Machine Learning and Insurance Claim Forecasts

Nikhil Bhandari
Rock Creek Analytics, LLC
www.RockCreekAnalytics.com
nikhil@rockcreekanalytics.com

September 30, 2020

Abstract

Insurance claims forecasting for extreme weather events that result in large scale destruction such as hurricanes, wildfires, floods, etc. is an important planning activity for insurance firms and any process improvements that can enhance the accuracy and quality of the forecasts should be welcome. This article provides an introduction to the forecasting methodology for insurance claims payouts and then discuss potential use of machine learning techniques to enhance the forecasting process.

Contents

1 Introduction

This article provides an introduction to forecasting insurance claims payouts. We focus specifically on the claims arising from weather events (*events*) that result in large scale destruction such as hurricanes, wildfires, floods, etc. We first provide a general overview of the traditional methodology and then discuss potential use of machine learning (ML) techniques to enhance the forecasting process.

While the examples and website references provided in this article are US-centric, the ideas presented herein are general and can be applied to all locations. In other regions and countries, the analyst will need to substitute the appropriate data sources for event data.

The remainder of the document is organized as follows. In the next section, we provide an overview of the general model structure and the elements that need to be considered within the modeling framework. In section III, we focus on where AI/ML techniques are most appropriate to use. In the final section, we discuss our conclusions.

Note that all the data used for illustrative figures in this article is simulated and does not reflect any actual event except where otherwise indicated.

2 Modeling Overview

A typical model to forecast the impact of a high impact weather event will involve the following steps (recognizing that specific situations require variations and/or adaptations to the general process).

- Determine the area of impact of the event.
- Determine the location of the properties affected or likely to be affected.
- For the affected properties, determine the likelihood of filing a claim.
- For the properties that are likely to file a claim, determine the payout amount.
- Run the process as a monte-carlo simulation to generate a distribution of the claim payouts.

In the following subsections we discuss each of these steps.

2.1 Property Location

It should be easy for the analyst to get the addresses of properties being insured. The addresses need to be properly geocoded (i.e., have a latitude and longitude for the address) which is generally not an issue with most newly issued policies. However, for older policies where the addresses have not been updated in a while there could be issues where the latitude and longitude of the property is not correctly determined.

The primary objective here is to convert the property addresses into latitudes and longitudes that can be used in the subsequent steps. To get the latitude and longitude of the addresses, any one of the publicly available geocoders can be used. The US Census Bureau [1] provides an excellent geocoder that is free to use. Other excellent commercial geocoders are provided by ESRI [2], CoreLogic [3], Google [4], etc. The geocoders can also help standardize the addresses which could be an additional benefit to some firms.

Note that none of the geocoders is perfect and each of these may have issues correctly identifying some of the addresses (especially if the address is not in a standard USPS format). The analyst has to either come up with an alternative way to identify the location of such policies or discard them from the analysis (not a recommended option).

2.2 Event Mapping

The next step in the process is to determine the area of influence of the event. This can be typically done using detailed maps and data provided by several US government agencies and/or university research groups. Some sources for different weather events include:

- *Precipitation* - the National Weather Service's Weather Prediction Center [6] provides precipitation forecasts for the country .
- *Hurricanes* - storm tracking, water height, wind speed and water inundation level information for hurricane events can be obtained from CERA [5].
- *Wildfires.* National Interagency Fire Center [7] provides data on perimeters of current wildfires.
- *Tornadoes.* National Weather Service provides tornado maps [8].

A typical map that shows the area of influence along with the event intensity level is shown in Figure 1. The figure shows the water inundation levels for a past hurricane event.

