**Ablation Report on House Sale Price Prediction using Machine Learning**

**<Student Name>**

**<University Name>**

**CSCE 5320 Scientific Data Visualization**

**<Professor Name>**

**14th of April, 2023.**

**Introduction**

The current task at hand is to write an Ablation report, which basically helps us to study the features that shall be used to train a machine learning model which can perform some task. The goal is to understand and note the change in performance of the model, as we leave out a feature for training and testing. How much of a difference would it make? Would you achieve the same results if you remove a column? And by achieving the same results by dropping a column we are feeding lesser training data, this achieving model optimization.

**Data Preprocessing and Normalization**

The current dataset at hand is a dataset that has sale values of houses, there are a total of 38 features within the train set. Out of these *81* Features, 38 Features are continuous and 43 features are categorical in nature. One of the primary observations that I made while simply understanding data with inbuilt pandas methods, the dataset contains a lot of null values! I had a choice to make here, I could either drop rows with null values, or fill these with some default values. Here’s what I did:

* I first noted the correlation for each continuous feature to understand their significance! This is because some of the features have very high range, and my hypothesis is the fact that features with higher range would be skewed towards having very high importance (co-relation) more so than they should, the observed correlations are as follows:

|  |  |
| --- | --- |
| Feature | Correlation |
| KitchenAbvGr | -0.13 |
| EnclosedPorch | -0.21 |
| MSSubClass | -0.08 |
| OverallCond | -0.07 |
| YrSold | -0.02 |
| LowQualFinSF | -0.02 |
| Id | -0.02 |
| MiscVal | -0.02 |
| BsmtHalfBath | -0.01 |
| BsmtFinSF2 | -0.01 |
| 3SsnPorch | 0.04 |
| MoSold | 0.04 |
| PoolArea | 0.09 |
| ScreenPorch | 0.11 |
| BedroomAbvGr | 0.16 |
| BsmtUnfSF | 0.21 |
| BsmtFullBath | 0.22 |
| GarageYrBlt | 0.26 |
| LotArea | 0.26 |
| HalfBath | 0.28 |
| OpenPorchSF | 0.31 |
| 2ndFlrSF | 0.31 |
| WoodDeckSF | 0.32 |
| LotFrontage | 0.34 |
| BsmtFinSF1 | 0.38 |
| Fireplaces | 0.46 |
| MasVnrArea | 0.47 |
| YearRemodAdd | 0.50 |
| YearBuilt | 0.52 |
| TotRmsAbvGrd | 0.53 |
| FullBath | 0.56 |
| 1stFlrSF | 0.60 |
| TotalBsmtSF | 0.61 |
| GarageArea | 0.62 |
| GarageCars | 0.64 |
| GrLivArea | 0.70 |
| OverallQual | 0.79 |
| SalePrice | 1.00 |
|  |  |

***Table Correlations of Continuous Features.***

* Usually a negative correlation shows inverse relationship between two variables, and a positive correlation depicts direct influence within two variables. As we can see close to 13 features! Have a very low correlation! Close to 0.0X this is not very useful in terms of training!
* This could possibly be because of the range of values within these features! One way to reduce the range to prevent skewing of significance (correlation) towards features with higher range is to limit this range.
* The easiest approach to this would be scaling the data. By subtracting the minimum value of the feature, and then dividing it by the range of the data (difference between the max and min of the same feature) we achieved a scaled value. This limits the range of each sample within continuous feature.
* The goal of the Regression task is to predict a numerical value that depicts the sale price of the house based on its features. However we know that categorical data within object form cannot be provided as input to Regression models. We can simply drop the categorical features but that would be very silly to do so.
* In order to work around this I’ve decided to one hot encode the 43 Features that are categorical in nature.
* With the data frames obtained from the above activities, I can now combine them to form my final normalized dataset, which will be used to train and validate the Regression Model.
* The final data frame contains of exact rows as the original set, and 300 features/columns as a result of transformations made above. These features help us capture most of the variance within the data, which is the absolute task of the Regression model, i.e. to capture most of the variance of the data.
* Additionally we are to find log values of the target variable, hence before using this to create training set and validation set, I transformed the target variable samples by replacing original values with their respective log values.

