Data Sloths Richter Earthquake Damage

By Achnaf | Carlo | Yuyang | Zhu Siyao

PROJECT FLOW

01

RESEARCH QUESTIONS

Problem Formulation to solve

02

EXPLORATORY DATA ANALYSIS

Data Preparation, Visualization, and Preprocessing

03

MACHINE LEARNING MODEL

7 ML Model Used

04

MODEL RESULT ANALYSIS

Performance Metrics Evaluation

Chosen Dataset

Nepal 2015 Earthquake

Retrieved from DRIVENDATA

Containing information on buildings' conditions, damage grade, and socio-economic statistics

Research Problem

Which proposed hotel buildings are safe enough to withstand earthquakes?

As consultants for hotel construction developers who are planning to reconstruct hotels after the earthquake, we predict damage grade based on building information and select the best plans to effectively help them make decisions.

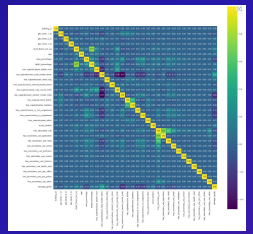
Problem Formulation

Whether buildings used as hotels are more prone to collapse than buildings used for other purposes?

Which proposed hotel buildings are safe enough to withstand earthquakes?

Hotel New World Collapsed in 1986

Correlation Matrix



