EARTHQUAKE DAMAGE MODELLING

G1 Team 2

Cerulean, Si Yun, Sean Jin, Genice, Xiao Yue

TABLE OF CONTENTS

Predict the level of damage to the building that was hit by the earthquake

Assumptions:
Constant earthquake magnitude of 7.8

PROBLEM STATEMENT

BUILDING DAMAGES

Major component of the losses

EARTHQUAKES

Leads to major economics losses

SAFETY

Safer building structures reduce fatality rates & losses

VULNERABILITY FUNCTION

GROUND MOTION INTENSITY

Others	**INDEPENDENT VARIABLE**
Ours	**CONSTANT**

DAMAGE LEVEL

Others	**CONTINUOUS VARIABLE**
Ours	**ORDINAL VARIABLE**

DATASETS

INPUT

23 attributes on building structures & mixture of binary, categorical and integer data types

OBSERVATIONS

760, 000 independent observations i.e. buildings

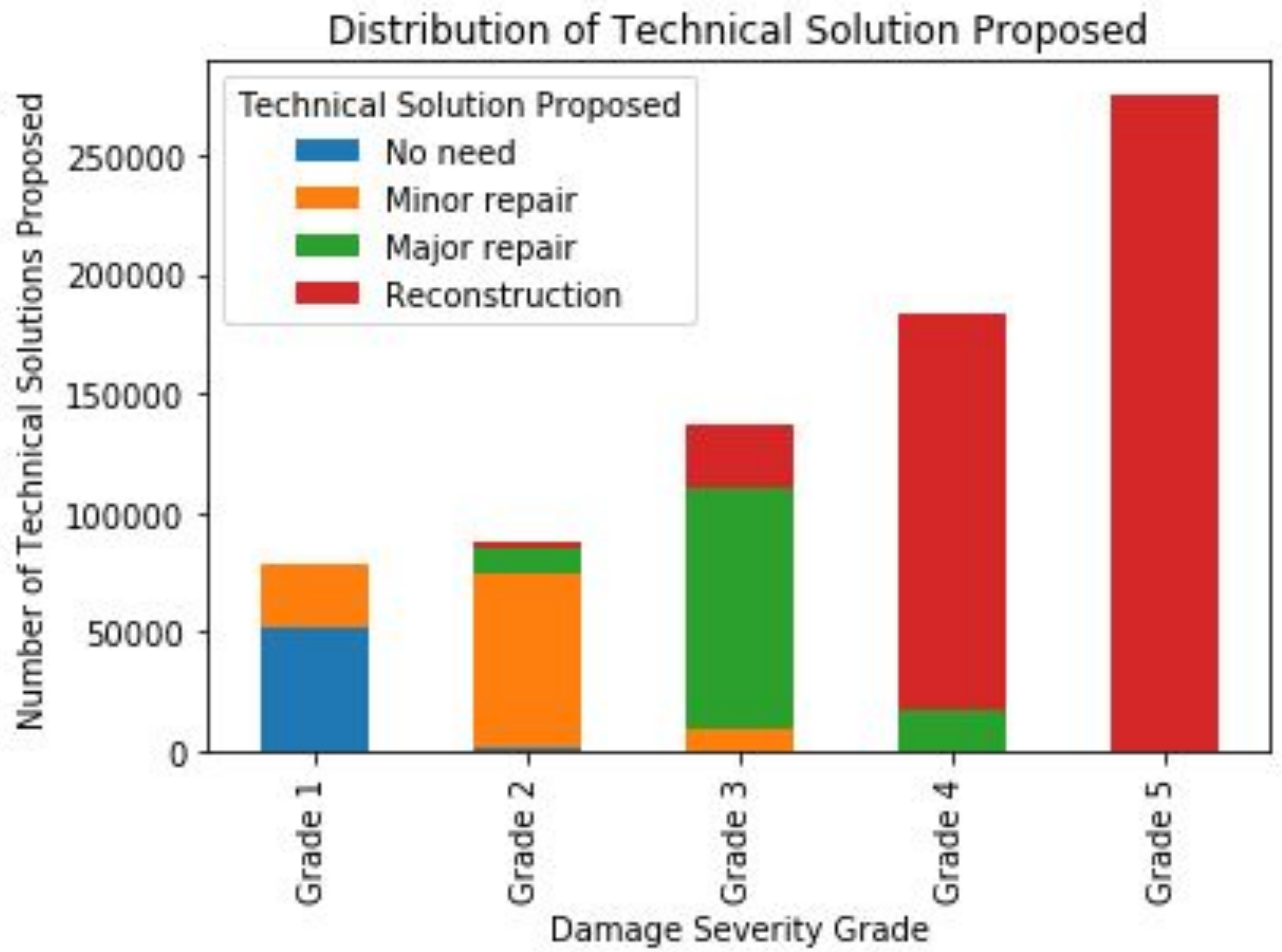
OUTPUT

3 Ordinal variables but focusing only on the 5 levels of damage

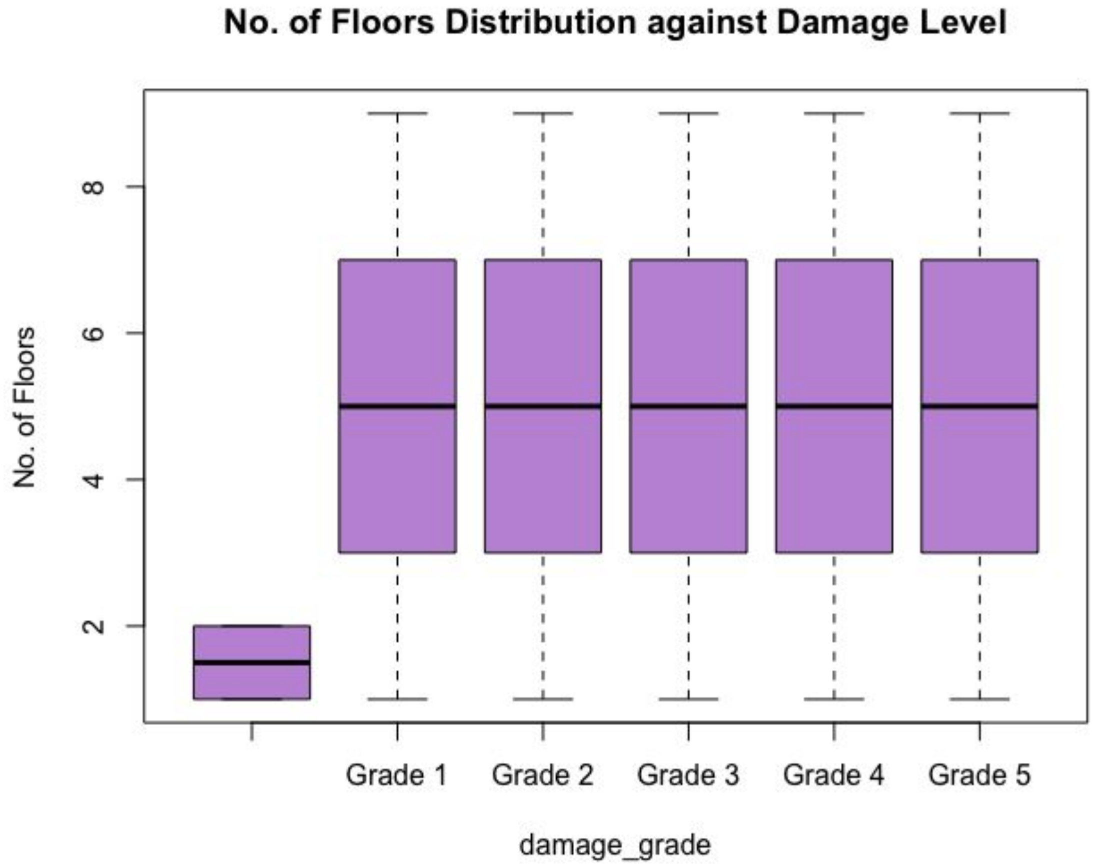
No Duplication

Imbalance

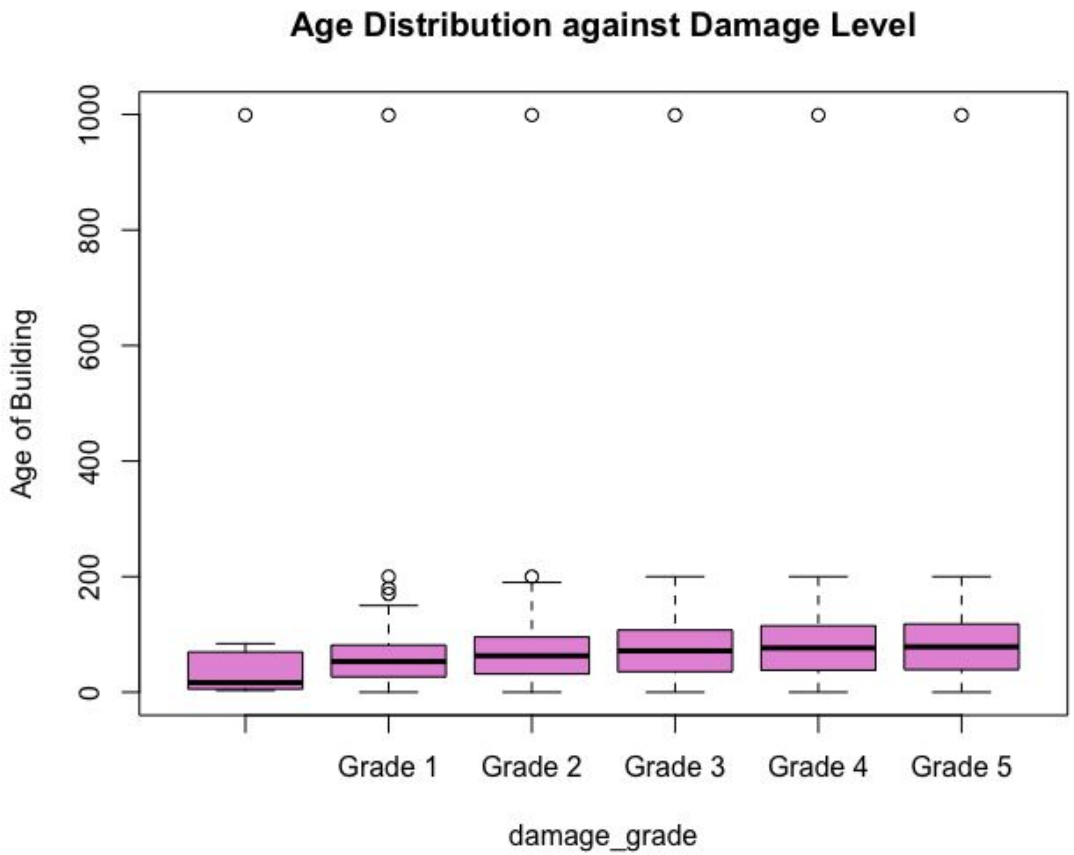
Minimal Missing Values



PRELIMINARY RESULTS

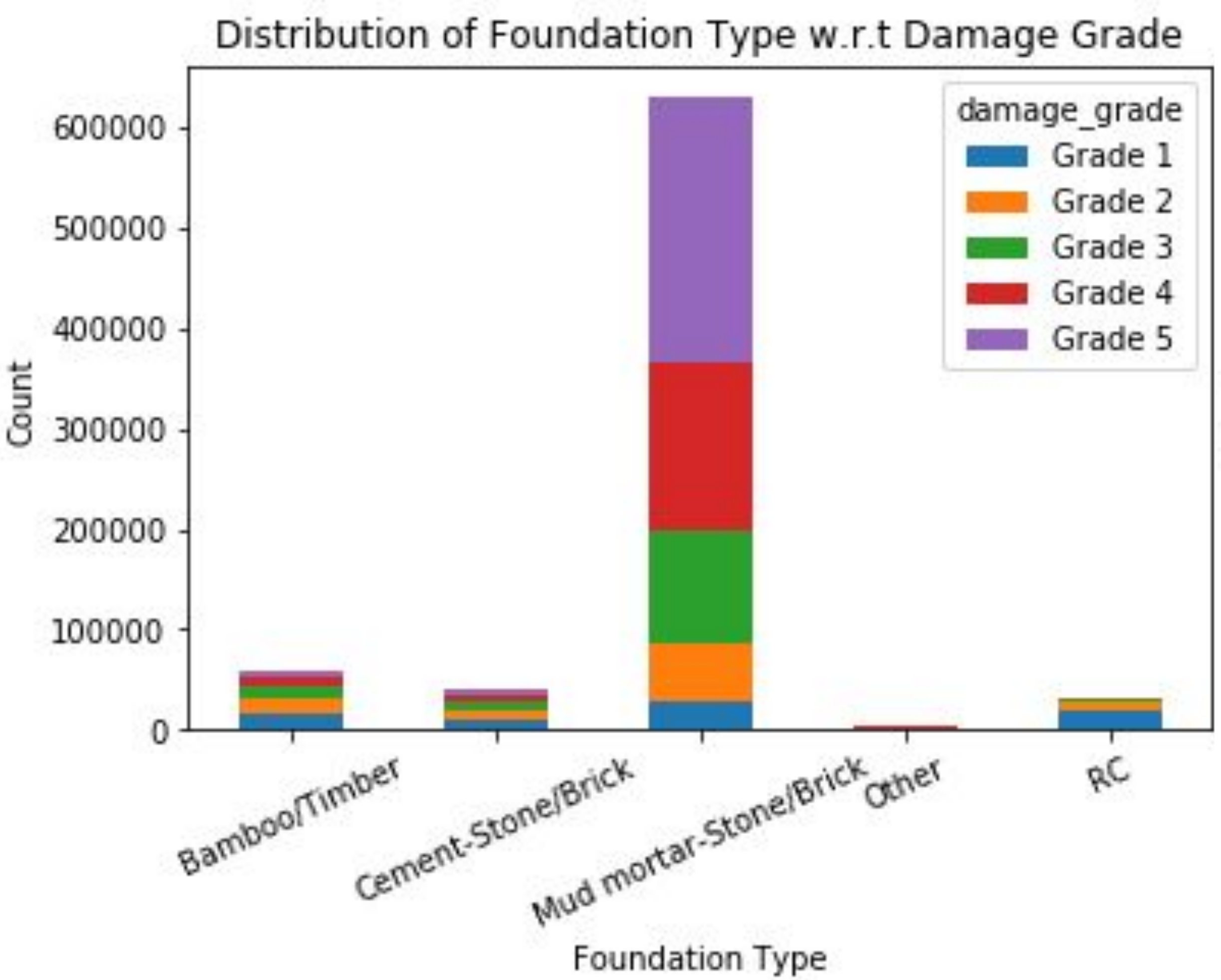


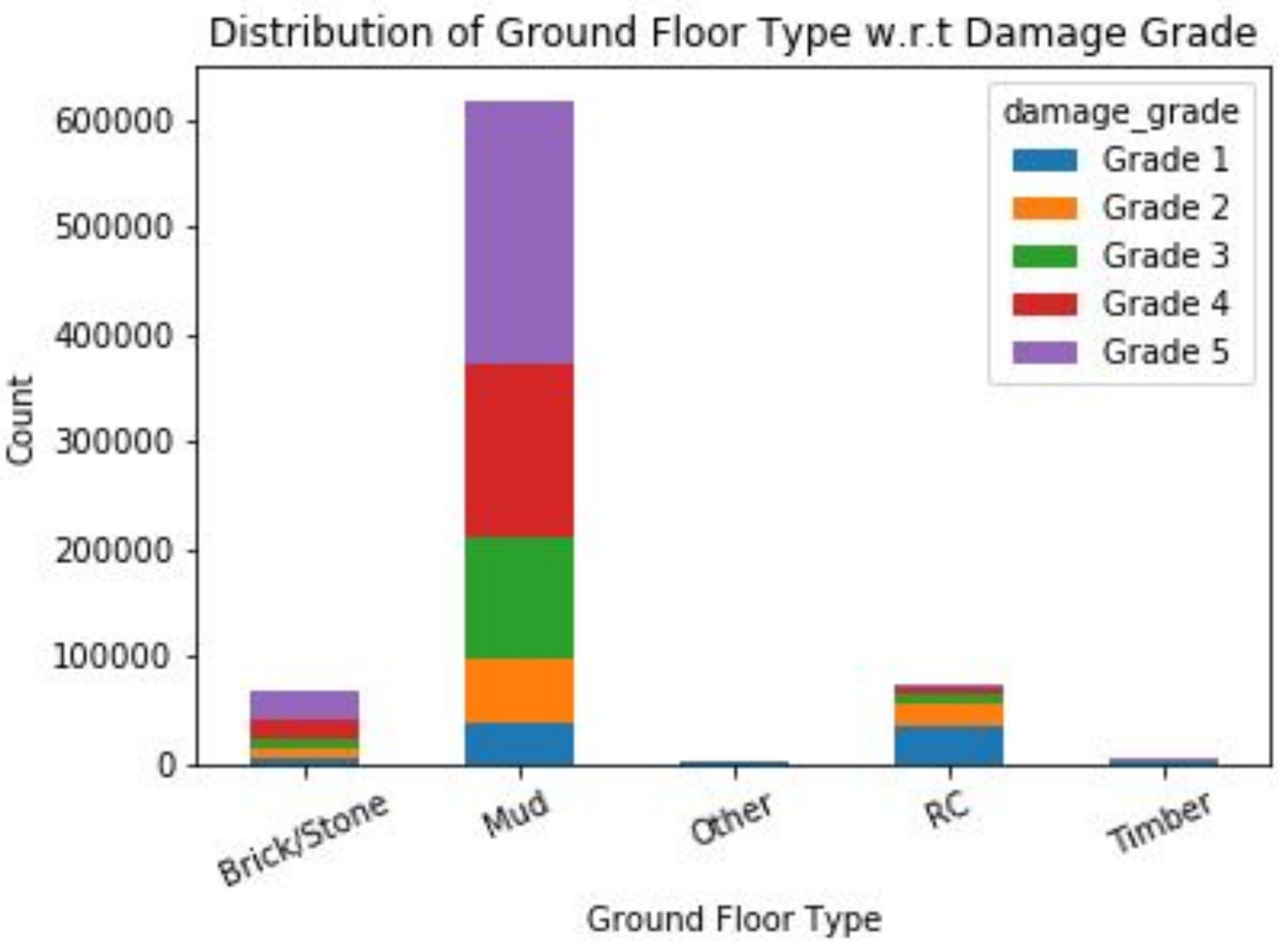
No Outliers



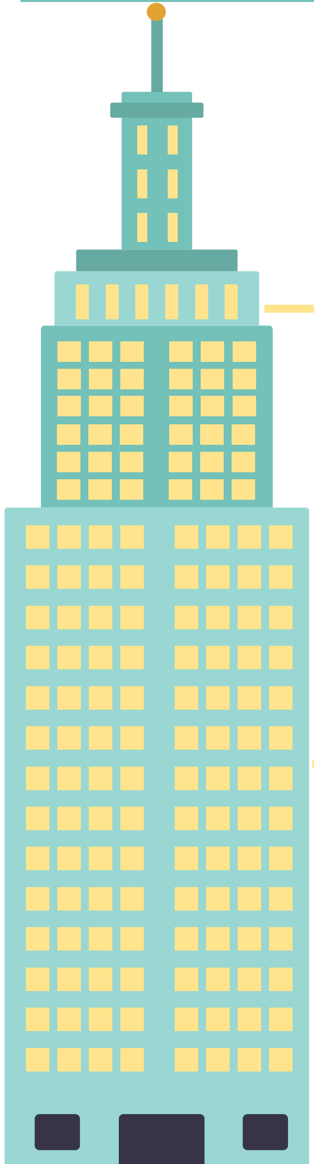
Presence of Outliers

PRELIMINARY RESULTS





TOOLS & RESOURCES



