

The background of the slide is a photograph of a server room. On both sides, there are long rows of server racks filled with electronic equipment. The perspective is looking down a central aisle. In the center of the image, there is a large, bright, white, fluffy cloud that appears to be floating or rising from the floor. The lighting is somewhat dim, with the cloud providing a bright focal point.

Web Mining Final Project

**New Jersey Institute of Technology
Spring 2020**

House Prices Predictor



Submitted By : Group 3

Members:
Haisam Ezmat
Pratik Parija
Sai Akhilesh Chunduri
Munazzam Mirza

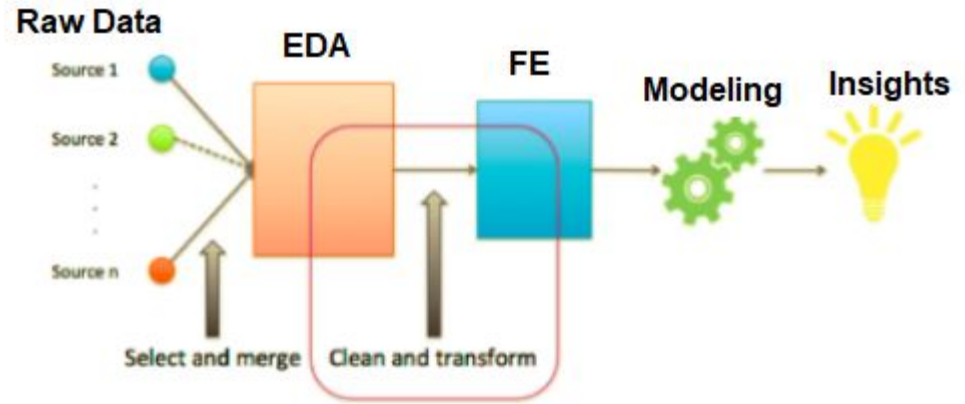
Prediction of House Price

We divided our project in 3 sections:

1: Exploratory Data Analysis (EDA)

2: Feature Engineering (FE)

3: Model Building - Training/Testing



Exploratory data analysis

Exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

We generated the following views.

- Graphs
- Histograms
- Bar Plots
- Density Plots



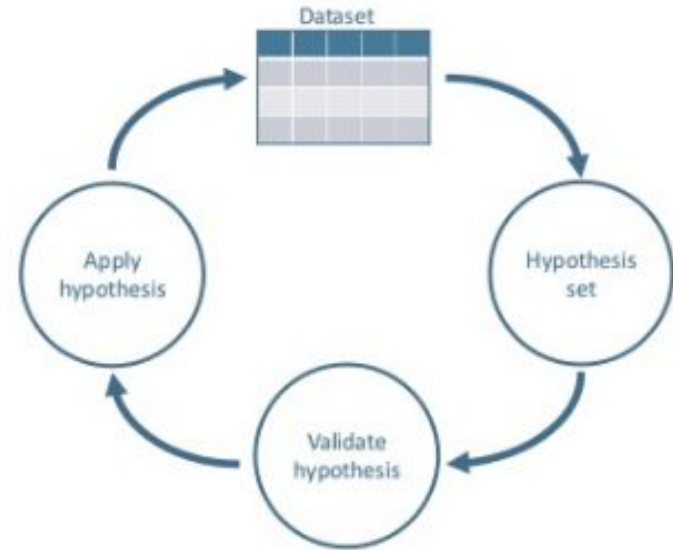
Feature Engineering

Feature engineering is the process of identifying and extracting predictive features from the data that can help us in making assumptions.

We analyzed the following three categories against Sale Price and created a Correlation Heatmap.

- Number of Bathrooms
- House Age
- Neighbourhood

Feature Engineering Cycle



Model Building

A machine learning algorithm has two types of parameters.

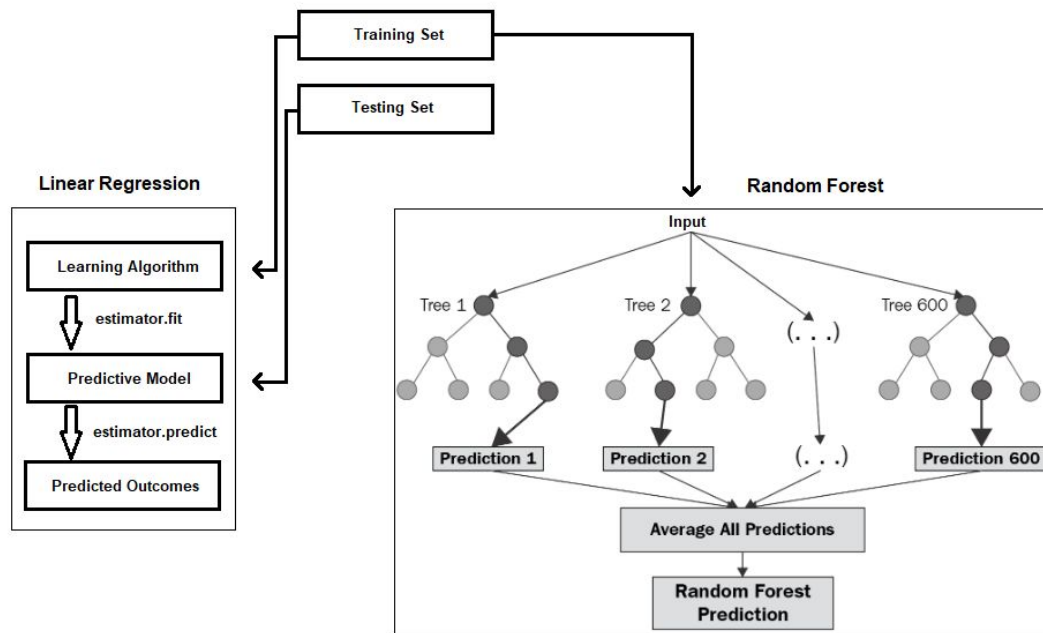
- 1- parameters that are learned through the training.
- 2- hyperparameters that we pass to the machine learning model.

