Post-fire Debris Flow Likelihood Prediction

Team Members: Alejandro Hohmann, Bhanu Muvva, Chunxia Tong
Jacobs School of Engineering, University of California, San Diego
ahohmann@ucsd.edu, bmuvva@ucsd.edu, chtong@ucsd.edu

Advisors: Daniel Roten, Ilkay Altintas
San Diego Supercomputer Center
d1roten@ucsd.edu, ialtintas@ucsd.edu

June 8, 2023

Abstract

Debris Flows are a distinct type of landslide that suddenly occur without warning. They are fast-moving channels of water and soil that carry large natural objects like boulders and trees, or human-made objects including cars. In the American West, Debris Flows have directly caused death and property damage. Debris Flows often occur after rain events and the burn scars left behind by wildfires increase their likelihood. Given the increasing frequency of extreme weather events, it is critical to predict Debris Flows and take precautionary action before they occur. This project builds upon prior research of predicting Debris Flows using additional geological features and more advanced machine learning techniques. The project also includes an intuitive interface for decision makers to access these probability estimates.

1 Introduction

Post wildfire flash flooding and debris flows are a realistic threat for homes and communities located within or along a wildland urban interface that has experienced a recent wildfire, which cause significant economic losses and human casualties.

What is a Debris Flow? Before fire and rain, Soil is trapped on steep rocky hills by vegetation, as shown in Figure 1. During summer's fire season, vegetation is burned, causing sediment to roll down steep hills. Within a few hours or days, channel bottoms are loaded with loose sediment, as shown in Figure 2. During an intense rain, the water and runoff move sediment in the steep channels, producing debris flows, as shown in Figure 3.

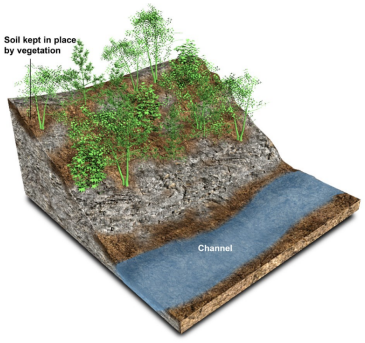


Figure 1: Before fire and rain

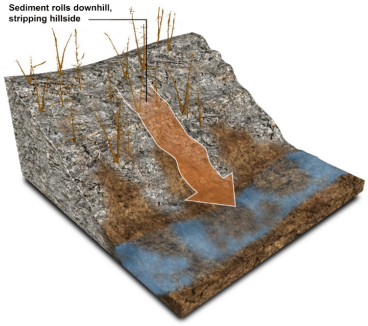


Figure 2: After fire

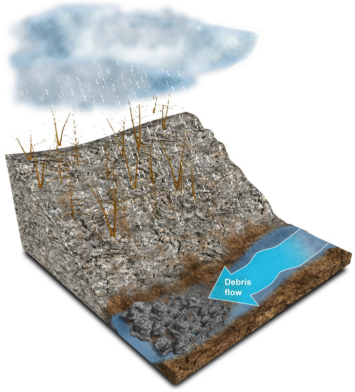


Figure 3: Rain and runoff

It is crucial to plan and prepare for this type of hazard to prevent and reduce the loss of life and property, and to develop community resilience.

Currently, prediction of post-fire debris flows is widely based on the use of power-law thresholds and logistic regression models. While these procedures have served with certain success in existing operational warning systems, in this study we investigate the potential to improve the efficiency of current predictive models with machine learning approaches.

The analysis is based on a database of post-fire debris flows recently published by the United States Geological Survey.

Results show that predictive models based on Neural Network improved performance with respect to the other models examined. In addition, new-feature-based Neural Network models demonstrated improvement in performance with generating new features, indicating a clear advantage regarding their ability to successfully assimilate new information. Complexity, in terms of variables required for developing the predictive models, is deemed important but the choice of model used is shown to have a greater impact on the overall performance.

2 Team Roles and Responsibilities

The team roles and responsibilities are defined separately for the three main aspects of the project, namely the data analysis, the modeling development, and the visualization.

2.1 Data Analysis

The data analysis portion involves revealing the correlation between features and the probability of debris flow. The project manager's responsibilities include leading team meetings, reviewing each team member's progress to ensure the project remains on track, and communicating progress to the project advisor. The solution architect helps guide the vision for using the developed library to analyze the identified datasets. The visualization and dashboard developer is in charge of creating impactful visualizations for communicating findings and designing a dashboard. The data analysis roles are assigned as follows:

- Project manager & Solution Architect: Alejandro Hohmann
- Visualization & Dashboard Developer: Bhanu Muvva
- Record Keeper: Chunxia Tong

2.2 Modeling Development

We developed new features and designed new models. The roles are assigned as follows:

- Integration Lead: Alejandro Hohmann
- New Features and Modeling: Bhanu Muvva
- Model adjusting: Chunxia Tong

2.3 Visualization

We created a visualization tool that integrated the prediction results with a map. This tool allows users to input rainfall data and view the corresponding landslide probability for any selected location on the map by simply moving the cursor.

- Data Integration.: Alejandro Hohmann
- Website Interface Design: Bhanu Muvva
- Website User Test: Chunxia Tong

3 Data Acquisition

The full data acquisition process involves identifying relevant data sources, collecting the data, designing the data pipeline, and implementing the data environment to realize the pipeline.

3.1 Data Sources

The data sources identified for this project include:

- The Staley et al. (2016) model
- The LANDFIRE Scott and Burgan fire behavior model
- The geological maps of US states:
- The USGS 3D elevation program:

A summary of the data sources is provided below in Table 1 and they are discussed further in the subsequent sections.

Table 1: Data Sources

Dataset Name	Source	Destination	Acquisition Notebooks, Code, Documents	Data Size	Other Notes, e.g., Confidentiality, Notes from data provider. Etc.
Staley et al. (2016)	https://pubs.er.usgs.gov/publication/ofr20161106	PostgreSQL DB	download using requests library and store to DB	275KB	rain and debris flow data for areas spanning seven states with a history of wildfires
LANDFIRE Scott and Burgan fire behavior model	https://landfire.gov/fbfm40.php	PostgreSQL DB	download using requests library and store to DB	3KB	used for deriving vegetation and fuel-related features
Geological map of US states	https://mrdata.usgs.gov/geology/state/	PostgreSQL DB	download using requests library and store to DB	415M	used for geographic boundaries of debris flow areas
USGS 3D elevation program	General link: https://www.usgs.gov/3d-elevation-program Parameterized link: https://prd-tnm.s3.amazonaws.com/StagedProducts/Elevation/13/TIFF/current/n35w119/USGS_13_n35w119.tif	PostgreSQL DB	download using requests library and store to DB	500MB	elevation data parameterized to include ⅓ arc-seconds (~10 meters) resolution

3.2 Data Collection

Once the relevant data sources are identified, the data is stored for continued access to support data exploration, model training, and visualization.

3.2.1 Features

The original features in Staley et al. (2016) is shown in Table 2.

Table 2: Features List

Feature Name	Description
Fire Name	Name of wildfire
Year	Year of wildfire occurrence
Fire_ID	Abbreviation of fire name
Fire_SegID	Concatenated fire abbreviation and unique segment ID generated during processing
Database	Database type: "Training" indicates data used to calibrate model equation, "Test" indicates data used to test model performance
State	State in which wildfire occurred
UTM_Zone	UTM zone containing majority of wildfire area
UTM_X	UTM X coordinate (Easting, in meters from zone origin)
UTM_Y	UTM Y coordinate (Northing, in meters from zone origin)
Response	Field-verified hydrologic response. 0 = no debris flow. 1 = debris flow
StormDate	Date of storm that produced the debris-flow response (in YYYY-MM-DD format)
GaugeDist_m	Distance (in meters) from rain gauge to documented response location
StormStart	Date and time (24-hour format, GMT) that storm began (in YYYY-MM-DD HH:MM format)
StormEnd	Date and time (24-hour format, GMT) that storm ended (in YYYY-MM-DD HH:MM format)
StormDur_H	Total duration of storm, in hours
StormAccum_mm	Total rainfall accumulation of storm, in millimeters
StormAvgI_mm/h	Average storm intensity, in milllimeters per hour
Peak_I15_mm/h	Peak 15-minute rainfall intensity of storm, in millimeters per hour
Peak_I30_mm/h	Peak 30-minute rainfall intensity of storm, in millimeters per hour
Peak_I60_mm/h	Peak 60-minute rainfall intensity of storm, in millimeters per hour
ContributingArea_km2	Contributing area of observation location, in square kilometers
PropHM23	Proportion of watershed burned at high or moderate severity and with gradients in excess of 23 degrees
dNBR/1000	Average differenced normalized burn ratio (dNBR) of watershed, divided by 1000
KF	Average KF-Factor (erodibility index of the fine fragments of the soil) of the watershed
Acc015_mm	Peak 15-minute rainfall accumulation of storm, in millimeters
Acc030_mm	Peak 30-minute rainfall accumulation of storm, in millimeters
Acc060_mm	Peak 60-minute rainfall accumulation of storm, in millimeters

