Predicting flood events in Louisiana

Springboard - Capstone 1

Jenny Rhee

Background

- Floods can be devastating, resulting in significant injuries, deaths, and economic damages.
- In August 2016, Southeast LA experienced prolonged rainfall for nearly 48 hours that resulted in catastrophic flooding and thousands of homes and businesses being submerged (denoted a 1000 year flood).
    - Approximately $10-15 billion in damages
- Even more minor floods can cause substantial damages.
- **Goal**: A model that can predict floods with forecast data

Client Profile

- The National Weather Service (NWS) has a mission to provide weather, water, and climate data, forecasts, and warnings for the protection of life and property and enhancement of the national economy.
- Ultimate vision: create a weather-ready nation - a society that is prepared for and responds to weather, water, and climate-dependent events
- This model will help stakeholders save lives and potentially minimize property damage by providing a tool to predict flood events.
- Can also be used to alert civilians and advise them on what steps to take

Data

- National Weather Service: storm data from 1950 to present (34 different storm events including various types of floods, personal injuries, damage estimates, etc.)
- NOAA's National Centers for Environmental Information: meteorological data (air temperature, precipitation, wind speed, etc.; stations in image on the right)

Data Cleaning - Storm Events

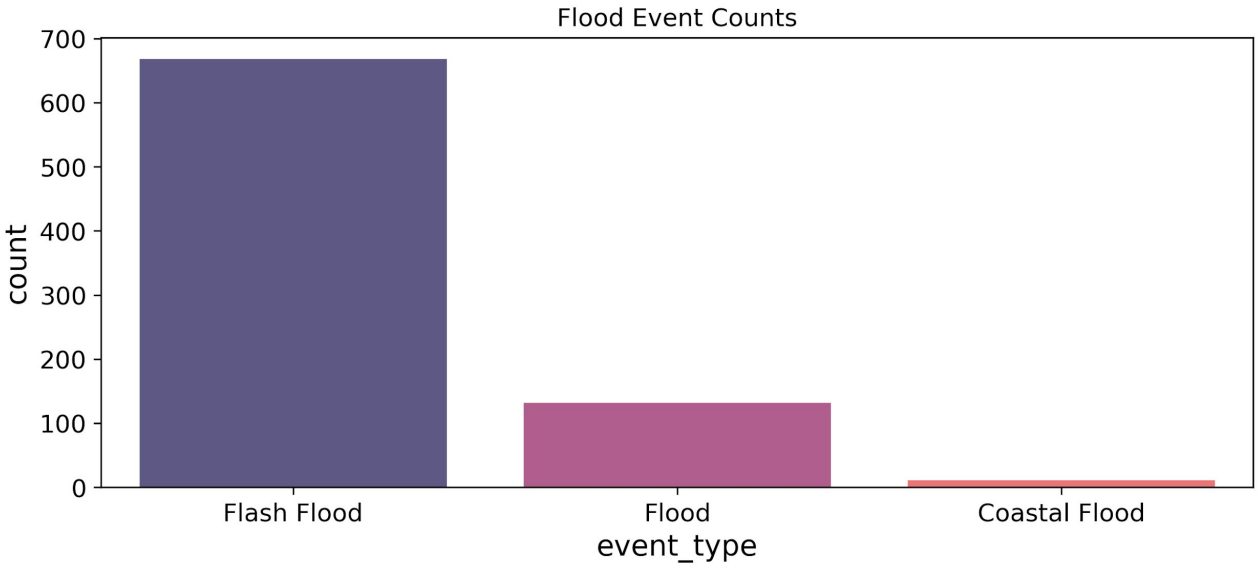
- Cleaned up parish names to be one of the distinct 64 parishes in LA
- Property damage
  - Strings with size abbreviation converted to numerical values (e.g., 100k → 100,000)