**Training and Validation**

Traditionally, *linear regression* model would be very useful for current dataset and task at hand, but I decided to choose *DecisionTreeRegressor* as my Machine Learning model, as Decision Trees are capable of capturing non-linear relationships between features and target variable, and they’re also less sensitive to outliers within data. These two criteria’s with the structure of dataset in mind, I decided to train a new instance of *DecisionTreeRegressor.*

Now that I had decided on what machine learning model would be the best fit for my dataset, I had to split the data into training set and validation set, with the objective of using the samples from validation set to obtain the model’s performance metrics.

I split the data into training and validation set/test set in 80/20 ratio, i.e. 80% of the final dataset we obtained from normalization will be used to train the model, while 20% will be kept aside for validation purposes.

Upon training the model and testing the newly trained model I achieved a great r2 score and other metrics as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | R2 score | MSE score | RMSE score |
| DecisionTreeRegressor | 0.996 | 0.000693 | 0.02633 |

***Table: Model Metrics.***

These metrics clearly show very high performance of our model, with very small error for the predicted values! This is great, this will allow us to understand the Ablation study for this model and this dataset!

**Ablation Study**

To introduce Ablation Study, we can simply say that it is used to evaluate the significance of individual features within the context of current task, ablation means removal as the name suggest very measure the change of performance by doing the same.

