**Predicting Whether a Homeowner Carrier Sufficient Flood Insurance**

**Introduction**

**Problem Statement & Benefit**

Most insurance carriers do not offer flood insurance for homes and other structures. Instead, the U.S. government offers flood insurance through the National Flood Insurance Program (NFIP). According to the NFIP, 90% of natural disasters involve flood but only 12-14% of households purchase flood insurance. Banks require households to purchase a flood policy for structures that sit within a flood plain, but banks do not require households to verify that the policy coverage contains suitable limits[[1]](#footnote-1) over time as owners remodel buildings or as inflation increases the cost to replace the structure.

Using claims data acquired from the NFIP, we will build a model to predict whether a policyholder is underinsured. Underinsured policies occur when a flood loss exceeds the limits specified in the flood insurance contract.

Our analysis will not predict the probability of a household incurring a flood loss. This problem would require a specific listing of flood policies and historical information about floods within each geographic area. Because most households acquire a loan to purchase a property, and because banks compel homeowners to carry flood insurance when risks are present, a model predicting probability of a loss feels unnecessary.

**Stakeholders & Proposal**

Our stakeholders are government agencies running the flood program and homeowners. Homeowners and government agencies can use model predictions to ensure a policyholder purchases enough limit to protect a policy holder from excessive losses. Policyholders will want to use the model predictions to understand their own policy’s risk of underinsurance. Government agencies can use predictions to develop marketing campaigns in areas with a higher percentage of underinsureds to influence policyholder to purchase higher limits.

A secondary stakeholder group is the United States taxpayers. Use of the model helps ensure policyholders purchase enough limit to avoid uncovered losses. In large events, it is common that the US Federal Government provides bail out payments to individuals for uninsured losses. Proper use of the model can reduce the frequency and severity of taxpayer-funded bailouts.

I would suggest pitching this project to the NFIP first. The stakeholder group is easy to reach with a small group of decision makers, and they have a motive to improve the profitability of the program through increase of premiums collected.

The NFIP publishes public data on flood losses for modeling and other purposes; no additional work is needed by the agency to begin our project. Depending on how the NFIP intends to implement the model, they may need to invest in applications to publish modeled results via the web or invest in marketing to make households that are potential underinsured aware of their policy deficiencies. Once a solution is built, the household stakeholder group would be engaged through consultation with NFIP employees or via the web (however the NFIP chose to publish the model). US taxpayers do not need influencing as they are secondary beneficiaries, but this would be a productive win for elected officials to tout the savings they are achieving for taxpayers.

**Data Source**

The NFIP publishes claims data on their website[[2]](#footnote-2). The data set includes loss information on each policyholder who incurs a loss during an event. The data set includes many additional policy attributes that will help us assess the total insured losses per unit of covered risk.

**Data Summary, Analysis, and Procedures**

**Data Preparation**

The model target is the underinsured indicator on each claim (1 if True; 0 if False). A claim is underinsured if the claim payment for the building loss is equal to or exceeds the policy limit for building limit. The underinsured field is not available in the NFIP data set, but we can calculate it by comparing the two fields noted (building loss amount and building coverage amount).

Because the NFIP will not pay more than a policy’s limit specifies, we cannot quantify the amount of underinsurance on each claim. Without the total insured loss, we can not suggest appropriate policy limits for underinsured policies. Our model will only seek to identify underinsured policies. If NFIP stakeholders want to model the amount underinsurance, internal data within the NFIP claims system could be obtained and incorporated into the pipeline as a future enhancement.

Beyond construction of the target field, the data required modifications to prepare for modeling. Several fields required changes to the input type, mostly to categorical variables. Fields with more than 30% of missing values were dropped. Fields not relevant to the model were also dropped from the pipeline. These include:

* 'id': Field is a randomized identifier of record for matching to NFIP internal database. Not usable or interpretable otherwise.
* 'asOfDate': Field is system timestamp of data refresh. Delete due to system key.
* 'longitude': Remove longitude field. We will be using city and state as factors; this is duplicative.
* 'latitude': Remove latitude field. We will be using city and state as factors; this is duplicative.
* 'reportedCity': This data lists all values as "Temporarily Unavailable" and is not meaningful.
* 'reportedZipcode': Remove as we will use state for our model. Zip code may be too granular to derive correlations given the sparsity of claims within each zip code (with Katrina and other mega-catastrophes as exceptions).
* 'censusTract': Another measurement of location, focusing on geolocation. We will use state and county for our model.
* 'dateOfLoss': 'dateOfLoss' overlaps with year of loss and is too granular given the sparsity of claims on a given day.