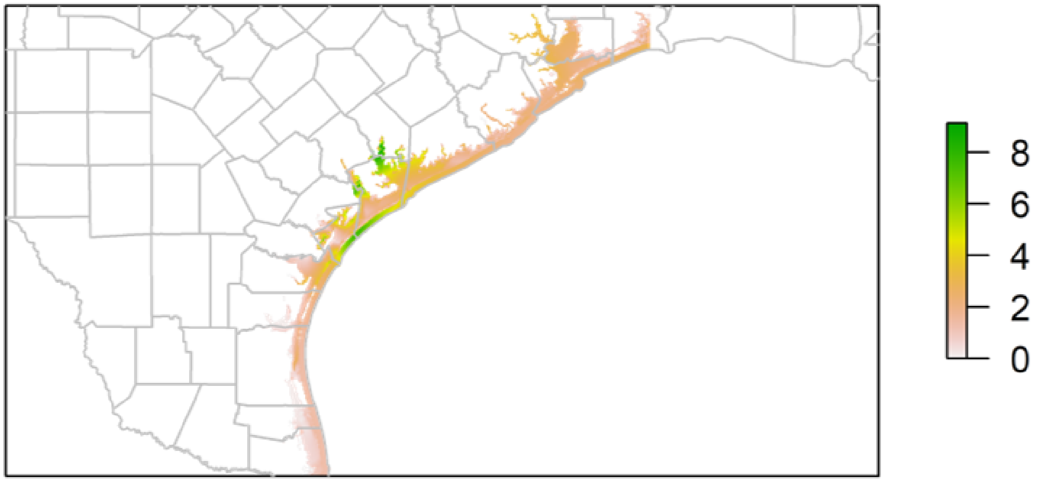

Figure 1: Maximum inundation levels (ft) for Hurricane Harvey.
Data source: CERA [5].

Once such a map is obtained, the property address can be superimposed using any GIS tool or a statistical or programming software. Figure 2 shows policy addresses on top of the map showing the water inundation levels. Using these two pieces of information, the analyst can then extract the event intensity level for each property location.

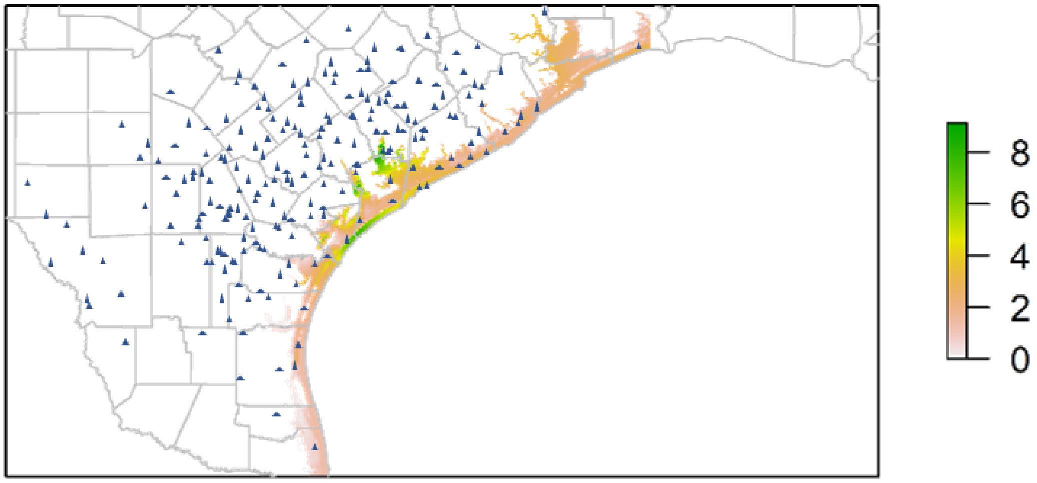

Figure 2: Policy locations with water inundation levels (ft) - illustrative.
The policy addresses shown as blue dots are simulated and do not correspond to any actual addresses.

2.3 Who will File the Claim

Most extreme weather events have one common theme that even within the event's area of influence, not every location is impacted similarly. Some of this is purely due to the characteristic of the event itself (for example, a hurricane generates storm surges that are dangerously high in some locations while moderate in some locations), some of this is due to the property location (for example, a property that is on top of a sloping road vs. the property that is at the bottom of the same road), and some of this is due to the property characteristics (for example, a property with good drainage vs another with bad drainage).

