HOUSE PRICE PREDICTION

CAPSTONE PROJECT – TEAM 4

MILESTONE 2 REPORT

Contents

OVERVIEW - DATA FOR MODELS BUILDING

- The dataset for the models building contains **21957 houses** and 26 features; Price, room_bed, room_bath and 23 others.

- We want to build the following regression models using train dataset and validate their performances on the test data.
  - ✓ Linear Regression Model (OLS)
  - ✓ Decision Tree Regressor Model
  - ✓ Random Forest Regressor Model
  - ✓ XGBoost Regressor Model

- We will evaluate the performance of the models using the following metrices.
  - ✓ RMSE (Root Mean Squared Error)
  - ✓ MAE (Mean Absolute Error)
  - ✓ R-squared (Coefficient of Determination)
  - ✓ Adjusted R-squared
  - ✓ MAPE (Mean Absolute Percentage Error)

- We will split the dataset into train and test to be able to evaluate the model that we will build on the train data.
  - ✓ Test data will cover 30% of the dataset
  - ✓ Train data will also cover 70%

OVERVIEW - DATA FOR MODELS BUILDING

- Before we proceed to build the models, we will drop 'yr_built', 'cid', 'yr_renovated', 'yr_sold', 'total_area' and 'dayhours'

- Why drop these variables:
  - ✓ cid - Not significant in the model building - they are just codes assigned to the houses.
  - ✓ renovation_status (new variable) has been created from yr_renovated.
  - ✓ yr_sold, yr_built & dayhour - yr_before_sold (new variable) has been created to replace these variables
  - ✓ total_area comprises; living_measure & lot_measure. Once were are adding these two variables to the model, total_area, which is their sum can be dropped.

- We have **20 features** for the models building:
  - ✓ Price – dependent variable.
  - ✓ 19 others – independent variables.

LINER REGRESSION MODEL (OLS)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	181785.73	106521.90	0.75	0.75	21.60

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	185359.29	108567.58	0.75	0.75	21.63

Inference

- The evaluation metrices for the test and train data are comparable. This indicates that there is no overfitting or underfitting
- The model has performed and generalized well on the test dataset.
- R-squared and adjusted R-squared (after considering the variables without multicollinearity) are 0.75. This is a good score.
- This indicates that 75% of the variability in the price of the houses are explained by independent variables in our model.
- RMSE – On average, the model's predicted prices on the test data differ from the actual prices by 185,359.29
- The MAE value is 108,567.58. This represents the average absolute difference between the model's predicted prices and the actual prices on the test data.
- MAPE - on average, the model's predictions deviate by 21.63% from the actual Prices.

ASSUMPTION OF LINER REGRESSION MODEL (OLS)

Assumption	How to test	Explanation
No multicollinearity in independent variables	VIF (Variance inflation factor)	Variables with VIF more than 5 were removed
There should be a linear relationship between dependent and independent variables	Plot residuals vs. fitted values and check the plot	We see no pattern in the plot below. Hence, the assumptions of linearity is satisfied
The residuals should be independent of each other	Plot residuals vs. fitted values and check the plot	We see no pattern in the plot below. Hence, the assumptions of independence is satisfied
Residuals must be normally distributed	Plot Q-Q plot	The residual terms are normally distributed
No heteroscedasticity, i.e., residuals should have constant variance	Use statistical test (like goldfeldquandt test): Null hypothesis : Residuals are homoscedastic Alternate hypothesis : Residuals have hetroscedasticity	P - value was 0.999999. Since p-value > 0.05 we can say that the residuals are homoscedastic

