Predicting Earthquake Damage in Nepal

Microsoft Professional Capstone : Data Science

Anurag Sareen, April 2018

Executive Summary

This document presents an analysis of the damage to buildings in Nepal due to a 7.8 magnitude earthquake that hit the region in April of 2015. The analysis was based on 10000 observations of buildings with various attributes. Key characteristics of the buildings were presented in the data and relevant feature set was created for training a model to predict the outcome of the severity of the damage caused based on key characteristics of the buildings.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between building characteristics and damage grade were identified. After exploring the data, a predictive model to classify damage grade from its features was created.

While many factors can help indicate the damage grade of a building, significant features found in this analysis were:

- `geo_level_1_id`, `geo_level_2_id`, `geo_level_3_id` geographic region in which building exists, from largest (level 1) to most specific sub-region (level 3). – buildings with less id values are most likely to get damaged than the buildings with greater id values
- `age` (type: int): age of the building in years. – new buildings were mostly damaged than older ones
- `area` (type: int): plinth area of the building in m2m2. – buildings with less area were more likely to be damaged than larger area
- `height` (type: int): height of the building in mm.
- `has_superstructure_mud_mortar_stone` (type: binary): flag variable that indicates if the superstructure was made of Mud Mortar - Stone.
- `has_superstructure_cement_mortar_brick` (type: binary): flag variable that indicates if the superstructure was made of Cement Mortar - Brick.
- `has_secondary_use` (type: binary): flag variable that indicates if the building was used for any secondary purpose. – buildings with secondary use are more likely to get damaged
- `has_secondary_use_agriculture` (type: binary): flag variable that indicates if the building was used for agricultural purposes. – buildings used for agriculture are more likely to get damaged
- `has_superstructure_rc_non_engineered` (type: binary): flag variable that indicates if the superstructure was made of non-engineered reinforced concrete.
- `has_superstructure_rc_engineered` (type: binary): flag variable that indicates if the superstructure was made of engineered reinforced concrete.
- `plan_configuration` (type: categorical): building plan configuration. Possible values: a779, 84cf, 8e3f, d2d9, 3fee, 6e81, 0448, 1442, cb88. – Plan configuration play a significant role in determining high risk buildings
- `foundation_type` (type: categorical): type of foundation used while building. Possible values: 337f, 858b, 6c3e, 467b, bb5f. - Foundation types play a significant role in determining high risk buildings
- `ground_floor_type` (type: categorical): type of the ground floor. Possible values: b1b4, b440, 467b, e26c, bb5f. Ground floor types play a significant role in determining high risk buildings
- `other_floor_type` (type: categorical): type of constructions used in higher than the ground floors (except of roof). Possible values: f962, 9eb0, 441a, 67f9. – Other floor types play a significant role in determining high risk buildings
- `position` (type: categorical): position of the building. Possible values: 3356, bfba, bcab, 1787. – Position plays a significant role in determining high risk buildings

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics.

Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 10000 observations are shown here:

Column	mean	std	min	25%	50% median	75%	max	DCount
building_id	9987.16	5800.801	1	4998.75	9963.5	15044.75	19999	10000
geo_level_1_id	7.1356	6.225567	0	2	6	10	30	31
geo_level_2_id	296.9303	279.3907	0	60	219	466	1411	1137
geo_level_3_id	2678.618	2520.664	0	606.75	1937.5	4158	12151	5172
count_floors_pre_eq	2.1467	0.736365	1	2	2	3	9	8
age	25.3935	64.48289	0	10	15	30	995	31
area	38.4381	21.26588	6	26	34	44	425	158
height	4.6531	1.792842	1	4	5	5	30	18
has_superstructure_adobe_mud	0.0897	0.285766	0	0	0	0	1	2
has_superstructure_mud_mortar_stone	0.7626	0.425511	0	1	1	1	1	2
has_superstructure_stone_flag	0.0299	0.17032	0	0	0	0	1	2
has_superstructure_cement_mortar_stone	0.019	0.136532	0	0	0	0	1	2
has_superstructure_mud_mortar_brick	0.0688	0.253126	0	0	0	0	1	2
has_superstructure_cement_mortar_brick	0.0725	0.259327	0	0	0	0	1	2
has_superstructure_timber	0.2561	0.4365	0	0	0	1	1	2
has_superstructure_bamboo	0.0877	0.282872	0	0	0	0	1	2
has_superstructure_rc_non_engineered	0.04	0.195969	0	0	0	0	1	2
has_superstructure_rc_engineered	0.0138	0.116666	0	0	0	0	1	2
has_superstructure_other	0.0141	0.117909	0	0	0	0	1	2
count_families	0.9846	0.423297	0	1	1	1	7	8
has_secondary_use	0.1086	0.311152	0	0	0	0	1	2
has_secondary_use_agriculture	0.0673	0.250553	0	0	0	0	1	2
has_secondary_use_hotel	0.0294	0.168933	0	0	0	0	1	2
has_secondary_use_rental	0.0064	0.079748	0	0	0	0	1	2
has_secondary_use_institution	0.0007	0.02645	0	0	0	0	1	2
has_secondary_use_school	0.0007	0.02645	0	0	0	0	1	2
has_secondary_use_industry	0.0008	0.028274	0	0	0	0	1	2
has_secondary_use_health_post	0.0002	0.014141	0	0	0	0	1	2
has_secondary_use_gov_office	0.0002	0.014141	0	0	0	0	1	2
has_secondary_use_use_police	0.0001	0.01	0	0	0	0	1	2
has_secondary_use_other	0.0053	0.072612	0	0	0	0	1	2

In addition to the numeric values, the earthquake damage observations include categorical features, including:

- **land_surface_condition** - 3 unique values; `'d502' '808e' '2f15'`
- **foundation_type** - 5 unique values; `'337f' '6c3e' '858b' '467b' 'bb5f'`
- **roof_type** - 3 unique values; `'7e76' 'e0e2' '67f9'`

