to have a linear trend with sales price. It shows better the quality, more the housing price is. We visualized this kind of relationship for the continuous features in order to understand the impacts of these features on the model. Further we wanted to understand the correlation between all the features in the dataset.



3 Model Exploration

3.1 1. Linear Regression

It is a form of regression that is a comparison between two variables and determines if there is a linear relationship between them. In our problem statement, since we need to predict housing price(Y) and analyze the correlation, given a series of house characteristics(80 different features), we decided to use regression models to establish relationships and predict the housing prices. The evaluation metrics we are using in order to understand performance were mentioned earlier and justified.

3.2 2. Decision Tree

Decision tree regression is another model that has been around for a long time. This model breaks down the dataset into subsets based on various features until leaf node(decision) is reached. It branches out into diverse splits and provides various possible outcomes.

3.3 3. Bagging(Ensemble)

Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. In this several subsets of datasets are chosen to come up with varied decision trees. An average of those predictions are used which is more dynamic and robust than a single decision tree.

3.4 4. Random Forest

It is an extension over bagging. It takes one extra step in addition to taking the random subset of data. It also takes the random selection of features rather than using all features to grow trees. For higher dimensionality data it performs very well.For the random forest, we have tried feature engineering by tuning a few of the hyper parameters.

- N_estimators
- Max_depth
- min_sample_leaf

3.5 5. Boosting

Boosting is another ensemble technique to create a collection of predictors. We fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.

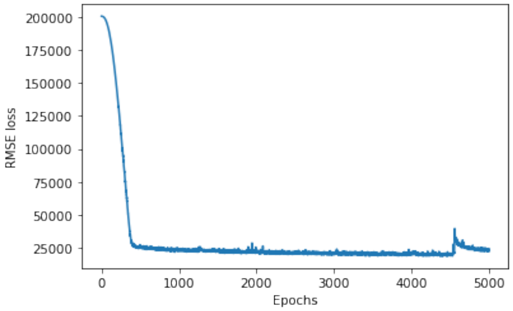
- XGboost
- GradientBoostingRegressor

3.6 6. SVM

For SVM, we tried 2 variations - Support Vector Classification, Support Vector Regression. We also used two types of kernels - linear and rbf.

3.7 7. Neural Network

In the housing price prediction problem we are dealing with a lot of categorical data. A key technique to making the most of deep learning model is to use embeddings for your categorical variables. This approach allows for relationships between categories to be captured. Since categorical data cannot be converted to float, the embeddings are formed as tensors. The continuous variables are stacked together and concatenated with categorical to as tensors train the model. Here a feed forward neural network is explored with the specifications, dropout=0.4, number of layers = 50-100, activation as ReLU. Finally RMSE loss is used as performance metric. The RMSE is decreasing with number of ephocs and becomes constant later.

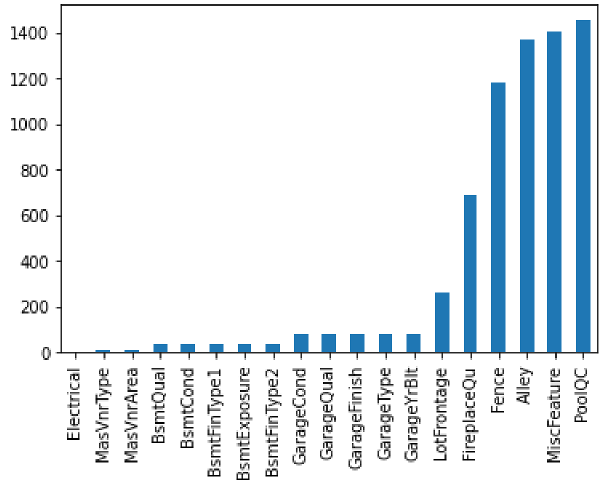


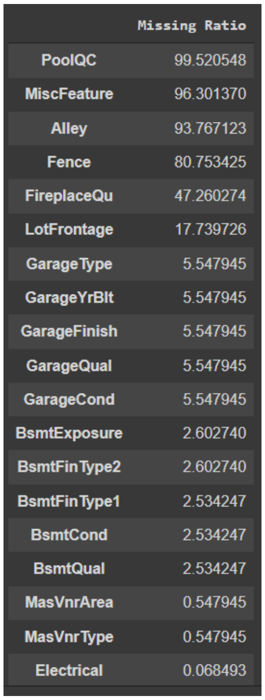
4 Data pre-processing and feature design

We wanted to assess the dataset - clean and preprocess it before running prediction models. We checked for below things-

- Identify missing data if any.
- How to handle these missing data
  - Remove data
  - Data imputation
- Encode the categorical data

1. We assessed the proportion of data that were missing across each feature in the data-set. Below is the ratio missing sorted in descending.





