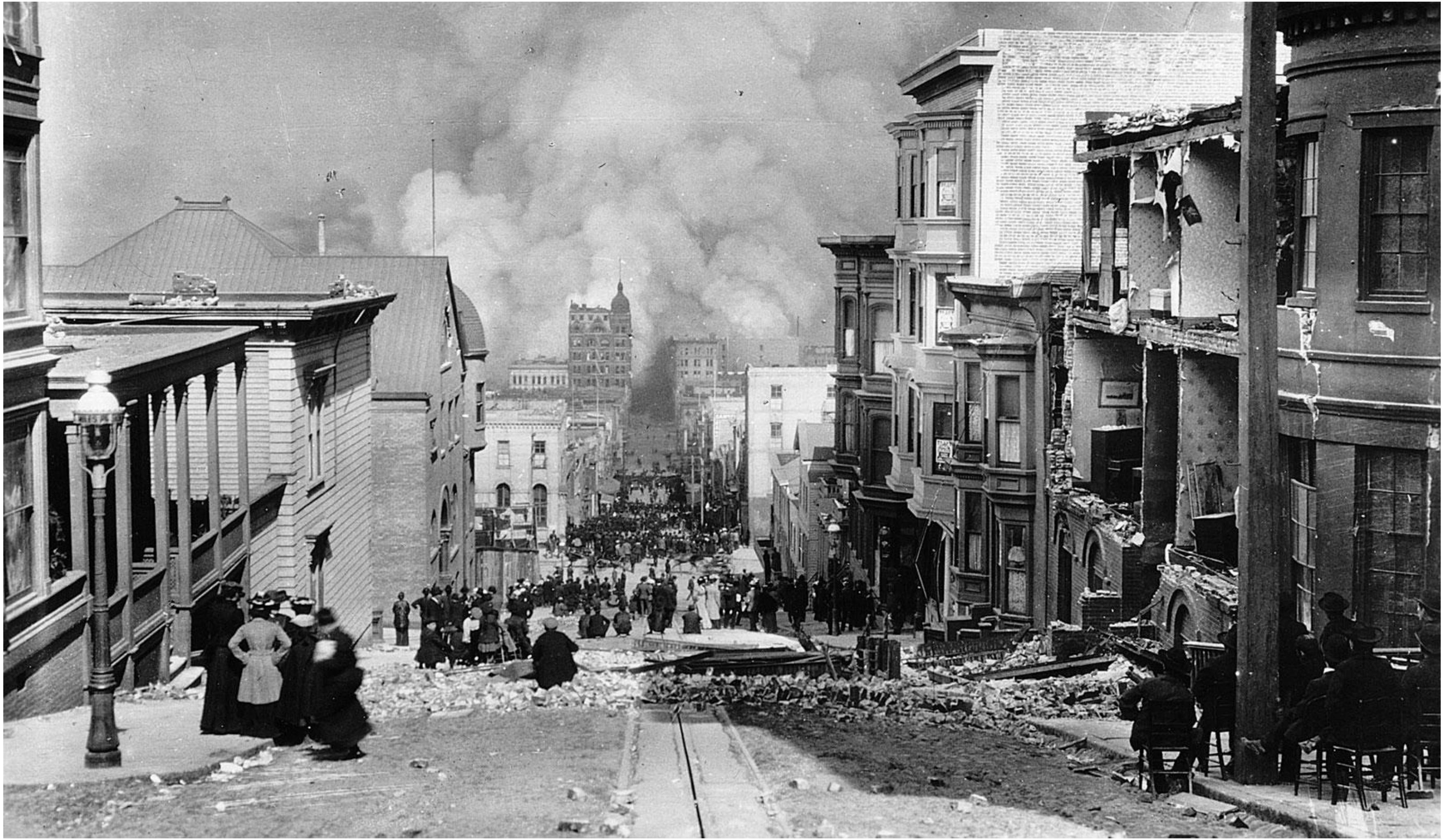
(measuring) Ground-Breaking ML Project

Richter Group:

Nithish, Alan, and Rushil



Problem and outline

- Modeling Earthquake Damage

  Can we classifying building damage during an Earthquake?

- Why this is important:

  The Government needs to disburse the **right** amount of money **quickly** to the casualties.

