Utilizing Machine Learning & Regression Analysis to Analyze Flood Insurance Data

Warda Aziz Khan, Hossein Daneshvar, Mahshid Jafar Pour, Anh Tran

1. Introduction

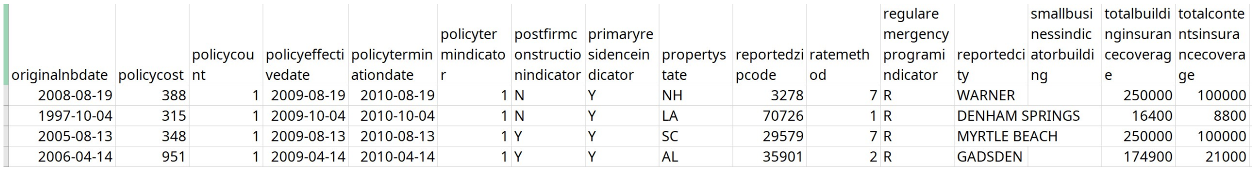
Damage caused by floods is not covered by a standard homeowner insurance. Instead, they are covered by policy issued by National Flood Insurance Program (NFIP), overseen by FEMA (Federal Emergency Management Agency). In other words, the market for providing flood insurance policies in the U.S. is almost exclusively backed by the 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. The top 10 most significant NFIP loss events are almost all hurricanes. 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 climate change, there is an enormous need to efficiently distribute financial risk.

During the last five decades, NFIP has earned more in premiums in most years than it has paid in claims. Take-up rates go higher after the occurrence of a disaster event. Rates are determined based on expected claims and flood zones on Flood Insurance Rate Map (FIRM). To calculate the premium, the current rating system considers the flood zone, the building occupancy type, the foundation type, the number of floors, the presence or not of a basement, whether the property is entitled to a subsidy, whether the property is a primary residence, prior claims, and the structure's elevation relative to the BFE (base flood elevation). The amount of coverage and the deductible will also affect the premium [1]. FEMA is now implementing risk rating 2.0. According to risk rating 2.0, premiums are calculated based on the features of individual property. There are caps beyond which the premiums cannot be increased during a certain year. Reason for policy cancellation is that people think of it as a wastage of money if they have not claimed for a few years. People living in high-risk areas keep the insurance policy for years [2].

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 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

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 [3] and the second one with 2 million claims that can be analyzed for claim amount based on similar features [4]. The screenshot of dataset and website are shown as below, in 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 has 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 have concentrated on a subset dataset which represents the data of Houston, TX which has the highest frequency in the dataset.



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

Figure 1: Snapshot of datasets with their associated references

3. Data Analysis (Progress so far):

The original Policy data is too big to analyze on personal computers (~12 GB). Therefore, we filtered the data to Houston only, using SQL and Spark. The data are exported as a parquet file for further cleaning in R. The data includes 45 columns representing for 44 predictor variables and 1 response variable. There are 2,029,540 rows in total. The total insurance premium of the policy will be predicted based on several main groups as flood zone rating, building elevation, structure, building type, location, insurance deductible, total insurance coverage, business type.

3.1. Data exploration through visualization:

Extensive data visualizations are made to get more sense of the data, similar to what is shown in Figure 2.

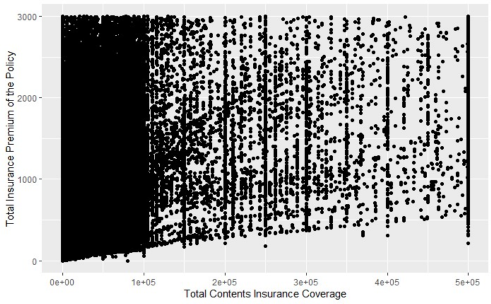


Figure 2: Snapshot of datasets with their associated references

3.2. General cleaning

The dataset must be cleaned before performing any analysis. Irrelevant columns and columns with too many NA values (>60% of rows), as well as few rows (less than 10 rows) with negative values in continuous columns were removed. Afterwards, dummy variables for categorical columns were applied. In addition, the column types were formatted as date and numeric values. In this work, the response, "Total Insurance Premium of the Policy", has a wide range of value with median = 348, mean = 543 and max = 96,900, as shown in Figure 3. However, the large value over 3,000 is only 0.1% of the whole dataset. Thus, any vales larger than 3,000 are treated as outlier and discarded from the dataset.