3.2.2 New Features

Table 3: New Features List

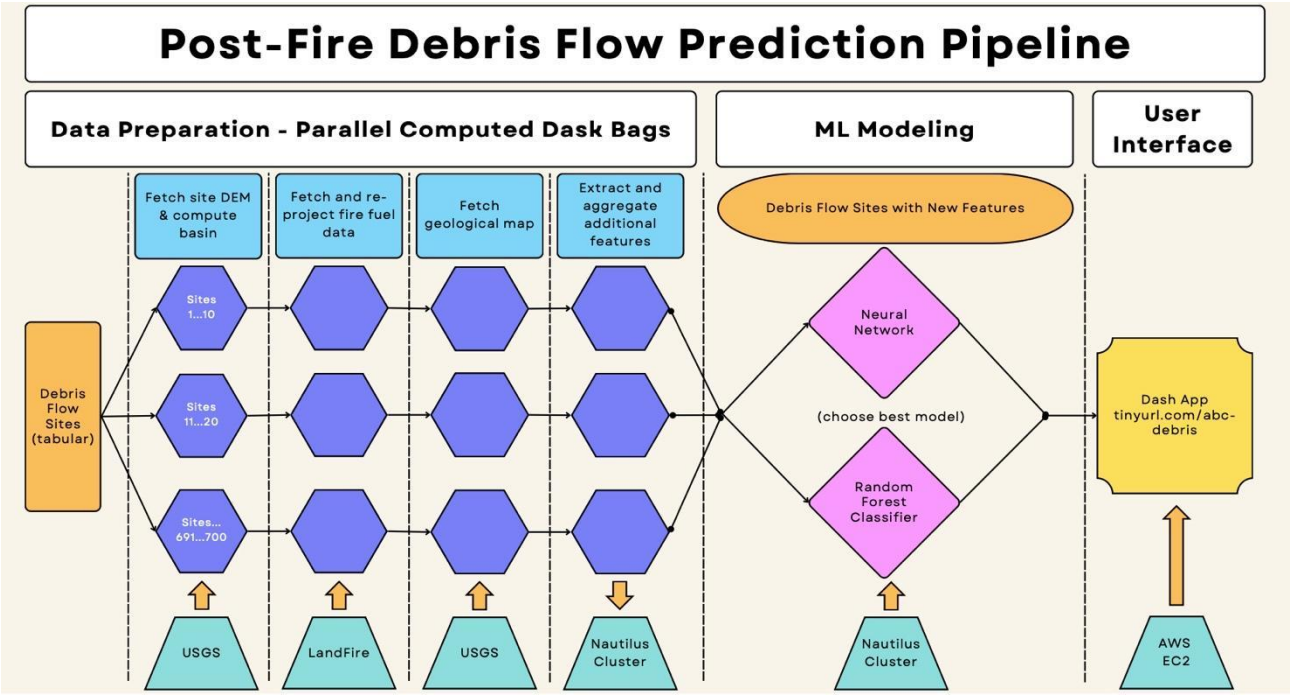
New Feature Name	Description
Site ID	Retrieves Staley data and adds site IDs. The result is stored in the Parquet file `staley16_debrisflow.parquet`.
Contributing region	Computes catchment area and fuel related features for each site. Reads `staley16_debrisflow.parquet` and writes `staley16_observations_catchment_fuelpars_v3.parquet` as well as `staley16_sites_catchment_fuelpars_v3.parquet`.
Rock type	Extracts rock types found within catchment area from geological map, and aggregates to fraction of Igneous, Metamorphic, Sedimentary and Unconsolidated rock types making up each catchment. Writes `staley16_observations_catchment_fuelpars_rocktype_v3.parquet`
Randomize storm data	Adds random noise to rain data due to measurements being up to 4km away from debris flow site.
Generate shape	Regenerated from geological map.
LNDS_RISKS	FEMA assessed Landslide risk data
Fire interval	Time between Last Fire and Debris flow

3.3 Data Pipeline

The data pipeline is illustrated in Figure 4 below. It shows description of the needs, approach, and data access and refresh frequency.

- The target variable of whether there is/isn't a debris flow is not something that is currently captured on a dynamic basis. Because of this, our data is static in nature. While still a big data problem, this simplifies our pipeline requirements because we can proceed with an initial query and create/aggregate the data to our needs without having to continually refresh.
- Python libraries needed include:
  - GeoPandas (for geospatial data)
  - xarray
  - dask (for parallelism)
  -

Figure 4: Data Pipeline



3.4 Data Environment

To ensure data integrity and accessibility, we implemented a distributed file system architecture with redundancy and replication mechanisms. This allowed us to store multiple copies of the data across different geographical regions, ensuring high availability and fault tolerance.

To facilitate efficient data retrieval and analysis, we employed AWS and Nautilus as our data warehousing solution. AWS and Nautilus provided a scalable and performant platform for querying and analyzing large volumes of data.

Throughout the project, we regularly monitored and maintained the data environment to ensure optimal performance and data integrity. This involved performing routine backups and applying patches and updates to the infrastructure.

Programming Language: Python
Cloud Providers: AWS and Nautilus
Python libraries used:

pandas==1.4.2	openpyxl==3.1.2	sklearn==1.0.2
torch==1.13.1	tabula==1.0.5	s3fs==2021.11.0
shapely==1.8.5	geopandas==0.12.2	pyarrow==11.0.0
pysheds==0.3.3	fastfuels==1.0.4	rioxarray==0.13.4
geojsoncontour==0.4.0	pygeos==0.13	plotly==5.14.1
xarray==2022.12.0	dask==2023.4.1	nodejs==16.19.0
dask==2023.4.1	folium==0.14.0	shap==0.41.0
dash==2.9.3	chardet==5.1.0	seaborn==0.12.2
jupyter-dash==0.4.2		

4 Data Preparation

4.1 Data Cleaning

The Staley et al. (2016) model data: https://pubs.er.usgs.gov/publication/ofr20161106

27 features:

Fire Name	Year	Fire_ID	Fire_SegID
Database	State	UTM_Zone	UTM_X
UTM_Y	Response	StormDate	GaugeDist_m
StormStart	StormEnd	StormDur_H	StormAccum_mm
StormAvgI_mm/h	Peak_I15_mm/h	Peak_I30_mm/h	Peak_I60_mm/h
ContributingArea_km2		PropHM23	dNBR/1000
KF	Acc015_mm	Acc030_mm	Acc060_mm

Dimensions before removing null values: (1550, 27)

	Fire Name	Year	Fire_ID	Fire_SegID	Database	State	UTM_Zone	UTM_X	UTM_Y	Response	...	Peak_I15_mm/h	Peak_I30_mm/h
0	Buckweed	2007	bck	bck_1035	Training	CA	11	368133.5165	3823231.989	0	...	3.2	2.0
1	Buckweed	2007	bck	bck_1090	Training	CA	11	367871.0165	3822984.489	0	...	3.2	2.0
2	Buckweed	2007	bck	bck_1570	Training	CA	11	367503.5165	3821741.989	0	...	3.2	2.0
3	Buckweed	2007	bck	bck_235	Training	CA	11	371108.5165	3824991.989	0	...	1.6	1.2
4	Buckweed	2007	bck	bck_363	Training	CA	11	370763.5165	3824576.989	0	...	1.6	1.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1545	Wallow	2011	wlw	wlw_47409	Test	AZ	12	660698.3581	3725248.835	0	...	14.0	8.0
1546	Wallow	2011	wlw	wlw_47535	Test	AZ	12	660178.3581	3725128.835	0	...	63.0	54.0
1547	Wallow	2011	wlw	wlw_47535	Test	AZ	12	660178.3581	3725128.835	0	...	29.0	16.0
1548	Wallow	2011	wlw	wlw_47535	Test	AZ	12	660178.3581	3725128.835	0	...	25.0	16.0
1549	Wallow	2011	wlw	wlw_47535	Test	AZ	12	660178.3581	3725128.835	0	...	14.0	8.0

1550 rows × 27 columns

Dimensions after removing null values: (1091, 27)

4.2 Generate New Features

4.2.1 Add Unique Site Identifier

- Download the raw Staley et al. (2016) Excel file
- Load it into a Pandas dataframe
- Convert all coordinates from UTM (different projections) to WGS84
- Stores the site location as Shapely point and represents the table as Geopandas dataframe
- Add a column with a unique site identifier for each debris flow record
- Save the result as parquet file

4.2.2 Extract Debris Flow Contributing Region

- Use the 1/3 arc second data from the USGS
- Compute the contributing region (catchment, basin) for the debris flow locations
- Add the additional fuel-related features

4.2.3 Extract Rock Type

- Collect the catchment area of each debris flow location
- Extract the fractional makeup of rock types (igneous, sedimentary, metamorphic)

4.2.4 Add Random Noise to Storm Data

Precipitation observations in the dataset were collected from rain gages up to 4 km away from the watersheds Staley et al., 2016, and only 123 unique precipitation values are present in the dataset. The assumption that the data are independent and identically distributed may thus not hold for this dataset.