- **ground_floor_type** - 5 unique values; `'b1b4' '467b' 'b440' 'e26c' 'bb5f'`
- **other_floor_type** - 4 unique values; `'f962' '441a' '9eb0' '67f9'`
- **position** - 4 unique values; `'3356' 'bfba' 'bcab' '1787'`
- **plan_configuration** - 9 unique values; `'a779' '8e3f' '84cf' '0448' 'd2d9' '6e81' '3fee' '1442' 'cb88'`
- **legal_ownership_status** - 4 unique values; `'c8e1' 'ab03' 'cae1' 'bb5f'`

Various queries were written to find out the properties of the data. It was observed that there were no null values in the dataset.

```
In [3]: print(df_train_values.isnull().sum())

building_id                               0
geo_level_1_id                            0
geo_level_2_id                            0
geo_level_3_id                            0
count_floors_pre_eq                       0
age                                       0
area                                      0
height                                    0
land_surface_condition                    0
foundation_type                           0
roof_type                                 0
ground_floor_type                         0
other_floor_type                          0
position                                  0
plan_configuration                        0
has_superstructure_adobe_mud              0
has_superstructure_mud_mortar_stone       0
has_superstructure_stone_flag             0
has_superstructure_cement_mortar_stone    0
has_superstructure_mud_mortar_brick       0
has_superstructure_cement_mortar_brick    0
has_superstructure_timber                 0
has_superstructure_bamboo                 0
has_superstructure_rc_non_engineered      0
has_superstructure_rc_engineered          0
has_superstructure_other                  0
legal_ownership_status                    0
count_families                            0
has_secondary_use                         0
has_secondary_use_agriculture             0
has_secondary_use_hotel                   0
has_secondary_use_rental                  0
has_secondary_use_institution             0
has_secondary_use_school                  0
has_secondary_use_industry                0
has_secondary_use_health_post             0
has_secondary_use_gov_office              0
has_secondary_use_use_police              0
has_secondary_use_other                   0
dtype: int64
```

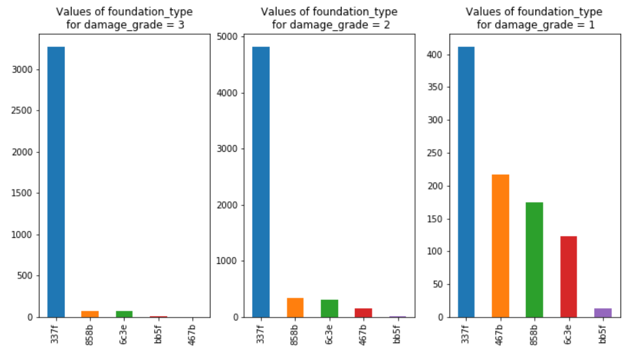
```
In [2]: df_train_values.dtypes

Out[2]: building_id                               int64
        geo_level_1_id                            int64
        geo_level_2_id                            int64
        geo_level_3_id                            int64
        count_floors_pre_eq                       int64
        age                                       int64
        area                                      int64
        height                                    int64
        land_surface_condition                   object
        foundation_type                          object
        roof_type                                object
        ground_floor_type                        object
        other_floor_type                         object
        position                                 object
        plan_configuration                       object
        has_superstructure_adobe_mud              int64
        has_superstructure_mud_mortar_stone       int64
        has_superstructure_stone_flag             int64
        has_superstructure_cement_mortar_stone    int64
        has_superstructure_mud_mortar_brick       int64
        has_superstructure_cement_mortar_brick    int64
        has_superstructure_timber                 int64
        has_superstructure_bamboo                 int64
        has_superstructure_rc_non_engineered      int64
        has_superstructure_rc_engineered          int64
        has_superstructure_other                  int64
        legal_ownership_status                   object
        count_families                          float64
        has_secondary_use                       float64
        has_secondary_use_agriculture             int64
        has_secondary_use_hotel                   int64
        has_secondary_use_rental                  int64
        has_secondary_use_institution             int64
        has_secondary_use_school                  int64
        has_secondary_use_industry                int64
        has_secondary_use_health_post             int64
        has_secondary_use_gov_office              int64
        has_secondary_use_use_police              int64
        has_secondary_use_other                   int64
        dtype: object
```

Bar charts were created to show frequency of these features, and indicate the following:

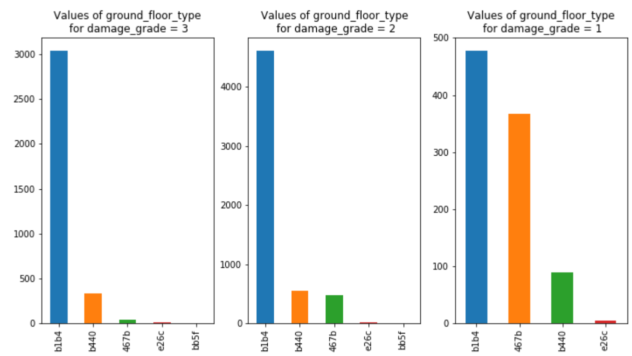
- The mean age of buildings with a damage_grade of 2 is higher than for buildings with a damage_grade of 1 and 3.
- Most of the buildings with a damage_grade of 1 are below average height.
- Most of the buildings with an above average area have a damage_grade of 2
- Most of the damage was done where the count_families was less than 2
- Damage was done to buildings where secondary use was agriculture and hotel
- Most of the buildings with damage grade of 2 and 3 has superstructure made of mud mortar stone
- Most buildings which has superstructure rc engineered were less damaged than the ones with rc non-engineered
- Buildings with superstructure of bamboo, timber, cement mortar brick, mud mortar brick, adobe mud, stone flag and cement mortar stone were comparatively less damaged.

Since land_surface_condition, foundation_type, roof_type, ground_floor_type, other_floor_type, position, plan_configuration and legal_ownership_status features had categorical values it was decided to covert these values to indicator values.