Data Preparation

Check Null and Duplicate Data

Merge Feature and Label on Training Data

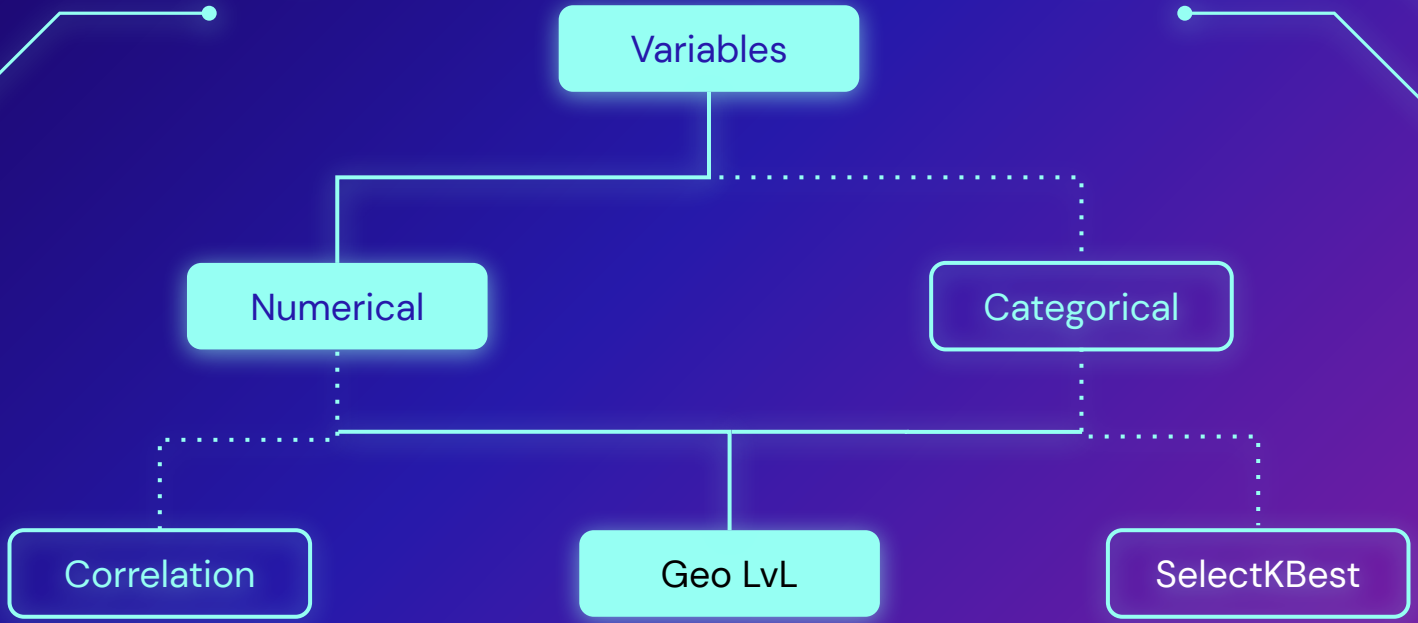
Encode Object Data to Categorical Data

Merge Numeric and Categorical Feature

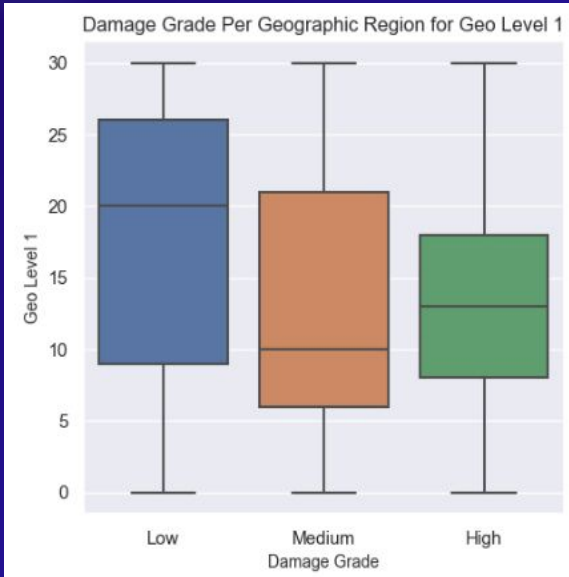
Clear Outliers

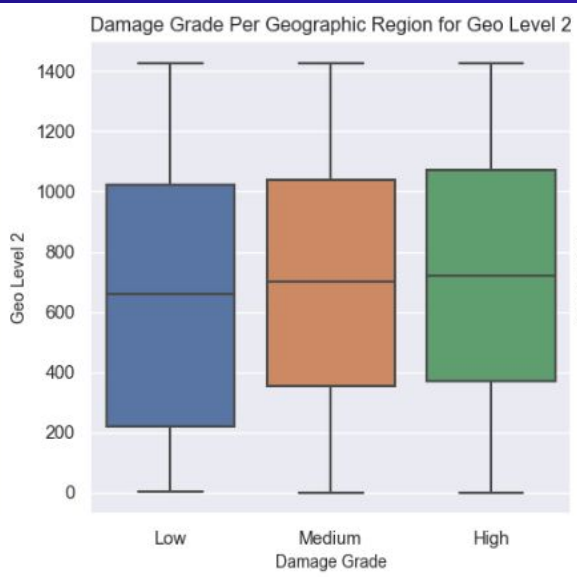
Data Scaling

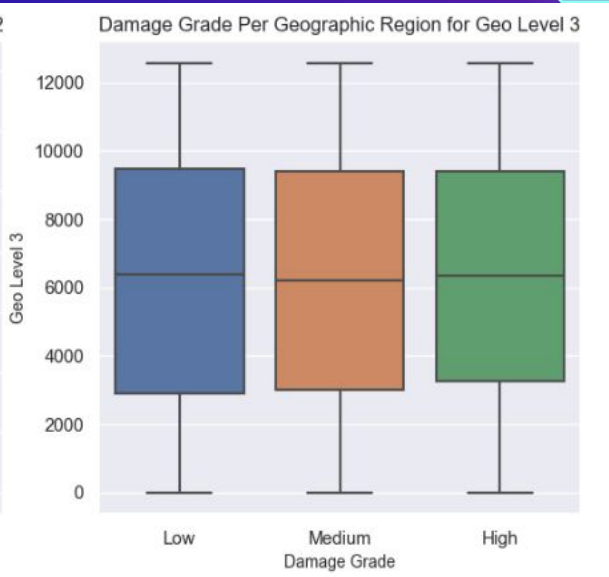
Data Visualization



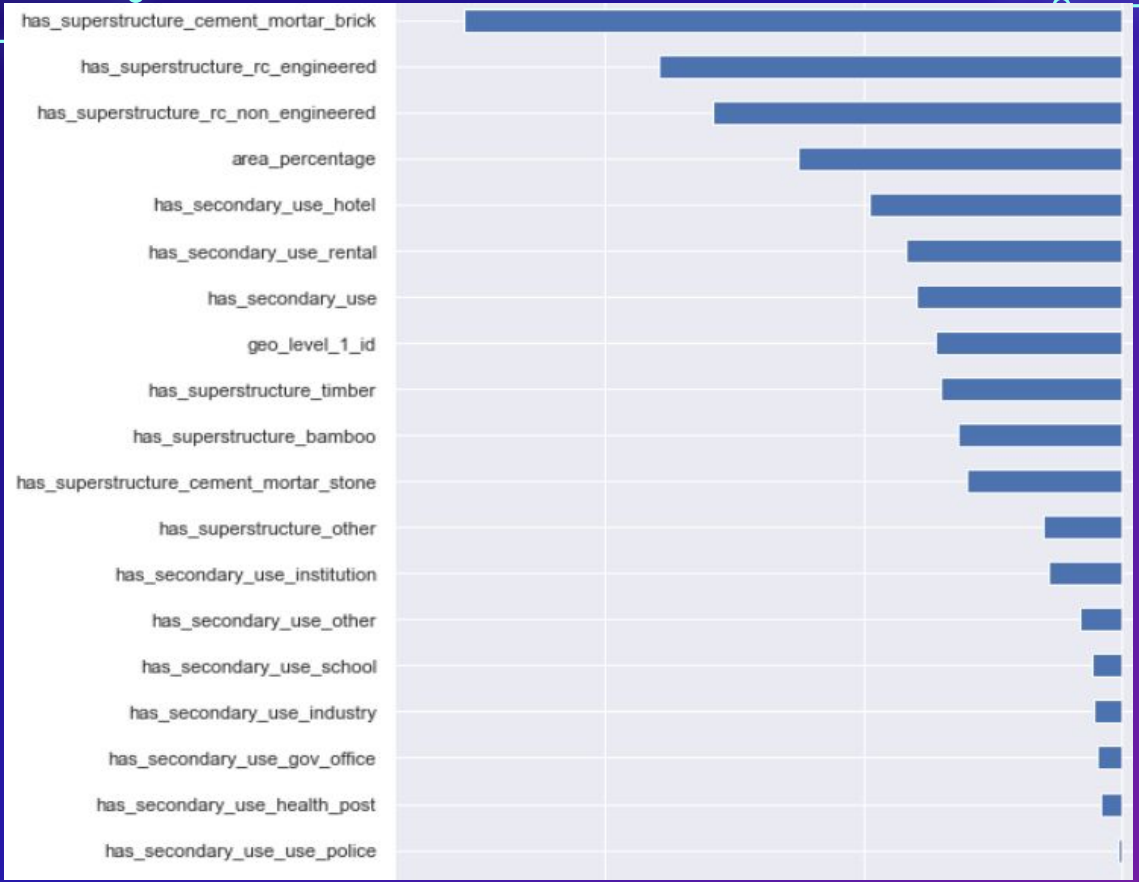
Geo LvL VS Damage Grade



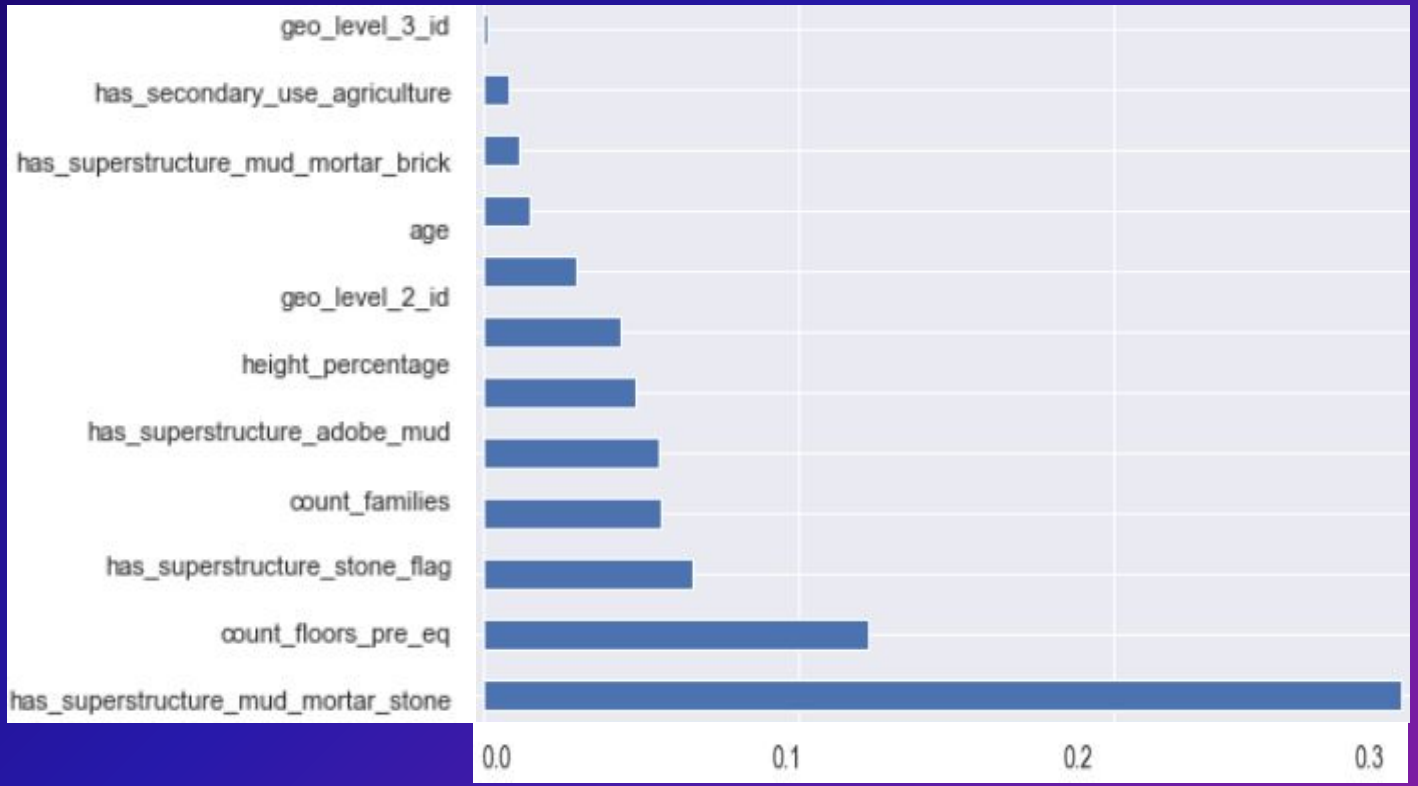




Negative Correlation



Positive Correlation



SelectKBest (TOP 10)