Random forests in particular tend to overfit to small-scale dependencies of debris flow likelihoods on precipitation data during training.

Nearby watersheds sharing the exact same i15 value may exhibit a similar debris flow response, and the ML algorithm effectively learns to assign a debris flow site to a group of similar sites based on the unique i15 record.

So, we add random noise to the precipitation features (i15, duration, total accumulation) before the test-train split. The amplitude of the noise is defined to range between -10% and 10% of the recorded value.

4.2.5 Extract geological age from debris flow basins

The data sources are the digitized versions of the Geological maps of the different states. There is no numerical age in these maps, only the unit name according to the geological time scale.

We extract statistics about the geological age of the rocks underlying the debris flow basins. The data is joined with a tabular representation of the geological time scale to obtain a numerical age range associated with each unit.

4.2.6 Landslide Risk

Preliminary investigations show that the regions in our data set are higher on the scale.

We generate the FEMA (Federal Emergency Management Agency) assessed Landslide risk data.

4.2.7 Calculate the Time between Last Fire and Debris Flow

The time between last fire and debris flow as a feature is a good parameter for the model and for users in the final data product.

5 Analysis Methods

The analysis methodology consists of initial data analysis to understand the data and model development for prediction.

5.1 Exploratory Data Analysis (EDA)

Once we had a robust data pipeline developed, we explored the data we had at our disposal to analyze the structure of the data and discover underlying trends.

5.1.1 Rainfall

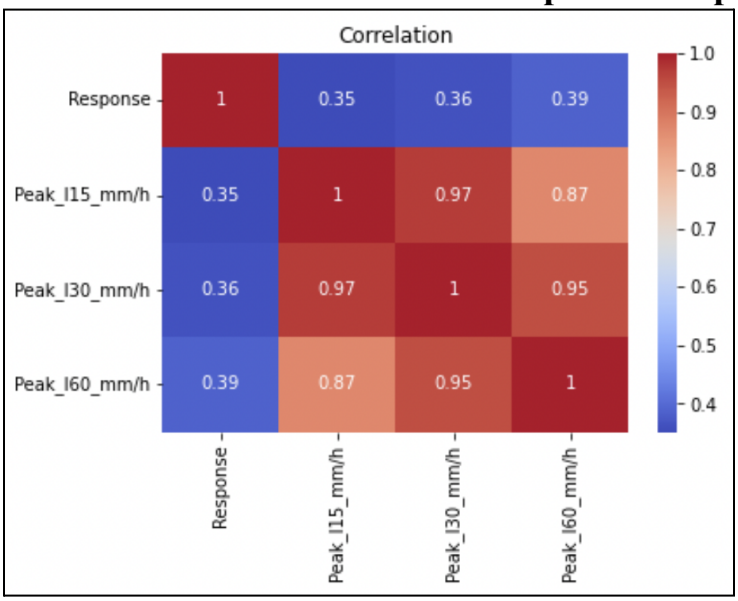
The following 6 features are related to rainfall:

peak_i15_mmh (Peak 15-minute rainfall intensity of storm, in millimeters per hour)
peak_i30_mmh (Peak 30-minute rainfall intensity of storm, in millimeters per hour)
peak_i60_mmh (Peak 60-minute rainfall intensity of storm, in millimeters per hour)

acc015_mm (Peak 15-minute rainfall accumulation of storm, in millimeters)
acc030_mm (Peak 30-minute rainfall accumulation of storm, in millimeters)
acc060_mm (Peak 60-minute rainfall accumulation of storm, in millimeters)

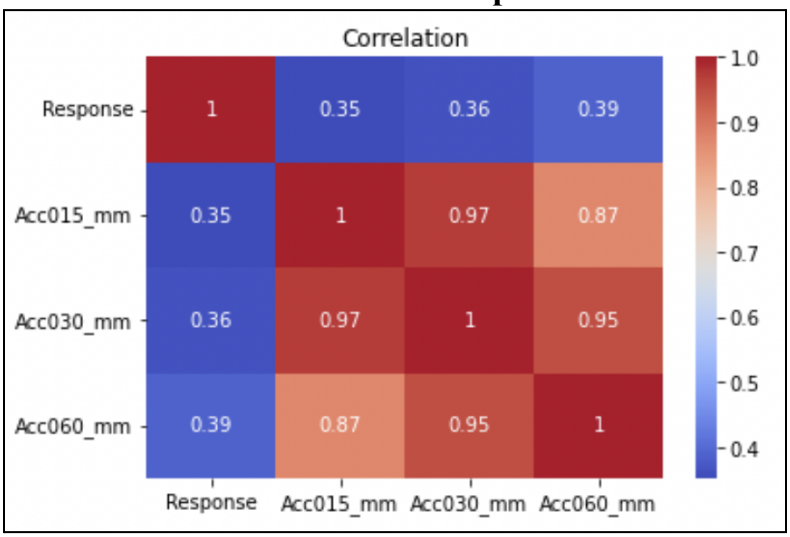
The correlation between debris flow response and the 15-min peak rainfall,30-min peak rainfall and 60-min peak rainfall is shown in Figure 5.

Figure 5: Correlation between debris flow response and peak rainfall



The correlation between debris flow response and the 15-min accumulation rainfall, 30-min accumulation rainfall and 60-min accumulation rainfall is shown in Figure 6.

Figure 6: Correlation between debris flow response and accumulation rainfall



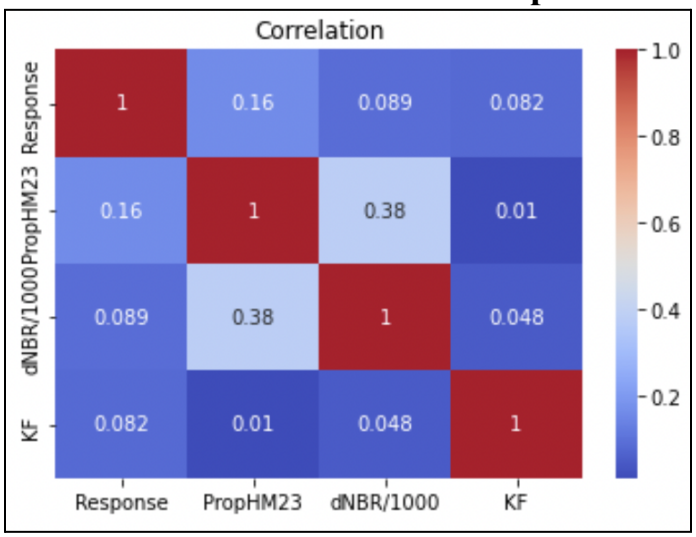
5.1.2 Fire Severity

The following 3 features are related to fire severity:

PropHM23 (Proportion of watershed burned at high or moderate severity and with gradients in excess of 23 degrees)
dNBR/1000 (Average differenced normalized burn ratio (dNBR) of watershed, divided by 1000)
KF (Average KF-Factor (erodibility index of the fine fragments of the soil) of the watershed)

The correlation between debris flow response and PropHM23, dNBR/1000 and KF is shown in Figure 7.