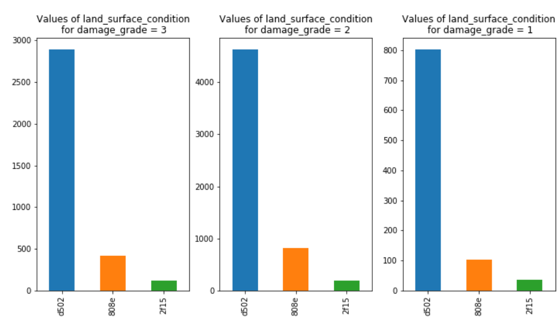
- Most of the damage was done where the foundation_type was 337f



- Most of the damage was done where the ground_floor_type was b1b4

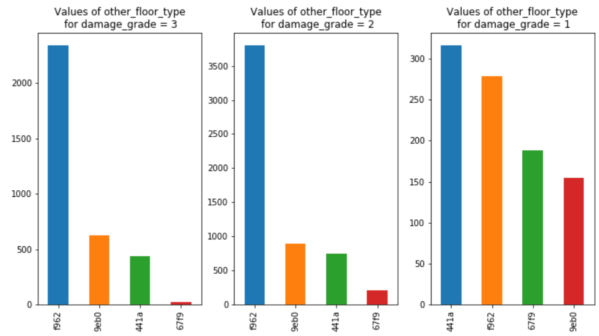


- Most of the damage was done where the land_surface_condition was d502

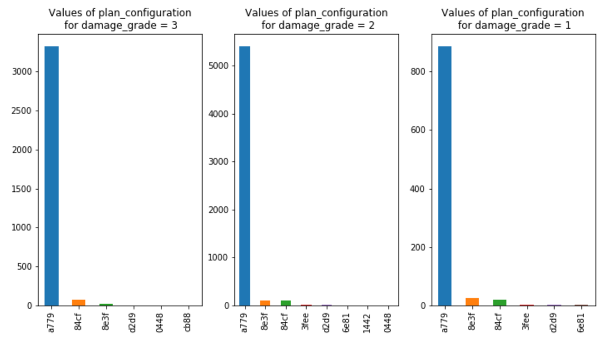


- Most of the damage was done where the legal_ownership_status was c8e1

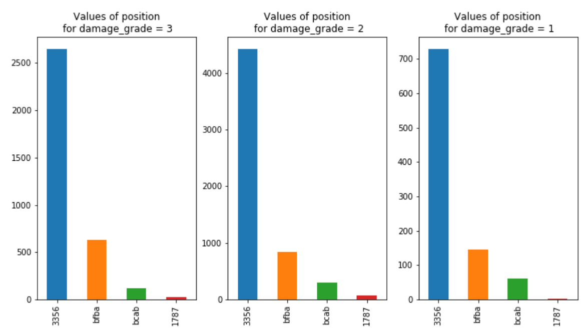
- Most of the damage was done where the other_floor_type was f962



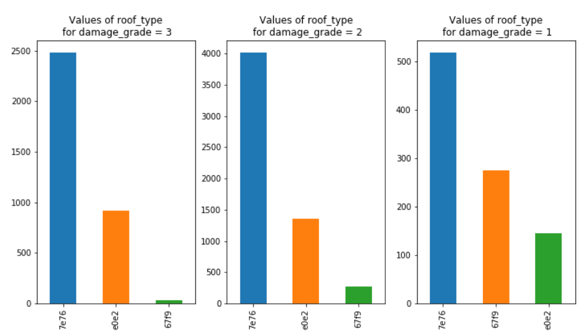
- Most of the damage was done where the plan configuration was a779



