# Executive Summary

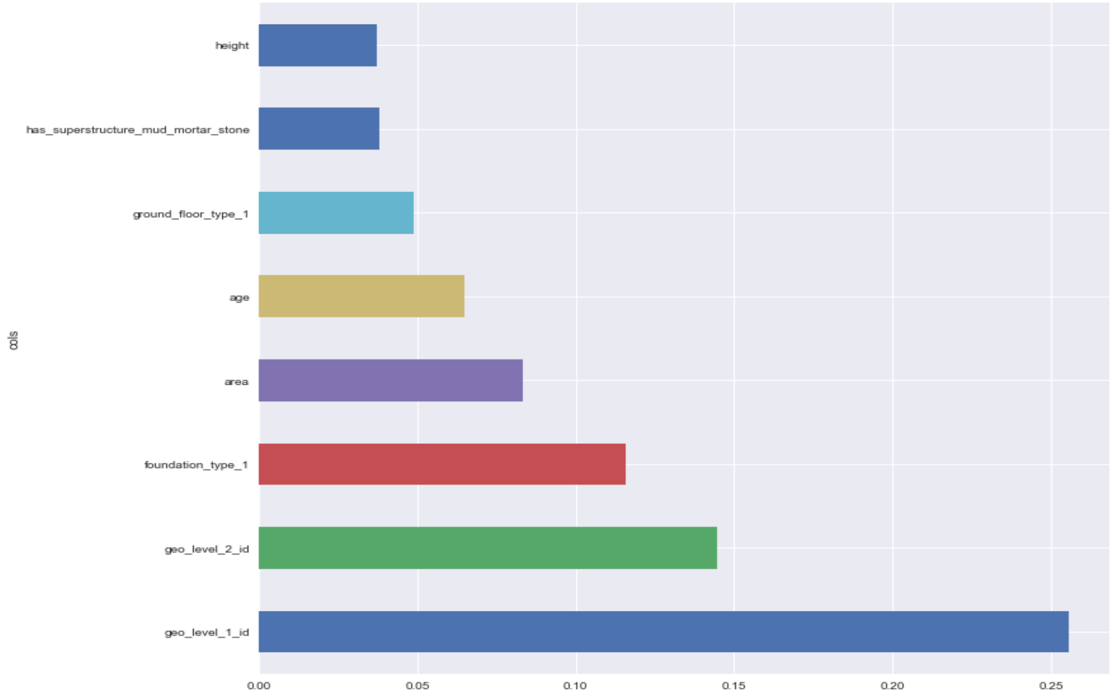
This report provides an analysis and recommendations based on the data related to the 2015 Gorkha Earthquake in Nepal. The data contained 20,000 samples collected by the Central Bureau of Statistics pertaining to the location and various attributes of structures.

Methods of analysis include miscellaneous data preparation and sci-kit learn's logistic regression. Sci-kit learns's random forest classifier was also used as a secondary analysis tool to aid in analysis, visualization, and specific recommendations.

Using the random forest classifier, the most important features in determining the damage grade of a structure from most to least important are:

- geographic region as determined by geo_level_1_id and geo_level_2 _id
- foundation type
- Area
- Age
- ground floor type
- superstructure type
- Height

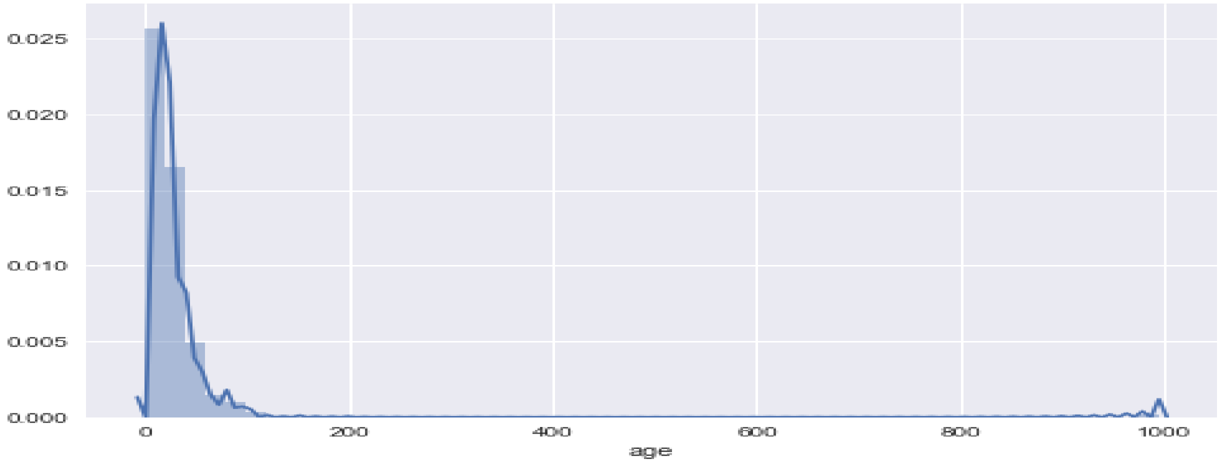
The graphs of these findings is shown below.