Data Preprocessing	Scikit-Learn Pandas Seaborn GeoPy OneHotEncoding Google Geolocation API
Modelling	XGBoost Mord DecisionTreeClassifier AdaBoostClassifier RidgeLogisticRegression ExtraTreesClassifier LogisticRegression MLPClassifier RandomForestClassifier LassoLogisticRegression ElasticNetRegression
Model Tuning	SMOTE RandomizedSearchCV GridSearchCV RFE Class Weight

COMPETITION LEADERBOARD

Richter's Predictor: Modeling Earthquake Damage

HOSTED BY DRIVENDATA

HOME PROBLEM DESCRIPTION ABOUT

User or team		Best public score	Timestamp	Trend (last 10)	# Entries
inoddy	1	0.7558	2019-12-31 09:54:30		64
seismicnz	2	0.7553	2019-12-19 09:47:12		27
Gillesvdw	3	0.7536	2019-05-08 20:01:38		15
SKOL_BOOMERS	4	0.7534	2019-10-23 17:56:55		6
DeereHunters	5	0.7528	2019-08-09 02:10:37		71

BASELINE MODEL

MODEL
Ordinal Logistic Regression

ENCODING METHOD
One Hot Encoding

F MEASURE
0.3381

	precision	recall	f1-score	support
1	0.56	0.28	0.37	15763
2	0.26	0.20	0.23	17451
3	0.26	0.10	0.14	27283
4	0.28	0.69	0.40	36769
5	0.54	0.28	0.37	55153
accuracy			0.34	152419
macro avg	0.38	0.31	0.30	152419
weighted avg	0.39	0.34	0.32	152419

```
Accuracy: 0.3381140146569653
Misclassification Rate: 0.6618859853430347
Precision: 0.3381140146569653