Data Cleaning - Meteorological

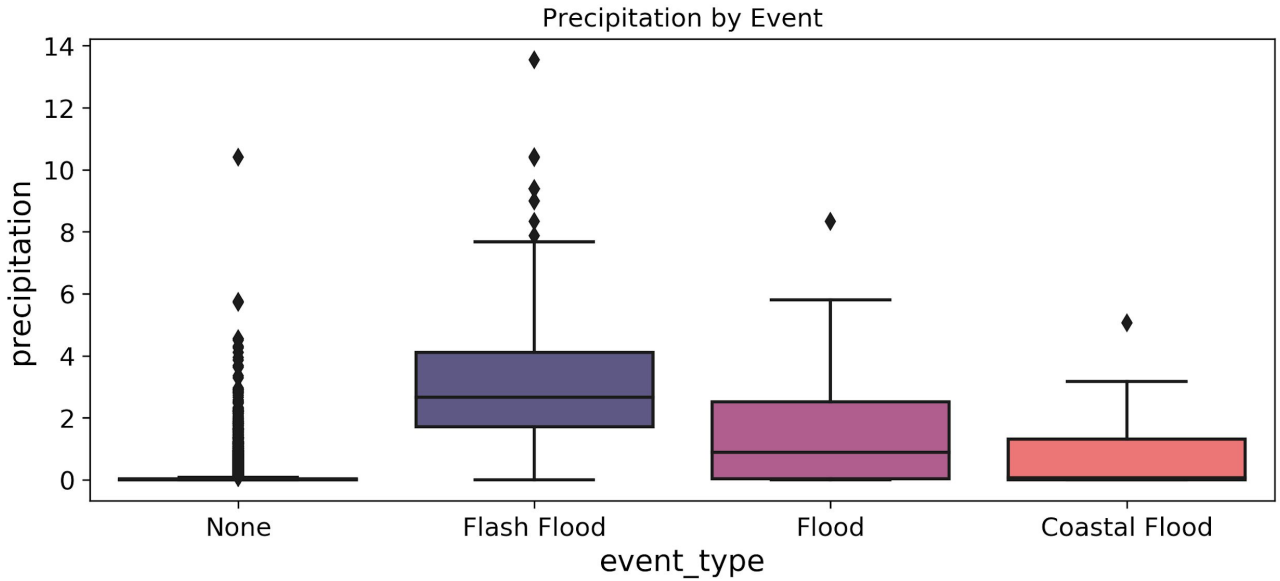
- 11 variables of interest: average daily wind speed, precipitation, max temperature, min temperature, fastest 2-minute wind speed, fastest 5-second wind speed
- Did not include parishes for each station -- API used to query for each parish based on coordinates of the station

Data Wrangling - Combined Data

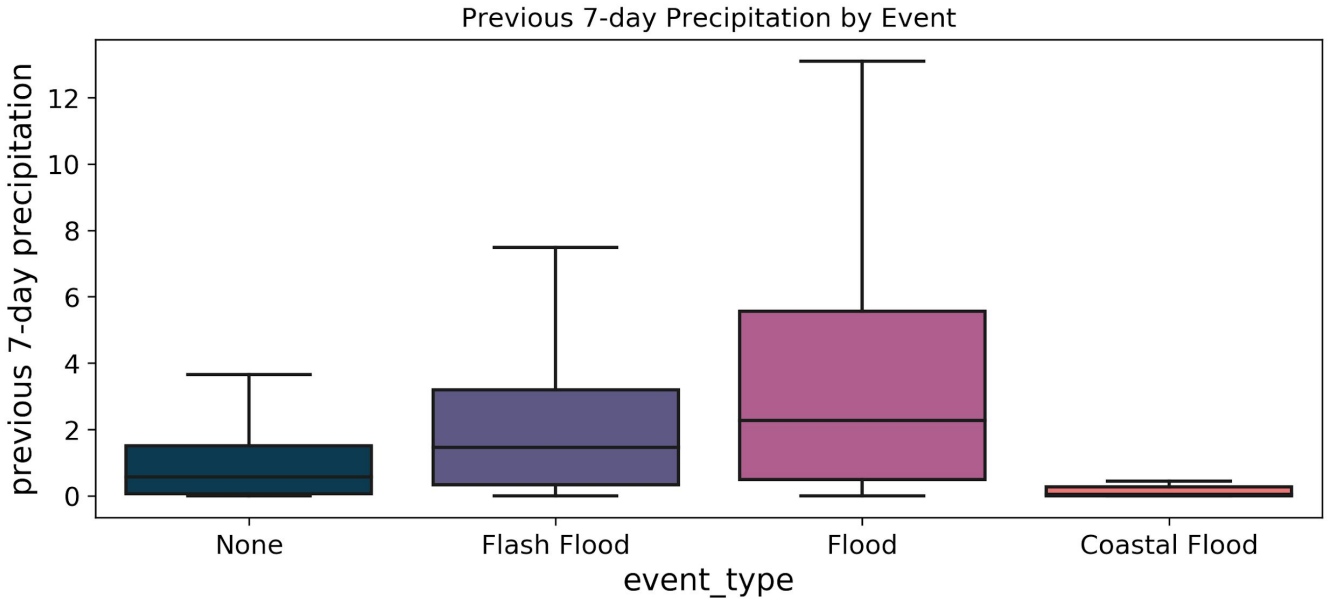
- For rows with null storm events data, a new event type called “None” was created -- no storm event happened on these days
- 2,202 rows with flood events (Coastal Flood, Flash Flood, and Flood), but only 812 rows with complete weather data
- 5,000 samples for days representing no severe storm event
- Feature engineering:
  - Previous 7-day total precipitation
  - Region of the state (Northwest, Northeast, Central, Southwest, Southeast)
  - Season (winter, spring, summer, fall)



