Data Newbies

DIHAN JANNATAN MUTAALIM

LEE YI FENG GLEN

NOTARIA MANAV BAIJU

RIASA FADHILLA MARTONO

Earthquake Database : Which factors affect the severity of the damage of a building after an earthquake?

Defined Problem

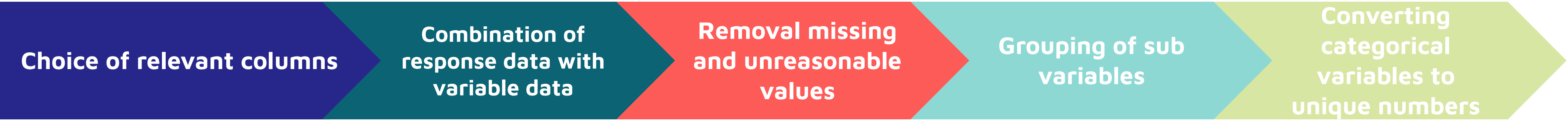
Which factors affect the severity of the damage of a building after an earthquake?

Predicting the severity of damage of a building after an earthquake will be a very useful tool to the civil engineering and architecture industry

Variables Considered

1. Age
2. Foundation type
3. Height
4. Count Floors
5. Land surface condition
6. Ground floor type
7. Roof type

Data exploration and Data preparation



Choice of relevant columns

Columns were chosen on the basis of the likelihood of the variables contained in the columns to affect damage grade

Combination of response data with variable data

Damage grade is combined with the other variables as it was not in the same file using .concat()

Removal missing and unreasonable values

Cleared out missing and unreasonable values using .dropna() from the dataset to ensure that they do not interfere with the classification of injury severity and skew our observations

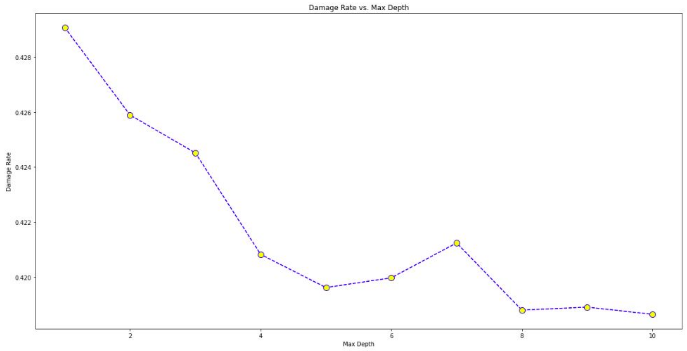
Grouping of sub variables

Grouped the different sub variables in ‘foundation_type’ to make the data easier to understand

Converting categorical variables to unique numbers

Converted the categorical variables to unique numeric values representing the different classes of the categorical variable using pd.factorize()

Decision Tree



●Optimal depth of the Tree was found using damage_grade and GridSearchCV

●Optimal depth is used to tune the Decision Tree

●The full Decision Tree was visualised using graphviz

●The Variable Importance was determined using Feature importance for decision tree models

●Most Influential Variable: *foundation_type_fact*

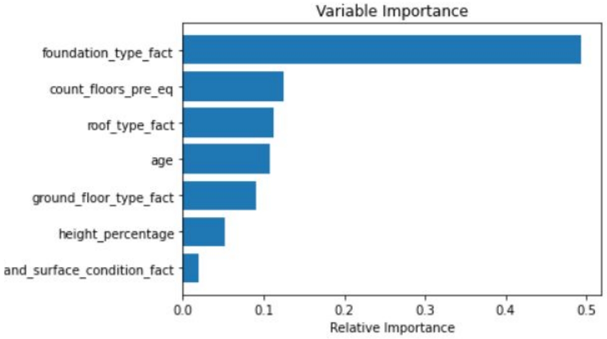
●Least Influential Variable: *land_surface_condition_fact*

Advantages:

- Trees are simple to understand, interpret and visualise.
- It is possible to validate a model using statistical tests.

Disadvantages:

- Decision tree learners create biased trees if some classes dominate.