DISTINCT SUB-VARIABLES

Gender		Gender_B	Gender_G
B		1	0
G	→	0	1
B		1	0
G		0	1
G		0	1

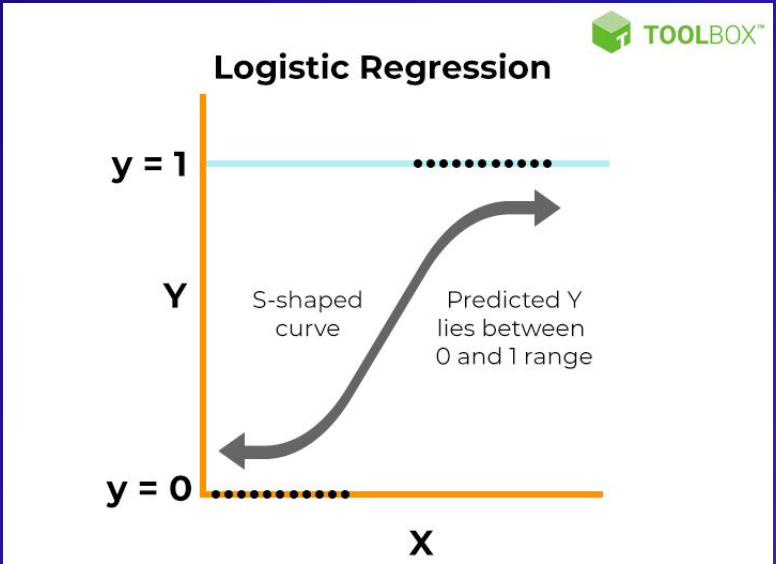
Feature	Score
ground_floor_type_v	32465.421066
roof_type_x	28048.595012
foundation_type_i	27929.304672
other_floor_type_s	18549.408221
foundation_type_w	8315.794578
other_floor_type_j	7422.919931
foundation_type_r	6391.952318
foundation_type_u	5494.248443
other_floor_type_q	5108.461280
ground_floor_type_f	3684.892346

Machine learning models

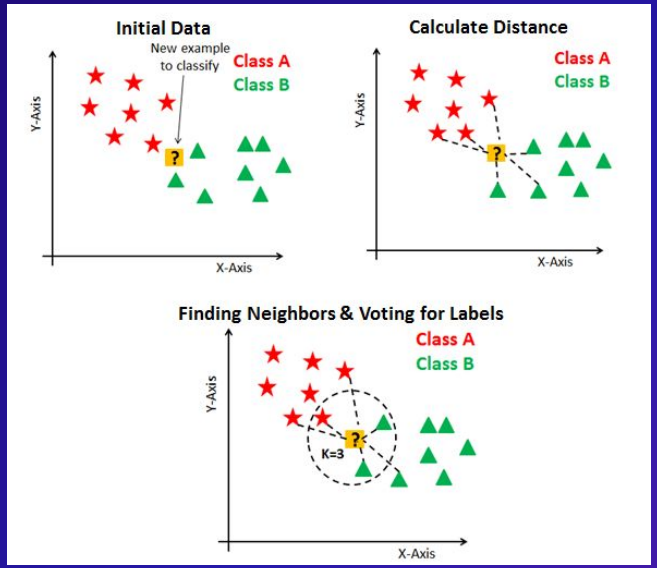
- To find out the most fit model, then use it for test_value prediction

- Decision Tree
- Logistic Regression
- KNN Model
- Linear Discriminant Analysis
- Random Forest
- Naive Bayes
- Extreme Gradient Boosting (XGB)