Figure 7: Correlation between debris flow response and fire severity



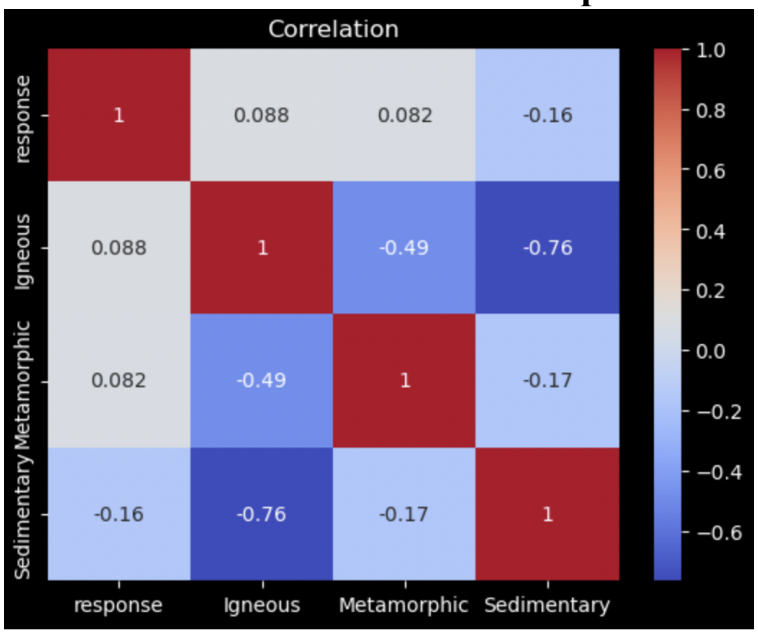
5.1.3 Rock Type

These are the rock type:
['Metamorphic', 'Sedimentary', 'Igneous and Sedimentary','Metamorphic and Sedimentary', 'Igneous','Igneous and Metamorphic', 'Melange', 'Unconsolidated', 'Water','Dam', 'Unconsolidated and Sedimentary', 'Tectonite']

We pick the following 3 features to explore correlation: **'Igneous', 'Metamorphic', 'Sedimentary'.**

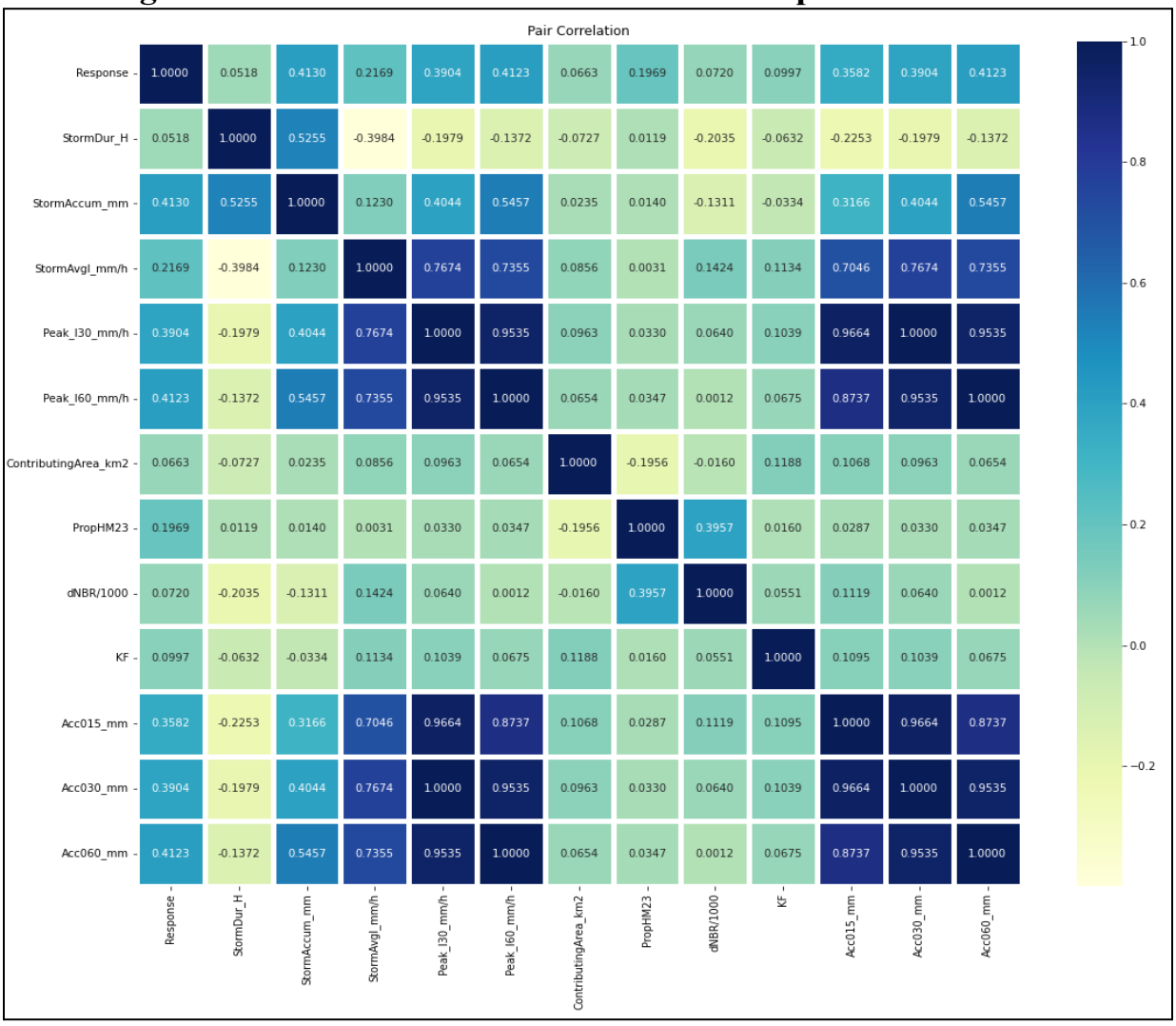
The correlation between debris flow response and Igneous, Metamorphic and Sedimentary is shown in Figure 8.

Figure 8: Correlation between debris flow response and rock type



5.1.4 Feature Correlation

Figure 9: Correlation between debris flow response and main features



5.2 Models

We compared two types of models: Logistic Regression (Staley '16) and Neural Network, applied to the problem of Post-fire Debris Flow Likelihood Prediction.

5.2.1 Staley Model

In Staley 2016 Model, there are 1550 observations across 716 sites in 7 states.

- ~20% of observations with a debris flow
- Drainage areas ranging from 0.2 - 8 km^2
- History of wildfire in each area between years 2000 - 2012
- Rain events between years 2000 - 2014
    - Rain gauges up to 4 km from Debris Flow sites
    - Collected by USGS, NOAA, local gov't
    - Rain has highest correlation with Debris Flow response

In Staley Model LR (Logic Regression), four features are used to predict the result ('Response').

- 15-minute rainfall accumulation
    - multiplied by subsequent features
- Proportion of watershed with slope > 23°
- Difference Normalized Burn Ratio
    - Change in landscape from pre-fire to post-fire
- Soil Erodibility Factor

Staley 2016 Model Performance Summary is shown in Table 4.

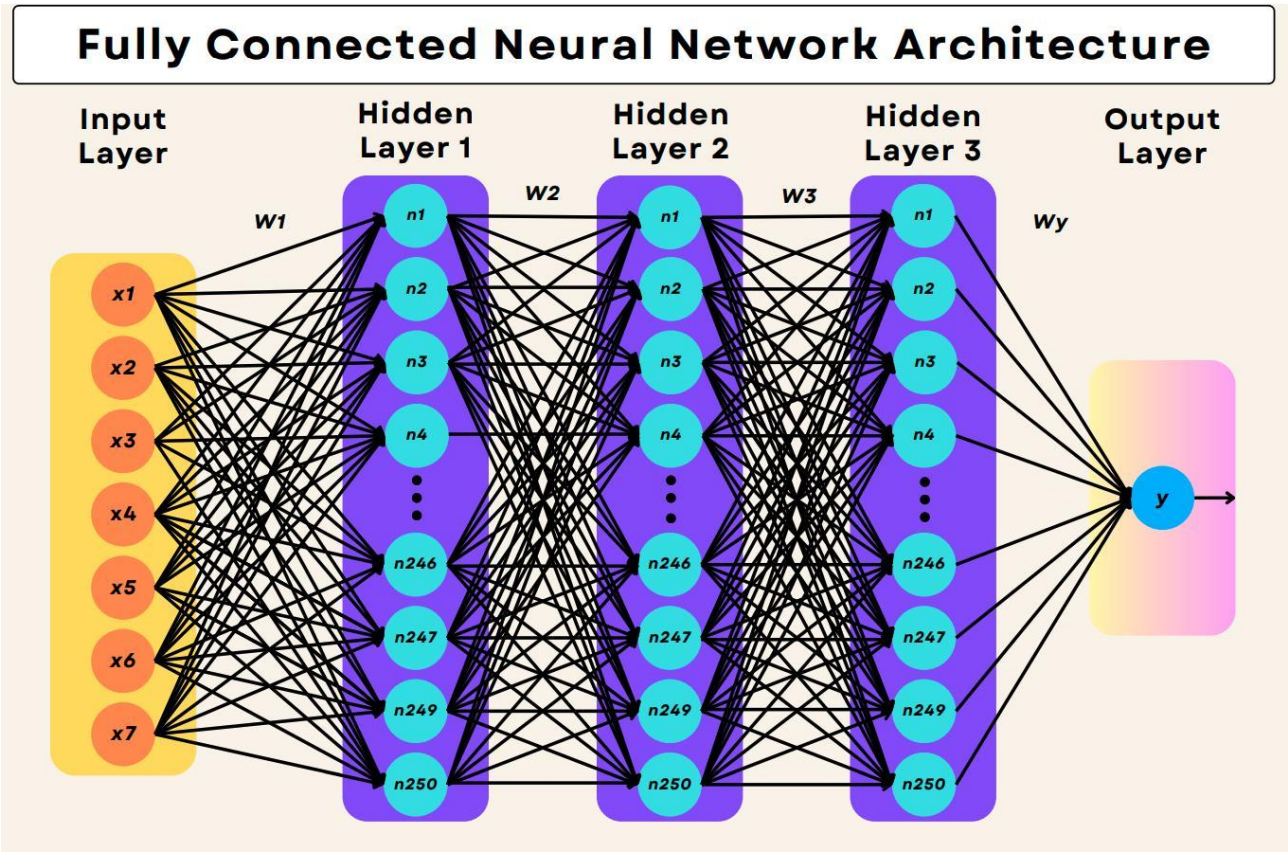
Table 4: Staley 2016 Model Performance Summary

Test Set Performance (IMW)	Logistic Regression (Staley)
Accuracy	0.6258
Precision	0.3544
Recall	0.7671
F1	0.4848
AUC	0.7178

5.2.2 Neural Network Model

The Architecture of Fully Connected Neural Network is shown in Figure 10.