- Data from Earthquake Damage Modelling Competition hosted on Drivendata.org

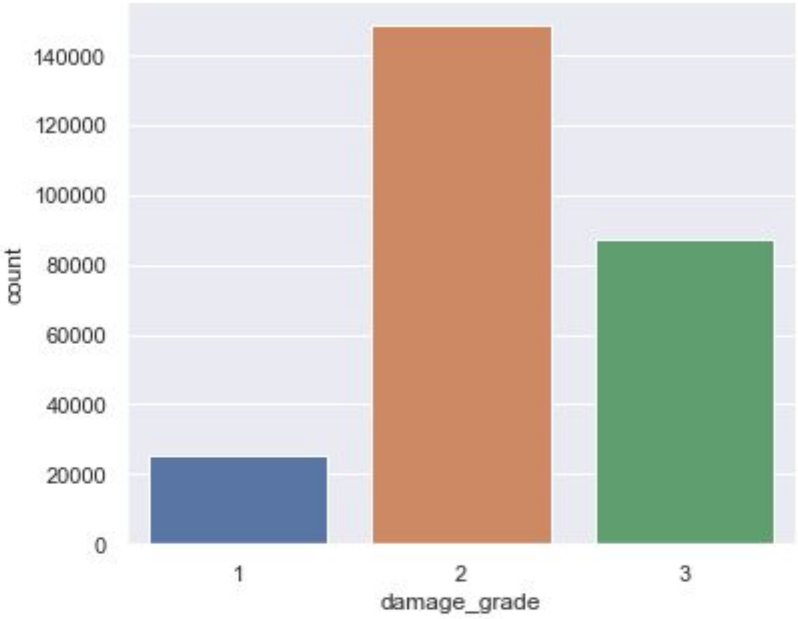
- Dataset size = 280 K Rows and 39 Features.

Target and Metrics

Competition on drivendata.org

- 1 represents low damage
- 2 represents a medium amount of damage
- 3 represents almost complete destruction

Trying to maximize F1 score with a micro average



Class Distribution

Features

Numeric

- Geo_level_1_id
- Geo_level_2_id
- geo_level_3_id
- Count_floors_pre_eq
- Age
- Area_percentage
- Height_percentage
- count_families

Binary

- has_superstructure_adobe_mud
- has_superstructure_mud_mortar_stone
- has_superstructure_stone_flag
- has_superstructure_cement_mortar_stone
- has_superstructure_cement_mortar_brick
- has_superstructure_bamboo
- has_superstructure_rc_non_engineered
- has_superstructure_rc_engineered

- has_secondary_use
- has_secondary_use_agriculture
- has_secondary_use_school
- has_secondary_use_use_police
- has_secondary_use_other

Categorical

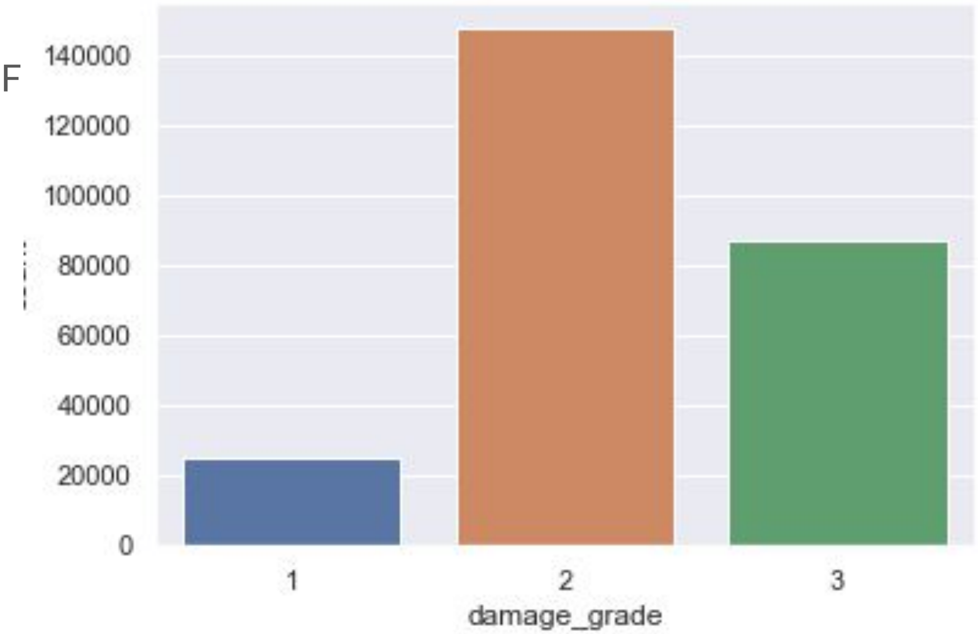
- Land_surface_condition
- foundation_type
- roof_type
- ground_floor_type
- other_floor_type
- position
- plan_configuration
- legal_ownership_status