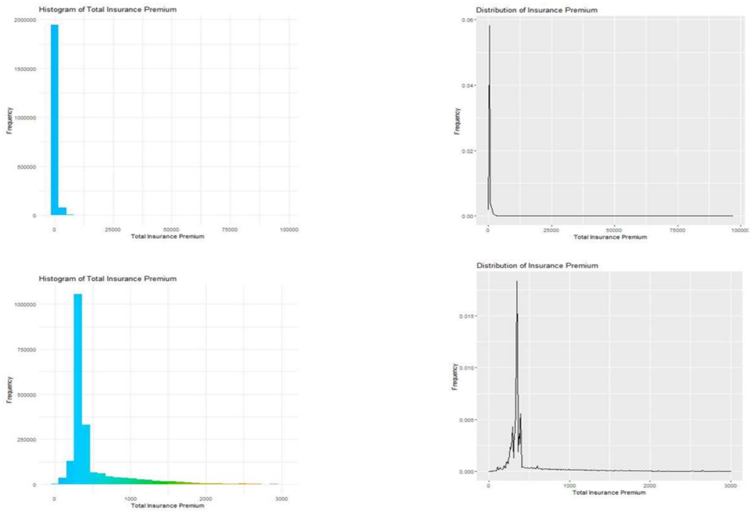


Figure 3: The histogram and distribution of response (Total Insurance Premium of the Policy): Top: original data, Bottom: after remove all the policy with the cost higher than 3,000.

3.3. Exploring relationships among variables

In this part, the correlation between numerical variables was examined to find out if there were highly correlated predictors in the dataset which might result in overfitting of the regression model. By removing these predictors, the model was easier to fit and interpret. To do that, several methods have been used, including cut-off value for both linear and Spearman correlation, linear regression, Ridge and Lasso. The correlations between all numeric variables are shown in Figure 4. It is observed that the crsdiscount and ratemethod has a strong correlation. In this case, a VIF cut-off = 5 is used to remove any correlated variables. We tried lasso with 5-fold cross-validation and stepwise regression for more feature selection. So far, the feature selection focus was on linear model. We will try non-linear models that may increase the model accuracy.

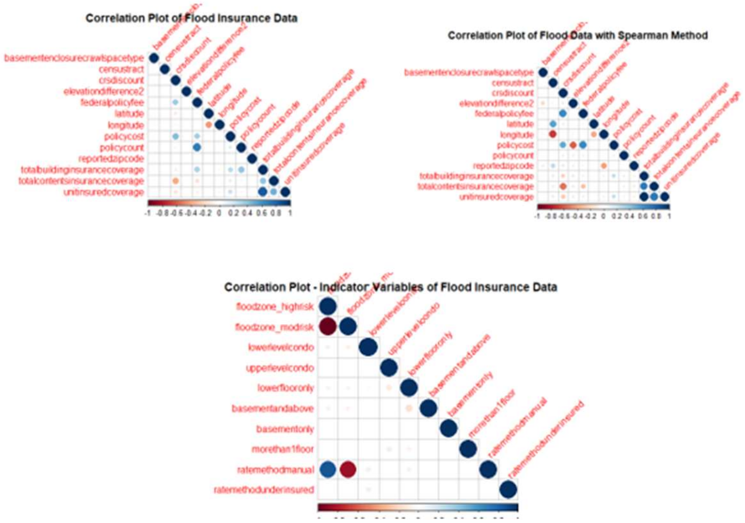
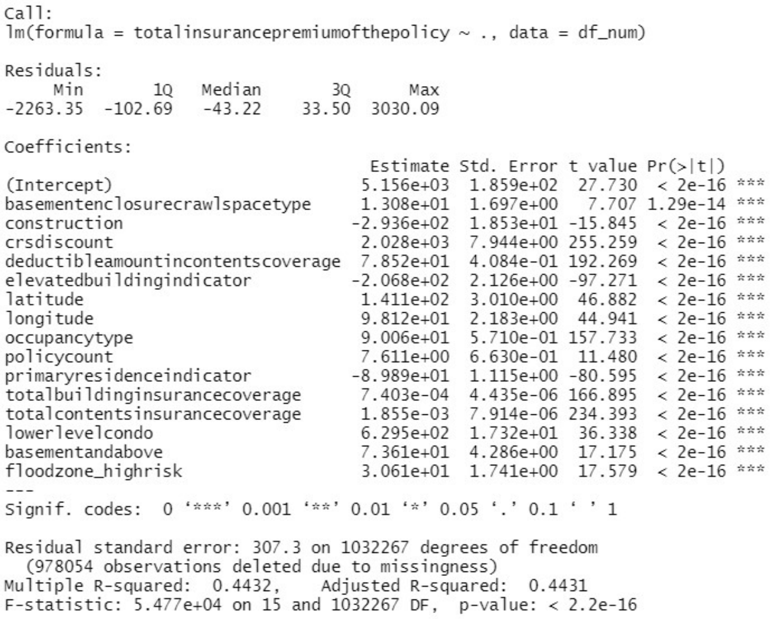