Recall: 0.3381140146569653
F-measure: 0.3381140146569653
```

TYPES OF ENCODING

#1

All One-Hot

#2

Mixture of One-Hot & Ordinal

#3

Mixture of One-Hot & Manual Ordinal

#4

Mostly Ordinal

#5

Mostly Manual Ordinal

#1. ONE-HOT ENCODING

One-Hot

	land_surface_condition	foundation_type	roof_type	ground_floor_type	other_floor_type	position	plan_configuration
394964	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
130914	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	Timber-Planck	Not attached	Rectangular
47552	Flat	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
91671	Steep slope	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
448138	Flat	Mud mortar-Stone/Brick	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Attached-1 side	Rectangular

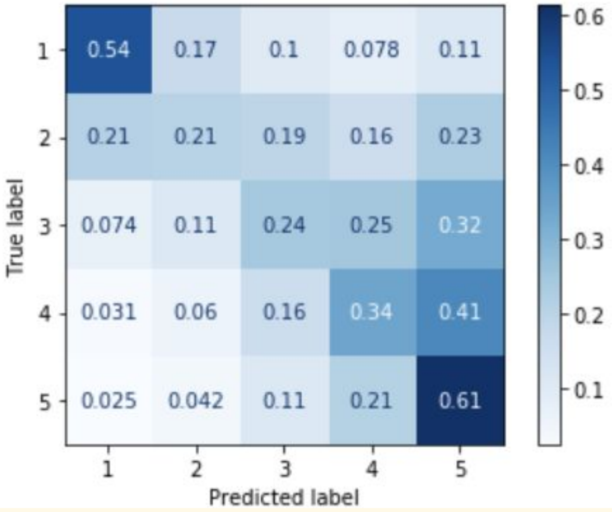
#1. RANDOM FOREST CLASSIFIER

ENCODING METHOD
One Hot Encoding

F MEASURE
0.4254

	precision	recall	f1-score	support
1	0.51	0.54	0.52	15763
2	0.26	0.21	0.23	17451
3	0.28	0.24	0.26	27283
4	0.35	0.34	0.35	36769
5	0.53	0.61	0.57	55153
accuracy			0.43	152419
macro avg	0.39	0.39	0.39	152419
weighted avg	0.41	0.43	0.42	152419

Accuracy: 0.42542596395462506
Misclassification Rate: 0.5745740360453749