Figure 10: Fully Connected Neural Network Architecture



Architecture and Training Parameters of the Fully Connected Neural Network:

Weight Initialization: Xavier Uniform
Activation Function: Rectified Linear Units (ReLu)
Output Function: Sigmoid
LR: 0.01
Dropout: 0.20
Epochs: 250
Loss Function: Binary Cross Entropy

5.2.3 Model Comparison

The Neural Network model achieved the highest Area Under Receiver Operating Characteristic Curve (AUC) as depicted in figure 11.3. Logistic Regression under random split actually performed worse than the SoCal training set.

Figure 11.1: AUC - ST16 - SoCal

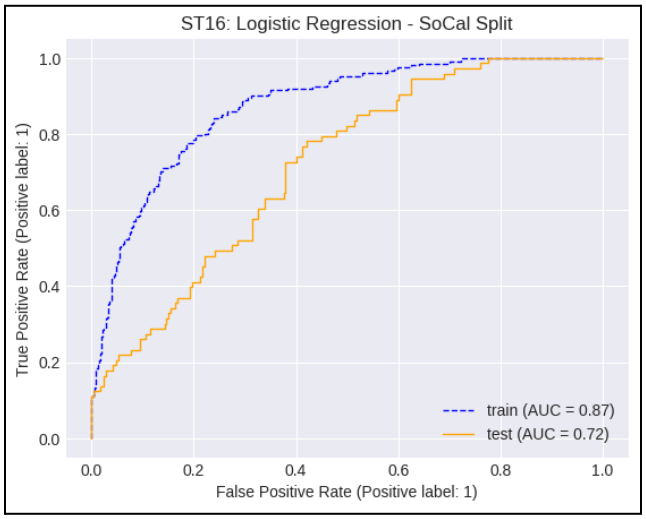


Figure 11.2: AUC - ST16 - Random Split

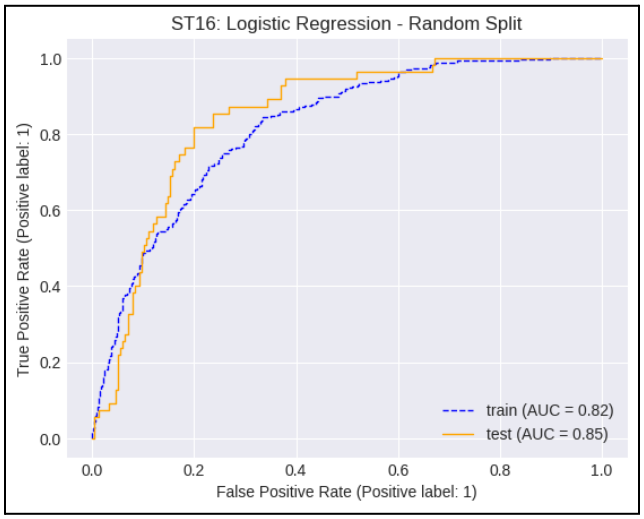
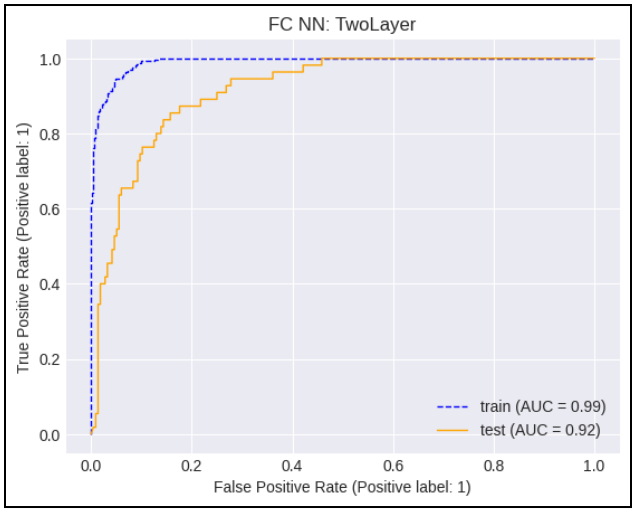


Figure 11.3: AUC - NN - Random Split



The Neural Network architecture beat the base model on *Accuracy, Precision, F1, and AUC* scores. However, Staley still outperformed on *Recall*, an important measure of the True Positive Rate. Decision makers would still need to be consulted to determine what is an acceptably balanced model. As these two compare, the Staley model will have many more False Positives which could undermine efforts by stakeholders to evacuate areas if their warnings are seen as overly cautious. Going overboard in the other direction, the team was able to train a model that had a *Recall* as high as 90%, but at the expense of *Precision*, meaning the *False Negatives* would mostly be eliminated but the *False Positives* greatly increased, making this hypothetical boy-cried-wolf scenario even more prevalent. Given that the Debris Flows occurred only 20% of the time in the original dataset means this decision would have to be carefully balanced.

Table 5: Model Performance Comparison

Test Set Performance	**Logistic Regression (SoCal)**	**Logistic Regression (RandomSplit)**	**Neural Network (RandomSplit)**
Accuracy	0.6258	0.8007	**0.8745**
Precision	0.3544	0.5200	**0.7143**
Recall	**0.7671**	0.2364	0.6264
F1	0.4848	0.3250	**0.6731**
AUC	0.7178	0.8476	**0.9217**

6 Findings & Reporting

6.1 Feature Findings

6.1.1 Rainfall

The heatmap of the correlation between debris flow response and peak rainfall shows similar correlation patterns with both peak rainfall and accumulation rainfall. This suggests that the correlation between rainfall and debris flow response is consistent.
However, when we focus on the 15-min rainfall, 30-min rainfall, and 60-min rainfall, we notice that as the accumulation time increases, the correlation becomes stronger. In other words, the longer the duration of rainfall accumulation, the marginally higher the correlation between debris flow response and rainfall.

The heatmaps visually depict the correlations between the features, and these observations provide insights into the relationships between debris flow response and different time intervals of peak rainfall and accumulation rainfall.

Many of the records do not have 30- or 60-min rainfalls. But logically, a record with 30 min rainfall has a 15 min rainfall. While longer rainfall does more strongly correlate with debris flow, the difference does not outweigh the lack of data. Additionally, the collinearity between these features is too high to simply include all three in the final predictive model. As such, the team chose to retain the 15-minute feature.

6.1.2 Fire Severity

The heatmap of the correlation between debris flow response and fire severity illustrates the correlation between debris flow response and fire severity. The magnitude of correlation with fire severity follows the order: PropHM23, dNBR/1000, and KF. The correlations of the latter two features are approximately half of the correlation observed with the first feature. This indicates that PropHM23 has the strongest correlation with debris flow response, while dNBR/1000 and KF exhibit weaker correlations.

6.1.3 Rock Type

This heatmap of the correlation between debris flow response and rock type illustrates the correlation between debris flow response and rock type. The heatmap displays the correlation values, where higher values indicate stronger correlations.

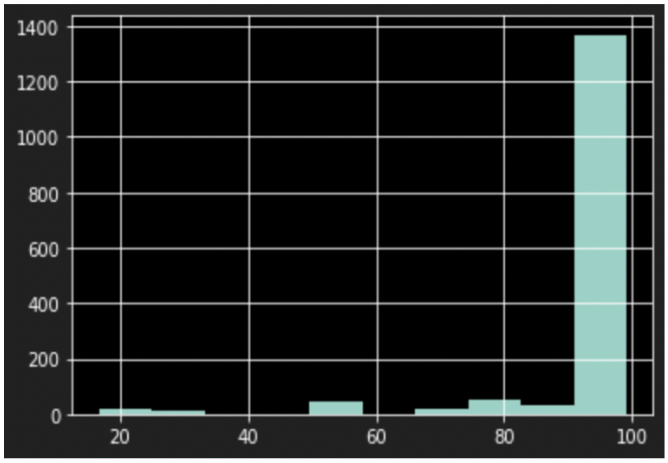
In particular, the rock type "Sedimentary" shows a negative correlation with debris flow response. This suggests that areas with sedimentary rock types are less prone to debris flows. On the other hand, the rock types "Igneous" and "Metamorphic" exhibit similar correlation values with debris flow response, indicating a moderate correlation. This implies that these two rock types have a comparable association with the occurrence of debris flows.