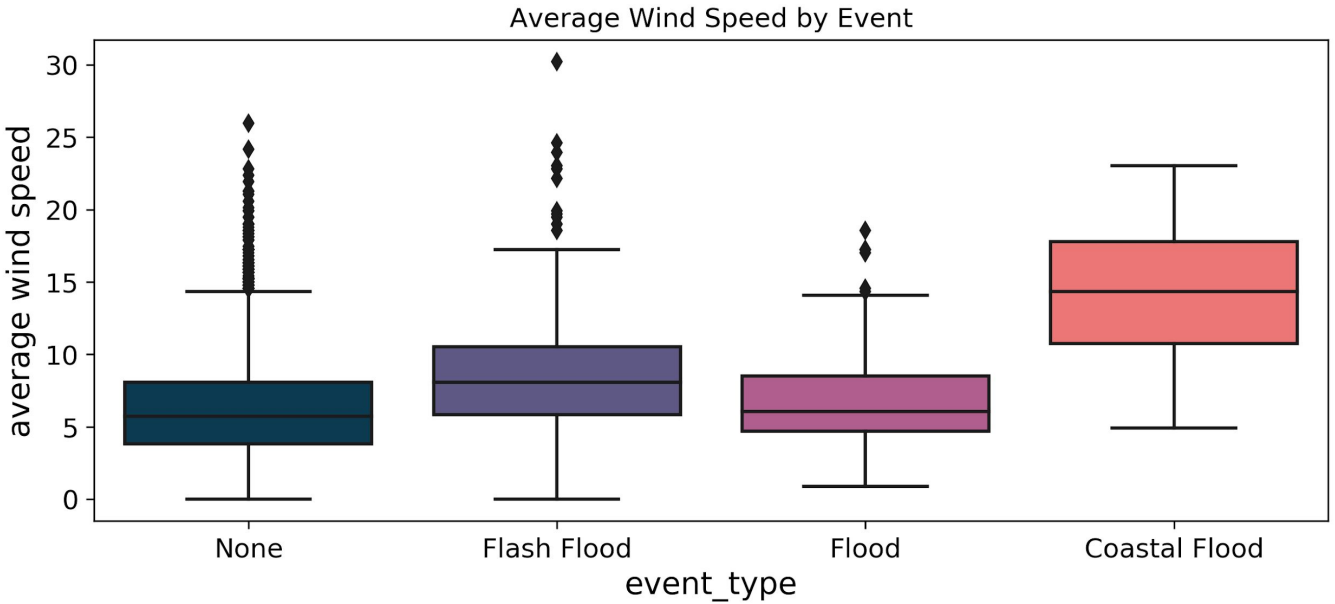
- Flash flood - life-threatening, rapid rise of water into a normally dry area beginning within minutes to multiple hours of the causative event
- Flood - any high flow, overflow, or inundation by water which causes damage
- Coastal flood - flooding of coastal areas due to vertical rise above normal water level caused by strong, persistent onshore wind, high astronomical tide, and/or low atmospheric pressure, resulting in damage, erosion, flooding, fatalities, or injuries