- Most of the damage was done where the position was 3356

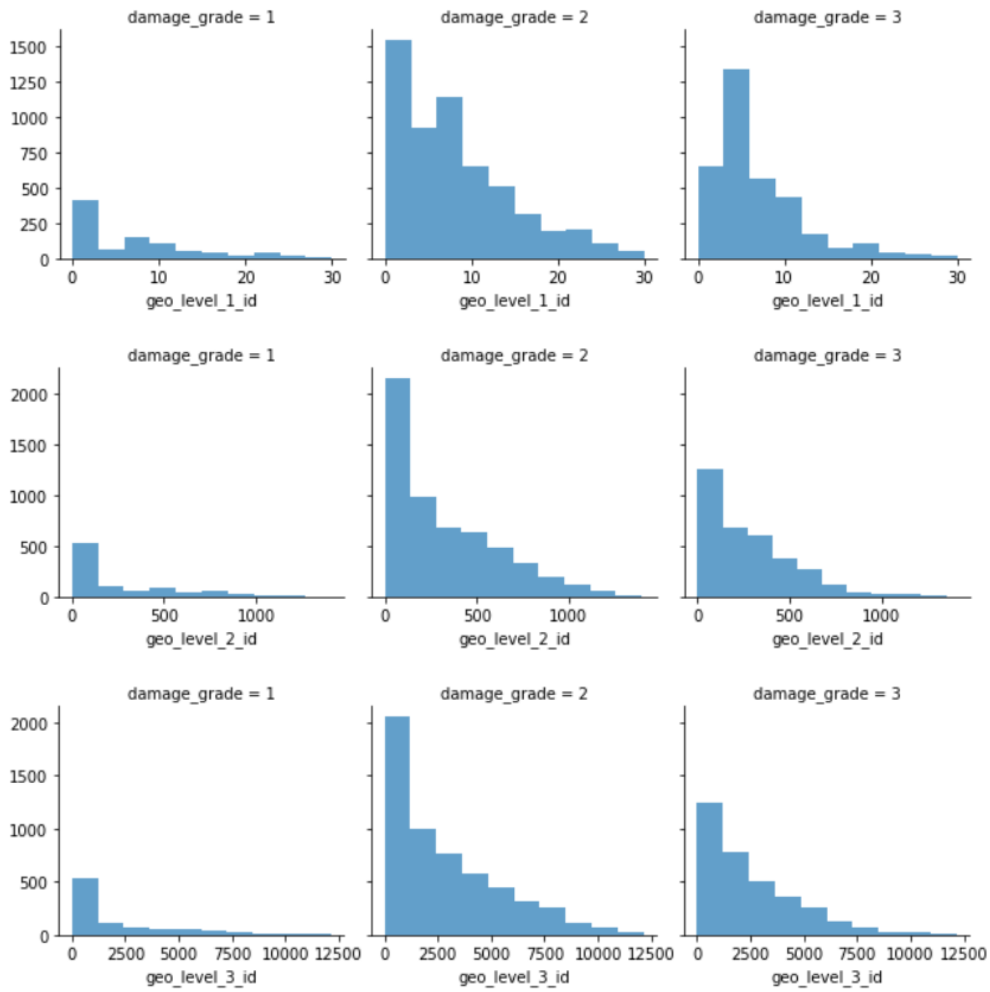


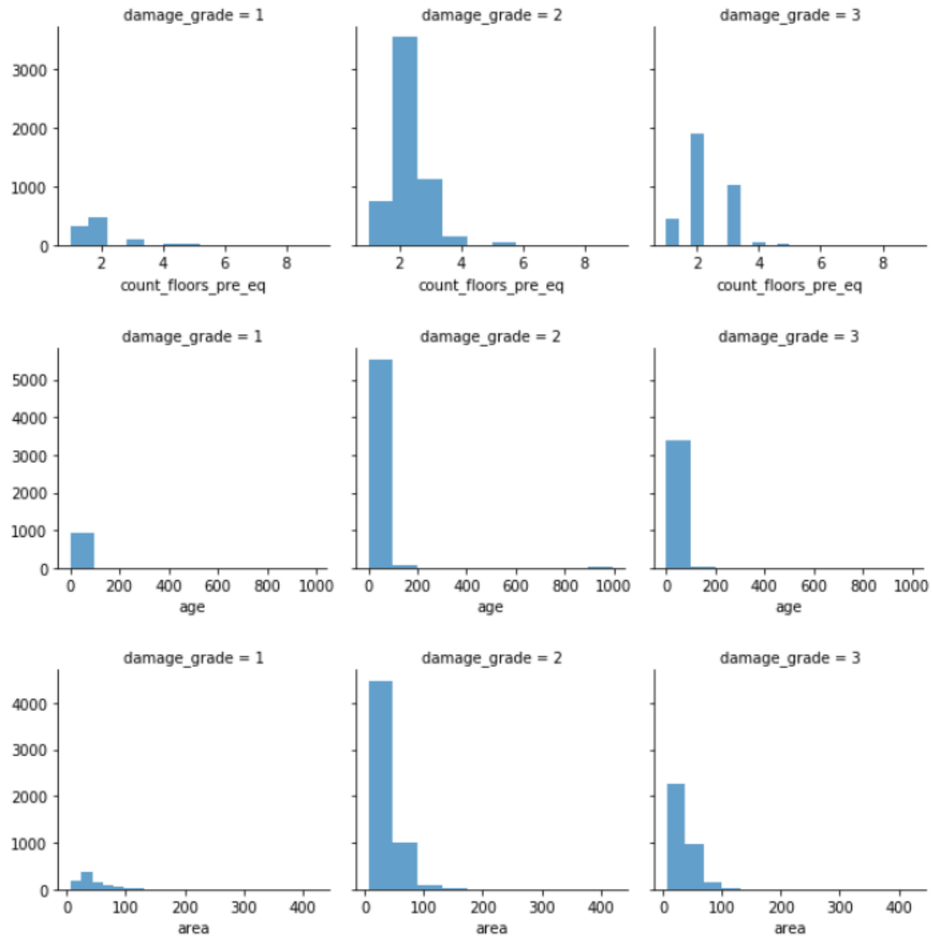
- Most of the damage was done where the roof_type was 7e76



- **land_surface_condition** - 3 unique values; `'d502' '808e' '2f15'`
- **foundation_type** - 5 unique values; `'337f' '6c3e' '858b' '467b' 'bb5f'`
- **roof_type** - 3 unique values; `'7e76' 'e0e2' '67f9'`
- **ground_floor_type** - 5 unique values; `'b1b4' '467b' 'b440' 'e26c' 'bb5f'`
- **other_floor_type** - 4 unique values; `'f962' '441a' '9eb0' '67f9'`
- **position** - 4 unique values; `'3356' 'bfba' 'bcab' '1787'`

- **plan_configuration** - 9 unique values; `'a779'  '8e3f'  '84cf'  '0448'  'd2d9'  '6e81'  '3fee'  '1442'  'cb88'`
- **legal_ownership_status** - 4 unique values; `'c8e1'  'ab03'  'cae1'  'bb5f'`