Precision: 0.42542596395462506
Recall: 0.42542596395462506
F-measure: 0.42542596395462506



#2. MIXTURE OF ONE-HOT & ORDINAL ENCODING

Ordinal

One-Hot

Ordinal

One-Hot

	land_surface_condition	foundation_type	roof_type	ground_floor_type	other_floor_type	position	plan_configuration
394964	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
130914	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	Timber-Planck	Not attached	Rectangular
47552	Flat	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
91671	Steep slope	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
448138	Flat	Mud mortar-Stone/Brick	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Attached-1 side	Rectangular

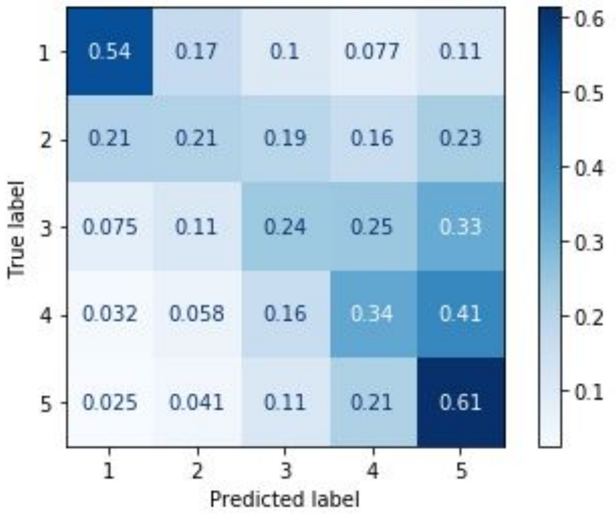
#2. RANDOM FOREST CLASSIFIER

ENCODING METHOD
Ordinal & One-Hot Encoding

F MEASURE
0.4258

	precision	recall	f1-score	support
1	0.51	0.54	0.52	15763
2	0.26	0.21	0.23	17451
3	0.28	0.24	0.26	27283
4	0.35	0.34	0.34	36769
5	0.53	0.61	0.57	55153
accuracy			0.43	152419
macro avg	0.39	0.39	0.39	152419
weighted avg	0.41	0.43	0.42	152419

```
Accuracy: 0.4258064939410441
Misclassification Rate: 0.5741935060589559
Precision: 0.4258064939410441
Recall: 0.4258064939410441
F-measure: 0.4258064939410441
```