	Missing Ratio
PoolQC	99.520548
MiscFeature	96.301370
Alley	93.767123
Fence	80.753425
FireplaceQu	47.260274
LotFrontage	17.739726
GarageType	5.547945
GarageYrBlt	5.547945
GarageFinish	5.547945
GarageQual	5.547945
GarageCond	5.547945
BsmtExposure	2.602740
BsmtFinType2	2.602740
BsmtFinType1	2.534247
BsmtCond	2.534247
BsmtQual	2.534247
MasVnrArea	0.547945
MasVnrType	0.547945
Electrical	0.068493

2. Now the above missing data needs to be handled.

- We have removed the top 5 features where more than 50% of the records do not have any data. The reason for removing it is we do not have enough data to do imputation here.It would affect the natural distribution of data.
- For others, we have taken mean, median, mode. These decisions were based on the statistics which did not affect the natural distribution of data.

3. Now in order to consider categorical variables, we need to encode them. We have used label encoder to redefine these categorical features and convert them into continuous form.

5 Performance validation

We used following evaluation metrics to compare and rate the models we tried.

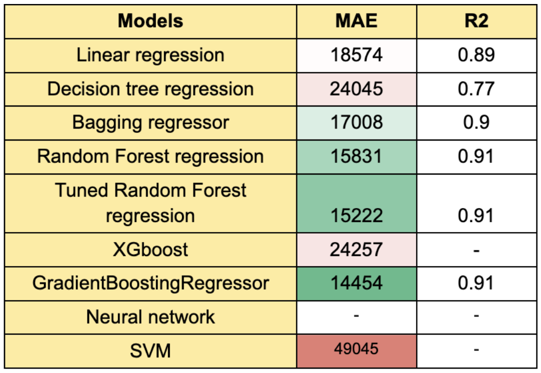
- Mean Absolute Error (MAE) :
  In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon.
- Mean Squared Error(MSE) :
  In statistics, the mean squared error (MSE) of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.
- R2 :
  R2 or R-squared metric denotes how well the regression model fits the observed data. Generally, a higher r-squared indicates a better fit for the model.
- Explained Variance Score :
  If $\hat{y}$ is the estimated target output, y the corresponding (correct) target output, and Var is Variance, the square of the standard deviation, then the explained variance is estimated as follows:

$$explained\_variance(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)} \tag{1}$$

6 Adaptation to under- and over-fitting

We observed over-fitting in Linear regression, SVM and Decision regressor. To combat that problem, we employed ensemble learning techniques - bagging, boosting and random forest. With ensemble learning, we got more generalized model and at a same time we were able to achieve good accuracy on testing data.

7 Results and Conclusion



Models	MAE	R2
Linear regression	18574	0.89
Decision tree regression	24045	0.77
Bagging regressor	17008	0.9
Random Forest regression	15831	0.91
Tuned Random Forest regression	15222	0.91
XGboost	24257	-
GradientBoostingRegressor	14454	0.91
Neural network	-	-
SVM	49045	-

- Gradient Boosting Regressor the model accuracy score was 90% for test and 96% for training. The model is not over fitting, performs well on both test and train data.
- We faced over-fitting issue for simpler models wherein training error was 0%. We used ensemble techniques to resolve the issue.
- RMSE for neural networks is higher than other models (187074). Neural Network does not perform well for this data.
- Even the best SVM has a MAE of 49045. Does not work well for this data.
- **Boosting technique** in ensemble learning helped to get good results in terms of performance
- We can conclude that ensemble learning has positive impact on model building and performance. They prove to great tool in the field of machine learning.

8 Citations and references

[1] https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
[2] https://www.kaggle.com/code/perotti/using-pytorch-on-hpp
[3] https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/
[4] https://www.fast.ai/2018/04/29/categorical-embeddings/

9 Contributions

Yukti Girdhar : Worked on data preprocessing and feature design and explored the neural network model.

Ashika Sethuraman : Contributed towards data exploration and modeling all the regression techniques.

Amruta Kulkarni : Worked on Support Vector Classification and Support Vector Regression. Checked performance metrics and over- fitting for models.