Competition Current

Competition leaderboard said baseline RF model is: 0.5815

Using only mode of 2: 0.5689

Current Leader is 0.7544



Pre Process Data

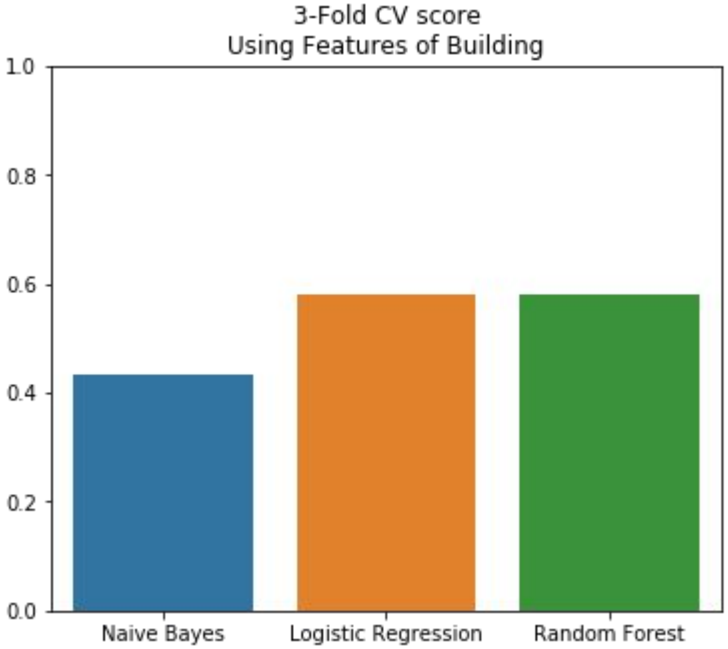
Got data into tidy form

Label encoded categoricals

Correlation matrices

Feature engineering

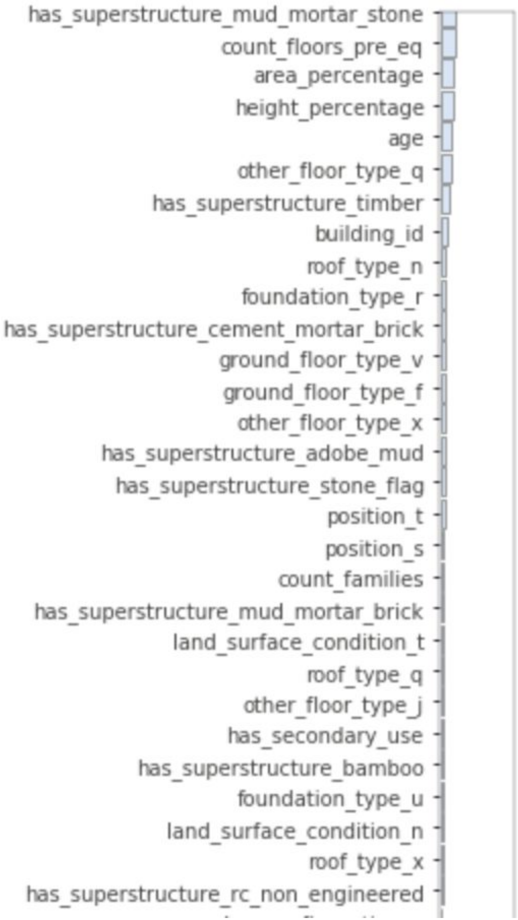
Baseline Model



Initial RF

Fit with all non-id features

Got an F1 micro score of 0.59



Improving Random Forest Model.