By analyzing this heatmap, we can gain insights into the relationship between different rock types and the likelihood of debris flow events. These findings can be valuable in assessing the susceptibility of specific rock types to debris flows and informing appropriate mitigation measures or land management strategies.

6.1.4 Landslide Risk

The team added a Landslide Risk parameter taken from FEMA to the potential feature set. As shown in Figure 12, the histogram demonstrates that areas with Debris Flows skewed towards high Landslide risk scores.

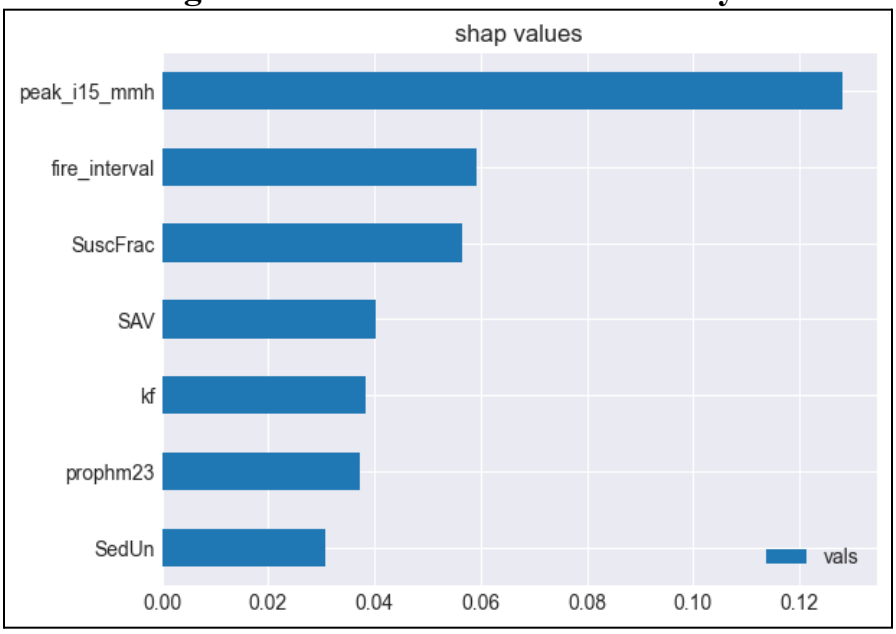
Figure 12: Histogram of landslide risk



6.2 Model Findings

In an effort to prevent overfitting, the team included subsets of features, ranked by their Shapely Additive Explanations (SHAP) values. The balanced model chosen by the team included seven final features (figure 13), of which four new additions.

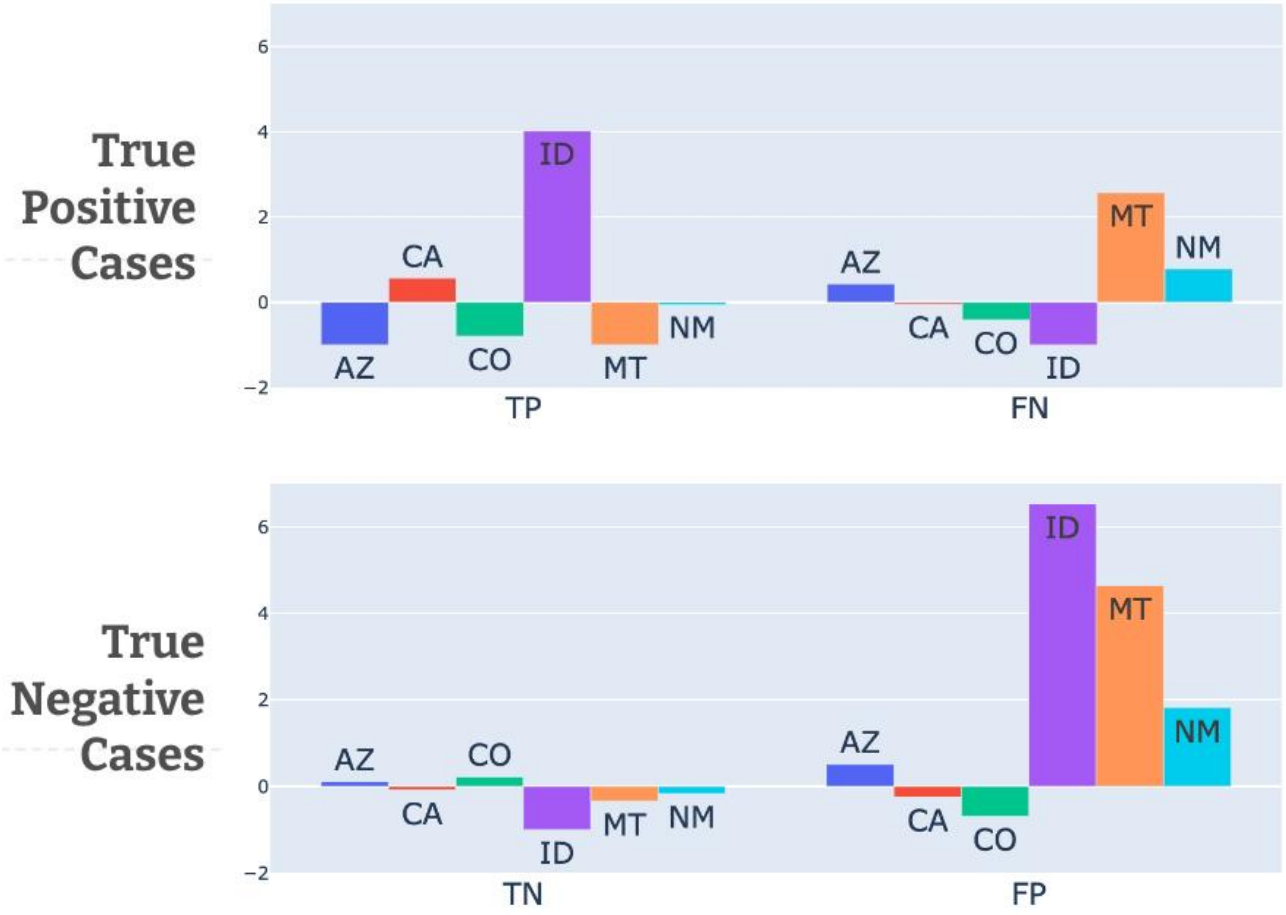
Figure 13: Final Features Ranked by SHAP



- **Rain:** rain in mm/h
- Fire Interval**:** time since wildfire
- SuscFrac**:** fraction of watershed covered by burn susceptible vegetation types
- SAV**:** Avg Surface Area to Volume across fuel categories
- **KF:** Fine Fragment Soil Erodibility of watershed
- **Prophm23:** Proportion of watershed with slope > 23°
- SedUn**:** fraction of watershed covered by sedimentary and unconsolidated rocks

Overall the NN model was able to yield a balanced set of performance metrics. Looking at where the model missed, the team created a confusion matrix by state. Because the sites are not evenly distributed by state (e.g. California makes up 55% of the records) the state confusion matrix was indexed against the number of records coming from each state. Overall the model did a great job of correctly classifying the Arizona, California, and Colorado sites which make up 93% of the records. The model particularly fared poorly in False Positive predictions in Idaho, Montana, and New Mexico which combined make up 7% of the total records.

Figure 14: State Confusion Index



6.3 User Interface

The user user interface is a browser application that simply requires an internet connection (when app is hosted). It includes point locations of debris flows within the data set and the catchment areas associated with these debris flows. The color channel represents the probability of a debris flow. Lighter white represents zero-to-little probability. Darker red represents higher-to-certain probability.

Interaction:

- The user can navigate by dragging the mouse in a familiar manner and zooming to any area with the mouse wheel.
- Input the rain amount in millimeters/hour
- Hover over catchment areas to view the probability of a debris flow at the given rain input
- Hover over the debris flow site to view the date of the last debris flow

The following figures demonstrate what a user would see in a specific area, Angeles National Forest, as the rain input is increased. Figure 15.1 represents a small amount of rain and the map demonstrates zero probability of a predicted debris flow. Figure 15.2 represents a higher rain amount and the map demonstrates moderately increased probability of debris flows. Figure 15.3 represents an even higher rain amount and the map demonstrates moderately high probability of debris flows in more catchment areas.