KNN

A model that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

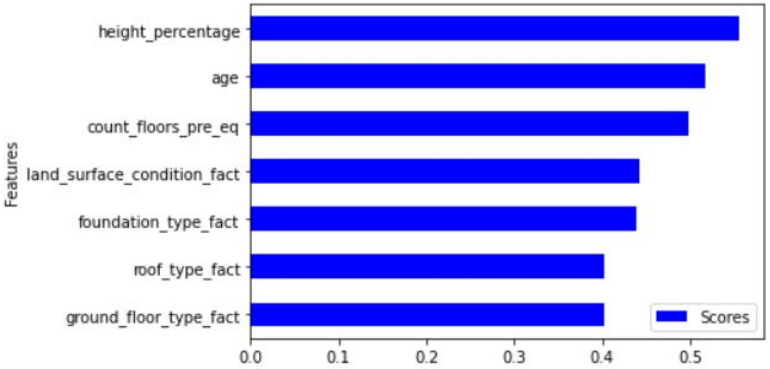
First hyperparameters tuning is to find optimal value of K and leaf size based on error rate, then we store into search space

Second hyperparameters tuning is to find optimal value of other parameters using Grid Search Cross Validation

Cross Validation Score is used as a metric for finding variable importance in the model

Adv: Easy implementation

Disadv: Does not work well with high dimensionality, does not work well with large dataset.



Random Forest

Consists of a large number of individual decision trees that operates as an ensemble

Outperform other individual models as it is a collection of a large number of relatively uncorrelated trees operating together

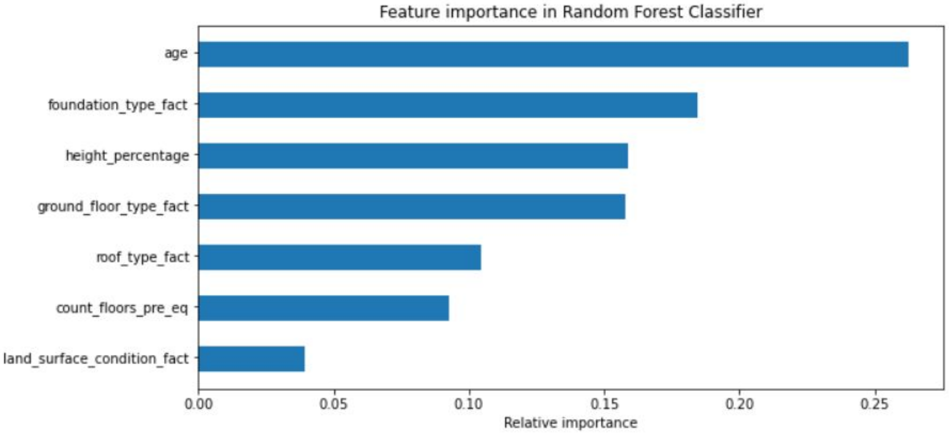
Adv: Reduces overfitting problem in decision trees and also reduces the variance and therefore improves the accuracy

Disadv: More complex and requires longer time

- GridSearch took a very long time with the amount of data

Confusion Matrix is used to show the performance of the model

Feature Importance is obtained and based on our model **age** is the most important parameter to predict a damage grade of a given building



Performance of models

Model	To predict	F - Score	Classification Accuracy
Decision Tree	Damage Grade 1	39%	58.06%
	Damage Grade 2	72%	
	Damage Grade 3	13%	
KNN	Damage Grade 1	37%	54.69%
	Damage Grade 2	66%	
	Damage Grade 3	33%	
Random Forest	Damage Grade 1	39%	57.56%
	Damage Grade 2	70%	
	Damage Grade 3	27%	

Conclusion

Foundation type is the most influential factor in determining the severity of the damage of a building after an earthquake

Age and Height also plays a major role based on the classification models

Land surface condition plays little to no role

Contribution of members

Data Cleaning : Lee Yi Feng

Decision Tree : Notaria Manav Baiju

KNN : Dihan Jannatan Mutaalim

Random Forest : Riasa Fadhilla Martono

Thank You!

QnA

- Propose Problem Statement “Which factors affect the severity of the damage of a building after an earthquake”
- Can we choose the variables, if yes how many?
  - Age
  - Foundation
  - Height and Count Floors
  - Land surface condition
  - Ground floor type
  - Roof type
- Meaning of abbreviation in dataset
- Can we change the dataset

Data Cleaning: Glen (Sunday)
Decision Tree: Manav
Random forest:Riasa
KNN:Riasa