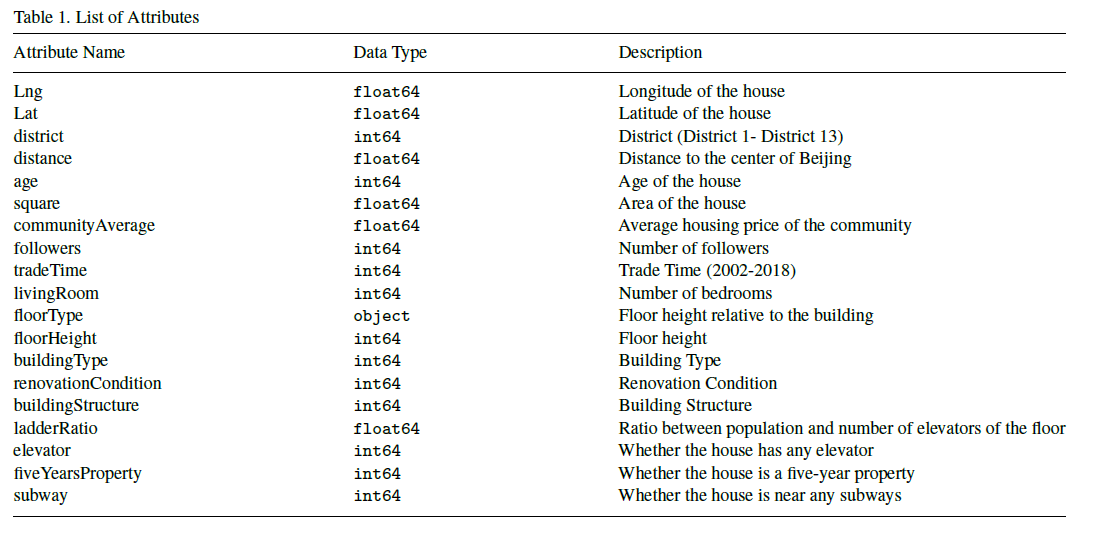
Class Activity 1

Read through the research article that has been provided to you which discusses the results of a specific supervised machine learning regression study, and answer the questions that follow:

1. What was the target variable, and which attributes were considered?

Housing price per square metre in Beijing



19 attributes present after data pre-processing (which were transformed in to 58 features after one-hot encoding)

1. What % training/test split was used? 80:20
2. What pre-processing was done on the data prior to training the model(s)?

* Variables with more than 50% missing data were removed from the dataset.
* The variable “Day on market” was removed because of 157, 977 missing data.
* Any observation which had missing values were also removed from the dataset
* Removal of outliers
* Other feature engineering processes which were performed included:
* Removal of attributes indicating the number of kitchens, bathrooms, and drawing rooms due to their ambiguity.
* Setting the number of living rooms (bedrooms were mistranslated to living rooms) in a range from 1 to 4.
* Adding attribute “distance” indicating the distance of the house from the center of Beijing.
* Replacing attribute “constructionTime” with attribute “age” by deducting the year that the house was constructed from the current year (2019).
* Setting minimum values for attributes “price” and “area”.
* Splitting the attribute “floor” into attributes “floorType” and “floorHeight”.
* Numerical values were standardized, while categorical values were one-hot-encoded. After being processed, the dataset included 58 features.

1. Which machine learning technique(s) were used?

Random Forest, Extreme Gradient Boosting, Light Gradient Boosting, Hybrid Regression, Stacked Generalization Regression

1. Which performance measures were used to evaluate the machine learning models?

Root Mean Squared Logarithmic Error (RMSLE)

1. Which model/s would you regard as having higher bias? Light Gradient Boosting (Stacked Generalization also has higher bias than some of the other models but has possibly struck the right balance between bias and variance)
2. Which model/s would you regard as being underfit? Light Gradient Boosting (Stacked Generalization has the lowest error on the test set, even though it has the second highest error on the training set)
3. Which model/s would you regard as being overfit? Random Forest (Hybrid Regression has second lowest error on the test set, so although it has higher variance, and it also seems to have struck a good balance between bias and variance)
4. Which model would you say has the best generalizability? Stacked Generalization Regression
5. Which ONE model would you select as the best performing? Stacked Generalization Regression (has slightly higher bias, but lowest variance), followed by Hybrid Regression (has slightly higher variance, but fairly low bias)