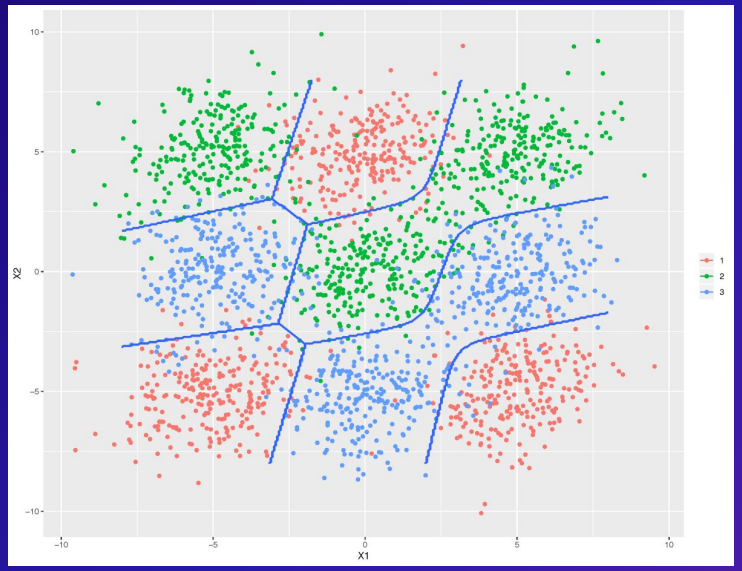
Logistic Regression



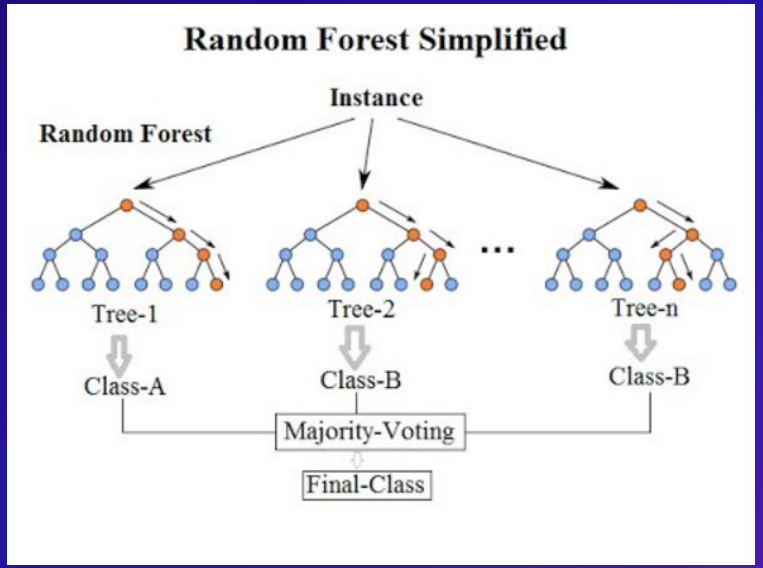
K-Nearest Neighbors



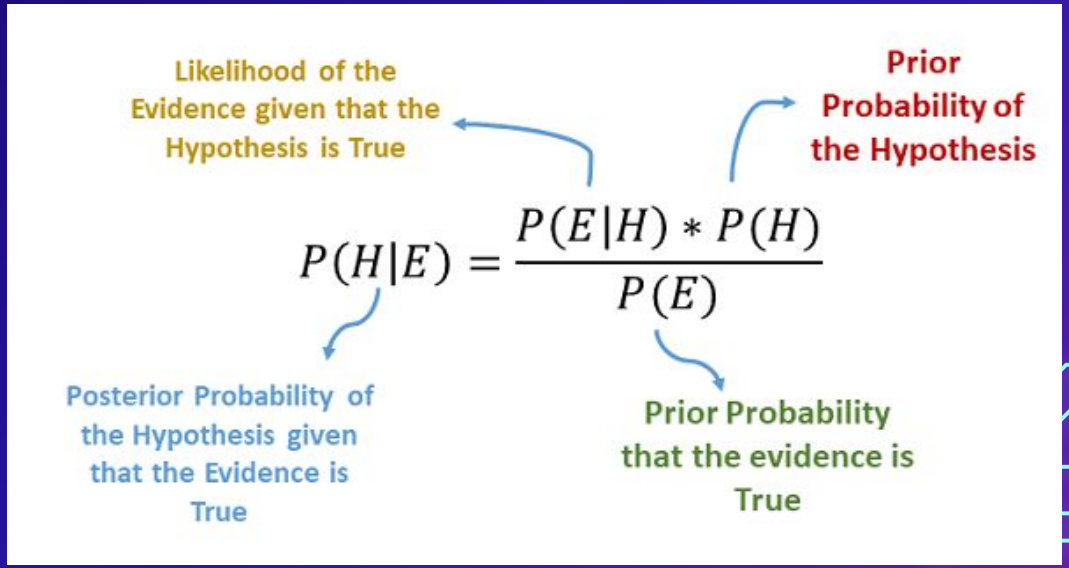
Linear Discriminant Analysis



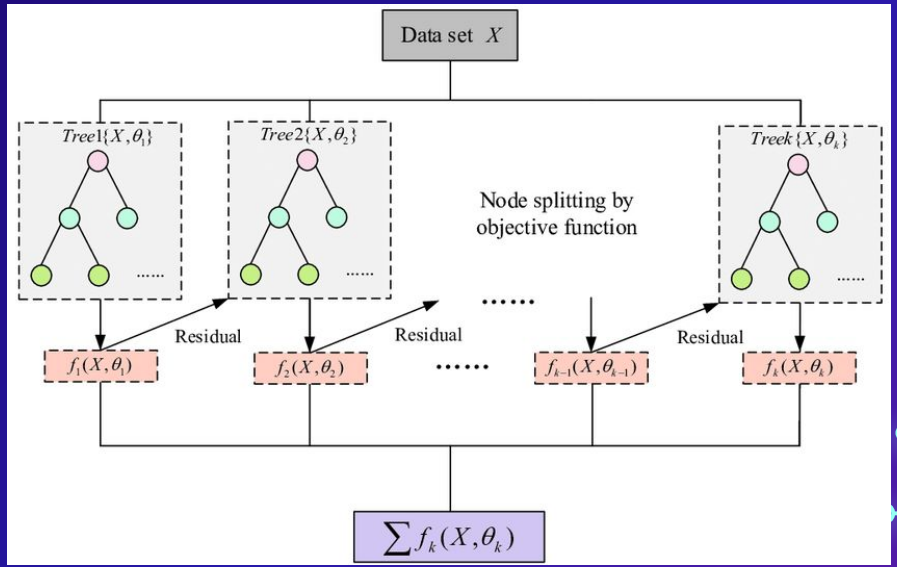
Random Forest



Naive Bayes


Likelihood of the Evidence given that the Hypothesis is True
Prior Probability of the Hypothesis
$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$
Posterior Probability of the Hypothesis given that the Evidence is True
Prior Probability that the evidence is True

XGBoost


Data set X
$Tree1\{X,\theta_1\}$
$Tree2\{X,\theta_2\}$
$Treek\{X,\theta_k\}$
Node splitting by objective function
Residual
Residual
Residual
$f_1(X,\theta_1)$
$f_2(X,\theta_2)$
$f_{k-1}(X,\theta_{k-1})$
$f_k(X,\theta_k)$
$\sum f_k(X,\theta_k)$

MODEL ANALYSIS

Performance Metrics

Our Performance Metrics

F1-Score (Micro)

Calculated using precision and recall of the test.