Figure 15.1: input 05mm

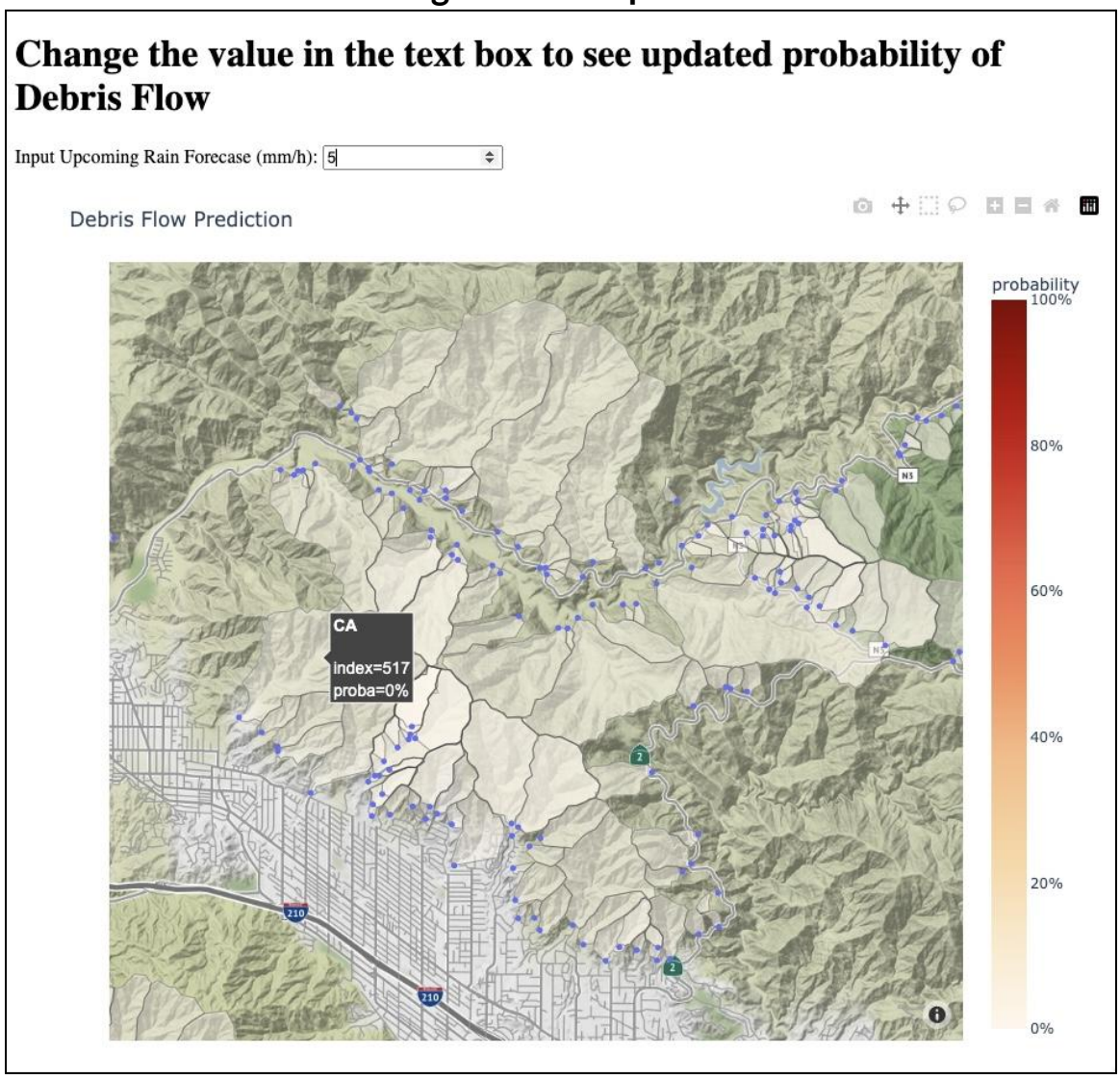


Figure 15.2: input 15mm

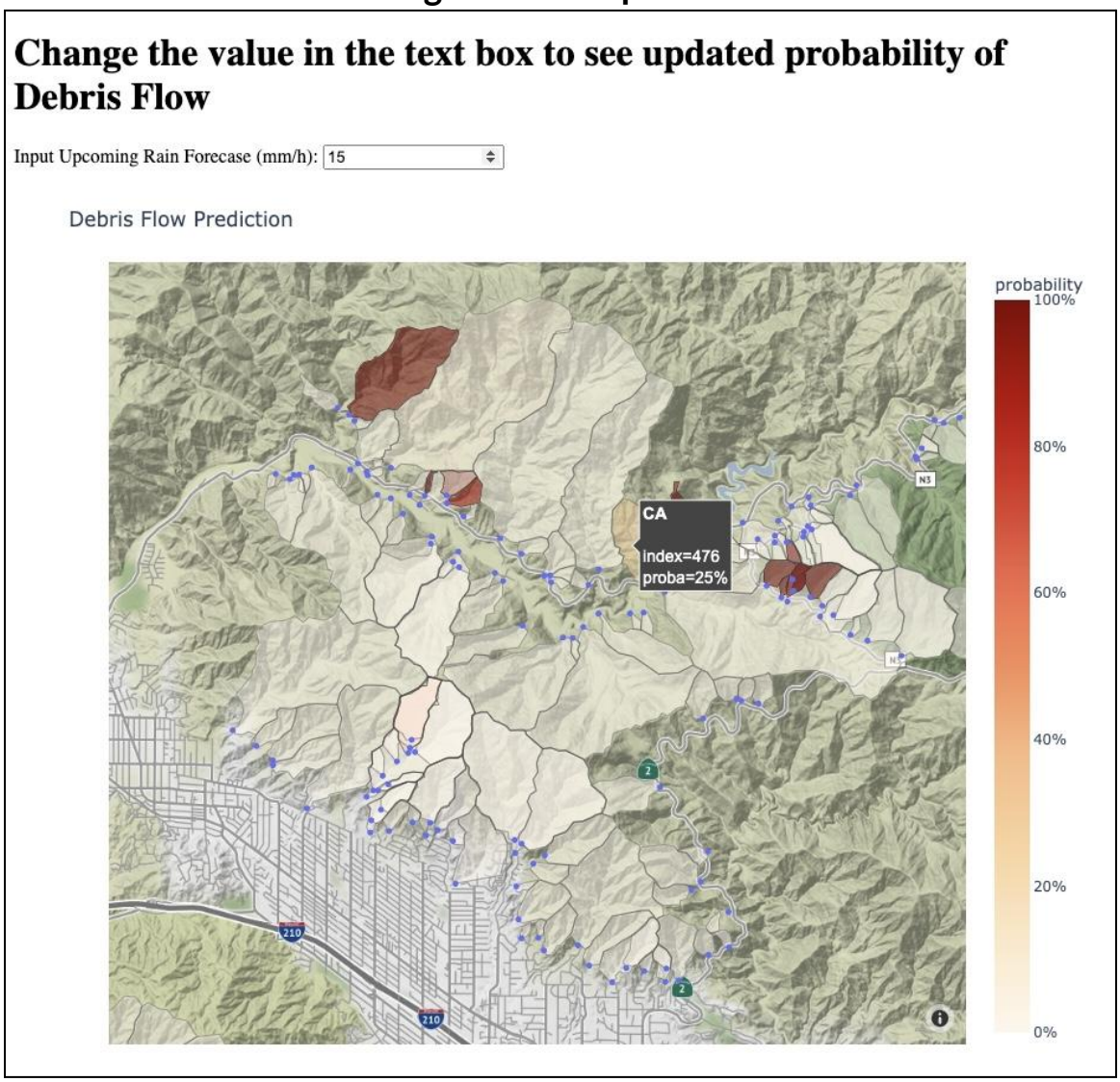
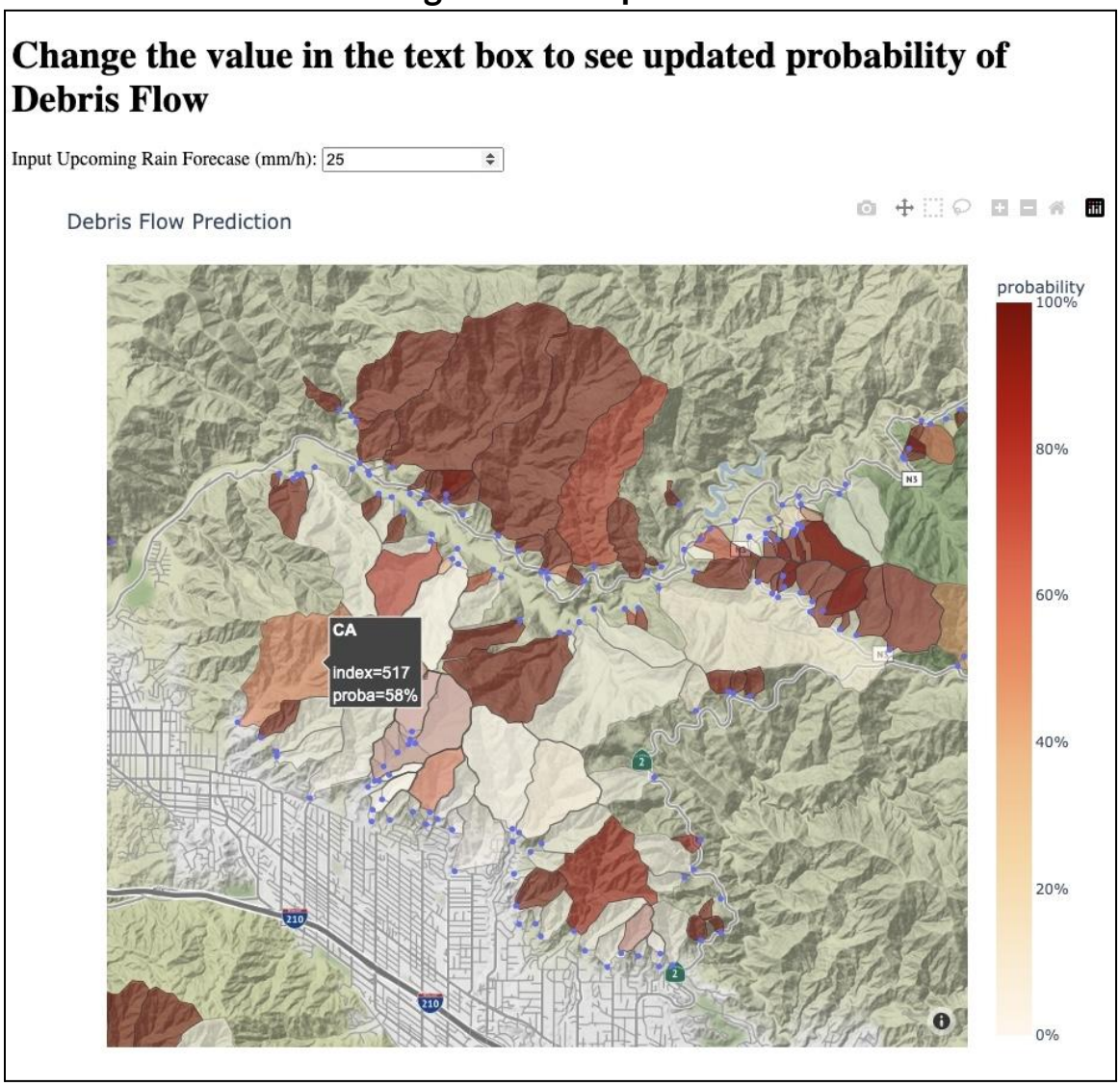