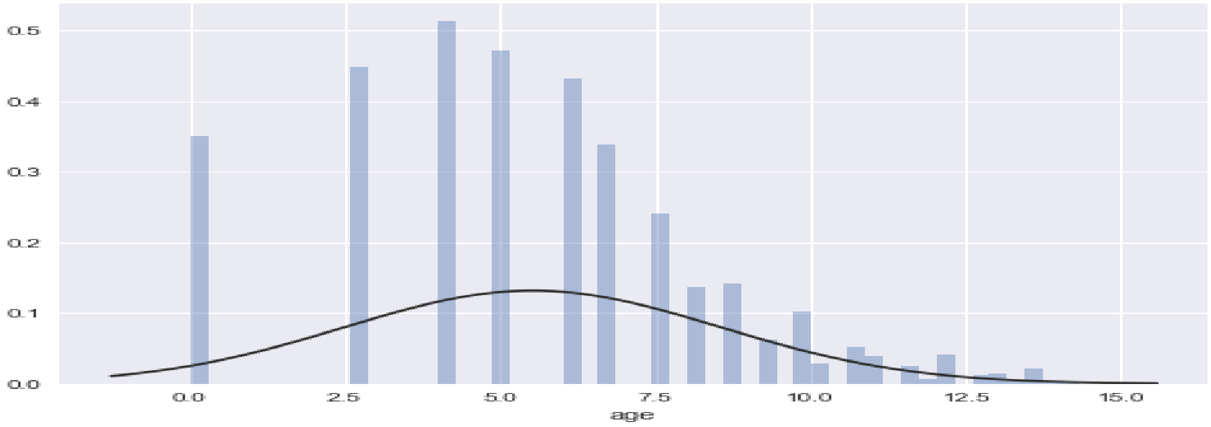
The report concludes that risk of earthquake damage of an individual building can be predicted with greater than 70% accuracy using the data contained in the data set and simple logistic regression.  Using random forest classification as outlined in the classification and recommendation section we can also determine the primary causes for a given damage grade prediction and make recommendations to mitigate the risk of a specific building and assist triage.

## Data

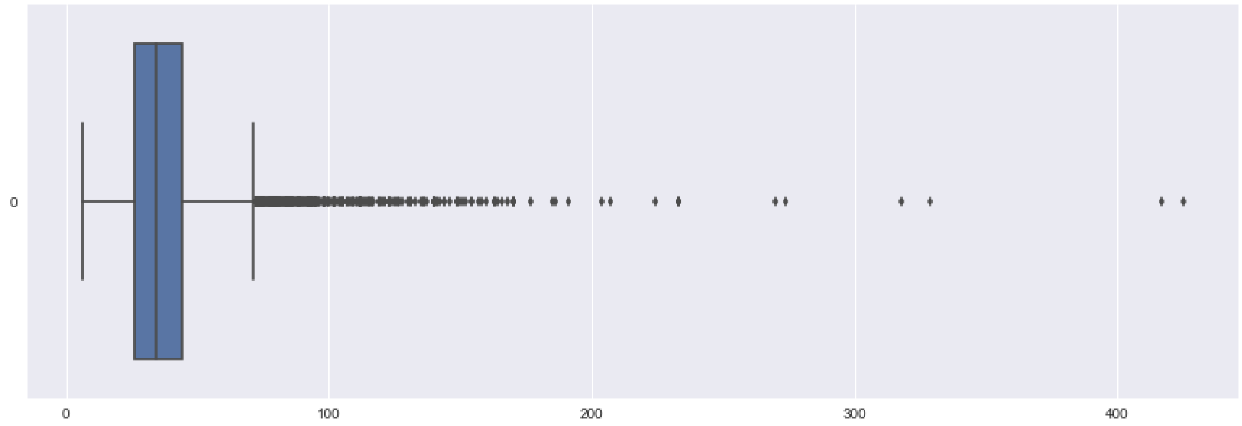
Initial data analysis revealed a few interesting features. The first action taken was to drop building_id, since it is random, and geo_level_3_id because most of the values apply to  only a few rows of data and could cause overfitting. Age has a maximum value of 995 years and a median of 15 years creating a highly skewed feature, as illustrated below.