Figure 4: Correlation between numeric and categorical variables (Top left: linear correlation for numeric; top right: spearman correlation for numeric data; bottom: linear correlation for categorical variables)



```
Call:
lm(formula = totalinsurancepremiumofthepolicy ~ ., data = df_num)

Residuals:
     Min       1Q   Median       3Q      Max 
-2263.35  -102.69   -43.22    33.50  3030.09 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        5.156e+03  1.859e+02  27.730  < 2e-16 ***
basementenclosurecrawlspacetype    1.308e+01  1.697e+00   7.707 1.29e-14 ***
construction                      -2.936e+02  1.853e+01 -15.845  < 2e-16 ***
crsdiscount                        2.028e+03  7.944e+00 255.259  < 2e-16 ***
deductibleamountincontentscoverage 7.852e+01  4.084e-01 192.269  < 2e-16 ***
elevatedbuildingindicator         -2.068e+02  2.126e+00 -97.271  < 2e-16 ***
latitude                           1.411e+02  3.010e+00  46.882  < 2e-16 ***
longitude                          9.812e+01  2.183e+00  44.941  < 2e-16 ***
occupancytype                      9.006e+01  5.710e-01 157.733  < 2e-16 ***
policycount                        7.611e+00  6.630e-01  11.480  < 2e-16 ***
primaryresidenceindicator         -8.989e+01  1.115e+00 -80.595  < 2e-16 ***
totalbuildinginsurancecoverage     7.403e-04  4.435e-06 166.895  < 2e-16 ***
totalcontentsinsurancecoverage     1.855e-03  7.914e-06 234.393  < 2e-16 ***
lowerlevelcondo                    6.295e+02  1.732e+01  36.338  < 2e-16 ***
basementandabove                   7.361e+01  4.286e+00  17.175  < 2e-16 ***
floodzone_highrisk                 3.061e+01  1.741e+00  17.579  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 307.3 on 1032267 degrees of freedom
  (978054 observations deleted due to missingness)
Multiple R-squared:  0.4432,	Adjusted R-squared:  0.4431 
F-statistic: 5.477e+04 on 15 and 1032267 DF,  p-value: < 2.2e-16
```

Figure 6: Rerun linear regression after Lasso

For the Linear Regression models, Stepwise, Ridge, Lasso were used to identify a subset of variables. Lasso coefficients are relatively small but they did not shrink to zero. Thus, all the selected variables have been kept. The linear regression model was then used to test the Lasso and showed that all variables were significant with the $R^2 = 0.44$, as shown in Figure 6. We also found that the Lasso result is similar to Ridge model.

4. Preliminary Results

After the cleaning step, the dataset includes 15 crucial variables and 1,987,995 datapoints which is ready for analysis. These variables represent for:

- Building structure: basementenclosurecrawlspacetype, construction, originalconstructiondate
- Building elevation: elevatedbuildingindicator
- Location: latitude, longtitude
- Building type: lowerlevelcondo, basementandabove
- Flood insurance deductible: deductibleamountincontentscoverage
- Total insurance coverage: crsdiscount, policycount, policyeffectivedate, totalcontentsinsurancecoverage, totalbuildinginsurancecoverage
- Business type: occupancytype, primaryresidenceindicator
- Rate method: floodzone_highrisk

4.1 Linear Regression model

The lm() function is used to fit a linear regression model on the data. Total insurance premium charge is used as dependent variable whereas all other features are used as explanatory variables to train the model. Upon building the model, we observe the regression coefficient values for each predictor. In this part, we split the data to by 70% for training and 30% for testing. The numerical variables are standardized. The reported coefficients of linear regression are shown in Figure 7.