#3. MIXTURE OF ONE-HOT & MANUAL ORDINAL ENCODING

Manual Ordinal

One-Hot

Manual Ordinal

One-Hot

	land_surface_condition	foundation_type	roof_type	ground_floor_type	other_floor_type	position	plan_configuration
394964	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
130914	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	Timber-Planck	Not attached	Rectangular
47552	Flat	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
91671	Steep slope	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
448138	Flat	Mud mortar-Stone/Brick	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Attached-1 side	Rectangular

#3. RANDOM FOREST CLASSIFIER

ENCODING METHOD
Manual Ordinal & One-Hot Encoding

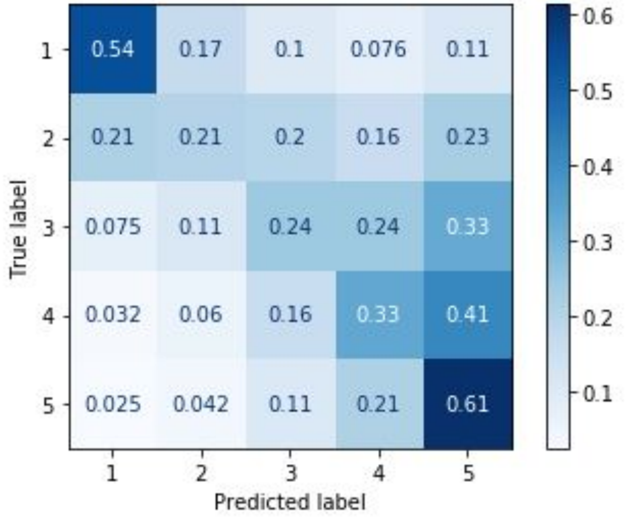
F MEASURE
0.4260

```
              precision    recall  f1-score   support

           1       0.51      0.54      0.53     15763
           2       0.26      0.21      0.23     17451
           3       0.28      0.24      0.26     27283
           4       0.35      0.33      0.34     36769
           5       0.53      0.61      0.57     55153

    accuracy                           0.43    152419
   macro avg       0.39      0.39      0.39    152419
weighted avg       0.41      0.43      0.42    152419

Accuracy: 0.4259901980724188
Misclassification Rate: 0.5740098019275812
Precision: 0.4259901980724188
Recall: 0.4259901980724188
F-measure: 0.4259901980724188
```



#4. MOSTLY ORDINAL ENCODING

Ordinal

One-Hot

	land_surface_condition	foundation_type	roof_type	ground_floor_type	other_floor_type	position	plan_configuration
394964	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
130914	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	Timber-Planck	Not attached	Rectangular
47552	Flat	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
91671	Steep slope	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
448138	Flat	Mud mortar-Stone/Brick	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Attached-1 side	Rectangular

#4. RANDOM FOREST CLASSIFIER

ENCODING METHOD
Mostly Ordinal Encoding

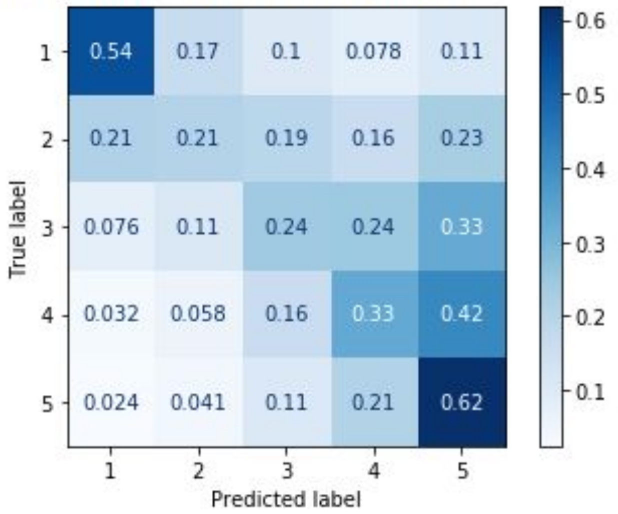
F MEASURE
0.4272

```
              precision    recall  f1-score   support

           1       0.51      0.54      0.52     15763
           2       0.26      0.21      0.23     17451
           3       0.28      0.24      0.26     27283
           4       0.35      0.33      0.34     36769
           5       0.53      0.62      0.57     55153

    accuracy                           0.43    152419
   macro avg       0.39      0.39      0.39    152419
weighted avg       0.41      0.43      0.42    152419

Accuracy: 0.42724988354470245
Misclassification Rate: 0.5727501164552975
Precision: 0.42724988354470245
Recall: 0.42724988354470245
F-measure: 0.4272498835447024
```



#5. MOSTLY MANUAL ORDINAL ENCODING

Ordinal

One-Hot

	land_surface_condition	foundation_type	roof_type	ground_floor_type	other_floor_type	position	plan_configuration
394964	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
130914	Moderate slope	Mud mortar-Stone/Brick	Bamboo/Timber-Light roof	Mud	Timber-Planck	Not attached	Rectangular
47552	Flat	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
91671	Steep slope	Bamboo/Timber	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Not attached	Rectangular
448138	Flat	Mud mortar-Stone/Brick	Bamboo/Timber-Heavy roof	Mud	TImber/Bamboo-Mud	Attached-1 side	Rectangular