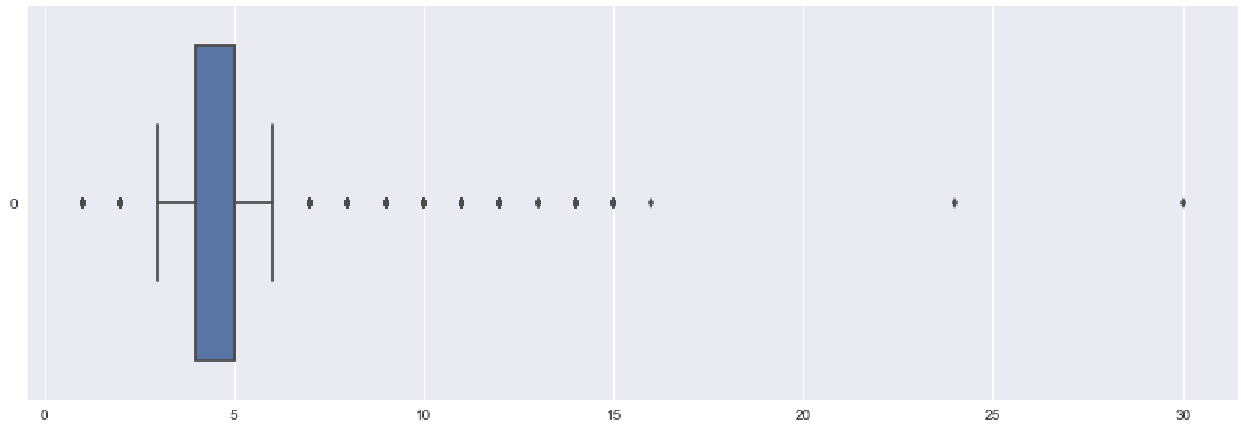


Ages above 110 years were treated as outliers and an age of 65, which was determined to have a similar damage grade distribution, was substituted. After normalization with Box-Cox, the distribution was much closer to normal, as illustrated below.



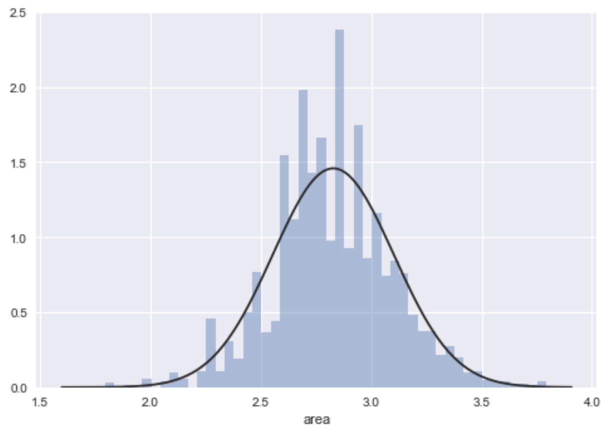
Other features with significant outliers included area(top) and height(bottom).

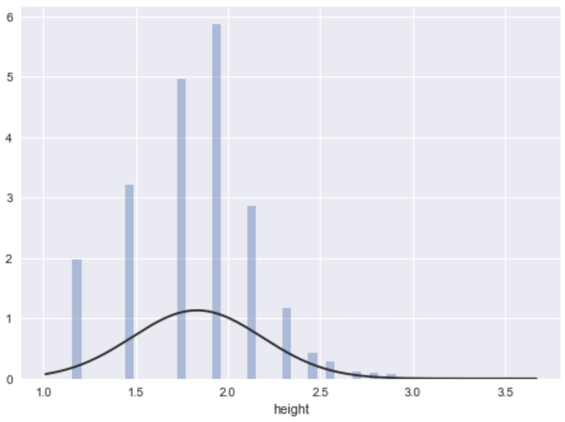




Area was fairly straightforward, all areas greater than 200 were assigned a value of 200.