**Data Analysis**

Flood losses have trended higher since the 1970’s. Within the 2000-2020 years, there have been several large flooding events caused by hurricanes. Spikes in the data are evident in the chart below.

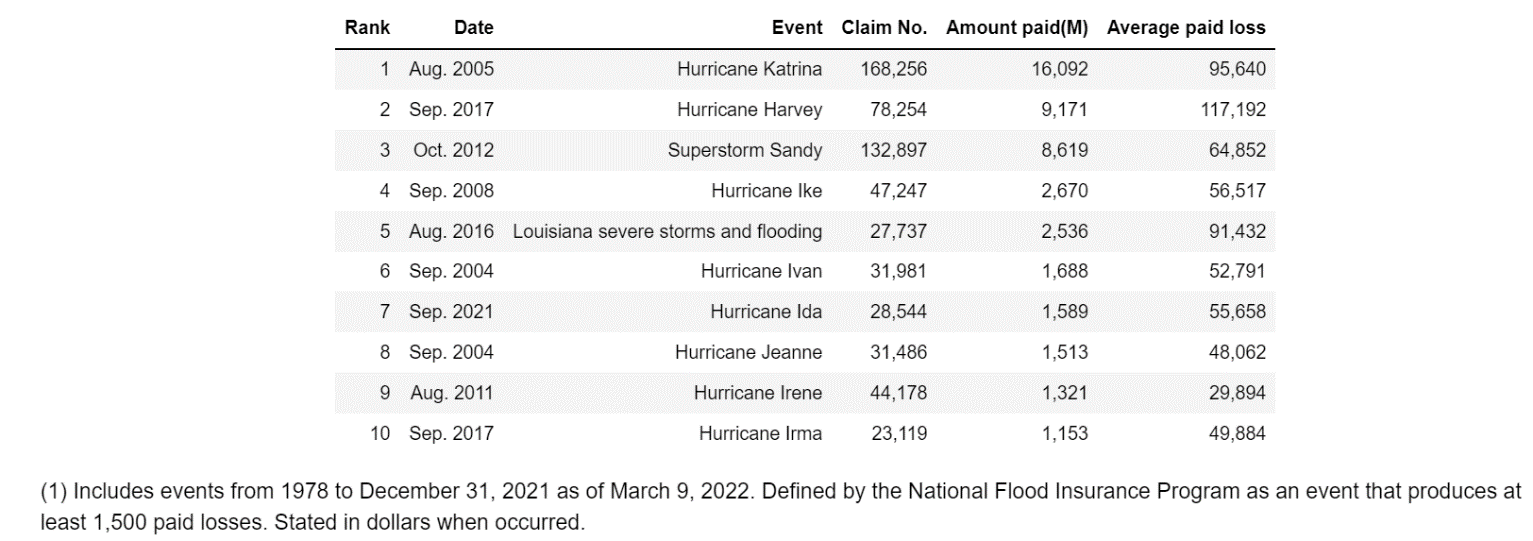
Exhibit 1: Flood Losses Increasing Over Time

Chart, line chart, histogram

Description automatically generated

Based on the information from the Insurance Institute, the large losses in 2005, 2012, and 2017 are the result of certain events. Large losses are rare (low probability of occurring over time). Additional exploration of underinsureds for these events is needed to ensure model predictions are not overly influenced by these rare events.

Exhibit 2: Several Significant Insurance Events Impact Flood Losses (1)



Hurricane Katrina in 2005 resulted in massive flooding that submerged most of the New Orleans area for weeks. This flooding event resulted in thousands of total property losses within New Orleans. The rarity and significance of this flood resulted in a large spike in underinsured losses. We will need to adjust for this in our modeling to avoid this rare event overinfluencing our target projections. This may require building a model without Katrina losses or balancing other underinsured losses to compensate for the larger percentage of Katrina claims. Exhibits below show spikes in underinsureds during 2005 and for the state of Louisiana.

Exhibit 3: Underinsured by Year

**Chart, line chart

Description automatically generated**

**Geographic Split of Underinsured Losses**  
Geographic location may influence the rate of underinsured because of differing trends in property inflation or economic status of homeowners. Geography variables may also indicate or pick-up on the presence of large flooding events, those that cause substantial damage likely to exceed homeowner policy limits. It will be important to assess causes of geographic results when building features.