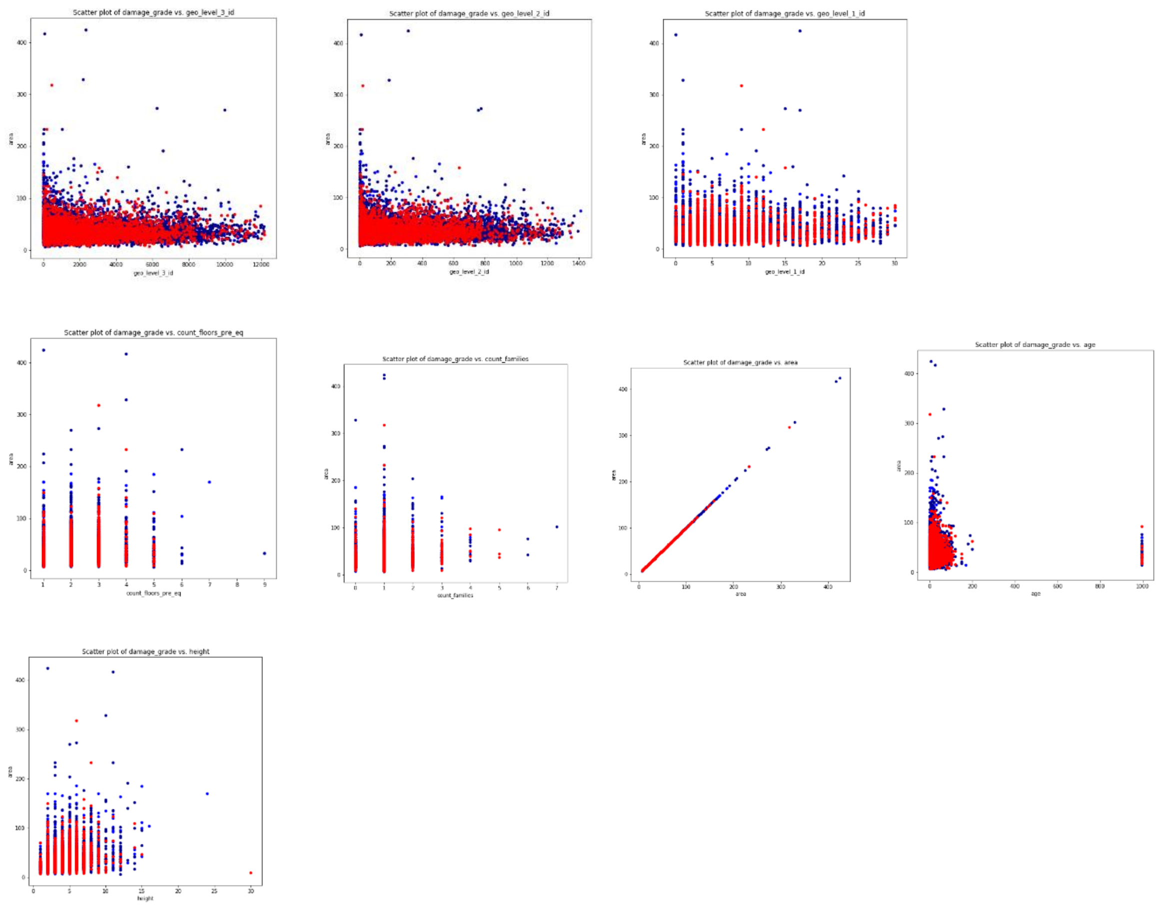

damage_grade = 1
damage_grade = 2
damage_grade = 3
count_floors_pre_eq
age
area

Correlation and Apparent Relationships

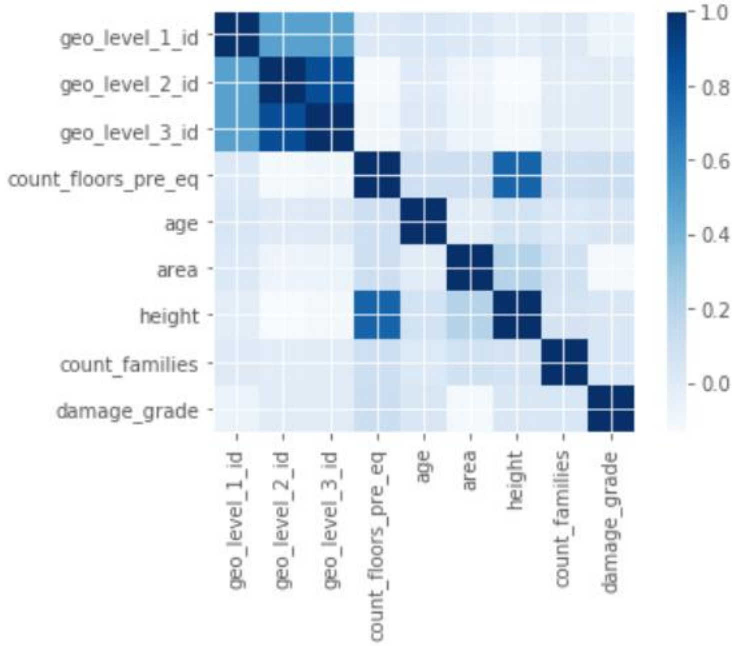
After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between Damage Grade and the other features. The categorical values were converted to indicator values and the following observations were recorded.

Numeric Relationships

The following scatter-plot matrix was generated initially to compare numeric features with one another. The key features in this matrix are shown here:



- There was a positive relation between geo_level_2_id and geo_level_3_id
- There was a positive relation between count_floors_pre_eq and height


geo_level_1_id
geo_level_2_id
geo_level_3_id
count_floors_pre_eq
age
area
height
count_families
damage_grade
1.0
0.8
0.6
0.4
0.2
0.0

Categorical Relationships

All categorical columns were converted to indicator values. Key feature matrix is shown below.

The correlation between the numeric columns was then calculated with the following results:

	geo_level_2_id	geo_level_3_id	count_floors_pre_eq	height	has_superstructure_mud_mortar_stone	has_superstructure_cement_mortar_brick
geo_level_2_id	1	0.870382836	-0.116854966	-0.131552688	0.352267285	-0.203370633
geo_level_3_id	0.870382836	1	-0.089244207	-0.108456127	0.296132128	-0.175352537
count_floors_pre_eq	-0.116854966	-0.089244207	1	0.77124901	-0.031195102	-0.088172898
height	-0.131552688	-0.108456127	0.77124901	1	-0.118450979	0.009787889
has_superstructure_mud_mortar_stone	0.352267285	0.296132128	-0.031195102	-0.118450979	1	-0.455778633
has_superstructure_cement_mortar_brick	-0.203370633	-0.175352537	-0.088172898	0.009787889	-0.455778633	1
has_superstructure_rc_non_engineered	-0.057156463	-0.048689249	0.014082725	0.094436123	-0.220728217	0.12791509
has_superstructure_rc_engineered	-0.094260241	-0.085546237	0.052101714	0.133340758	-0.207985064	0.115679924
has_secondary_use	-0.0307938	-0.031280681	0.052241153	0.079551914	-0.071143512	0.056102903
has_secondary_use_agriculture	0.022759979	0.019316486	0.000146845	-0.021270105	0.057944435	-0.061248681
foundation_type-337f	0.19443943	0.178281891	0.134488408	0.00495806	0.532385375	-0.38165555
foundation_type-467b	-0.128600123	-0.122336184	0.039057846	0.161826482	-0.343606048	0.238468891
roof_type-67f9	-0.175205092	-0.160517268	0.028528213	0.166192251	-0.422184634	0.424382432
ground_floor_type-467b	-0.200115409	-0.181077215	-0.079958306	0.062683653	-0.461016076	0.549397768
other_floor_type-441a	-0.040786391	-0.038768699	-0.652418051	-0.520932979	-0.202189864	0.22763091
other_floor_type-67f9	-0.142559636	-0.130252786	0.168954359	0.303721028	-0.347012776	0.330466741

