Utilizing Machine Learning & Regression Analysis to Analyze Flood Insurance Data

Warda Aziz Khan[*a], Hossein Daneshvar[*b], Mahshid Jafar Pour[*c], Anh Tran[*d]

**a: An Electrical Engineering from University of Engineering and Technology Lahore (2010); Currently taking 4th course in the Georgia Institute of Technology OMSA program.*

**b: PhD in Structural Engineering with diverse experience in different industries. Currently taking 4th course in the Georgia Institute of Technology OMSA program.*

**c: PhD in Petroleum Engineering with experience in Oil & Gas and Mining industry. Currently taking 4th course in the Georgia Institute of Technology OMSA program.*

**d: PhD in Materials Engineering with 10 years of working experience in the field of thin films/ coatings. Currently taking 5th course in the Georgia Institute of Technology OMSA program.*

1. Introduction

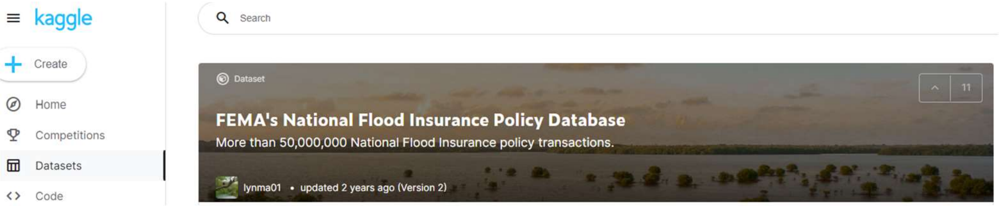
The market for providing flood insurance policies in the U.S. is almost exclusively backed by the National Flood Insurance Program (NFIP). The NFIP is only available for policies purchased within participating communities, and partners with private insurance companies to distribute flood insurance policies to homeowners and businesses. As flooding is the primary vector of economic damages inflicted on local communities as demonstrated by the 2016-2019 hurricane seasons and given the projected increase in destructive flooding due to of climate change, there is an enormous need to efficiently distribute financial risk due to climate change.

Here are some research questions we are trying to answer in this project:How does flood zone, elevation, property state, no. of floors and built year affect insurance cost, coverage and premium? What is the correlation between different variables? What are the most important variables in premium estimation? Can we predict premium based on the variables available?

2. Dataset

2.1 Datasets and/or Potential Data Sources

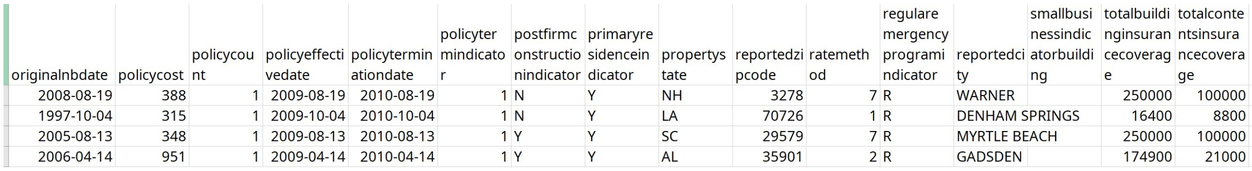
The dataset for flood insurance is available for analysis, development, visualization, or transparency through the website of "FEMA's National Flood Insurance Policy Database" which includes more than 50,000,000 National Flood Insurance policy transactions. In addition, two other data sets can be used if needed. The first one relates to the major events and the second one with 2 million claims that can be analyzed for claim amount based on similar features. The screenshot of dataset and website are shown as below, in the Figure 1. This data is collected over 12 years. We can refer to the second dataset with the list of major events. Also, the third dataset with 2 million claims that can be analyzed for claim amount based on similar features. Due to a large size of the dataset with 50 million rows approximately, we will concentrate on a subset dataset which represents for a particular state or city. For example, we will focus on the data of Houston which has the highest frequency in the dataset.



FEMA's National Flood Insurance Policy Database

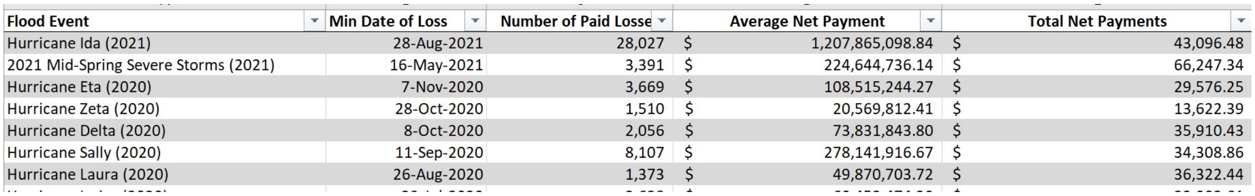
More than 50,000,000 National Flood Insurance policy transactions.

lynma01 • updated 2 years ago (Version 2)