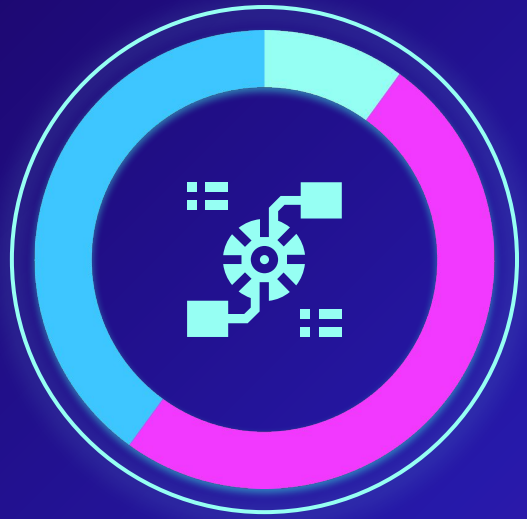
Cohen's Kappa Score

compares an observed accuracy with an expected accuracy (Random chance)

Accuracy comparison

	Algorithm	CV F1 Score	F1 Score (Train)	Cohen Kappa (Train)	F1 Score (Test)	Cohen Kappa (Test)
0	Logistic Regression	0.592223	0.592663	0.975613	0.591242	0.461877
1	Decision Tree	0.654438	0.986503	0.975626	0.653697	0.379044
2	K-Nearest Neighbors	0.702164	0.790084	0.563495	0.705407	0.369198
3	Linear Discriminant Analysis	0.587194	0.587501	0.173455	0.585824	0.171779
4	Naive Bayes	0.431251	0.431645	-0.001376	0.430308	0.000344
5	Extreme Gradient Boosting	0.713369	0.744108	0.975613	0.726804	0.461877
6	Random Forest	0.726390	0.986488	0.510700	0.713988	0.478843