This results in a situation where not everyone within the event's area of influence will file a claim. However, the likelihood of filing a claim will typically increase with the increase in the intensity of the event and this increase usually stabilizes at some point. Thus, if a storm generates a large amount of rainfall, it is likely that more claims will be filed in locations with higher inundation levels.

It is therefore important to have a good model of who is likely to file a claim and how does the likelihood change with the intensity of the event. This is typically done as a statistical analysis of past events.

The model can be a simple cross-tab that calculates the share of folks filing a claim for a specific intensity level - such a model is represented in Figure 3.

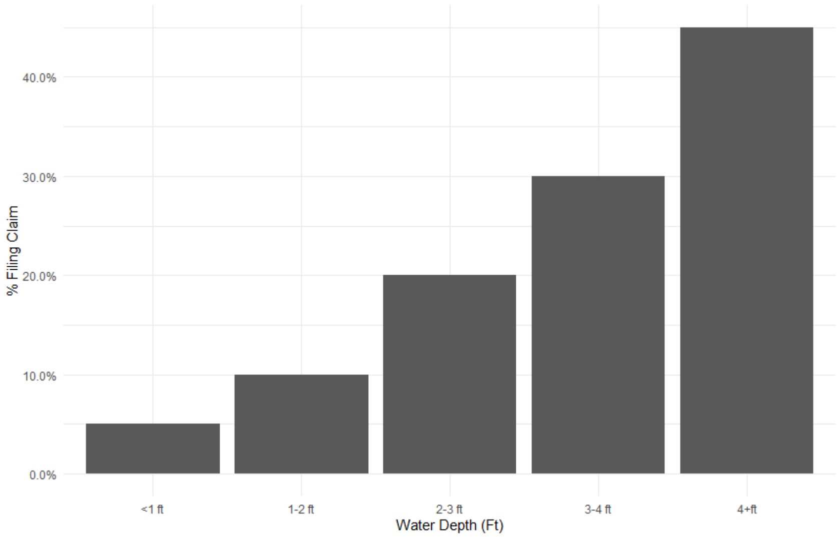

Figure 3: Likelihood of filing a claim for a flooding event (illustrative).

A more sophisticated model will be a logistic regression / classification model of the form:

$$p = \frac{1}{1+\exp^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots)}}$$

where p is the share, x_i are the predictors, and β_i are the parameters of the model. With sufficient data, a logistic classification model can be estimated for different past events individually or a general model can be estimated that considers multiple past events.

The advantage of this model is that it allows the analyst to understand the relationships between the different predictors and perhaps also help the insurance firm to proactively work with policy holders to make appropriate changes to their property to limit/reduce the impact of a future severe weather event. Other typical econometric classification models such as the probit regression and linear discriminant analysis can also be considered.

2.4 Payout Distribution

At this point based on the prior steps, the analyst should know the location of each of the policies in the event's area of influence, the intensity level associated with the event, and the likelihood of filing a claim. The final piece of the puzzle is to figure out how much will be paid to the policy holder if they were to file a claim. This can be accomplished by undertaking a statistical analysis of prior events. The objective of such statistical analyses of past events is to determine a payout distribution.

Typically, some of the claims are denied, some are paid the total insured value (in cases of complete destruction of the property) and the remaining claims are paid somewhere between zero and the total insured value. Depending on the event, this could be a smooth distribution, or it could be a situation where there are large spikes at the ends (i.e., large number of denied claims and a large number of full payout) and a smooth distribution for the intermediate range. The analyst can look at several past events and determine if there is a consistent pattern of the payout. Figure 4 shows an example of such an analysis where event's payout follows a log-normal distribution pattern.