For height, all values greater than 20 were assigned a value of 20. All values equal to one were considered to be errors due to the fact that an average adult would not be able to stand inside. These values were assigned a value of 2. Box-Cox normalization was used to further normalize these features as shown below.

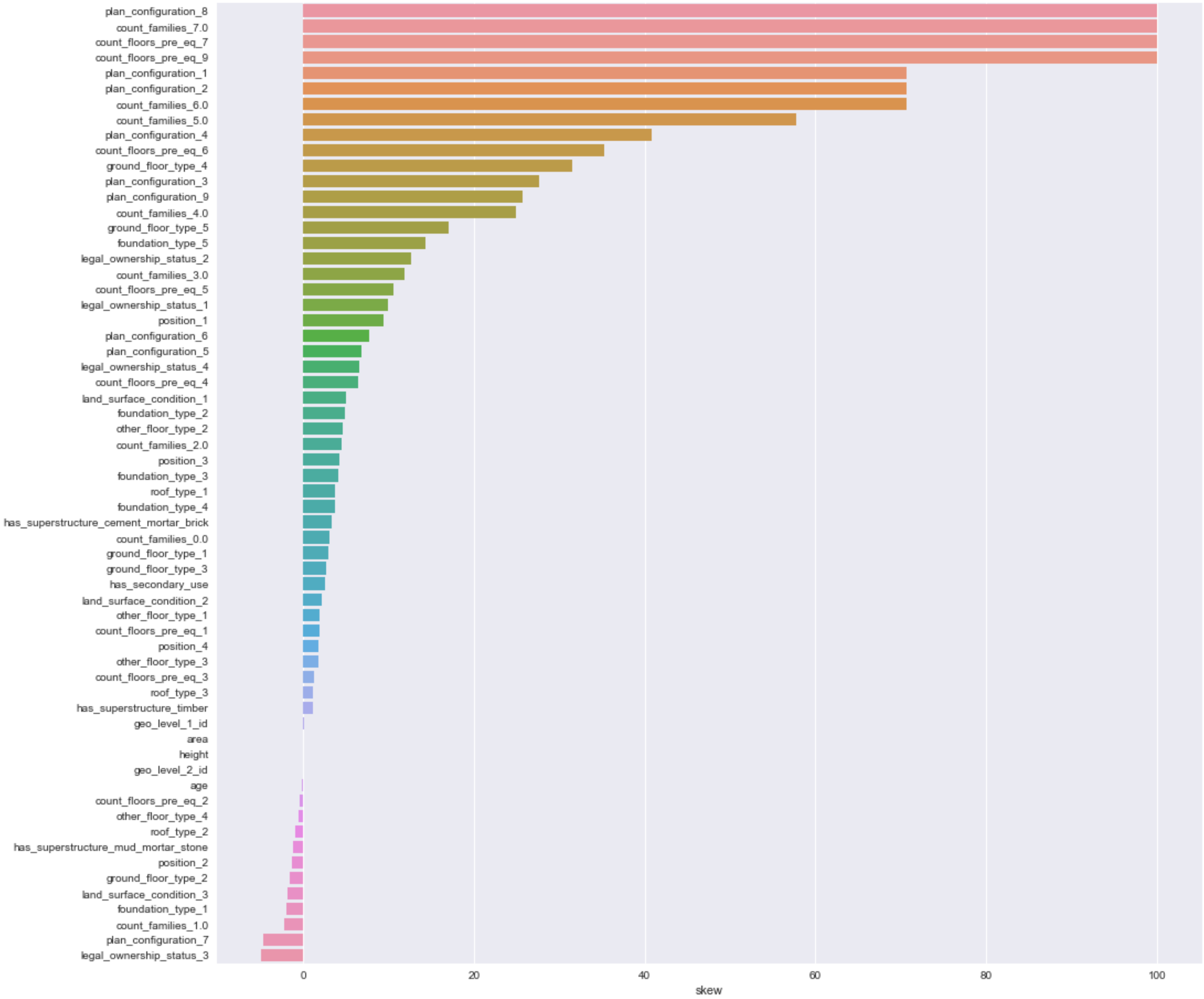




No null values were found.  No outliers were found using statsmodels' OrdinaryLeastSquares.

