# MACHINE LEARNING – TERM 3

- SESHATHRI S

# PROBLEM STATEMENT

The problem statement of this project is to determine the price of the house using the data set provided.

The dataset contains multiple factors that contributes to the increase/decrease of the house price.

In this project, we will be exploring the dataset and build models using machine learning algorithms to predict the price of the house

# GIVEN DATASET

1460 records

81 columns (one target and 80 independent variable)

Contains both categorical and continuous variables
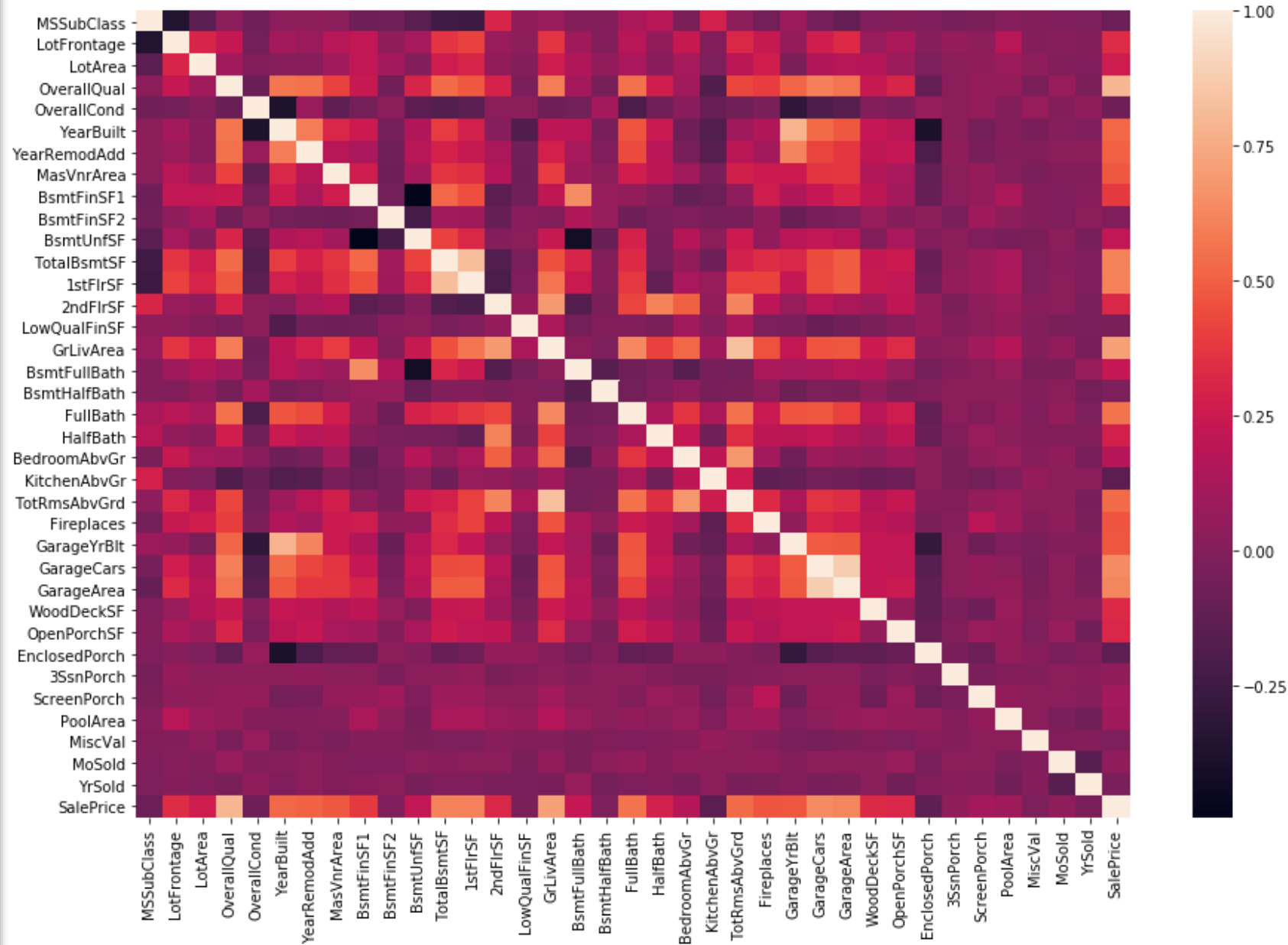
# PRE PROCESSING DATA

Removed columns that have less than 1000 records

Replaced missing categorical value with mode value

Replaced missing continuous value with mean value

Encoded the categorical values

# HEAT MAP ON CONTINUOUS DATA

# TRAINING MODELS

- Decision tree and Random forest algorithms are used

- Multiple parameters like the test size, scaled data, Grid Search, Randomized Search

- Multiple test sizes are used (0.15, 0.2, 0.25, 0.3, 0.35, 0.4)

# RESULTS

48 models were generated

Results were compared with MAE, MSE, RMSE, MSLE and R-squared values

Time taken for each model to train and test were also calculated

# CONCLUSION

- The random forest model with the test size of 0.15 has the highest accuracy score of 92%

- The same model has an accuracy score of 88% for other test sizes (0.2, 0.25, 0.3)

| | Scaled | Test size | Time taken | MAE | MSE | RMSE | MSLE | R2 |
|---|---|---|---|---|---|---|---|---|
| Random Forest | False | 0.15 | 00:00:00.312882 | 16010.574429 | 5.182287e+08 | 22764.636424 | 0.020405 | 0.920040 |
| Random Forest | False | 0.20 | 00:00:00.232549 | 17467.646918 | 7.915036e+08 | 28133.673734 | 0.024228 | 0.889020 |
| Random Forest | False | 0.25 | 00:00:00.223355 | 17862.902740 | 7.506323e+08 | 27397.668715 | 0.022894 | 0.888186 |
| Random Forest | False | 0.30 | 00:00:00.245639 | 18773.155708 | 8.432271e+08 | 29038.372223 | 0.029526 | 0.881906 |
| Random Forest | True | 0.15 | 00:00:00.248706 | 0.097774 | 2.057914e-02 | 0.143454 | 0.000127 | 0.878796 |

Top 5 accurate models