- Flash floods had the highest average precipitation.
- Days with no storm events had the lowest average precipitation.
- One-way ANOVA:
  - F-value = 2565.98
  - $p < 0.001$



- Coastal floods had the lowest average previous 7-day precipitation.
  - These events are driven by wind, tide, or low atmospheric pressure rather than purely by precipitation.
- Floods had the highest average previous 7-day precipitation.
- One-way ANOVA:
  - F-value = 125.78
  - $p < 0.001$

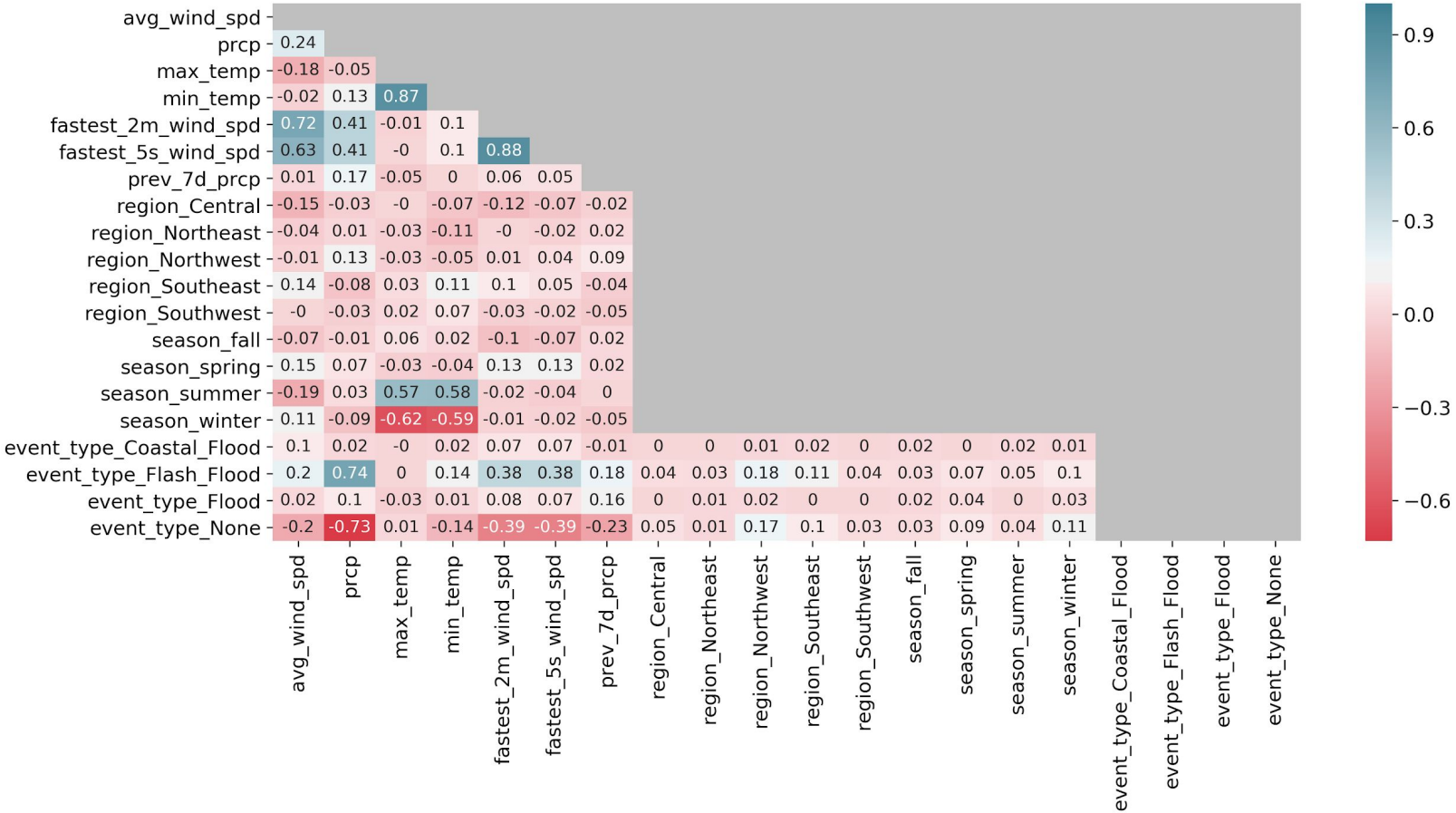


- Coastal floods had the highest average wind speed.
  - Again, wind is one of the driving factors of coastal flooding.
- There were no statistical differences in wind speed between floods and days with no events.
- One-way ANOVA:
  - F-value = 102.07
  - $p < 0.001$