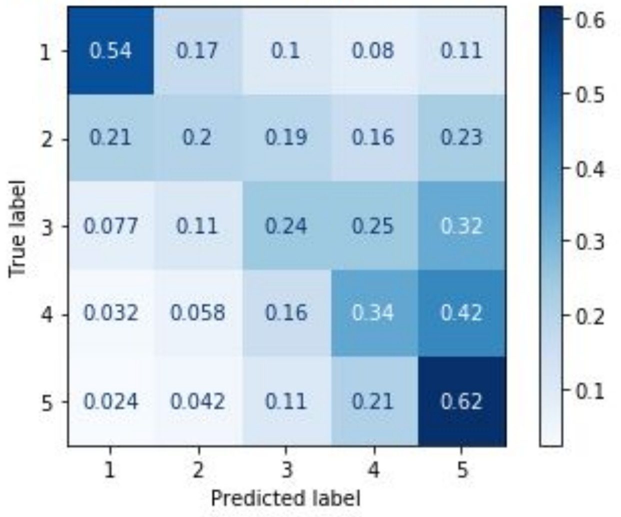
#5. RANDOM FOREST CLASSIFIER

ENCODING METHOD
Mostly Manual Ordinal Encoding

F MEASURE
0.4274

	precision	recall	f1-score	support
1	0.51	0.54	0.52	15763
2	0.26	0.20	0.23	17451
3	0.29	0.24	0.26	27283
4	0.36	0.34	0.35	36769
5	0.53	0.62	0.57	55153
accuracy			0.43	152419
macro avg	0.39	0.39	0.39	152419
weighted avg	0.41	0.43	0.42	152419

```
Accuracy: 0.42742702681424233
Misclassification Rate: 0.5725729731857576
Precision: 0.42742702681424233
Recall: 0.42742702681424233
F-measure: 0.42742702681424233
```

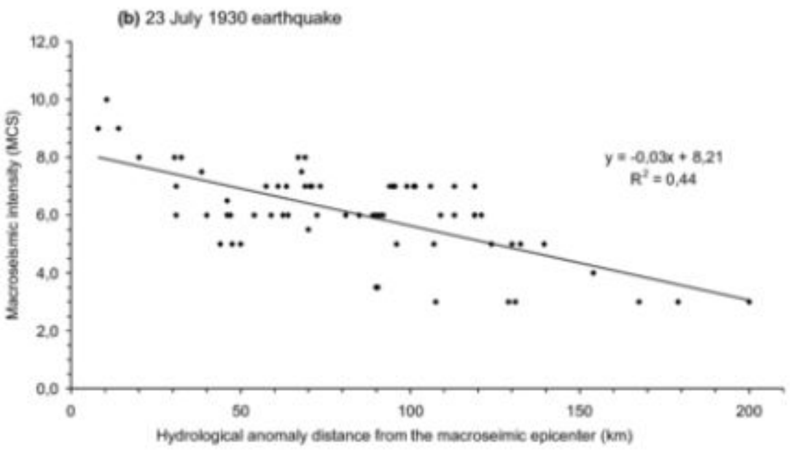


FEATURE ENGINEERING - DISTANCE

- Created a Google Geolocation API account
- Called the API using Python, using municipals + the string “Nepal” as inputs
- Got the outputs in JSON, parsed to extract the latitude and long values
- Used GeoPY to get our distance between the epicentre and respective points

FEATURE ENGINEERING - INTENSITY

- We used **intensity against distance** to get an estimation of the intensity of the Earthquake at the distance from the epicentre
- Shows a **MCS at epicenter of 8.21**, which correlates to a **magnitude of around 6.3**. Assuming linearity between MCS and magnitude, the Ghorka **earthquake of magnitude 7.8** would **score 10.2 on the MCS**. As such, we use that value as our intercept
- By using **y=-0.03x+10.2**, where x is our distance from the epicentre, we can get our new magnitude



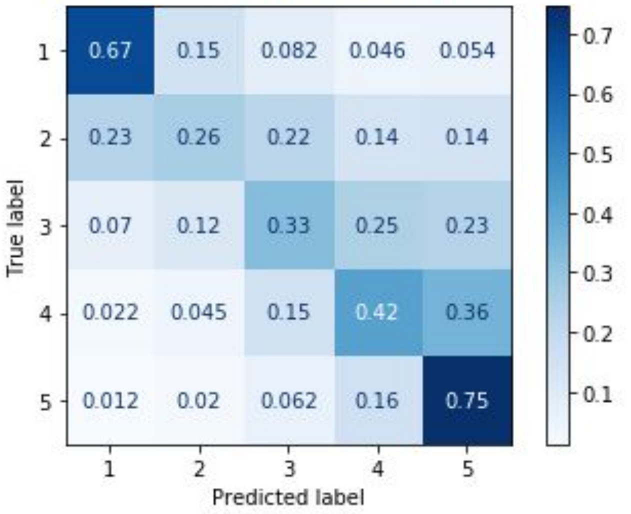
FE INTENSITY - RANDOM FOREST CLASSIFIER

FEATURE ENGINEERING
Intensity

F MEASURE
0.5340

	precision	recall	f1-score	support
1	0.59	0.67	0.63	13537
2	0.35	0.26	0.30	14523
3	0.38	0.33	0.35	23075
4	0.45	0.42	0.44	31780
5	0.65	0.75	0.70	49342
accuracy			0.53	132257
macro avg	0.48	0.49	0.48	132257
weighted avg	0.52	0.53	0.52	132257

Accuracy: 0.5340208835827215
Misclassification Rate: 0.4659791164172785
Precision: 0.5340208835827215
Recall: 0.5340208835827215
F-measure: 0.5340208835827215



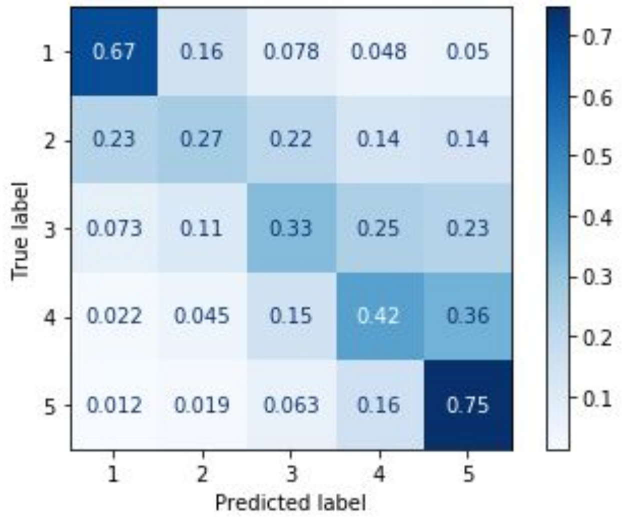
FE DISTANCE - RANDOM FOREST CLASSIFIER

FEATURE ENGINEERING
Distance

F MEASURE
0.5346

	precision	recall	f1-score	support
1	0.59	0.67	0.62	13537
2	0.35	0.27	0.30	14523
3	0.39	0.33	0.36	23075
4	0.45	0.42	0.43	31780
5	0.65	0.75	0.70	49342
accuracy			0.53	132257
macro avg	0.49	0.49	0.48	132257
weighted avg	0.52	0.53	0.52	132257