Exhibit 4: Underinsured Claims by Geography

**A picture containing shape

Description automatically generated**

Exhibit 5: Underinsured Percent of Total Claims, by Geography

**Chart, histogram

Description automatically generated**

**Model Preparation**

Model building required categorical variables to be one-hot encoded. Additionally, features were converted using StandardScaler from the scikit-learn library to ensure all distributions were standard normal.

Underinsured claims represent 12% of our data set, which constitutes an unbalanced class. Applying SMOTE, a synthetic rebalancing procedure, we reweighted the underinsured attribute to balance classes. SMOTE is an oversampling technique that generates new observations by randomly selecting an observation from the minority class, finding its nearest neighbors, then selecting a point in between the selected observation and the nearest neighbor.

The business question requires a classification model. Logistic regression, random forest, and AdaBoost are common algorithms for solving classification problems. Flooding losses are complex due to the related nature of a building and geography. An expensive building near the ocean is more likely to suffer a large loss than the same building miles away from water. The Random Forest and AdaBoost algorithms may better recognize the multi-faceted nature of our variables. A decision tree algorithm is also an option for classification problems. Since we are running both Random Forest and AdaBoost algorithms, both of which are more robust and better at predicting results, we will not run a decision tree model for this problem.

Government agencies will prefer the model identify all underinsureds, even if the model includes some false positives. The recall, true positives divided by all positive observations, is appropriate to optimize the model for this audience. Policyholders may want a more balanced model – buying enough insurance to cover potential losses while not paying more for a policy than necessary. Losses can potentially be a larger financial consequence, so we will use recall as the principal metric. Some consideration will be given to precision (true positives divided by all predicted positives) to ease policyholder concerns that the model is not just telling everyone to purchase more insurance.

**Model Results**

Logistic Regression performed best on recall with a score of 0.74, but this model appears to overpredict underinsureds as the precision is low (0.26) and the F1 score is 0.39. AdaBoost and Random Forest performed nearly as good as logistic regression on the stakeholder’s target metric of recall. Both achieved a 72% score for recall. AdaBoost performed marginally better than Random Forest on our secondary metric, precision, with a score of 0.32 (versus 0.31 for Random Forest).

We used GridSearch to optimize parameters. GridSearch used the recall statistic for optimizing performance, but the best model did not achieve a better recall score than AdaBoost or the originally run Random Forest model. If precision and recall were equally important (see F1 score of 0.53), we would use the optimized model run from this process.

The GridSearch process used ‘recall’ to find the best model. This process yielded a lower recall score on our target variable (underinsured =1, or true underinsured) than non-optimized models. Given the run times of 7-8 hours to find our best model, it is not practical to rerun this process with new hyperparameters. However, we should consider retuning our model before publishing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Recall** | **Precision** | **F1** | **Accuracy** |
| **Logistic Regression** | 0.74 | 0.26 | 0.39 | 0.73 |
| **Random Forest I** | 0.72 | 0.31 | 0.44 | 0.78 |
| **AdaBoost** | 0.72 | 0.32 | 0.44 | 0.79 |
| **Random Forest w/ GridSearch Optimization** | 0.64 | 0.46 | 0.53 | 0.87 |

**Conclusion**

The models selected seems useful for predicting whether a policyholder is underinsured. Geography, purchased limits, building characteristics, and other variables accurately predict 72% of underinsureds. Results appear significant given that underinsureds represent only 12% of our sample. Government agencies and policyholders should feel comfortable using the model to provide guidance on insurance purchases.

As noted above, I would recommend one additional GridSearch optimization process to tune the AdaBoost or retune the Random Forest model to see if better results can be achieved.

Beyond our original business problem, the NFIP should consider additional, more informative questions for future study. Obtaining internal data on the total loss (including losses assessed but not paid by the NFIP), the government could predict the amount of underinsurance for each loss. This information would inform how much additional limit a policyholder should purchase. We should also perform additional work to identify important features in our model. Identifying features that predict underinsurance would provide the NFIP and policyholders information about how they mitigate underinsured losses. Key features also inform a policyholder about why they need additional insurance, which might improve marketing efforts to these individuals.

1. A limit is an amount specified in an insurance contract that dictates the maximum amount an insurer will pay under that contract for a covered loss. For flood policies, contracts will specify one limit to cover damage to the building and another limit to cover loss of contents held within the structure. [↑](#footnote-ref-1)
2. Link to data: <https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v1> [↑](#footnote-ref-2)