All object data types were converted to int8 data type.

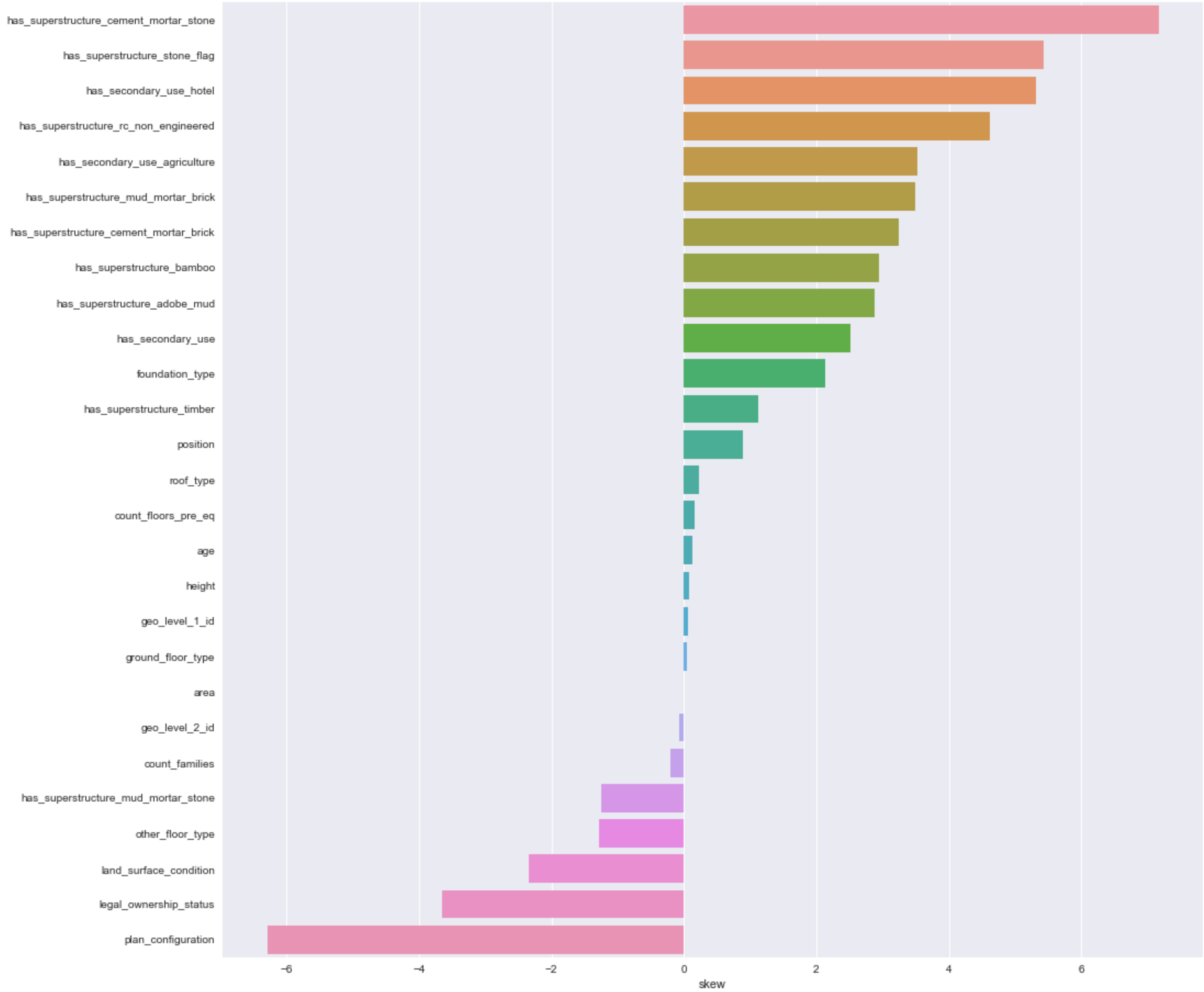
There were several features that were very skewed. Below is a chart representing the skew as determined by Scipy's skew function.