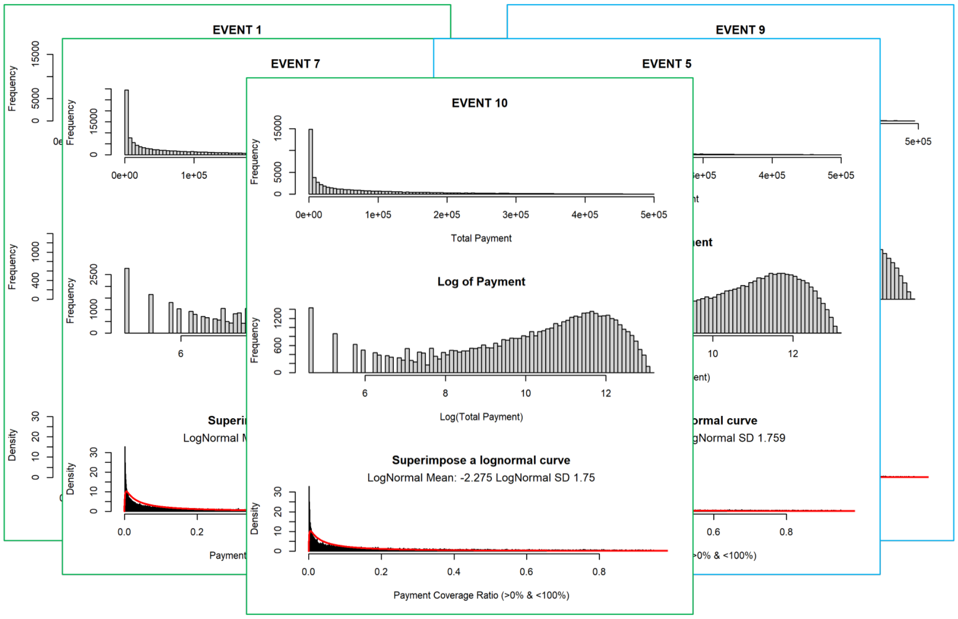

Figure 4: Estimating payout distribution using data from a number of past events (illustrative).

2.5 Estimating Total Payout

The final step in the process is to combine all the different elements discussed earlier to estimate the total payout for the event. This is typically done as a monte-carlo simulation with multiple iterations.

In each iteration, for each policy in the event's area of influence, the event intensity level for the policy address is used to determine if the policy will file a claim or not, and if it does file a claim then what the payout amount will be. The total payout amounts are added up for the iteration. Once a sufficient number of iterations are performed, one can create a total payout distribution as shown in Figure 5.

