PROJECT REPORT ON HOUSING DATA

The housing data contains the two types of datasets.

Train.csv and Test.csv of rows and columns numbers are (1460, 81) for train.csv

And (1459, 80) for test.csv.

Train data contains 81 columns with target variable called SalePrice.

We have to work on the target data of train dataset.

Simultaneously, we also worked on test data .

The test contains 80 columns with no target variable.

It contain 80 columns without the columns name saleprice.

Also, both dataset contain lots and lots of nan values (Missing Values) .

We used heatmap to check null values in dataset.

We used mean and mode value to fill missing values with column names.

Dropped the unwanted data and columns to clean the dataset with duplicate values.

Simultaneously, we did the same thing on test dataset.

Further we did EDA process on housing dataset.

EDA (Exploratory Data Analysis)

Also drop some nan values and unwanted from train and test dataset.

Also we handle the categorical feature in dataset.

The categorical feature contains 39 columns.

After this we apply one hot encoding on categorical feature.

We also did some normalization on some columns.

We did countplot on MSSubclass, Saleprice.

We combine the train data and test data to handle the categorical features.

We also checked the skew ness of data i.e. 1.8919117627933302.

We combine all the columns in final dataframe by using One hot encoder.

After This process we used model Evaluation on dataset.

MODEL EVALUATION

After this we used in RandomForest classifier to get best accuracy model.

```
We got accuracy around 0.9950773558368495
```

This is the best model to work on the house data.

The accuracy of R*2 is :

```
R^2 is:
 0.9999367794767109
```

Conclusion

We get the best model and the better accuracy of dataset i.e.

```
R^2 is:
 0.9999367794767109
```

.