BEST MODEL PERFORMANCE



1

EXTREME GRADIENT BOOSTING

F1-Score: 0.7268
Cohen's Kappa: 0.4619

2

RANDOM FOREST

F1-Score: 0.7140
Cohen's Kappa: 0.4789

3

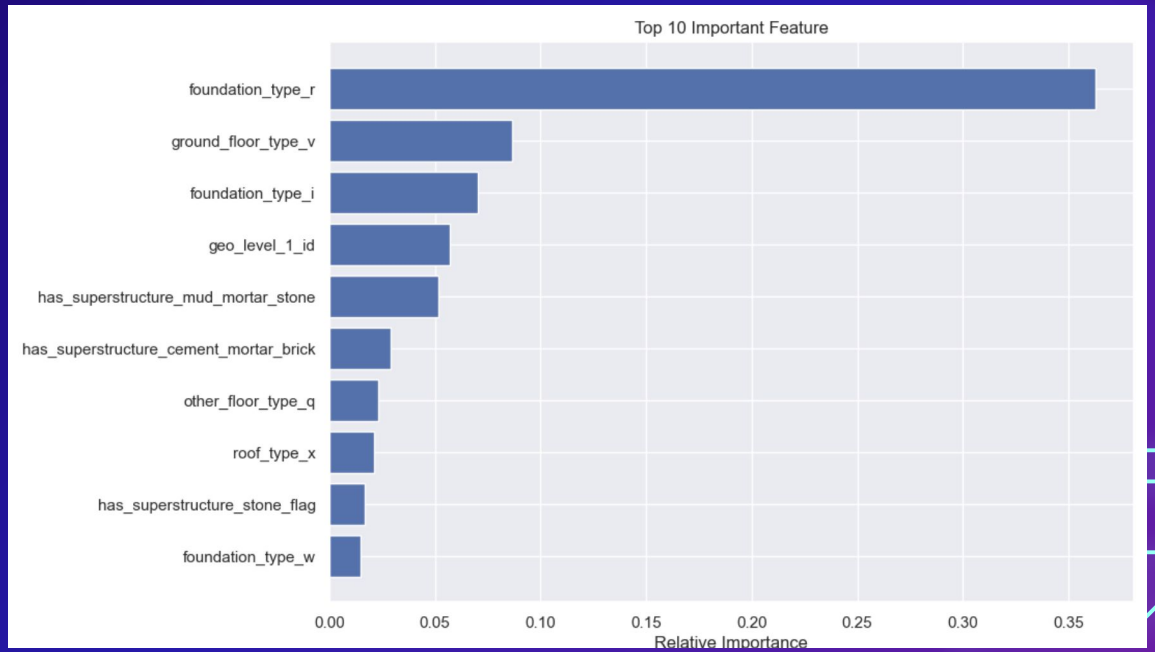
K-NEAREST NEIGHBORS

F1-Score: 0.7054
Cohen's Kappa: 0.3692

Feature Importances

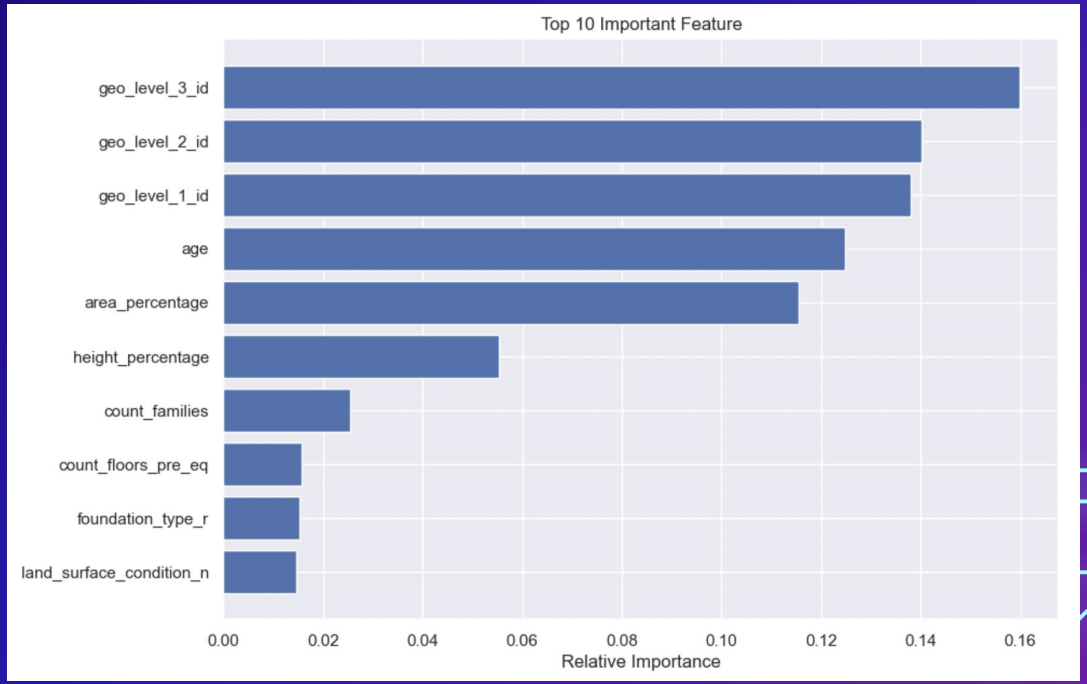
Feature Importances

Extreme Gradient Boosting



Feature Importances

Random Forest

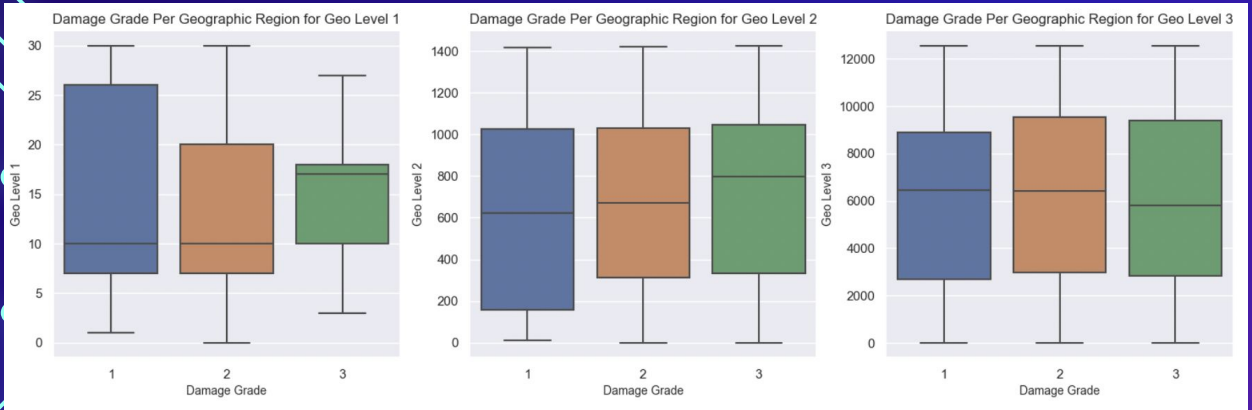


Feature Importances

K-Nearest Neighbors

Deploying Model to Test Labels

Assume that these hotels do not exist!



building_id	Geo Level 1	Geo Level 2	Geo Level 3	Age	Potential Damage Grade
379498	17.0	834.0	11920.0	0.0	2
897228	17.0	658.0	10741.0	0.0	2
931531	18.0	1317.0	11887.0	0.0	1
736534	7.0	545.0	9356.0	0.0	1
988469	8.0	463.0	6973.0	0.0	1

Deployment Result

Safe Enough

building_id	Geo Level 1	Geo Level 2	Geo Level 3	Age	Potential Damage Grade
931531	18.0	1317.0	11887.0	0.0	1
736534	7.0	545.0	9356.0	0.0	1
988469	8.0	463.0	6973.0	0.0	1
383689	26.0	1050.0	8829.0	0.0	1
265814	26.0	1401.0	2557.0	0.0	1

Data Dimension: 210 rows, 55 columns

Not Safe Enough

building_id	Geo Level 1	Geo Level 2	Geo Level 3	Age	Potential Damage Grade
711500	7.0	1265.0	4059.0	0.0	3
251354	8.0	463.0	8236.0	0.0	3
480465	17.0	1313.0	1349.0	0.0	3
971408	7.0	52.0	1819.0	0.0	3
336685	17.0	566.0	8505.0	0.0	3