Correlation Matrix

- Three methods used to find correlation coefficients depending on the feature types
  - Pearson’s correlation - continuous vs. continuous [-1, 1]
  - Point biserial correlation - categorical vs. continuous [-1, 1]
  - Cramer’s V - categorical vs. categorical [0, 1]

Correlation Matrix



Flood Classification Overview

- Modeling task: classify whether or not a flood event occurred
- Decided to drop coastal floods because they were too different from other flood events with a small sample size (n = 11)
- Compared performance of models with standardized data (sklearn.preprocessing.scale) and scaled data (MinMaxScaler)
- Comparison of two classification algorithms - logistic regression and random forest
- Hyperparameter tuning used for both models with GridSearchCV

A note on metric choice

- F-β score is a way of measuring accuracy in a model by taking into account both precision and recall.
  - β can be configured to give more weight to precision or recall.
- Decided to optimize for β = 2, which puts twice as much emphasis on recall
- The advantage was that the model will find more true flood events.
- The disadvantage was that the model will more likely label an event as a flood event when this may not be the case.
- This tradeoff was more ideal than the alternative because we would rather not miss any true flood events in order to mitigate potential damages if possible.

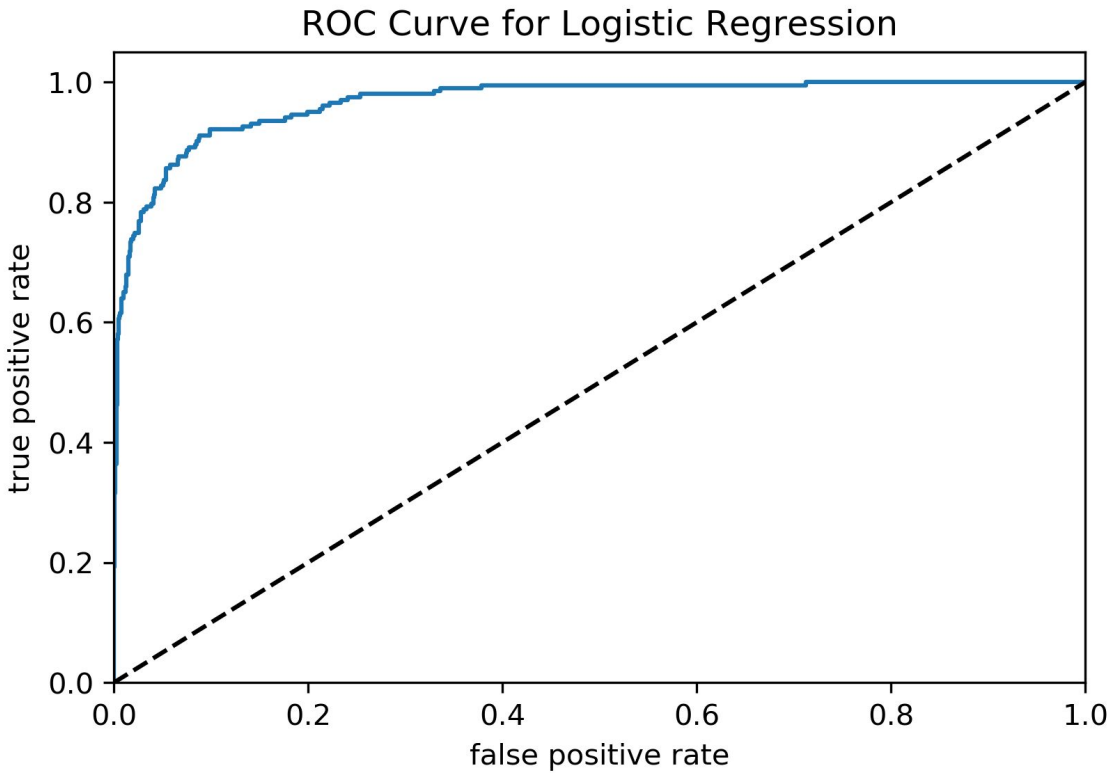
Logistic Regression

- Scoring function: ROC-AUC
- The unprocessed data performed the best and was used to train the final model.