Fig - Linearity of variables & Independence of error terms assumptions

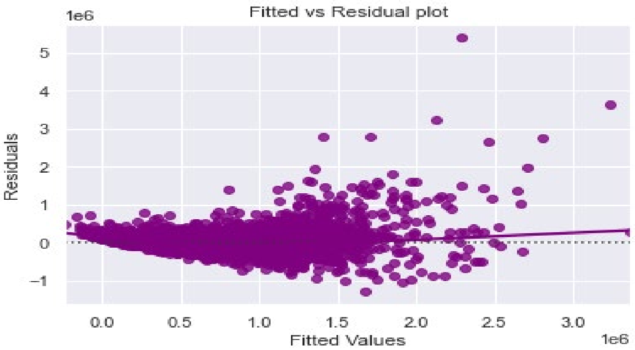
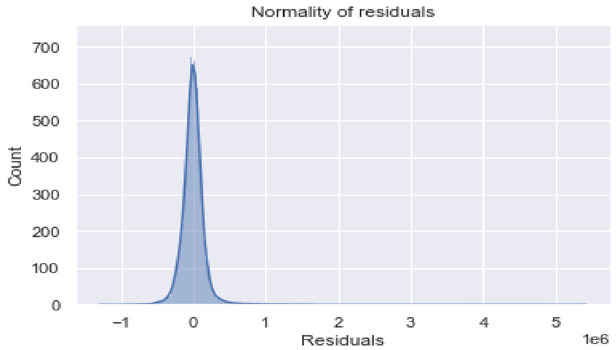


Fig - Normality of error terms assumptions



DECISION TREE REGRESSOR MODEL(DEFAULT)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	3041.06	126.72	1.00	1.00	0.04

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	181846.99	102022.04	0.76	0.76	18.62

- Minimal errors on the training set, each price has been predicted correctly.
- The model has performed very well on the training set.

- The decision tree model is overfitting the data as expected and not able to generalize well on the test set.
- We will have to prune the decision tree.

DECISION TREE REGRESSOR MODEL(TUNNED)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	119688.59	63209.73	0.89	0.89	11.60

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	164115.60	88583.43	0.81	0.80	15.88

- The performance of the model after hyperparameter tuning has become generalized.
- The model can explain 80% of the variability in the prices with the independent variables in the test data within 15.88% error margin

RANDOM FOREST REGRESSOR MODEL(DEFAULT)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	48483.71	25932.75	0.98	0.98	4.90

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	131914.36	70290.35	0.87	0.87	12.94

- The metrices are not comparable. Adj. R-squared for Train and Test data are 0.98 and 0.87, respectively.
- The model is overfitting on the train dataset. Let's try to reduce overfitting and improve the performance by hyperparameter tuning.

RANDOM FOREST REGRESSOR MODEL(TUNNED)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	135397.34	81995.00	0.86	0.86	16.12

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	156704.65	88947.29	0.82	0.82	16.62

- After tunning the model, the model performance has generalized.
- We have Adj. R-squared 0.86 for train and 0.82 for test.
- On the average, the model’s predicted prices deviated from actual prices by 16.62% on the test and 16.12% on train data

XGBOOST REGRESSOR MODEL(DEFAULT)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	56998.02	40549.90	0.98	0.98	9.03

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	126688.00	70253.64	0.88	0.88	12.90

- The model is overfitting on the train data – Adj. R-square is 0.98 and that of Test data is 0.88
- Let’s try to reduce the overfitting and improve the performance of the model

XGBOOST REGRESSOR MODEL(TUNNED)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	79639.41	54700.92	0.95	0.95	11.47

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	124680.15	70297.02	0.89	0.89	13.07

- After tunning the model, the model performance has generalized.
- We have Adj. R-squared 0.95 for train and 0.89 for test.
- There is big difference between RMSE for train data and RMSE for Test data.

MODELS PERFORMANCE COMPARISON

Model/Metrics	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Linear Regression_Train	181,785.73	106,521.90	0.75	0.75	21.60
Linear Regression_Test	185,359.29	108,567.58	0.75	0.75	21.63
Decision Tree_Train	3,041.06	126.72	1.00	1.00	0.04
Decision Tree_Test	181,846.99	102,022.04	0.76	0.76	18.62
Tunned Deccion Tree_Train	119,688.59	63,209.73	0.89	0.89	11.60
Tunned Deccion Tree_Test	164,115.60	88,583.43	0.81	0.80	15.88
Random Forest_Train	48,678.37	25,945.45	0.98	0.98	4.92
Random Forest_Test	132,558.54	70,525.86	0.87	0.87	12.98
Tunned Random Forest_Train	135,397.34	81,995.00	0.86	0.86	16.12
Tunned Random Forest_Test	156,704.65	88,947.29	0.82	0.82	16.62
XGBoost_Train	56,998.02	40,549.90	0.98	0.98	9.03
XGBoost_Test	126,688.00	70,253.64	0.88	0.88	12.90
Tunned XGBoost_Train	79,639.41	54,700.92	0.95	0.95	11.47
Tunned XGBoost_Test	124,680.15	70,297.02	0.89	0.89	13.07