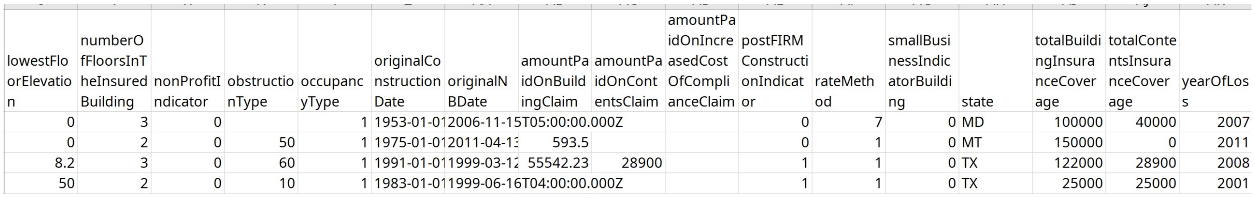
originalnbdate	policycost	policycount	policyeffectivedate	policyterminationdate	policytermindicator	postfirmconstructionindicator	primaryresidenceindicator	propertystate	reportedzipcode	ratemethod	regularemergencyprogramindicator	reportedcity	smallbusinessindicatorbuilding	totalbuildinginsurancecoverage	totalcontentsinsurancecoverage
2008-08-19	388	1	2009-08-19	2010-08-19	1	N	Y	NH	3278	7	R	WARNER		250000	100000
1997-10-04	315	1	2009-10-04	2010-10-04	1	N	Y	LA	70726	1	R	DENHAM SPRINGS		16400	8800
2005-08-13	348	1	2009-08-13	2010-08-13	1	Y	Y	SC	29579	7	R	MYRTLE BEACH		250000	100000
2006-04-14	951	1	2009-04-14	2010-04-14	1	Y	Y	AL	35901	2	R	GADSDEN		174900	21000

Ref: https://www.kaggle.com/lynma01/femas-national-flood-insurance-policy-database



Flood Event	Min Date of Loss	Number of Paid Losse	Average Net Payment	Total Net Payments
Hurricane Ida (2021)	28-Aug-2021	28,027	$ 1,207,865,098.84	$ 43,096.48
2021 Mid-Spring Severe Storms (2021)	16-May-2021	3,391	$ 224,644,736.14	$ 66,247.34
Hurricane Eta (2020)	7-Nov-2020	3,669	$ 108,515,244.27	$ 29,576.25
Hurricane Zeta (2020)	28-Oct-2020	1,510	$ 20,569,812.41	$ 13,622.39
Hurricane Delta (2020)	8-Oct-2020	2,056	$ 73,831,843.80	$ 35,910.43
Hurricane Sally (2020)	11-Sep-2020	8,107	$ 278,141,916.67	$ 34,308.86
Hurricane Laura (2020)	26-Aug-2020	1,373	$ 49,870,703.72	$ 36,322.44

Ref: https://nfipservices.floodsmart.gov/reports-flood-insurance-data



lowestFloorElevation	numberOfFloorsInTheInsuredBuilding	nonProfitIndicator	obstructionType	occupancyType	originalConstructionDate	originalNBDate	amountPaidOnBuildingClaim	amountPaidOnContentsClaim	amountPaidOnIncreasedCostOfComplianceClaim	postFIRMConstructionIndicator	rateMethod	smallBusinessIndicatorBuilding	state	totalBuildingInsuranceCoverage	totalContentsInsuranceCoverage	yearOfLoss
0	3	0		1	1953-01-01	2006-11-15T05:00:00.000Z				0	7	0	MD	100000	40000	2007
0	2	0	50	1	1975-01-01	2011-04-1	593.5			0	1	0	MT	150000	0	2011
8.2	3	0	60	1	1991-01-01	1999-03-1	55542.23	28900		1	1	0	TX	122000	28900	2008
50	2	0	10	1	1983-01-01	1999-06-16T04:00:00.000Z				1	1	0	TX	25000	25000	2001

Ref: https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v1

Figure 1: Snapshot of datasets with their associated references

2.2 Data Wrangling

A typical data of "flood insurance claim" is shown in Figure 1. It includes several important attributes as elevated building indicator, flood zone, latitude, longitude, no. of floors of an insured building, state, total building insurance coverage, total content insurance coverage, insurance premium, insurance cost and original construction date. For each claim, several information can be achieved including data cleaning:

- scaling and normalizing;
- remove/ impute invalid data/ missing data/ outliers;

- utilize dummy variables for categorical variables if needed;
- parsing dates must be employed;
- data visualization to explore if there is any correlation (both linear and Spearman) between variables.

3. Methodology

The main aim of this work is to predict the premium based on historical claim data of users which can efficiently help estimate the premium in advance. To fulfil this task, the data firstly are groups into clusters based on location and claim frequency. Then a regression model is used to predict the probability of flood incident from clustering result. Finally, the model is validated for its performance. Details of each step are described as below:

- Feature selection: to ensure selected features are uncorrelated: use Lasso, ElasticNet regression; cluster similar features with their correlations as a distance measure
- Explore time plots and determine the periods where policy cost increases with time. (Possibility of an event/flood), validate with second dataset that includes major floods
- Regression analysis (both linear regression and log transformation if needed):
  1) Estimate policy cost and total building insurance coverage based on different features. Some that seem important are flood zone, elevation, property type, whether the building is under construction or not, insurance deductible, insurance coverage, policy rating method
  2) Estimate total insurance on the content based on different features: flood zone, elevation, insurance coverage, location of contents, obstruction type
- Feature Engineering: explore possibility of making new features that are more relevant in modeling.
- Classification: to classify locations with and without property damage
- Prediction: KNN, SVM or Random Forest algorithms. Used to predict the amount of property damage from a particular flash food event
- Validation: k-fold cross validation is used to evaluate the performance of models.

4. Anticipated Discoveries or Conclusions

We can comment on the most important variables affecting the insurance cost, coverage or premium e.g., flood zone, elevation, property state, no. of floors and built year. Also, the total number of claims in the country or in each state can be calculated, after major flood events. This information can provide valuable information for the insurance companies on how to revise their policies to maximize their profits; also help the insurance clients to make a data-driven decision when purchasing an insurance policy.