data	ROC-AUC testing score
No preprocessing	0.969673
`MinMaxScaler`	0.969586
`scale`	0.969665

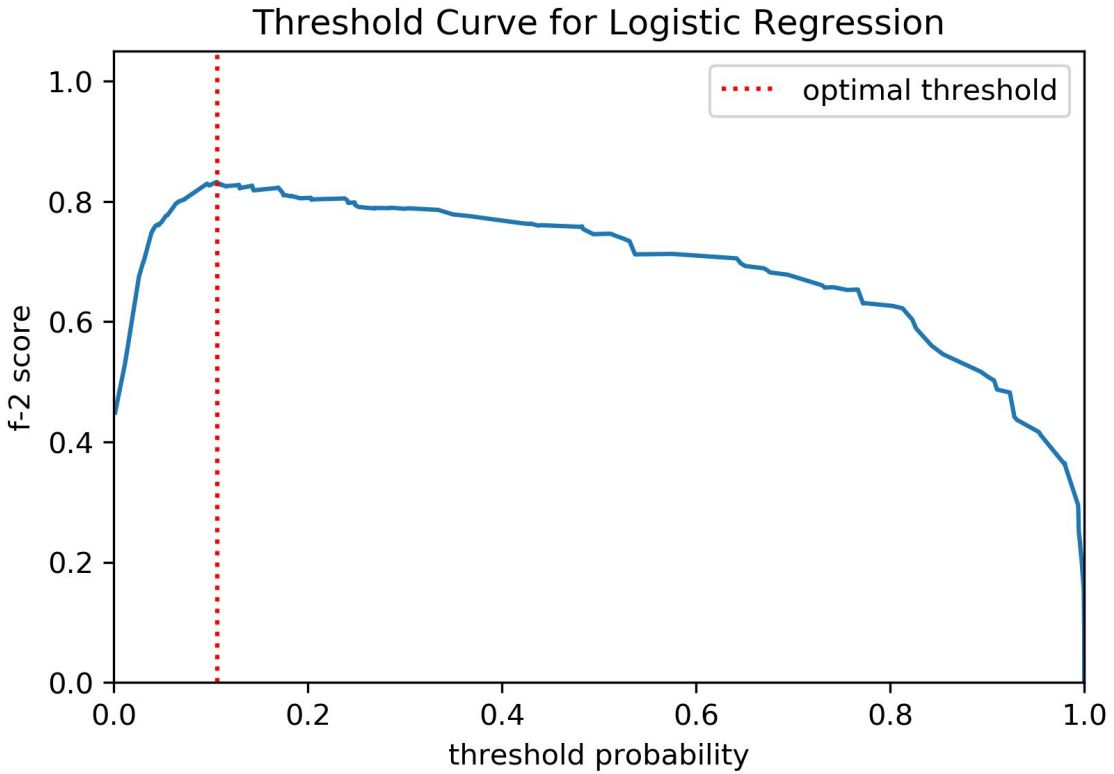
Logistic Regression - final model

- Hyperparameter tuning:
  - Cs = 59.948
  - Penalty = L2
- ROC-AUC score = 0.9676



Logistic Regression - thresholding

- Final step - choosing what threshold probability should be used to optimize F-2 score
- Optimal threshold = 0.106
- Best F-2 score = 0.833



Logistic Regression - confusion matrix & classification report

	no flood	flood
no flood	1138	109
flood	19	184

	precision	recall	f1-score	support
0	0.96	0.91	0.95	1247
1	0.63	0.91	0.74	203
accuracy			0.91	1450
macro avg	0.81	0.91	0.84	1450
weighted avg	0.93	0.91	0.92	1450