The following features were found to contain greater than 95% of the same value and were dropped to prevent overfitting:

has_secondary_use_use_police ,has_secondary_use_gov_office ,has_secondary_use_health_post , has_secondary_use_school ,has_secondary_use_institution ,has_secondary_use_industry , has_secondary_use_other ,has_secondary_use_rental, has_superstructure_rc_engineered , has_superstructure_other.

Box-Cox normalization was used on the remaining features to further reduce skew, as shown below. Note that the highest levels are now near seven, while prior to dropping the sparse features and normalization the highest levels were near 100. Also note that the features age, area, and height have very little skew.



The next step taken was to create dummy features from features whose values are not ordinal and/or the values do not have any meaning in relation to other values. The features that met these criteria were:

geo_level_1_id, geo_level_2_id, count_floors_pre_eq, count_families, land_surface_condition, foundation_type, roof_type, ground_floor_type, other_floor_type, position, plan_configuration, legal_ownership_status

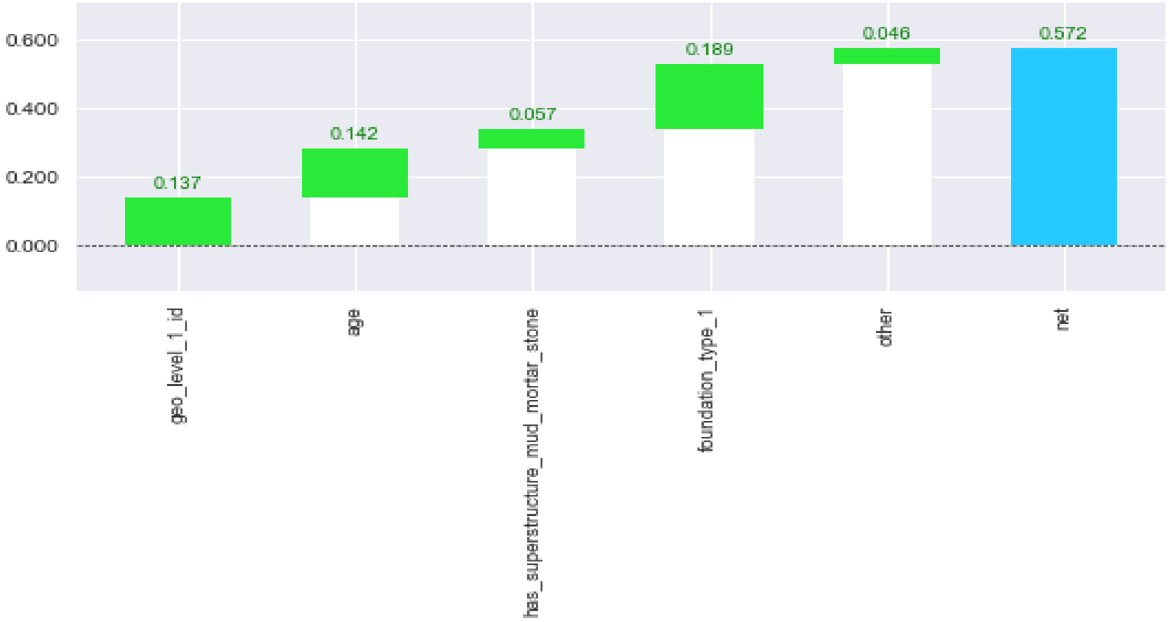
The result of this was a data set of 1339 features. At this point the data is very sparse but the data that remains is important for our models.

## Classification and Recommendations

Logistic regression was used to predict the damage grade. This achieved an f1 score of .7046 on the test data. Additionally, Logistic Regression is not compute-intensive, so models can be retrained quickly whenever new data is available. This model takes about a minute to train on a second generation I5 laptop.

Random Forest Classifier was used to create the recommendations and visualizations that can be used to evaluate individual structures. The following are examples of the feature contributions for specific structures from each damage grade.

