PROJECT NAME
HOUSING - PRICE PREDICTION PROJECT

SUBMITTED BY
MRS.VARSHA V. SHINDE

FLIP ROBO

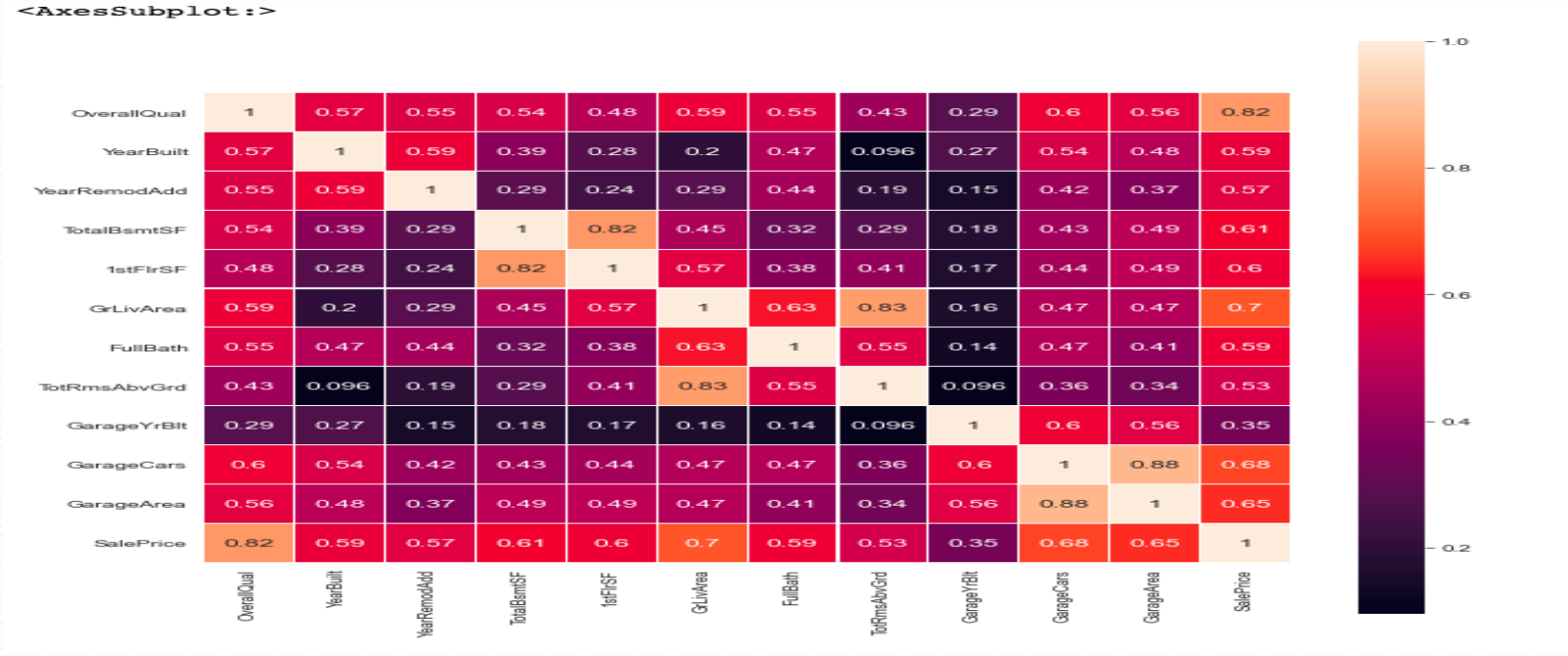
TOPICS

- ✓Problem statement
- ✓EDA STATMENT & VISUALIZATION
- ✓STEPS & ASSUSMPTIONS USED TO COMPLTE PROJECT MODEL DHASBOARD
- ✓FINALIZE MODEL AND CONCLUSION

Problem statement

- Prices of real estate properties are sophisticatedly linked with our economy. Despite this, we do not have accurate measures of housing prices based on the vast amount of data available.
- Therefore, the goal of this project is to use machine learning to predict the selling prices of houses based on many economic factors. A systematic method can be built to derive a layered knowledge graph and design a structured Deep Neural Network (DNN) based on it. Neurons in a structured DNN are structurally connected, which makes the network time and space efficient; and thus, it requires fewer data points for training.
- The structured DNN model has been designed to learn from the most recently captured data points which allows the model to adapt to the latest market trends. To demonstrate the effectiveness of the proposed approach, we can use a case study of assessing real properties in small towns

EDA STATEMENT & VISUALIZATION



From the above heat map, garage space, general living area, and overall quality metric are highly correlated with our target variable.

Continue....

Exploratory Data Analysis (EDA). By conducting explanatory data analysis, we obtain a better understanding of our data. This yields insights that can be helpful later when building a model, as well as insights that are independently interesting

In this report, we describe our approach to these steps and the results that we obtained.

In order to understand our data, we first perform exploratory data analysis. This will provide us with insights that will be useful in building prediction models, as well as insights that may be of interest to stakeholders. As part of the Exploratory Data Analysis we aim to: • Look into the relationship between each variables and annual house price percentage change, and identify any patterns.

For example, between the year of construction of a house and its annual percent price change. We will also analyse relationships between the features. This may reveal that certain features are redundant and this would help the subsequent analysis.

STEPS & ASSUMPTION

Logistic Regression

It assumes that there is minimal or no multi collinearity among the independent variables.

It usually requires a large sample size to predict properly.

It assumes the observations to be independent of each other

Linear Regression

There should be a linear relationship.

There should be no or little multicollinearity.

Decision Trees

Initially, whole training data is considered as root.

Records are distributed recursively on the basis of the attribute value.

Naive Bayes

The biggest and only assumption is the assumption of conditional independence.

CONCLUSION

- This paper investigates different models for housing price prediction. Three different types of Machine Learning methods including Random Forest, XGBoost, and LightGBM and two techniques in machine learning including Hybrid Regression and Stacked Generalization Regression are compared and analyzed for optimal solutions.
- Even though all of those methods achieved desirable results, different models have their own pros and cons. The Random Forest method has the lowest error on the training set but is prone to be over fitting. Its time complexity is high since the dataset has to be fit multiple times.
- The XGBoost and LightGBM are decent methods when comparing accuracy, but their time complexities are the best, especially LightGBM. The Hybrid Regression method is simple but performs a lot better than the three previous methods due to the generalization.
- Finally, the Stacked Generalization Regression method has a complicated architecture, but it is the best choice when accuracy is the top priority. Even though Hybrid Regression and Stacked Generalization Regression deliver satisfactory results, time complexity must be taken into consideration since both of them contain Random Forest, a high time complexity model.

CONTINUE..

- Stacked Generalization Regression also has K-fold cross-validation in its mechanism so it has the worst time complexity. Further research about the following topics should be conducted to further investigate these models, especially the combinations of different models:

  a] The coupling effect of multiple regression models.

  b] The “re-learn” ability of machine learning models.

  c] The combination of Machine Learning and Deep Learning methods.

  d] The driven factors for the good performance of tree-based models.

  e] The faster ways to fit complex models.