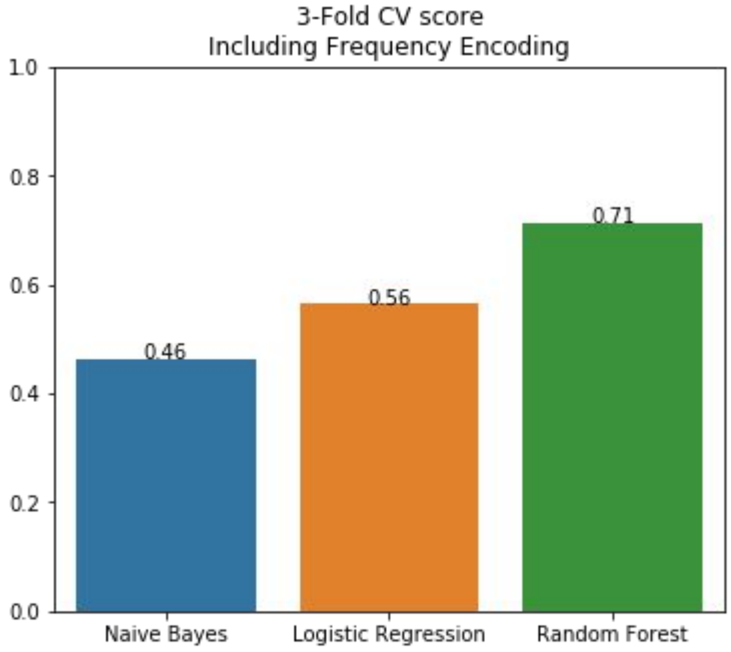
- Earthquake has an epicenter and shock waves ripple across the earth’s surface from this point.
- Can we use geographical ID features to proxy distance from epicenter?
- Can we utilise the pattern between the target classes ?

Feature engineering

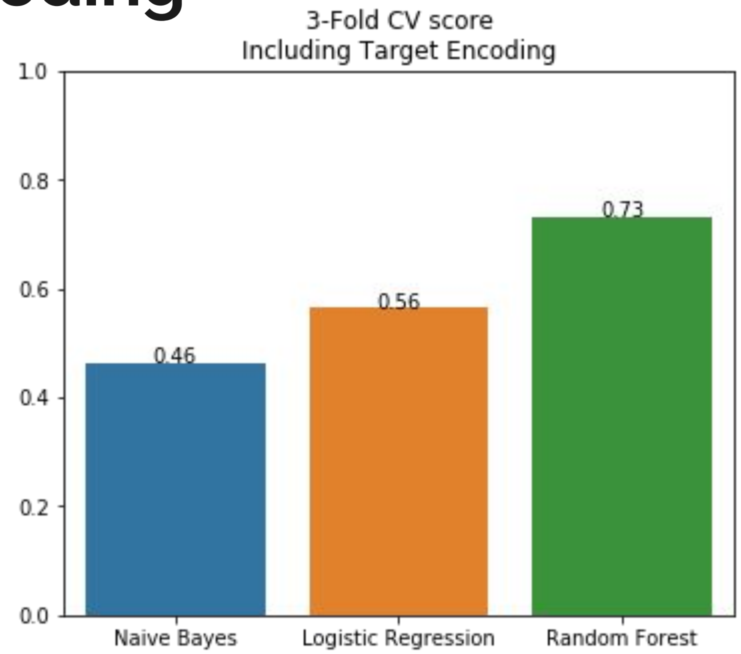
Focused on Geographic Region ID’s

- Target and Frequency Encoding
- Had to account for test and train geo id discrepancies, especially in geo id 2 and 3

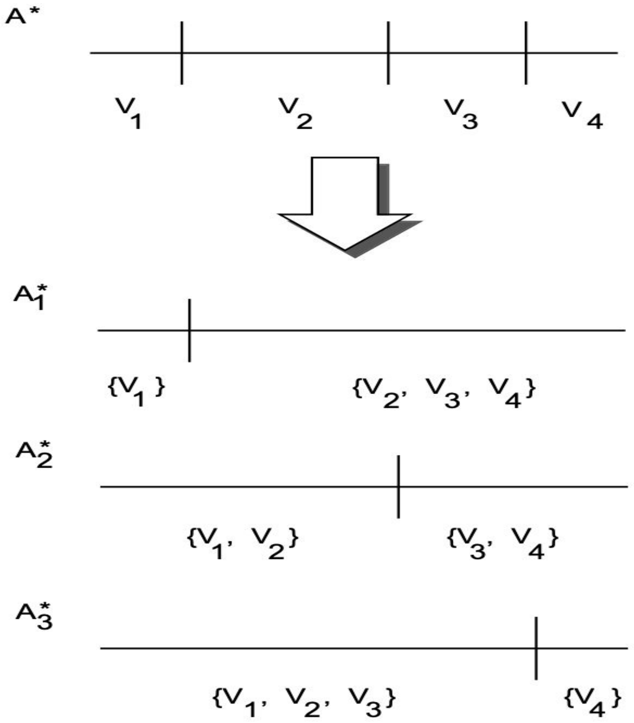
Frequency Encoding helps!