Accuracy: 0.5345804002812705
Misclassification Rate: 0.4654195997187295
Precision: 0.5345804002812705
Recall: 0.5345804002812705
F-measure: 0.5345804002812705



MODELS

		F1 SCORE
01	Baseline (Ordinal Logistic Regression)	0.3381
02	XGBoost Classifier	0.5618
03	Random Forest Classifier	0.5340
04	Adaboost-Decision Tree	0.5276
05	Extra Trees Classifier	0.5207
06	Multi-Layer Perceptron	0.5179

MODELS

		F1 SCORE
07	Adaboost Classifier	0.4693
08	Lasso Regression	0.4505
09	Ridge Regression	0.4500
10	Elastic Net Regression	0.4238
11	One-Versus-Rest Logistic Regression	0.4227
12	Multinomial Logistic Regression	0.4149

SELECTED MODELS FOR TUNING

- EXTRA TREES CLASSIFIER
- XGBOOST CLASSIFIER
- RANDOM FOREST CLASSIFIER
- ADABOOST-DECISION TREE

MODEL TUNING - EXTRA TREES CLASSIFIER

Feature Engineering

0.5207

Feature Selection (using RFE)

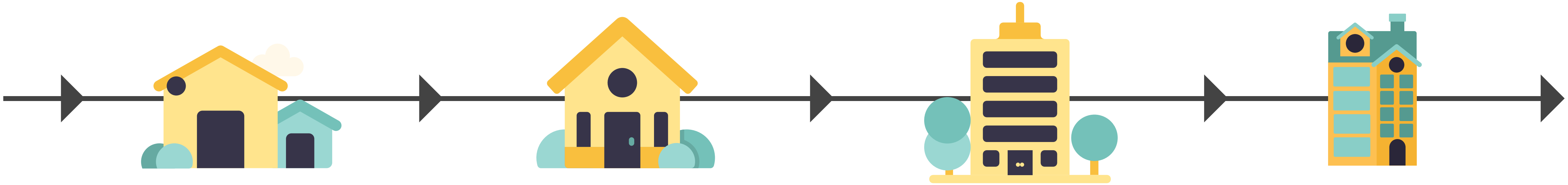
Ran but failed

SMOTE

0.5209

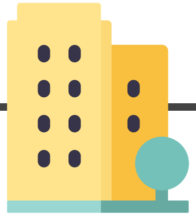
Class Weights

0.5193



SMOTE & Class Weight

0.5208



Randomized SearchCV

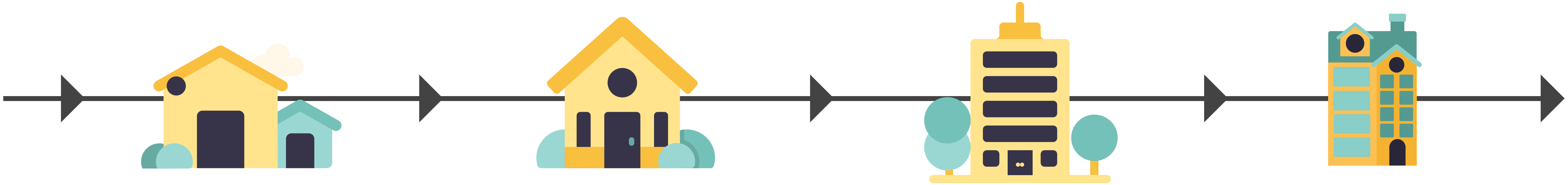
Ran but failed

Finalised F Score

0.5209

MODEL TUNING - RANDOM FOREST CLASSIFIER

Feature Engineering	Feature Selection	SMOTE	Class Weights
0.5340	0.5467	0.5132	0.5138



SMOTE & Class Weight

0.5130

Randomized SearchCV

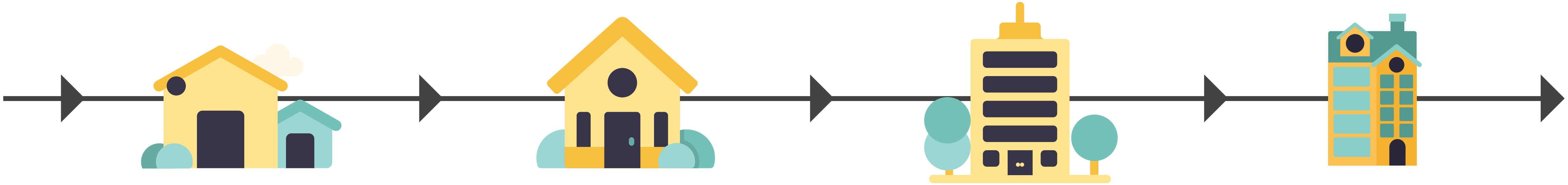
0.5442

Finalised F Score

0.5467

MODEL TUNING - XGBOOST CLASSIFIER

Feature Engineering	Feature Selection	SMOTE	Class Weights
0.5618	0.5467	0.5596	0.5413