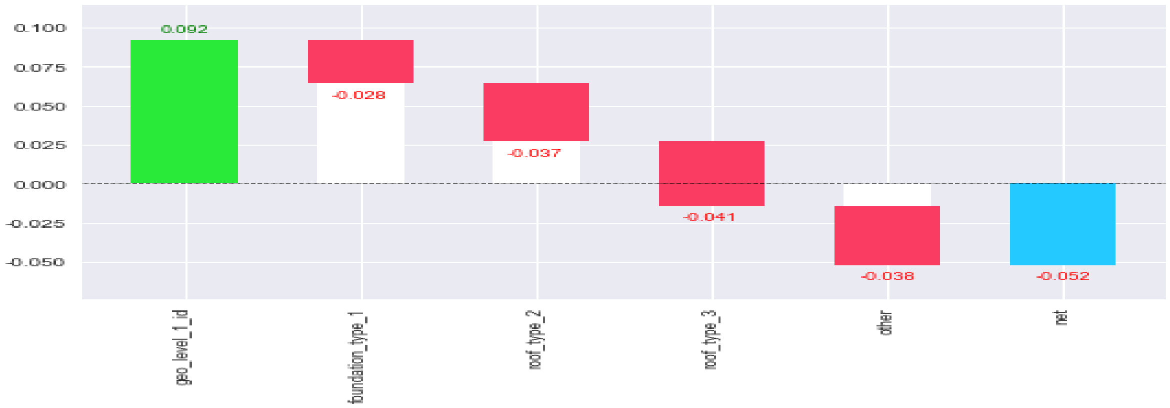
### Damage Grade 1, Building 4218:



This example shows building 4218 which predicted to be damage grade 1. Looking at this building we can see that the geo_1_id is 1, this must have been a lesser impacted region as this contributed to a lower damage grade. The age of this building was 0 years old, a new building, again contributing to the lower damage grade. It does not have a mud mortar or stone superstructure, this is a binary feature, so it can contribute or take away from a value. It does not have foundation type 1(337f), another negative feature. Other is the sum of the remaining features.

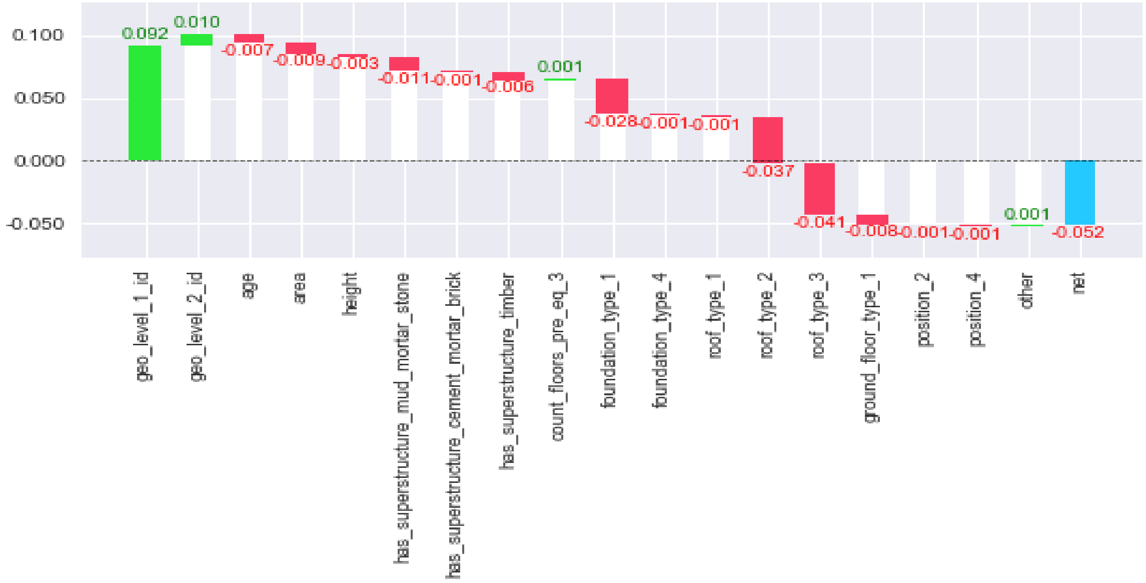
This report would not recommend any changes be made to building 4218 to prevent damage due to a similar earthquake. It should be considered low risk for triage.