Target Encoding

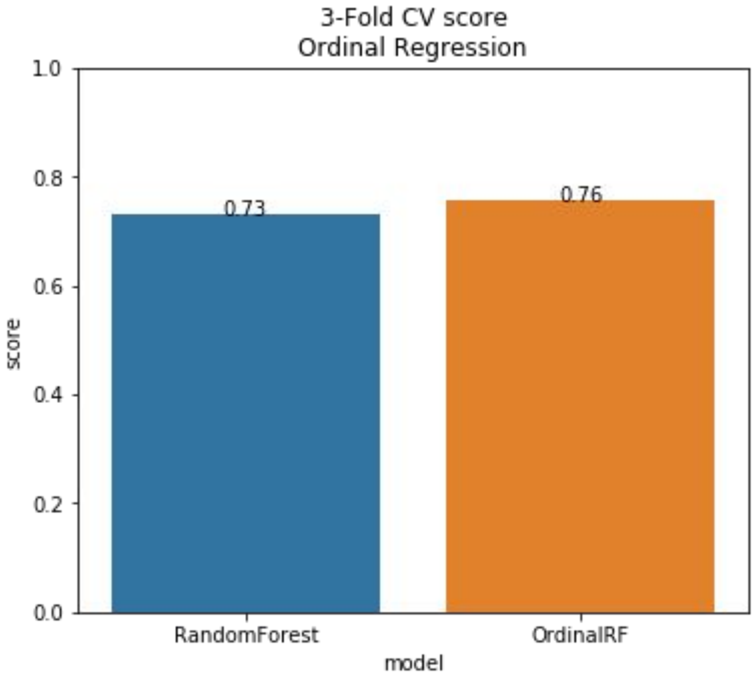


Ordinal Regression Intuition

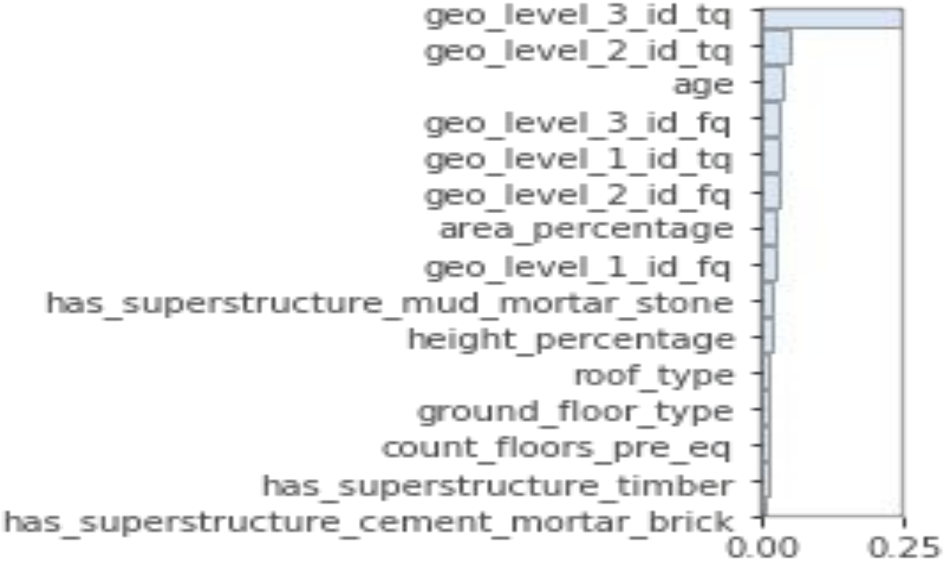


https://www.cs.waikato.ac.nz/~eibe/pubs/ordinal_tech_report.pdf

Ordinal Regression Improvement



Feature Importances (top 15)



Conclusion

- Current model is able to predict the right label 75% of the times. Government would be able to disburse the funds faster.
- Geographical Features are the most important for improving the accuracy.
- Ordinal Regression marginally improves the accuracy of the model.

Conclusion

- Current model is able to predict the right label 75% of the times. Government would be able to disburse the funds faster.
- Geographical Features are the most important for improving the accuracy.
- Ordinal Regression marginally improves the accuracy of the model.

Limitations

- Never before seen Geo Ids in test dataset cannot have frequency or target encoding.
- Target encoding for geographical level with low number of observations is not reliable.

Ranked within Top 5 % on the leaderboard!