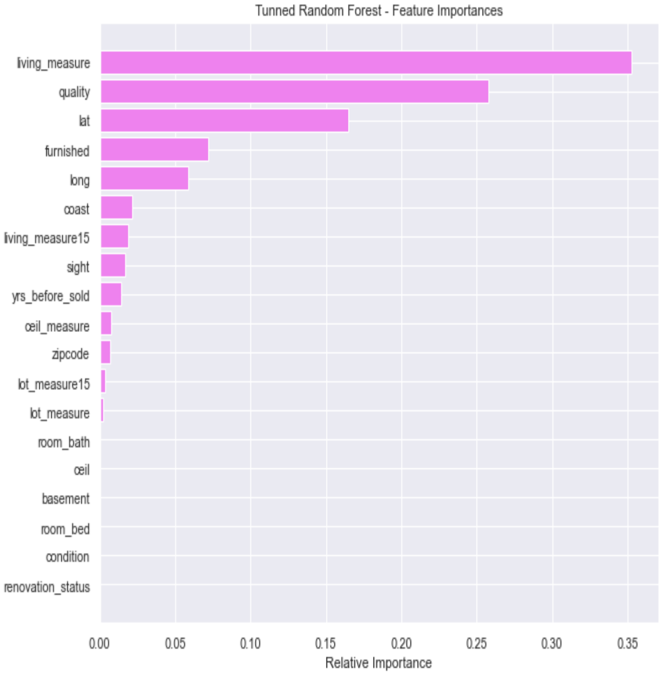
Tunned Random Forest is the best performing model.

THE BEST MODEL PERFORMANCE

Tunned Random Forest Regression Model

- The evaluation metrices for the test and train data are comparable. This indicates that there is no overfitting or underfitting
- The model has performed and generalized well on the test dataset.
- R-squared and adjusted R-squared for the train data is 0.86 and that of the test is 0.82. This is a good score.
- This indicates that 82% of the variability in the price of the houses are explained by independent variables in our model on the test data.
- RMSE – On average, the model’s predicted prices on the test data differ from the actual prices by 156,704.65
- The MAE value is 88,947.29. This represents the average absolute difference between the model's predicted prices and the actual prices on the test data.
- MAPE - on average, the model's predictions deviate by 16.62% from the actual Prices on the test data.

Fig - Important Features in the model



The top five variables in the model that contributed largely to the prediction of the house prices:

- **Living_measure**
- **Quality**
- **Furnished**
- **Lat & long (location of the house)**

BUSINESS INSIGHT AND RECOMMENDATION

- We have built a random forest regressor model that is able to explain approximately 82% of the variability in the prices by the independent variables in our model within an average error margin of 16.62%

- The top five contributing independent variables to our model predictions are as follows:
  living_measure
  quality
  furnished
  Location (lat & long)

- From the visualization analysis, we observed a high positive correlation between price and living_measure. This indicates that, the larger the square footage of the living area the higher the price.

- Furnished – We observed that furnished houses have higher prices than non furnished houses in the dataset

- Quality is another significant factor, high-quality houses (graded 9 - 13) command higher prices than lower quality houses. Notably, high-quality houses are furnished, while low and moderate-quality houses (graded 3 - 8) are non-furnished

- Additional information of the houses can be collected to gain better insights. Information such as: location crime rates, school quality, proximity to public transportation/train station, house in gated community, presence of a security system, etc. These may appeal to certain buyers and can significantly influence house prices.

END OF REPORT