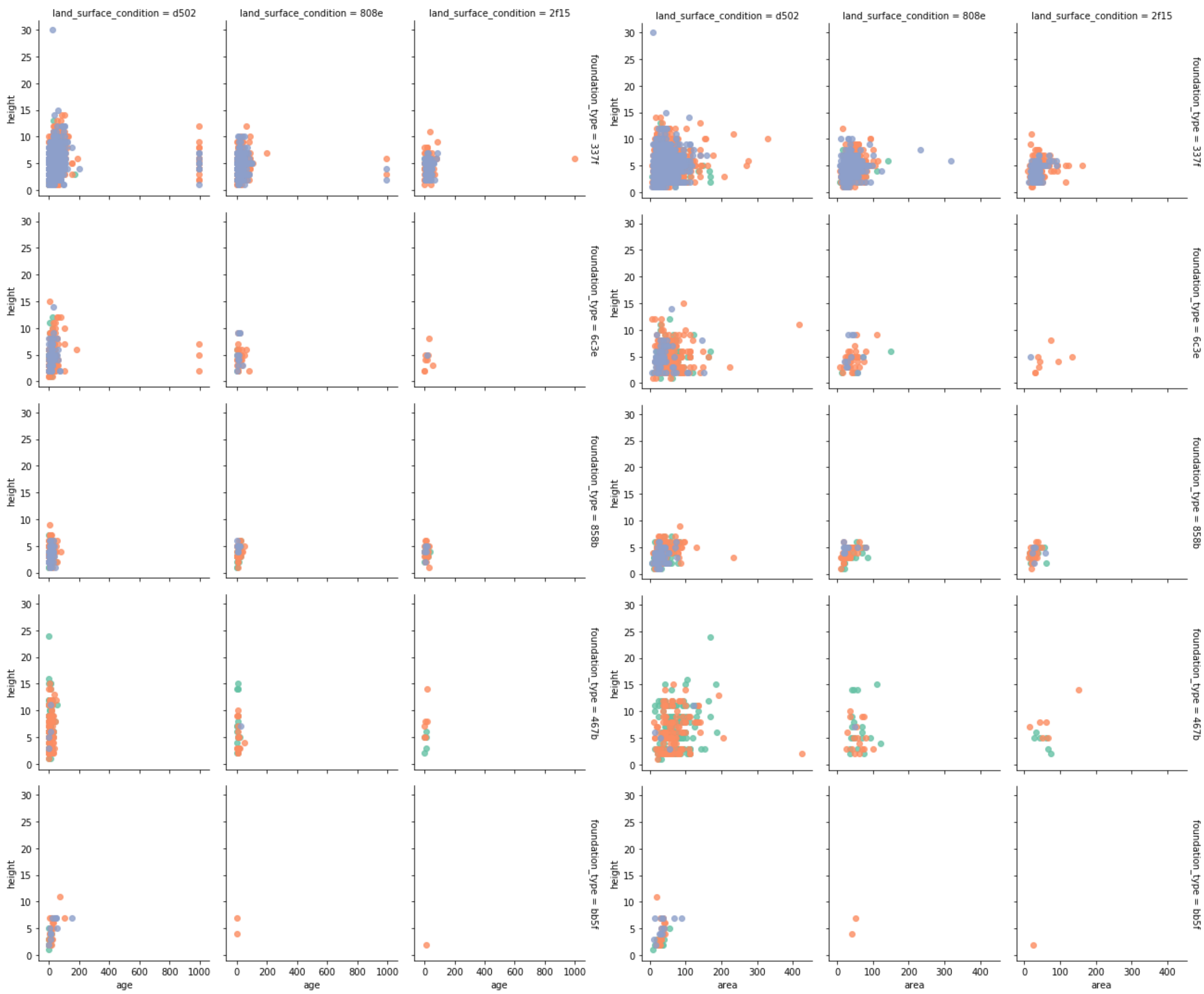
	has_superstructure_rc_non_engineered	has_superstructure_rc_engineered	has_secondary_use	has_secondary_use_agriculture	foundation_type-337f	foundation_type-467b	roof_type-67f9	ground_floor_type-467b	other_floor_type-441a	other_floor_type-67f9
geo_level_2_id	-0.057156463	-0.094260241	-0.0307938	0.022759979	0.19443943	-0.128600123	-0.175205092	-0.200115409	-0.040786391	-0.142559636
geo_level_3_id	-0.048689249	-0.085546237	-0.031280681	0.019316486	0.178281891	-0.122336184	-0.160517268	-0.181077215	-0.038768699	-0.130252786
count_floors_pre_eq	0.014082725	0.052101714	0.052241153	0.000146845	0.134488408	0.039057846	0.028528213	-0.079958306	-0.652418051	0.168954359
height	0.094436123	0.133340758	0.079551914	-0.021270105	0.00495806	0.161826482	0.166192251	0.062683653	-0.520932979	0.303721028
has_superstructure_mud_mortar_stone	-0.220728217	-0.207985064	-0.071143512	0.057944435	0.532385375	-0.343606048	-0.422184634	-0.461016076	-0.202189864	-0.347012776
has_superstructure_cement_mortar_brick	0.12791509	0.115679924	0.056102903	-0.061248681	-0.38165555	0.238468891	0.424382432	0.549397768	0.22763091	0.330466741
has_superstructure_rc_non_engineered	1	0.010848358	0.091126648	-0.022242239	-0.288620767	0.509955566	0.458495293	0.377450679	0.054596566	0.400175991
has_superstructure_rc_engineered	0.010848358	1	0.093707472	-0.028354222	-0.270810256	0.538390709	0.455141232	0.345602227	0.044112295	0.396581537
has_secondary_use	0.091126648	0.093707472	1	0.769587395	-0.095039989	0.167451754	0.140531715	0.12622735	-0.030155738	0.167484172
has_secondary_use_agriculture	-0.022242239	-0.028354222	0.769587395	1	0.034202836	-0.040146579	-0.051212104	-0.061992887	-0.05111846	-0.035881414
foundation_type-337f	-0.288620767	-0.270810256	-0.095039989	0.034202836	1	-0.465907245	-0.460883179	-0.500777644	-0.2003663	-0.39800298
foundation_type-467b	0.509955566	0.538390709	0.167451754	-0.040146579	-0.465907245	1	0.711442447	0.578669394	0.081995737	0.621438332
roof_type-67f9	0.458495293	0.455141232	0.140531715	-0.051212104	-0.460883179	0.711442447	1	0.667232859	0.118099668	0.710530471
ground_floor_type-467b	0.377450679	0.345602227	0.12622735	-0.061992887	-0.500777644	0.578669394	0.667232859	1	0.206478195	0.545970633
other_floor_type-441a	0.054596566	0.044112295	-0.030155738	-0.05111846	-0.2003663	0.081995737	0.118099668	0.206478195	1	-0.087273453
other_floor_type-67f9	0.400175991	0.396581537	0.167484172	-0.035881414	-0.39800298	0.621438332	0.710530471	0.545970633	-0.087273453	1

