Predicting Earthquake Damage with Light Gradient-Boosted Machines

Wenjie Duan
Hao Xu
Collin Prather

Group : what, Wen(jie), where, Hao, Collin

Dataset info

Number of variables	39
Number of observations	260601
Missing cells	0 (0.0%)
Duplicate rows	0 (0.0%)

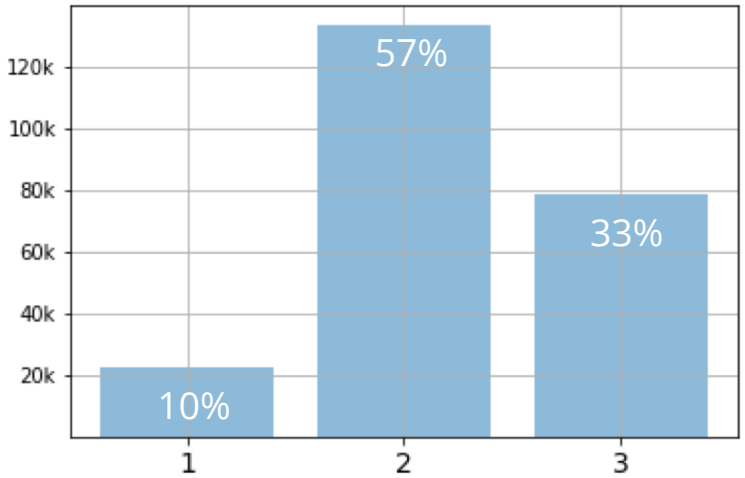
Target variable is `damage_grade`:

- `1`: low damage
- `2`: moderate damage
- `3`: complete destruction

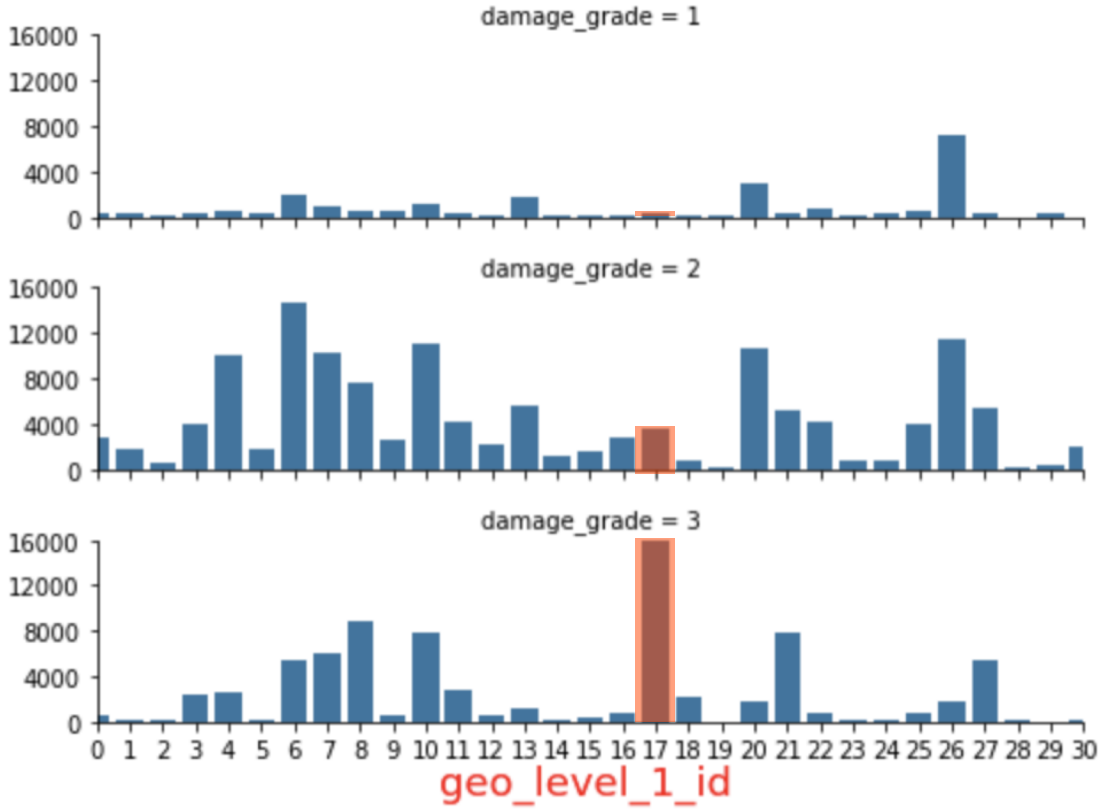
Type of features:

- Where is the building?
  - `Geo_level_1_id`
  - `Geo_level_2_id`
- How is the building?
  - `Age`
  - `Height`
  - `Surface`
  - `Foundation`
  - ...