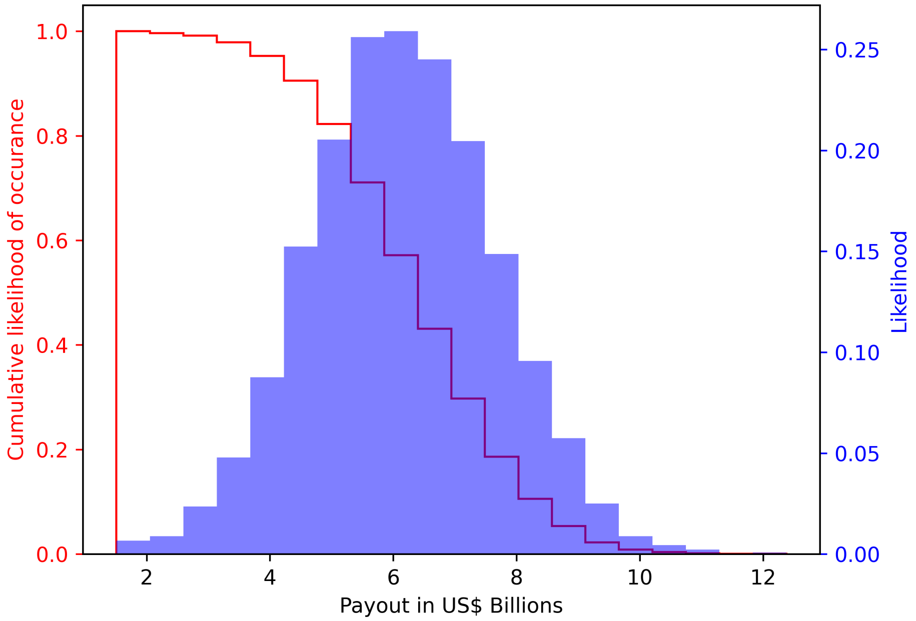

Figure 5: Total claim payout estimate for a hypothetical event.
The simulation assumes an event with 100,000 claims filed on average (distributed normally with a standard deviation of 25,000 claims, and 25,000 minimum number of claims for each iteration), an average coverage amount of $300,000, maximum coverage of $1,000,000, 8% of claims are denied, 10% of claims receive 100% of coverage amount, and intermediate claim payouts follow a log-normal distribution. The model is run for 2,500 iterations.

3 Use of ML Techniques

The process for estimating claims payouts described in the previous section relies heavily on two key relationships. One is the relationship of the intensity of the event to the likelihood of filing a claim and the other is the distribution of the payout amounts. In the traditional process as outlined earlier this is done via standard econometric and statistical modeling. In this section, we explore how non-traditional machine learning techniques can be used to enhance the estimation of these relationships. Note that in this article, we will not address the techniques used for event mapping as they are quite complex subjects and not within the scope of this article.

For the purposes of this article, we would distinguish between traditional econometric modeling (e.g., regression, logit, etc.) with the non-traditional machine learning techniques (e.g., decision trees, neural networks, etc.) Within the context of a typical problem of understanding the relationship between multiple variables, there is no clear demarcation as to which model is an econometric model and which is a machine learning model - the distinction is often ”cultural” and depends on the training background of the analyst (i.e., whether the analyst is trained primarily in economics or in computer science). In our view, the key difference between the two is the objective of the modeling task itself. The other important fact is that ML techniques can be applied to non-econometric situations as well (e.g., image recognition, speech recognition, etc.) - one can view the ML approach as a much larger set of potential techniques that can be applied to econometric problems as well as non-econometric problems.

The typical objective of standard econometric modeling is to understand which of the independent variables are important in explaining the behavior of the dependent variable, and how the model can help explain an underlying theoretical model. In this approach, one would often prefer a model where the estimated coefficients of the independent variables are in conformance with the expectations over another model with a better fit but where the coefficients of the independent variables are not in conformance with the expectations.

The focus of the ML techniques, as applied to econometric problems, is more on the speed of execution, ease of codability, minimizing the number of independent variables and the ability to ”learn on the fly.” These techniques are quite useful and appropriate in situations with very fast flowing data where the model needs to provide a result quickly and also adapt itself continuously.

In the following sub-sections we will discuss the ML techniques that can be helpful in enhancing the claim payout estimation process.

3.1 Claim Filing and Event Intensity

As discussed earlier, the typical model for claim filing is a classification model that uses the event intensity as the key independent variable. Other variables that can be included are property characteristics.

The appropriate machine learning techniques to reflect the relationship between claim filing, event intensity and property characteristics include random forest decision tree (see Figure 6) or a neural network based algorithm (see Figure 7). These ML techniques allow for non-linear relationships to be estimated. The models have to be ”trained” using past data sets and models can be improved by ”learning” as new data comes in.