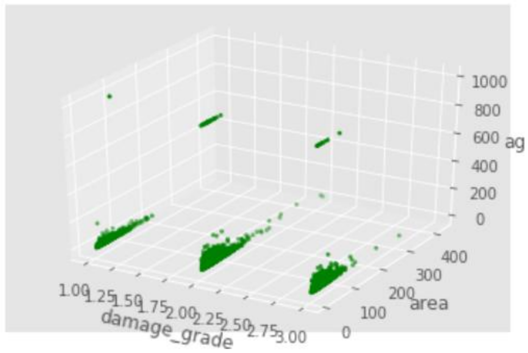
- There was a positive relation between has_superstructure_mud_mortar_stone and foundation_type-337f
- There was a positive relation between has_superstructure_cement_mortar_brick and ground_floor_type-467b
- There was a positive relation between has_superstructure_rc_non_engineered and foundation_type-467b
- There was a positive relation between has_superstructure_rc_engineered and foundation_type-467b

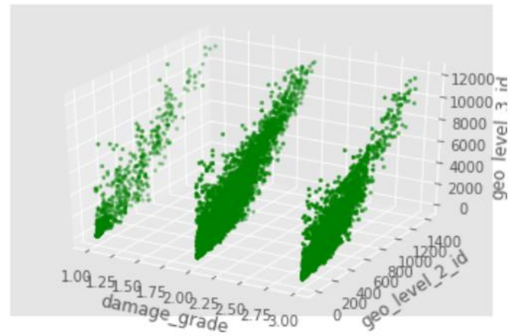
These correlations validate the plots by showing a negative correlation between other_floor_type-441a and height, other_floor_type-441a and count_floors_pre_eq, ground_floor_type-467b and foundation_type-337f and moderate to strong positive correlations for the other numeric features as shown in the above correlation matrix

Multi-faceted Relationships

Apparent relationships between damage grade and individual features are helpful in determining predictive heuristics. However, relationships are often more complex, and may only become apparent when multiple features are considered in combination with one another. To help identify these more complex relationships, some faceted plots were created. Categorical columns such as foundation_type, land_surface_condition etc. were converted to indicator columns and plotted







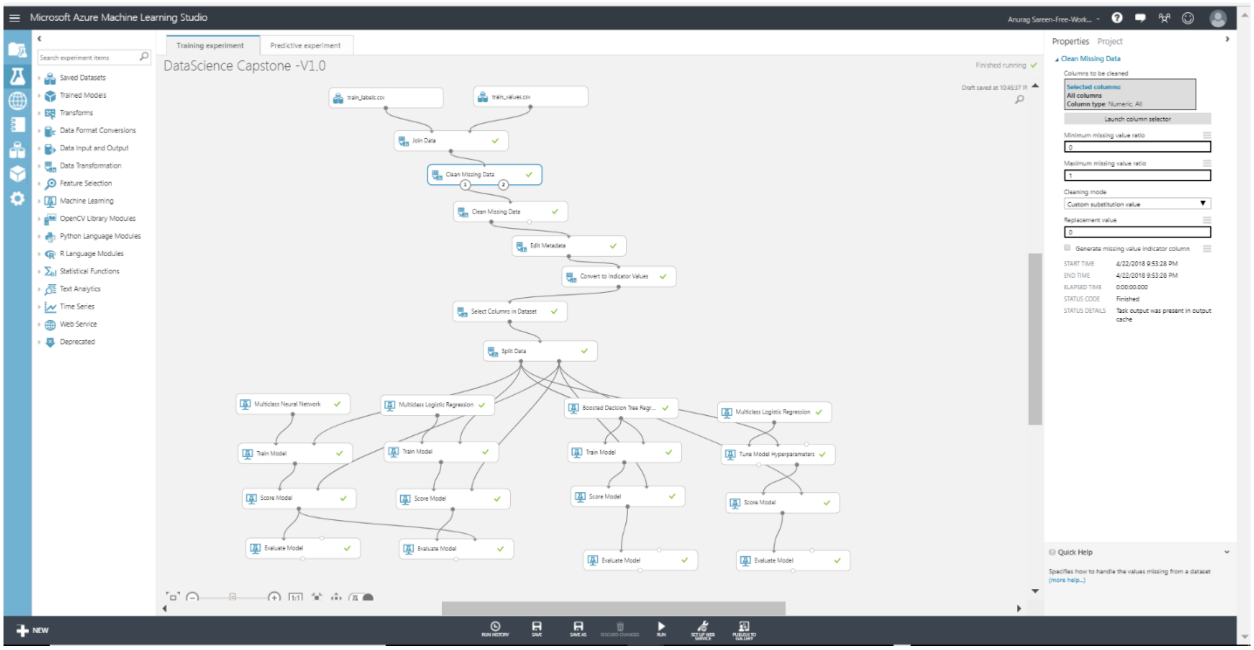
3d plots were also created to show relationship between various features. From these plots, it can be seen that damage was mostly done to buildings with small area and newer buildings. It was also observed that most of the damage was done to regions with lower geo level ids

Classification of Buildings Based on Damage Grade

Multiclass Logistic Regression with Tune Model Hyperparameters

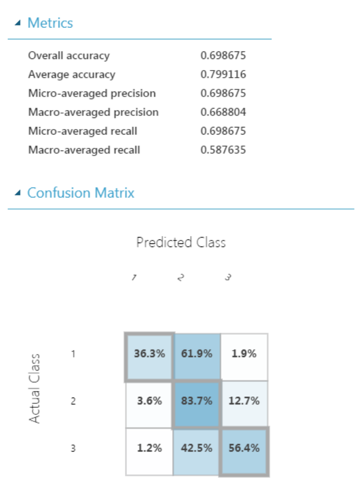
Training Experiment

Various Classification and Regression Models, like Multiclass Neural Network, Multiclass Logistic Regression, Boosted Decision Tree Regression etc. were used to train the model.