Figure 15.3: input 25mm



7 Solution Architecture, Performance & Evaluation

7.1 Deployment

- In order to grant public access to the dash app, we need to deploy the app to a server
- We decided to use Amazon EC2 service for this purpose
- We launched a t2.xlarge server with Ubuntu OS , 4 Cores and 32 GB RAM.
- Once the instance is up and running, follow below steps to launch the dash app
- SSH into the server and clone the git repo
- Repo has all the libraries required to create a virtual environment in requirements.txt file and the code base
- Install the libraries and then run the dash app
- Dash app will be rendered on port 8050 of the EC2 server
- This app can now be accessed from EC2 IP & Port 8050.

7.2 Scalability & Robustness

Some of our joins were conducted using SiteIDs created during the data preparation process. However, each of the records includes geospatial data, both the lat/lon of the historical debris flows as well as the lat/lon shapes (both in EPSG:4326 coordinate system). Using geopandas, a user could hypothetically complete a spatial join with new data to this dataset given it includes this same geospatial attribute. This allows additional records to systematically be created for more areas that are not part of the original dataset.

The data preparation steps are designed to be run in parallel using available CPUs. The code is written so that the parallel processing is distributed to as many cores as reasonable without overtaxing the machine. The parallelism will flex up or down automatically by detecting the systems available compute power.

7.2.1 Database Robustness

The data fits into memory and a database is not needed for the feature set. The dataset includes more features than were included in our final model. A future user could continue to experiment with the 19 total features (original and created) available in the final data file. While the data is organized in tabular fashion, it does include the geospatial attributes for each catchment. The team did experiment with uploading the data as-is to a PostGIS database and this is easily achievable. If more sites were generated, they could follow the same format as the current data and simply upload sequentially without issue.

7.2.2 Model Scalability

Models were trained on the Southern California and Intermountain West regions. The final model (named `*TwoLayer_250_epochs_optimized_roc_auc_score*` in the app) was tuned specifically to perform well on new data within these same regions. A future data scientist could feel confident in model performance if they gather data for these same regions and infer the probability of a debris flow using the current model. However, the authors make no claim about model efficacy on regions outside of the dataset. For example, adding sites from Northern California, which may have different geological features outside the distribution of the Staley data, may yield unexpected and unreasonable results.

7.2.3 Compute Scalability

The model training was designed to run on both CPUs and GPUs without additional configuration by a future user. The code systematically checks for GPU and attaches if available. Training a single model with a predefined set of hyperparameters can be reasonably done on CPU, though with increased training time vs GPU, as demonstrated by Table 6. However, to harness the power of the grid search and programmatically select the optimal hyperparameters, it is highly recommended to complete the training steps on a server with available GPUs. The difference in a single training run partially depends on the underlying model complexity (e.g 1 vs 3 hidden layers), but the GPU can save hours in a single training run, let alone the hundreds of training runs the team performed.

Table 6: CPU vs GPU training time

Processing Unit	Features / Hidden Layers / Hidden Nodes	Learning Rate	Epochs	Wall Time
CPU*	13 / 3 / 250	0.001	1000	0 h 1 m 12.0 s
GPU**	13 / 3 / 250	0.001	1000	0 h 0 m 06.9 s

*CPU: Apple M1 Pro
**GPU: NVIDIA GeForce GTX 1080 Ti

8 Conclusions

There are many individual factors that influence the likelihood of a debris flow. Taking all of these factors, feeding them to a Neural Network, and hoping for the best is a poor strategy. This will lead to an overly complex model that perfectly fits the training data. However, by feature engineering, hyperparameter tuning, and optimizing on the appropriate metrics, it's possible to improve upon the Staley '16 Logistic Regression model currently deployed by the USGS. In addition to creating a model that outperforms Staley, it's also important to deliver the model to end-users in a fast and accessible manner. By building a mapping tool with a simple interface and an adjustable rain parameter, end-users can quickly utilize model predictions to make informed decisions about evacuation procedures.

References

[1] Staley et al (2016) https://pubs.er.usgs.gov/publication/ofr20161106
[2] https://landfire.gov/fbfm40.php
[3] https://mrdata.usgs.gov/geology/state/
[4] https://www.usgs.gov/3d-elevation-program
[5] https://prdtnm.s3.amazonaws.com/StagedProducts/Elevation/13/TIFF/current/n35w119/USGS_13_n35w119.tif
[6] Jessica Block https://nwschat.weather.gov/lsr/#LOX,MTR,EKA,SGX,STO,REV,HNX,MFR,VEF/202301040800/202301080759/1100
[7] Daniel Roten, Jessica Block, Daniel Crawl, Jenny Lee and Ilkay Altintas, Machine Learning for Improved Post- fire Debris Flow Likelihood Prediction, IEEE Big Data 2022 December 17-20, Osaka, Japan.
[8] Efthymios I. Nikolopoulos, Elisa Destro, Md Abul Ehsan Bhuiyan, Marco Borga, and Emmanouil N. Anagnostou, Evaluation of predictive models for post-fire debris flow occurrence in the western United States, Natural Hazards and Earth System Sciences, 2018.

Appendices

A. DSE MAS Knowledge Applied to the Project

This project required applying knowledge and skills gained from most classes in the UCSD DSE MAS program. The project made heavy use of Python programming, web scraping, data wrangling, relational databases, machine learning, and data visualization. The courses most heavily utilized in this project include:

- Python for Data Analysis (DSE 200)
- Data Management Systems (DSE 201)
- Data Integration & ETL (DSE 203)
- Machine Learning (DSE 220)
- Scalable Data Analysis (DSE 230)
- Data Visualization (DSE 241)

B. Link to UCSD Library Archive for Reproducibility

This project is archived at the UCSD Library Digital Collections to ensure results are reproducible and ensure future projects can build upon this work. Digital Object Identifier: https://doi.org/10.6075/J0PG1RZH

The archive includes the following:

- This report
- The presentation slides
- The project poster
- Team's Github Repository: https://github.com/gojandrooo/DSE-Capstone