Distribution of `damage_grade`



Feature Engineering



Models & Evaluation Metric

Model	SVC	MLP Classifier (neural network)	Random Forest	Gradient-Boosted Machines
Pros	Outputs the separating hyperplane with automatic correction for unbalanced classes	Able to learn very flexible decision boundary	Avoids overfitting, counteracts class imbalances	Benefits of both gradient-based and tree-based learning
Cons	Poor performance with high dimensional data	Computationally expensive, many hyperparameters	Many hyperparameters	More sensitive to overfitting
F_{micro}	0.33	0.69	0.73	0.74

$$F_{micro} = \frac{2 \cdot P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}}$$

$$P_{micro} = \frac{\sum_{k=1}^{3} TP_k}{\sum_{k=1}^{3} (TP_k + FP_k)}, \quad R_{micro} = \frac{\sum_{k=1}^{3} TP_k}{\sum_{k=1}^{3} (TP_k + FN_k)}$$

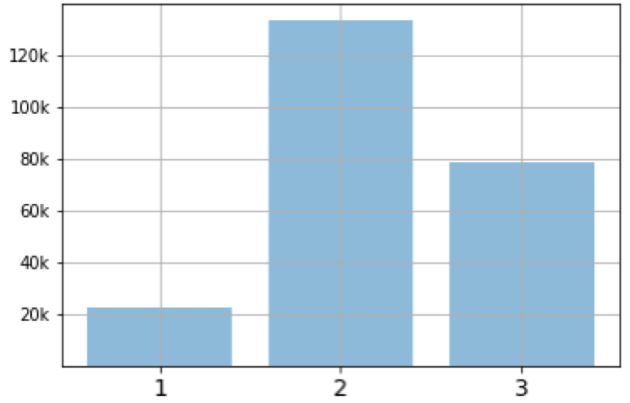
Deal with Imbalanced Dataset

1. SMOTE Oversample -- Find Max probability

Probability of Class-1	Class-2	Class-3	
[0.279,	0.308,	0.411]	-> 3
[0.301,	0.298,	0.399]	-> 3
[0.287,	0.372,	0.340]	-> 2

Start threshold tuning
from the distribution of data

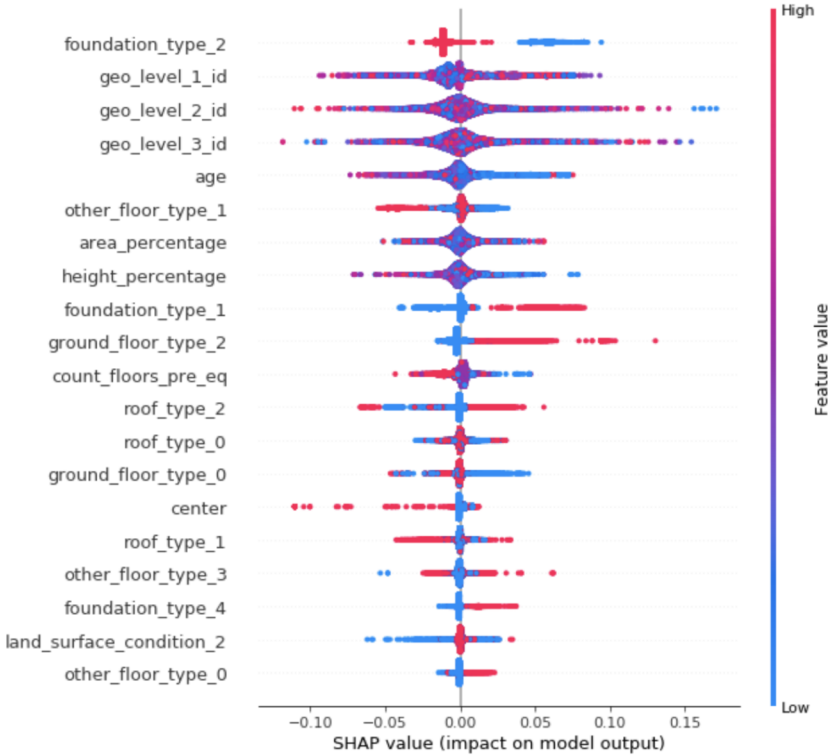
Distribution of `damage_grade`



1. Tuning Probability Threshold -- Based on rules

Probability of Class-1	Class-2	Class-3	
[0.084,	0.548,	0.366]	-> 3
[0.186,	0.518,	0.295]	-> 1
[0.098,	0.614,	0.287]	-> 2

LGBM SHAP Feature Importances



Conclusion

- We ranked **top 10%** in this Drivendata Competition.
- We solved this **multi-classification** problem with **4** ML algorithms, and **GBM** really stands out.
- Our **feature engineering**(find the “accident **center**”) method is very useful in exploiting the geographic id (location) info.