For the current task, I’ll use the leave one out approach, i.e. remove one feature per iteration and note the resulting values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | R2 score | MSE score | RMSE score | Change |
| All Features | 0.996 | 0.000693 | 0.02633 | Original |
| MSSubClass | 0.996 | 0.000590 | 0.02428 | Insignificant |
| MSZoning | 0.998 | 0.000318 | 0.01783 | Slight Increase |
| LotFrontage | 0.998 | 0.000318 | 0.01783 | Slight Increase |
| LotArea | 0.994 | 0.000955 | 0.03090 | Slight Decrease |
| Street | 0.994 | 0.000972 | 0.00097 | Slight Decrease |
| Alley | 0.998 | 0.000317 | 0.01779 | Slight Increase |
| LotShape | 0.995 | 0.000819 | 0.02863 | Slight Decrease |
| LandContour | 0.993 | 0.001140 | 0.03375 | Slight Decrease |
| Utilities | 0.996 | 0.000617 | 0.02484 | Insignificant |
| LotConfig | 0.997 | 0.000513 | 0.02264 | Slight Increase |
| LandSlope | 0.993 | 0.001184 | 0.03440 | Slight Decrease |
| Neighborhood | 0.996 | 0.000566 | 0.02379 | Insignificant |
| Condition2 | 0.993 | 0.001208 | 0.03475 | Slight Decrease |
| Condition1 | 0.991 | 0.001626 | 0.04033 | Slight Decrease |
| BldgType | 0.996 | 0.000651 | 0.02550 | Insignificant |
| HouseStyle | 0.996 | 0.000719 | 0.02682 | Insignificant |
| OverallQual | 0.998 | 0.000297 | 0.01722 | Slight Increase |
| OverallCond | 0.996 | 0.000616 | 0.02481 | Insignificant |
| YearBuilt | 0.997 | 0.000494 | 0.02223 | Slight Increase |
| YearRemodAdd | 0.994 | 0.000942 | 0.03069 | Slight Decrease |
| RoofStyle | 0.997 | 0.000416 | 0.02038 | Slight Increase |
| RoofMatl | 0.995 | 0.000785 | 0.02802 | Slight Decrease |
| Exterior1st | 0.995 | 0.000903 | 0.03004 | Slight Decrease |
| Exterior2nd | 0.998 | 0.000345 | 0.01857 | Slight Increase |
| MasVnrType | 0.996 | 0.000731 | 0.02703 | Insignificant |
| MasVnrArea | 0.993 | 0.001124 | 0.03352 | Slight Decrease |
| ExterQual | 0.990 | 0.001690 | 0.04110 | Slight Decrease |
| ExterCond | 0.993 | 0.001182 | 0.03437 | Slight Decrease |
| Foundation | 0.992 | 0.001369 | 0.03700 | Slight Decrease |
| BsmtQual | 0.996 | 0.000580 | 0.02408 | Insignificant |
| BsmtCond | 0.994 | 0.001052 | 0.03242 | Slight Decrease |
| BsmtExposure | 0.997 | 0.000456 | 0.02135 | Slight Increase |
| BsmtFinType1 | 0.994 | 0.000943 | 0.03070 | Slight Decrease |
| BsmtFinSF1 | 0.997 | 0.000391 | 0.01977 | Slight Increase |
| BsmtFinType2 | 0.995 | 0.000870 | 0.02949 | Slight Decrease |
| BsmtFinSF2 | 0.995 | 0.000914 | 0.03019 | Slight Decrease |
| BsmtUnfSF | 0.997 | 0.000457 | 0.02136 | Slight Increase |
| TotalBsmtSF | 0.997 | 0.000398 | 0.01996 | Slight Increase |
| Heating | 0.997 | 0.000440 | 0.02097 | Slight Increase |
| HeatingQC | 0.998 | 0.000263 | 0.01622 | Slight Increase |
| CentralAir | 0.994 | 0.000944 | 0.03073 | Slight Decrease |
| Electrical | 0.995 | 0.000928 | 0.03047 | Slight Decrease |
| 1stFlrSF | 0.995 | 0.000899 | 0.02998 | Slight Decrease |
| 2ndFlrSF | 0.994 | 0.001095 | 0.03308 | Slight Decrease |
| LowQualFinSF | 0.997 | 0.000389 | 0.01972 | Slight Increase |
| GrLivArea | 0.998 | 0.000321 | 0.01790 | Slight Increase |
| BsmtFullBath | 0.996 | 0.000564 | 0.02374 | Insignificant |
| BsmtHalfBath | 0.997 | 0.000540 | 0.02323 | Slight Increase |
| FullBath | 0.998 | 0.00020 | 0.01414 | Slight Increase |
| HalfBath | 0.998 | 0.000314 | 0.01773 | Slight Increase |
| BedroomAbvGr | 0.994 | 0.001021 | 0.03195 | Slight Decrease |
| KitchedAbvGr | 0.997 | 0.000429 | 0.02072 | Slight Increase |
| KitchenQual | 0.998 | 0.000368 | 0.01914 | Slight Increase |
| TotRmsAbvGrd | 0.997 | 0.000390 | 0.01974 | Slight Increase |
| Functional | 0.996 | 0.000602 | 0.02453 | Insignificant |
| Fireplaces | 0.997 | 0.000414 | 0.02034 | Slight Increase |
| FireplaceQu | 0.998 | 0.001031 | 0.32111 | Slight Increase |
| GarageType | 0.997 | 0.000481 | 0.021930 | Slight Increase |
| GarageYrBlt | 0.997 | 0.000601 | 0.024521 | Slight Increase |
| GarageFinish | 0.997 | 0.000802 | 0.028318 | Slight Increase |
| GarageCars | 0.997 | 0.000656 | 0.025604 | Slight Increase |
| GarageArea | 0.994 | 0.001007 | 0.031737 | Slight Decrease |
| GarageQual | 0.995 | 0.000877 | 0.029616 | Slight Decrease |
| GarageCond | 0.995 | 0.000770 | 0.027746 | Slight Decrease |
| PavedDrive | 0.993 | 0.001123 | 0.033506 | Slight Decrease |
| WoodDeckSF | 0.995 | 0.000852 | 0.029184 | Slight Decrease |
| OpenPorchSF | 0.997 | 0.000424 | 0.020595 | Slight Increase |
| EnclosedPorch | 0.997 | 0.000520 | 0.022799 | Slight Increase |
| 3SsnPorch | 0.995 | 0.000831 | 0.028824 | Slight Decrease |
| ScreenPorch | 0.993 | 0.001177 | 0.034304 | Slight Decrease |
| PoolArea | 0.993 | 0.001125 | 0.033547 | Slight Decrease |
| PoolQc | 0.996 | 0.000565 | 0.023764 | Insignificant |
| Fence | 0.998 | 0.000361 | 0.018996 | Slight Increase |
| MiscFeature | 0.994 | 0.001011 | 0.031793 | Slight Decrease |
| MiscVal | 0.991 | 0.001524 | 0.039038 | Slight Decrease |
| MoSold | 0.997 | 0.000519 | 0.022771 | Slight Increase |
| YrSold | 0.997 | 0.000502 | 0.022396 | Slight Increase |
| SaleType | 0.994 | 0.001091 | 0.033032 | Slight Decrease |
| SaleCondition | 0.996 | 0.000569 | 0.023851 | Insignificant |

***Table Ablation Report for DecisionTreeRegressor without Hyperparams***