The models were compared for accuracy and following metrics and confusion matrix was generated for the most accurate model – Multiclass Logistic Regression with Tune Model Hyperparameters.

The model was created using the Multiclass Logistic Regression with Tune Model Hyperparameters and trained with 65% of the data. Testing the model with the remaining 35% of the data yielded the following results:



Metrics

Overall accuracy	0.698675
Average accuracy	0.799116
Micro-averaged precision	0.698675
Macro-averaged precision	0.668804
Micro-averaged recall	0.698675
Macro-averaged recall	0.587635

Confusion Matrix

Actual Class \ Predicted Class	1	2	3
1	36.3%	61.9%	1.9%
2	3.6%	83.7%	12.7%
3	1.2%	42.5%	56.4%

The Confusion Matrix shows the various percentages of accurately predicted damage grade 1, 2 and 3. The model accurately predicted 36.3% of damage grade 1, 83.7% of damage grade 2 and 56.4% of damage grade 3

Predictive Experiment

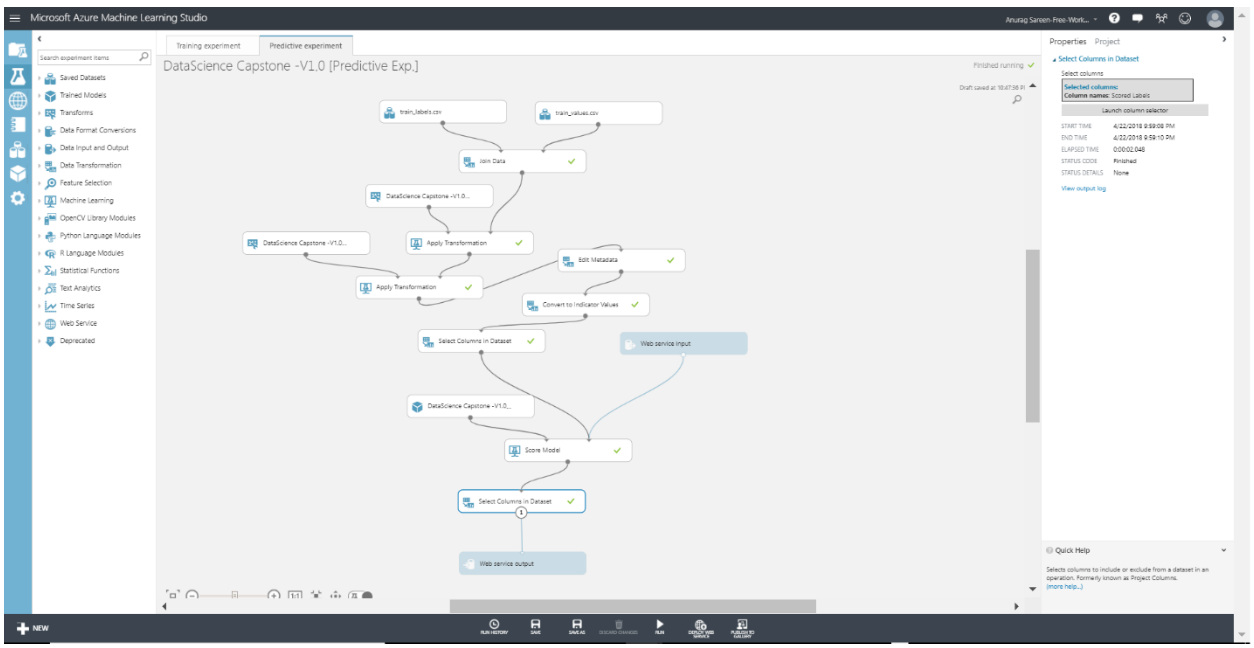
A webservice with the following input parameters was created

Input columns:

Column names:
building_id,geo_level_1_id,geo_level_2_id,geo_level_3_id,count_floors_pre_eq,age,area,height,has_superstructure_adobe_mud,has_superstructure_mud_mortar_stone,has_superstructure_stone_flag,has_superstructure_cement_mortar_stone,has_superstructure_mud_mortar_brick,has_superstructure_cement_mortar_brick,has_superstructure_timber,has_superstructure_bamboo,has_superstructure_rc_non_engineered,has_superstructure_rc_engineered,has_superstructure_other,count_families,has_secondary_use,has_secondary_use_agriculture,has_secondary_use_hotel,has_secondary_use_rental,has_secondary_use_institution,has_secondary_use_school,has_secondary_use_industry,has_secondary_use_health_post,has_secondary_use_gov_office,has_secondary_use_use_police,has_secondary_use_other,land_surface_condition-2f15,land_surface_condition-808e,land_surface_condition-d502,foundation_type-337f,foundation_type-467b,foundation_type-6c3e,foundation_type-858b,foundation_type-bb5f,roof_type-67f9,roof_type-7e76,roof_type-e0e2,ground_floor_type-467b,ground_floor_type-b1b4,ground_floor_type-b440,ground_floor_type-bb5f,ground_floor_type-e26c,other_floor_type-441a,other_floor_type-67f9,other_floor_type-9eb0,other_floor_type-f962,position-1787,position-3356,position-bcab,position-bfba,plan_configuration-0448,plan_configuration-1442,plan_configuration-3fee,plan_configuration-6e81,plan_configuration-84cf,plan_configuration-8e3f,plan_configuration-a779,plan_configuration-cb88,plan_configuration-d2d9,legal_ownership_status-ab03,legal_ownership_status-bb5f,legal_ownership_status-c8e1,legal_ownership_status-cae1

Output columns:

Column names: Scored Labels



The webservice was used to predict the damage grade level based on input feature set. The best score in completion was 0.7001.

EVALUATION METRIC

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

The metric used for this competition is the micro-averaged F1 score.

Conclusion

This analysis has shown that the risk of damage to buildings can be confidently predicted from its characteristics. In particular, the geographic region in which the building exists, the age of the building, the plinth area of the building, the height of the building, position of the building, foundation type, ground floor type, other floor type, non-engineered/engineered reinforced concrete have a significant effect on the risk of buildings getting damaged due to earthquake.