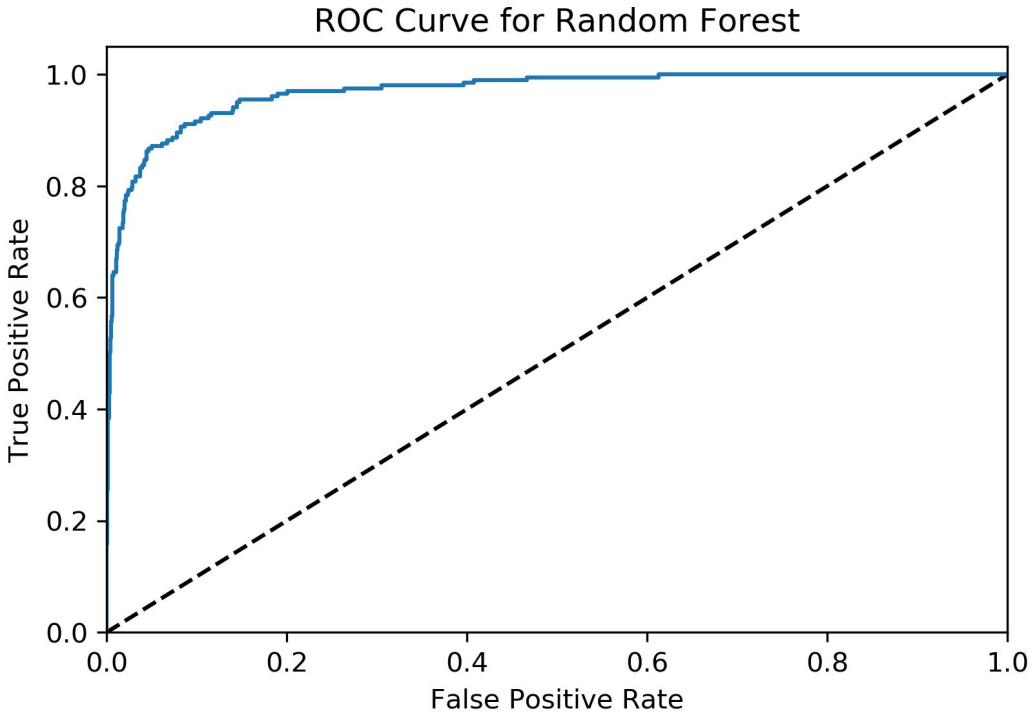
Random Forest

- Scoring function: mean accuracy
- The unprocessed data performed the best and was used to train the final model.

data	mean accuracy testing score
No preprocessing	0.9503
`MinMaxScaler`	0.9448
`scale`	0.9497