Damage Grade 2, Building 7484:

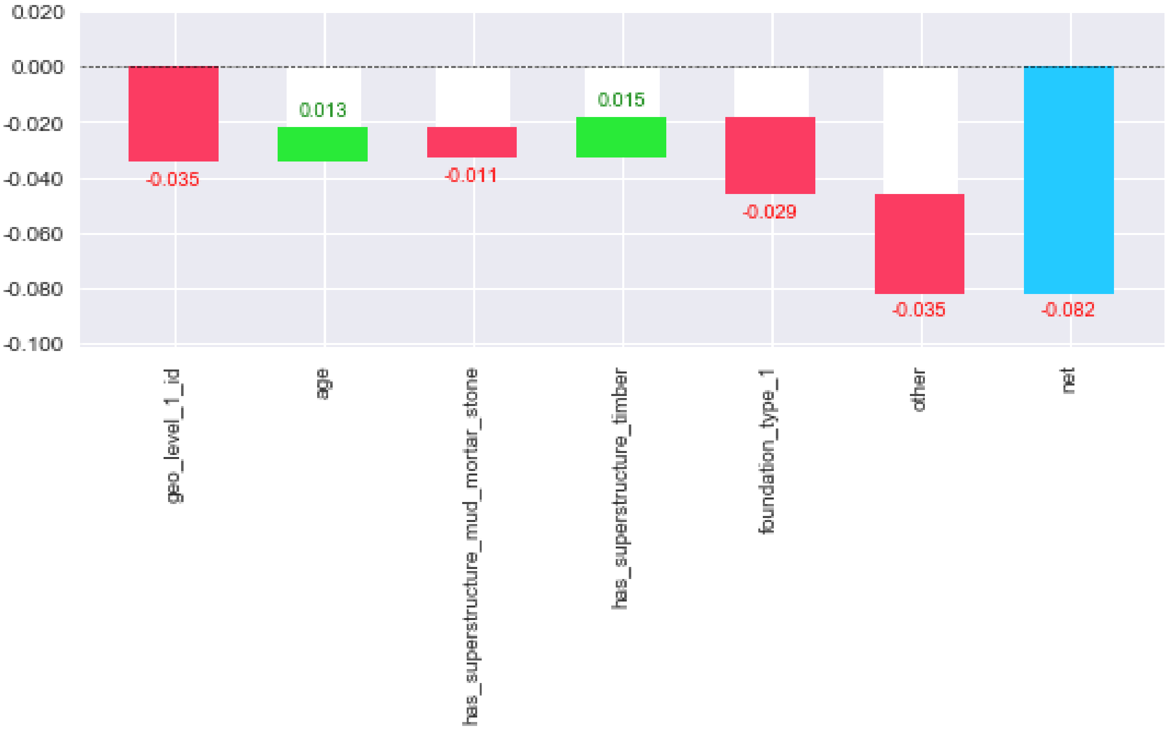


Building 7484 is interesting because it is in the geo_level_1_id of 10 which appears to be lightly hit region, yet the damage grade was 2. The primary contributing factors to its predicted damage grade of 2 are its foundation type_1 which is type 337f, it does not have roof type 2, but it does have a roof type of 3, e0e2. It also has several smaller contributions that add up to a significant increase in damage grade. A more detailed report, as shown below, reveals several features that make up the "other" category.

This report would recommend these structural issues be addressed to reduce the damage grade of 2 despite its low risk geo level. It should be considered a medium priority for triage.



Damage Grade 3, Building 6668:



Building 6668 was a damage grade 3 building yet it had some positive features. The geo_level_1_id is 5, a very large contributor to the damage grade. The age of the building was 5, which was a plus. Contrary to building 4218 the superstructure is mud, mortar, or stone so it takes a hit, but the fact that the superstructure also contain timber negates this feature. Again, we have foundation type 1(337f) which greatly contributes to this buildings damage grade and a primary reason for the damage grade of building 7484 despite its advantageous geographic region.

This report would recommend a closer look at all buildings in this geo level of 5 and all buildings with foundation type 337f should be considered at risk and a priority for triage.

## Conclusion

This analysis has successfully demonstrated not only the ability to quickly predict the damage grade using logistic regression but also to make specific recommendations as to why a building is categorized in specific damage grade using random forest classifier. Specific actions can be taken to reduce the risk based on geographic region and/or structural issues for at risk buildings. Low risk buildings can be categorized as low risk during initial triage and high-risk buildings can be prioritized.