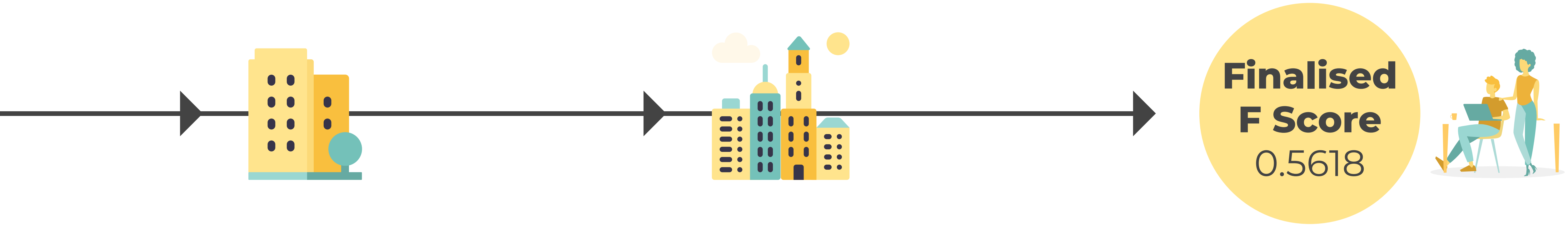


SMOTE & Class Weight

Ran but failed

Randomized SearchCV

Ran but failed



MODEL TUNING - ADABOOST-DECISION TREE

Feature Engineering

0.5276

Feature Selection

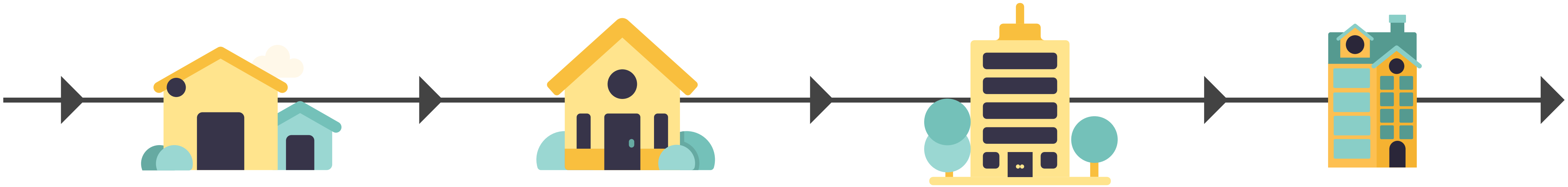
0.5286

SMOTE

0.5192

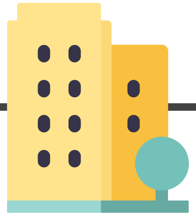
Class Weights

0.5086



SMOTE & Class Weight

0.4966



Randomized SearchCV

Ran but failed

Finalised F Score

0.5286

		F1 SCORE	% Difference to Baseline
00	Baseline (Ordinal Logistic Regression)	0.3381	-
01	XGBoost Classifier	0.5618	+ 59.0%
02	Random Forest Classifier	0.5467	+ 54.7%
03	Adaboost-Decision Tree	0.5286	+ 49.6%
04	Extra Trees Classifier	0.5209	+ 47.5%

Conclusion

- Dataset is hard to do modelling on
- **Feature engineering** improves F-score of our models the most
- **Feature selection** contributes significantly in model improvement as well

FUTURE WORKS - CONSIDERATIONS

#1

GridSearchCV & RandomizedSearchCV

#2

LightGBM package for Gradient Boosting Model

BETTER ACCURACY

FASTER TRAINING SPEED & HIGHER EFFICIENCY

CAPABLE OF HANDLING LARGE-SCALE DATA

REFERENCES

1.Zitzman, L., 2019. How Earthquake-Proof Buildings Are Designed. [Blog] *BigRentz*, Available at: <https://www.bigrentz.com/blog/earthquake-proof-buildings> William Harris
2."How Earthquake-resistant Buildings Work" 13 September 2011. HowStuffWorks.com. <https://science.howstuffworks.com/engineering/structural/earthquake-resistant-buildings.htm>
3.DrivenData. 2020. *Richter's Predictor: Modeling Earthquake Damage*. [online] Available at: <https://www.drivendata.org/competitions/57/nepal-earthquake/>
4.Eq2015.npc.gov.np. 2020. *Municipality Codes*. [online] Available at: <http://eq2015.npc.gov.np/docs/#/location_codes?id=municipality-codes>
5.GitHub. 2020. *SVM-Hyper-Parameter-Tuning-Using-Gridsearchcv*. [online] Available at: <https://github.com/clareyan/SVM-Hyper-parameter-Tuning-using-GridSearchCV/blob/master/SVM_Iris.ipynb>
6.Lightgbm.readthedocs.io. 2020. *Welcome To Lightgbm'S Documentation! — Lightgbm 2.3.2 Documentation*. [online] Available at: <https://lightgbm.readthedocs.io/en/latest/>
7.Liu, C., 2020. SVM Hyperparameter Tuning using GridSearchCV. [Blog] *towards data science*, Available at: <https://towardsdatascience.com/svm-hyper-parameter-tuning-using-gridsearchcv-49c0bc55ce29>
8.Data Science Stack Exchange. 2020. *Parameters In Gridsearchcv In Scikit-Learn*. [online] Available at: <https://datascience.stackexchange.com/questions/13410/parameters-in-gridsearchcv-in-scikit-learn>
9.Scikit-learn.org. 2020. *Sklearn.Model_Selection.Gridsearchcv — Scikit-Learn 0.22.2 Documentation*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
10.Scikit-learn.org. 2020. *Sklearn.Svm.SVC — Scikit-Learn 0.22.2 Documentation*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
11.Scikit-learn.org. 2020. *Sklearn.Feature_Selection.RFE — Scikit-Learn 0.22.2 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html>
12. Imbalanced-learn.readthedocs.io. 2020. *Imblearn.Over_Sampling.SMOTE — Imbalanced-Learn 0.5.0 Documentation*. [online] Available at: <https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html>
13.Marketplace. 2018. *Why Bamboo And Earth Are Better Than Steel And Concrete After A Himalayan Earthquake - Marketplace*. [online] Available at: <https://www.marketplace.org/2018/04/25/why-bamboo-and-earth-are-better-steel-and-concrete-after-himalayan-earthquake/>
14.Mishra, G., 2020. *Performance Of Various Types Of Buildings During Earthquake*. [online] The Constructor. Available at: <https://theconstructor.org/earthquake/performance-buildings-types-earthquake/2224/>
15.Scikit-learn.org. 2020. *Sklearn.Utils.Class_Weight.Compute_Class_Weight — Scikit-Learn 0.22.2 Documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html>
16.Scikit-learn.org. 2020. *Decision Tree Regression With Adaboost — Scikit-Learn 0.22.2 Documentation*. [online] Available at: <https://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_regression.html>
17.Scikit-learn.org. 2020. *sklearn.ensemble.ExtraTreesClassifier — Scikit-Learn 0.22.2 Documentation*. [online] Available at <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html>
18.Rikunet, 2019. *Modelling rating data correctly using ordered logistic regression*. Available at: http://rikunert.com/ordinal_rating
19.Fabian Pedregosa-Izquierdo. Feature extraction and supervised learning on fMRI: from practice to theory. PhD thesis. Available at: https://pythonhosted.org/mord/
20.Bhattacharyya, S., 2018. Ridge and Lasso Regression: L1 and L2 Regularization [Blog] *towards data science*, Available at: <https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b>
21.Scikit-learn.org. 2020. *sklearn.linear_model.LogisticRegression — Scikit-Learn 0.22.2 Documentation.* [online] Available at <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html>
22.Stackoverflow. Getting distance between two points based on latitude/longitude. Available at: https://stackoverflow.com/questions/19412462/getting-distance-between-two-points-based-on-latitude-longitude

THANK You!
Questions?