Architectural Decisions Document (ADD)

The Lightweight IBM Cloud Garage Method for Data Science

Project:

Predict the level of building damage caused by the Nepal Earthquake of 2015

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

The datasets have been downloaded from <https://www.kaggle.com/mullerismail/richters-predictor-modeling-earthquake-damage>

The goal is to predict the level of damage to buildings caused by the 2015 Gorkha earthquake in Nepal. The ordinal variable damage\_grade, which represents a level of damage to the building that was hit by the earthquake. There are 3 grades of the damage:

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

### Justification

The dataset was collected through surveys by the Central Bureau of Statistics that work under the National Planning Commission Secretariat of Nepal. This survey is one of the largest post-disaster datasets ever collected, containing valuable information on earthquake impacts, household conditions, and socio-economic-demographic statistics.

## Enterprise Data

### Technology Choice

We will use Jupyter Notebooks on IBM Watson to do data exploration and run models.

### Justification

Jupyter Notebooks allow both code and documentation and can be shared with others.

## Streaming analytics

### Technology Choice

The static data set is provided from the source that does not require real-time streaming.

### Justification

Not applicable.

## Data Integration

### Technology Choice

No data integration applied.

### Justification

The data set used for this project comes from two csv files from the kaggle site. All available information can be obtained by joining the two data frames on the common column, building\_id.

## Data Repository

### Technology Choice

The downloaded csv files are uploaded to IBM Cloud Object Storage connected to the Watson Studio project.

### Justification

The Watson Studio on IBM Cloud provides all the functions required for notebook creation, data storage and PySpark processing.

## Discovery and Exploration

### Technology Choice

An ETL notebook was created to explore the data using with Pandas data frames and MatplotLib and Seaborn libraries. I will use matplotlib.pyplot, seaborn and other Python libraries to visualize the features of the buildings affected by the earthquake and test data distributions and interdependecies between variables. Plots and graphs used include:

* Histograms
* Boxplots
* Correlation heatmaps

### Justification

Python and its libraries (pandas, matplotlib, seaborn) provide many tools for data exploration and visualization.

## Actionable Insights

### Technology Choice

Features Engineering

Use binning on column, age (age of building) as values range from values range from 1 to 995. Binning helps to remove outlier effect and impact on model.

Extract top 25 features and see if prediction accuracy increases.

Model Building

Use two supervised classification algorithms, Logistics Regression and Random Forest and a Deep Learning classification algorithm, Feedforward Neural Networks.

Since this is a Multi-class classification we’ll use a MulticlassClassificationEvaluator that will evaluate the predictions using the f1 metric, which is a weighted average of precision and recall scores.

### Justification

* The PySpark Machine Learning libraries (MLlib and ML Packages) provide a host of algorithms that can be used for model building, testing and evaluation.
* Three different classification algorithms (two supervised and 1 deep learning) are built to compare performance. Eventually logistic regression with class weighing (to handle imbalance) and hyperparameter tuning gives the best F1-score.
* The top 25 features did not increase prediction accuracy so the full feature set is retained in the model building.
* F1 metric is used as the model performance indicator as it balances between precision and recall scores and focuses on reducing both false positives and false negatives.

## Applications / Data Products

### Technology Choice

The results of these project will be as follows.

* Submission to Kaggle.
* PPT presentation with modeling results
* Video presentation of modeling results
* Jupyter Notebooks

### Justification

This is a academic exercise so no actual deployment is needed.

## Security, Information Governance and Systems Management

### Technology Choice

Not applicable.

### Justification

Not applicable.