The MinMax of the linear regression which indicates the accuracy rate of each row is 62%. With a large dataset with nearly 2 million datapoints and some outliers, the model accuracy of 62% is a good initial attempt. It is observed that nearly all variables (except construction, elevatedbuildingindicator and primaryresidenceindicator) are positively related to the total premium insurance charge, indicating the insurance chare would increase with these factors. It is also noted that the “lowerlevelcondo” strongly affects the policy cost.

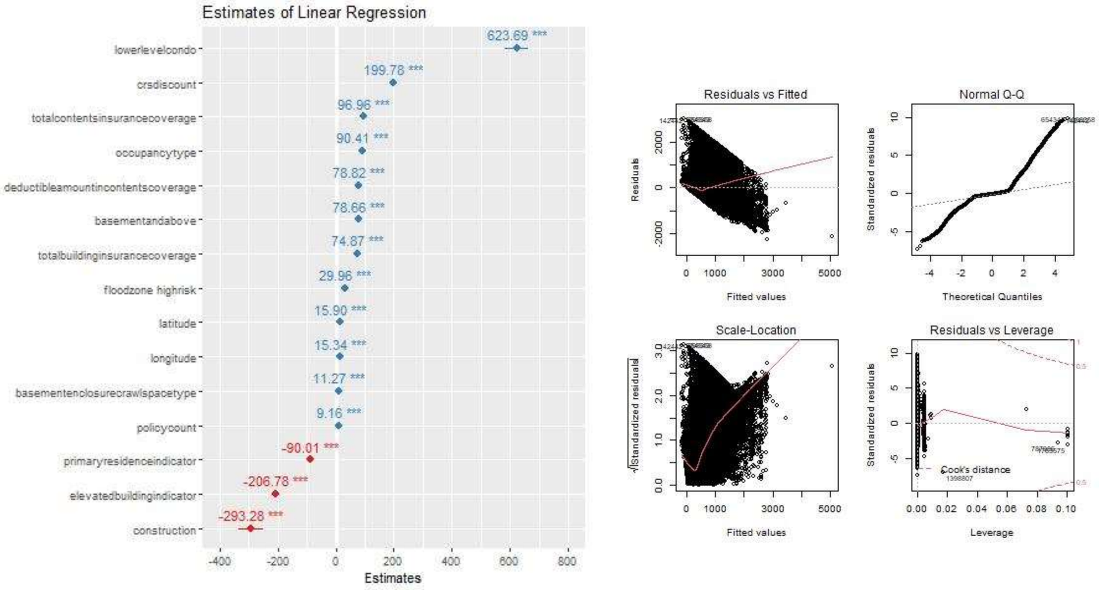
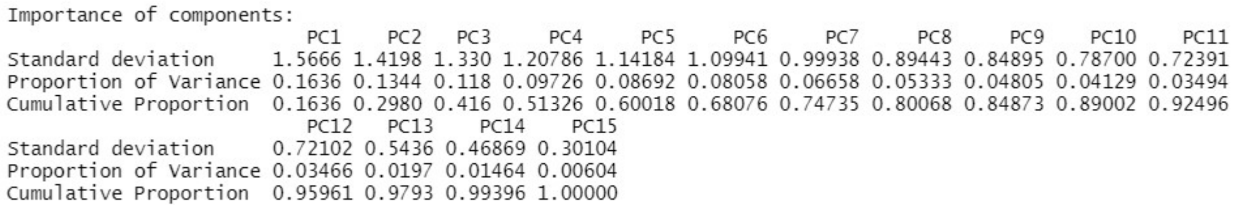


Figure 7: Coefficients of Linear Regression and Residuals Plot of Linear Regression