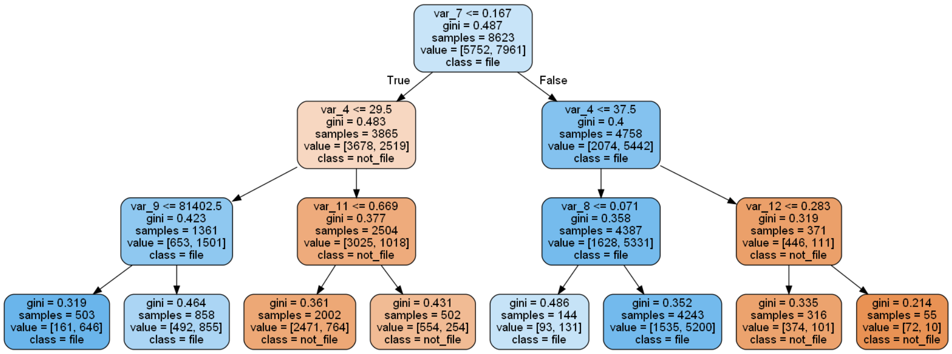

Figure 6: Random forest decision tree (illustrative).

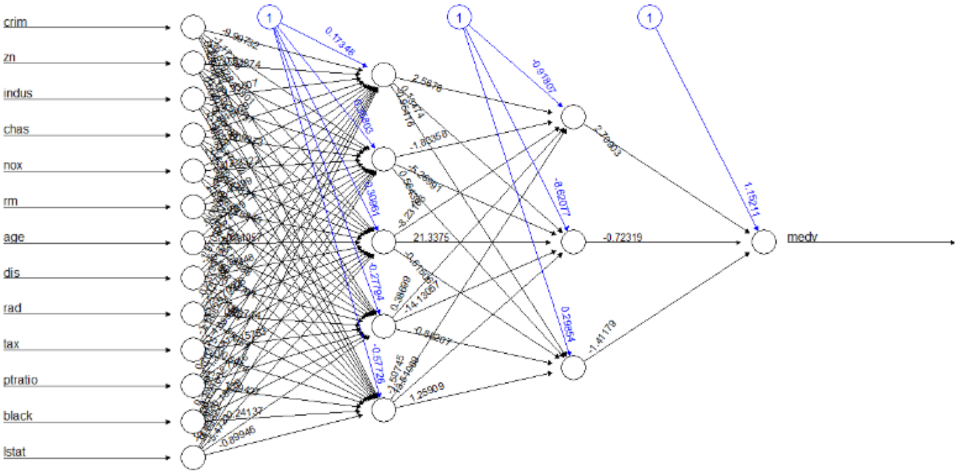

Figure 7: Neural network model (illustrative).

3.2 Estimating the Payout Distribution

In machine learning terms, estimating the payout distribution is essentially a pattern recognition problem. Some event types have well recognized payout distribution patterns that can be easily fitted with a standard statistical distribution (such as normal or log-normal) while there are some events where the payout does not follow a standard distribution. In such cases machine learning techniques such as neural networks, clustering methods (hierarchical, k-means, correlation), etc. can be quite helpful.

4 Closure

Insurance claims forecasting for extreme weather events is an important planning activity for insurance firms and any process improvements that can enhance the accuracy and quality of the forecasts should be welcome. We believe that machine learning techniques provide some areas of consideration in the overall forecasting methodology.

References

[1] *US Census Geocoder*. https://geocoding.geo.census.gov/geocoder/.

[2] *ArcGIS World Geocoder*. https://www.esri.com/en-us/arcgis/products/arcgis-world-geocoder.

[3] *PxPoint Geocoder*. https://www.corelogic.com/Products/PxPoint-Geocoder.aspx.

[4] *Google Maps Platform*. https://cloud.google.com/maps-platform.

[5] *CERA - Coastal Emergency Risks Assessment*. https://cera.coastalrisk.live/.

[6] *National Weather Service - Weather Prediction Center*. https://www.wpc.ncep.noaa.gov/kml/kmlproducts.php#qpf.

[7] *National Interagency Fire Center - Wildfire Perimeters*. https://data-nifc.opendata.arcgis.com/datasets/wildfire-perimeters.

[8] *National Weather Service - New Tornado Map*. https://www.weather.gov/gsp/newTornadoMap.