Once the model is identified, the next step is to tune its hyperparameters to obtain the best predictive power possible.

We evaluated two models to identify the right model with the least RMSE value.

Model Building

Figure below shows an overview of our model.

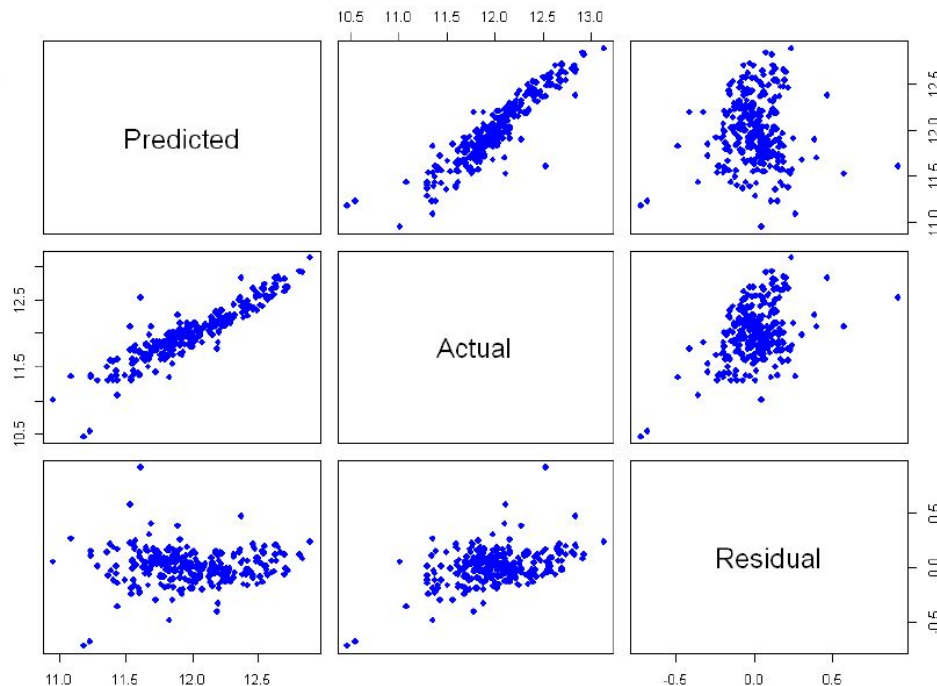


Linear Regression

	ME	RMSE	MAE	MPE	MAPE
Test set	0.007261273	0.1538444	0.1075528	0.04271564	0.9029266

Linear Regression model means estimating the values of the coefficients used in the representation with the data that we have available.

RMSE values < 0.1 are very satisfactory.

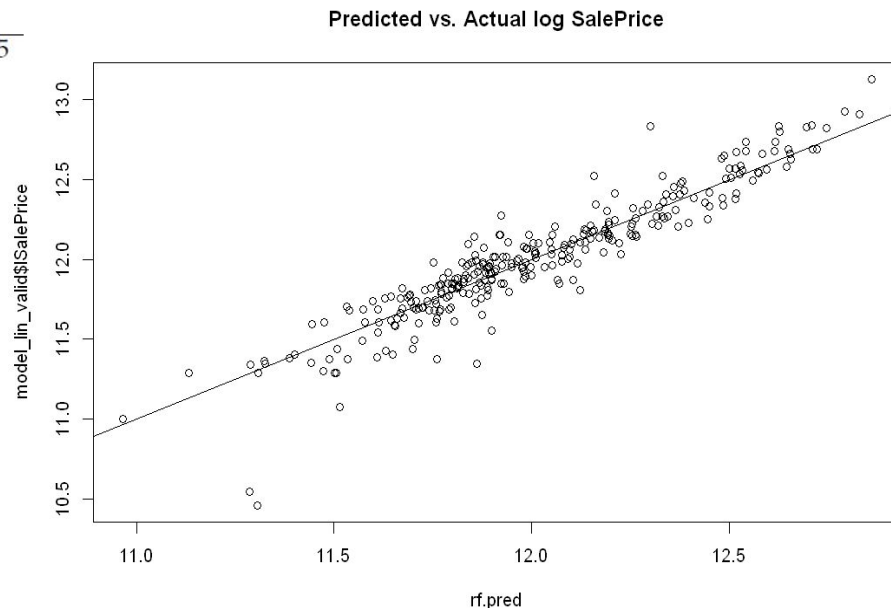


Random Forest

	ME	RMSE	MAE	MPE	MAPE
Test set	-0.0004281985	0.1384497	0.09486207	-0.02296528	0.7980985

Random Forest is more flexible and easier to apply. It is one of the most popular algorithms because of its simplicity.

Random Forest showed lower RMSE value.



Conclusion

In this project, we predicted Sale Price from our dataset using two different algorithms, **Linear Regression** and **Random Forest** in R programming language.

After comparing the performance metrics of both the algorithms we conclude that Random Forest performed better and gave more accurate results.