4.2. PCA

There are 15 variables after cleaning dataset, Principal Component Analysis (PCA) is used to examine if we can reduce the number of variables as well as avoid multicollinearity. Function prcomp() with scaled and na exclude is used for that purpose. As per output, it is observed that the first principal component, PC1, explains only 16.36% of the variability, and the second principal component explains 13.44%. Together, the first 8 and 10 components explain around 80% and 90% of the variability respectively (based on "Cumulative proportion").



```
Importance of components:
                          PC1    PC2   PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     1.5666 1.4198 1.330 1.20786 1.14184 1.09941 0.99938 0.89443 0.84895 0.78700 0.72391
Proportion of Variance 0.1636 0.1344 0.118 0.09726 0.08692 0.08058 0.06658 0.05333 0.04805 0.04129 0.03494
Cumulative Proportion  0.1636 0.2980 0.416 0.51326 0.60018 0.68076 0.74735 0.80068 0.84873 0.89002 0.92496
                          PC12   PC13    PC14    PC15
Standard deviation     0.72102 0.5436 0.46869 0.30104
Proportion of Variance 0.03466 0.0197 0.01464 0.00604
Cumulative Proportion  0.95961 0.9793 0.99396 1.00000
```

To determine the number of principal components required to represent the predictors, screeplot is used. Screeplot represents the variation of each PC captured from the data. We look for the "elbow point", where the amount of variance significantly drops. With a good selection of the number of PCs, we can build a model without losing any important information. It was observed from Figure 8 that the elbow started at 8 PCs and then 10 where the "Cumulative Proportion" ranged from around 80% to 90%. It was also observed from ggbiplot graph that predictors distributed mostly on positive side of the most two PCs: PC1 and PC2. This indicates that the total insurance premium policy will substantially rise with the variables.

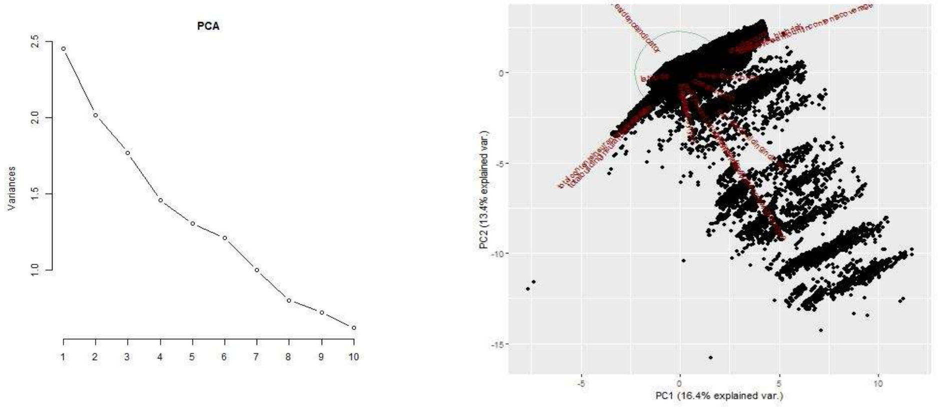


Figure 8: Screeplot (left) and ggbiplot (right)

5. Anticipated discoveries or conclusions

It is noted that the dataset after significant cleaning procedure still have large NA values in "deductibleamountincontentscoverage" with 769,951 NA values and "basementandabove" variables with 371232 NA values. Currently, we are working to finalize how to impute these missing values, as omitting them reduces the data size. PCA results also suggest that there is still some correlation between the variables that needs to be addressed. The linear model with the current feature selection has low R^2 (0.44). In the next step, we will model using other methods including random forest, SVM, or nonlinear transformation which may help with the accuracy. Combination of PCA and other linear regression algorithms will be in our scope as well.

6. References

1. Karapiperis, Dimitris & Kunreuther, Howard & Lamparelli, Nick & Maddox, Ivan & Kousky, Carolyn & Surminski, Swenja & Dolese, Ned & Patel, Paresh & Larkin-Thorne, Sonja. (2017). Flood Risk and Insurance. 10.13140/RG.2.2.27243.13608.

2. Congressional Research Service: Informing the legislative debate since 1914. (2021). National Flood Insurance Program: The Current Rating Structure and Risk Rating 2.0. https://crsreports.congress.gov

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

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