Data Dimension: 28 rows, 55 columns

Conclusion

- Geographic Level (3,2,1) are the utmost important features in predicting damage grade
- Successful in helping hotel developers to reconstruct after the earthquake
- Not guaranteedly accurate, there is limitation due to the lack of knowledge and other missing features or consideration to take into account

Contribution of Members

Achnaf Habibullah : Exploratory Data Analaysis, Extreme Gradient Boosting, K-Nearest Neighbors, Deployment

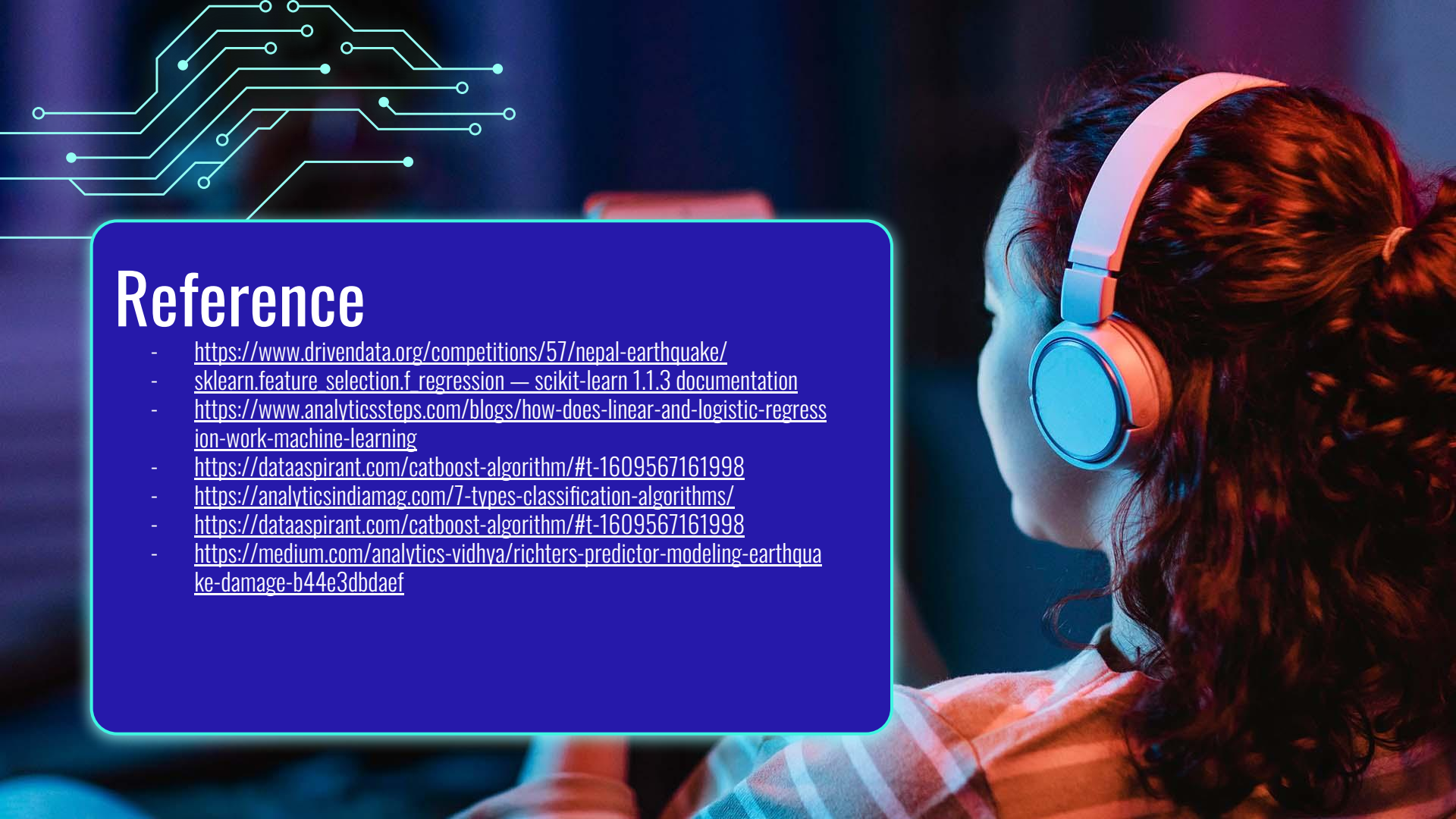
Carlo Lee : Logistic Regression, Decision Tree

Jin Yuyang : Random Forest, Naive Bayes

Zhu Siyao : K-Nearest Neighbors, Linear Discriminant Analysis

THANKS

Do you have any questions?



Reference

- https://www.drivendata.org/competitions/57/nepal-earthquake/
- sklearn.feature_selection.f_regression — scikit-learn 1.1.3 documentation
- https://www.analyticssteps.com/blogs/how-does-linear-and-logistic-regression-work-machine-learning
- https://dataaspirant.com/catboost-algorithm/#t-1609567161998
- https://analyticsindiamag.com/7-types-classification-algorithms/
- https://dataaspirant.com/catboost-algorithm/#t-1609567161998
- https://medium.com/analytics-vidhya/richters-predictor-modeling-earthquake-damage-b44e3dbdaef