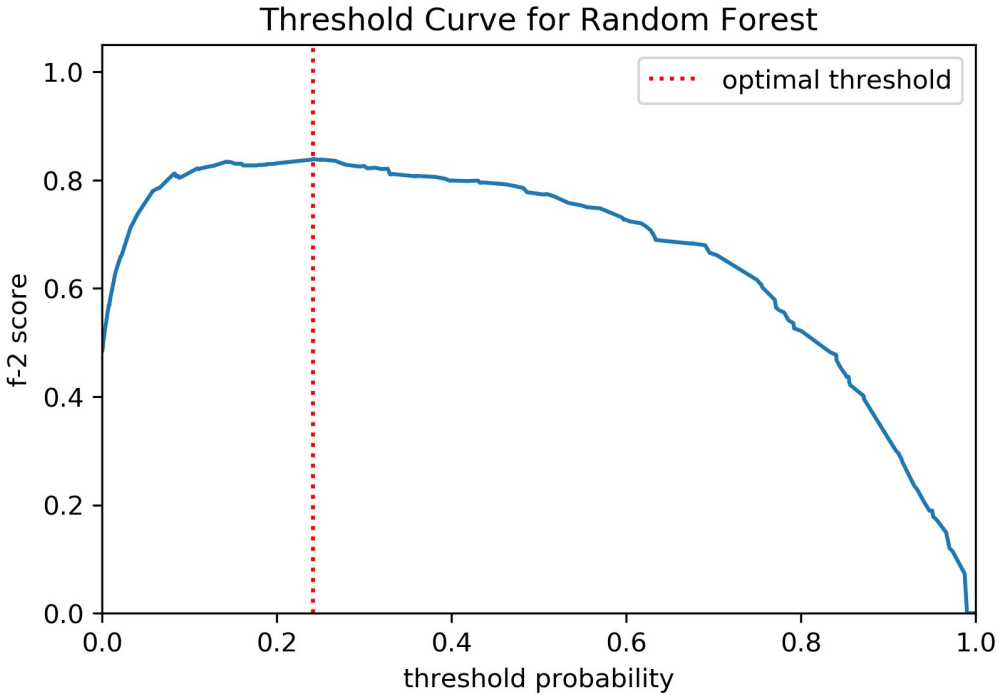
Random Forest - final model

- Hyperparameter tuning:
  - min_samples_leaf = 0.001
  - min_samples_split = 0.001
- ROC-AUC score = 0.9696
- The random forest model performed better than logistic regression (ROC-AUC = 0.9676).



Random Forest - thresholding

- Final step - choosing what threshold probability should be used to optimize F-2 score
- Optimal threshold = 0.241
- Best F-2 score = 0.839



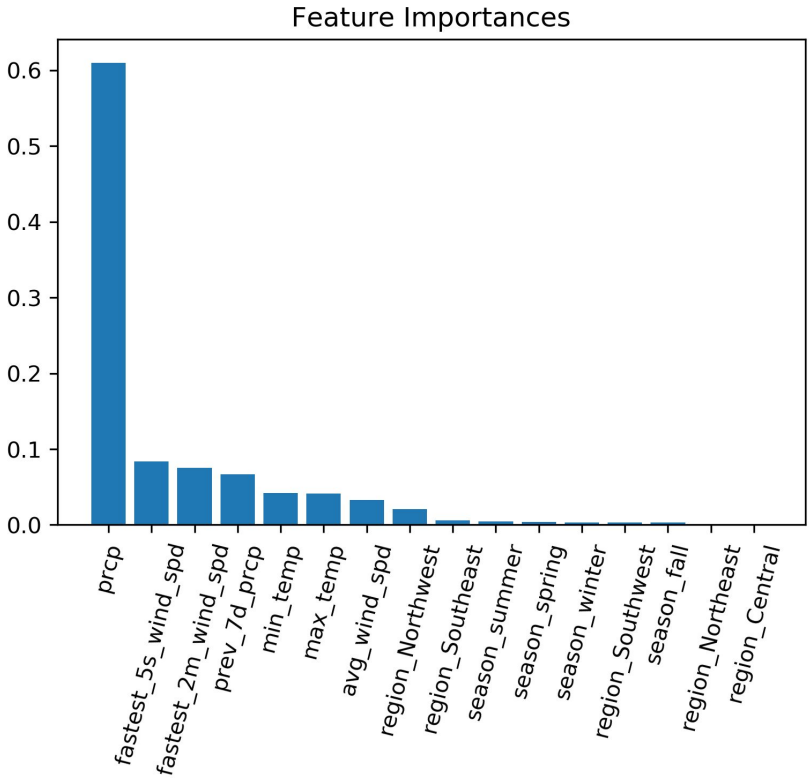
Random Forest - confusion matrix & classification report

	no flood	flood
no flood	1186	61
flood	27	176

	precision	recall	f1-score	support
0	0.98	0.95	0.96	1247
1	0.74	0.87	0.80	203
accuracy			0.94	1450
macro avg	0.86	0.91	0.88	1450
weighted avg	0.94	0.94	0.94	1450

Random Forest - feature importances

- As expected, precipitation was by far the most important feature (0.610 vs. 0.084 for fastest 5-second wind speed in second place).
- The least important features were seasonal and regional.



Recommendation

- Random forest performed the best with F-2 score = 0.839 vs. logistic regression’s F-2 score = 0.836
- Implement a random forest classification model and run the model multiple times daily with updated forecast data
- Forecast data will likely become more accurate as the day continues
- Model can be monitored or actions can be taken based on the predictions throughout the day
- NWS can alert potentially affected civilians to move valuables to higher elevations, stock up on essentials, and/or make plans to evacuate.

Future Direction

- Deploying the model to production and automating the model to run at an ideal frequency (i.e., as often as forecast data updates)
- Further data exploration in order to find other data sources or feature engineering to strengthen the model
- Web application for easy access to the model