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

Flooding is the most catastrophic natural disaster, both in the United States (US) and worldwide, with global damages exceeding $1 trillion since 1980 [1]. Climate change and progressive development of flood-prone areas may increase these losses by up to a factor of 20 by the end of the century [2]. In the US, 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.

There are numerous studies focused on quantifying future flood losses by translating the physical phenomenon of flooding, mainly the depth of inundation, to its economic impacts in terms of dollars of damage. This translation requires relationships between flood water depth and the resulting asset damages, referred to as depth–damage functions or curves [3 & 4], which are substantially scatter and uncertain in most cases [5 to 8]. As a result, it is necessary to have a second look at the desired economic response and different variables which might play role. Specifically, there is not much research performed in terms premium calculation and the parameters that may affect it. Uncertainty in flood losses has been called the main bottleneck in flood risk studies, an obstacle that may be remedied using large-scale premiums flood insurance data [3].

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 [9]. 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 [10].

Improvements in quality and the quantity of input data, in addition to the advancements in analytic techniques and computational power have largely improved the precision of physical flood models. The same can be applier to insurance premium models. Here are some research questions we are trying to answer in this project: How does flood zone, elevation, property state, no. of floors and other features 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. Kaggle has summarized the data with description of what each column represents. Therefore, we will not repeat data description here. Due to the large size of the dataset with 50 million rows approximately, we have concentrated on a subset of data which represents the data of Houston, TX which has the highest frequency in the dataset. The size of the selected dataset and the modelling approaches applied validates the empirical relationships found. The screenshot of dataset and website are shown in Figure 1.

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

In addition, two other data sets were considered as backups. The first one relates to the major events [11] and the second one with 2 million claims that can be analyzed for claim amount based on similar features [12]. This data is collected over 12 years. Although, we considered the second dataset with the list of major events, also, the third dataset with 2 million claims initially to understand the problem, to keep the scope of the project manageable, we did not used those in the final analyses.

3. Data Analysis:

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 (totalinsurancepremium). There are 2,029,540 rows in total. The total insurance premium of the policy will be predicted based on several main variables such 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.

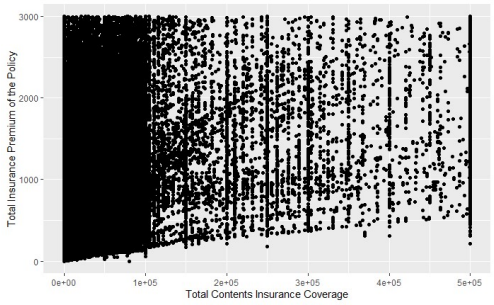

Figure 2: Snapshot of datasets with their associated references

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.

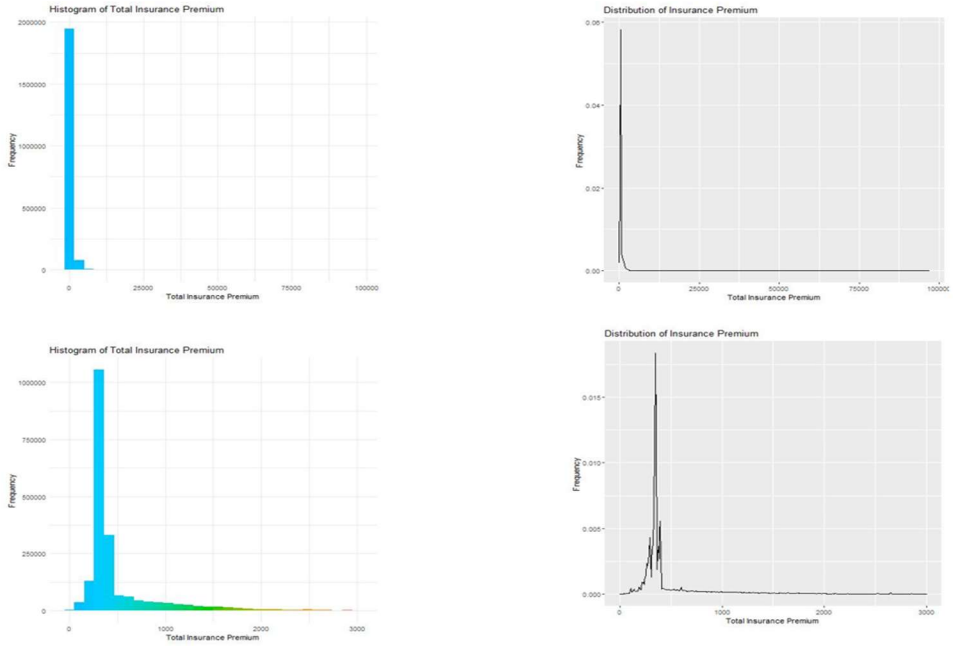

Figure 3: The histogram and distribution of response (Total Insurance Premium of the Policy): Top: original data, Bottom: after remove all the policy with the cost higher than 3,000.

3.2.1 Imputing Missing Values:

After the previous steps of exploratory data analysis and data cleaning, there are still columns that have missing values and removing those rows are not an option, as they could potentially affect the results. We tried “locationofcontents” column imputation with mode and 0 to see which method gives better results. Both imputation with 0 and mode will be covered in this study.

3.3 Feature Selection and Feature Engineering

The flood insurance dataset is big and has more than 40 columns. Many of the columns are categorical variables that were transformed to several dummy variables. Therefore, multi-dimensionality and big data size are the two main challenges. To reduce multi-dimensionality several steps are followed, after data cleaning, that are describe here. We need to select the most relevant and the right number of features to include in the final model. Too many features can produce an overfit model and too little can cause the model to perform poorly as important features may not be included to explain some variations of the target variable.

3.3.1 Exploring Relationships among Variables

For the first step of feature selection, linear and Spearman correlation matrices were run, as shown in the Figure 4 and Figure 5, respectively. Upon inspecting the data, there is non-linear dependency between the features. A cut-off of 0.6 is chosen and highly correlated features are dropped. This includes removal of the following features: “ratemethod_prp”, “crsdiscount”, “ratemethod_manual”, “censustract” and “totalcontentsinsurancecoverage”. Even after this step, linear regression shows some feature with NA as a coefficient. NA as a coefficient in a regression indicates that the variable in question is linearly related to the other variables. Therefore, “policytermindicator,mobilehomeortrailer”, “floodzone_undetermined”, and “occupancytype_business” are removed.

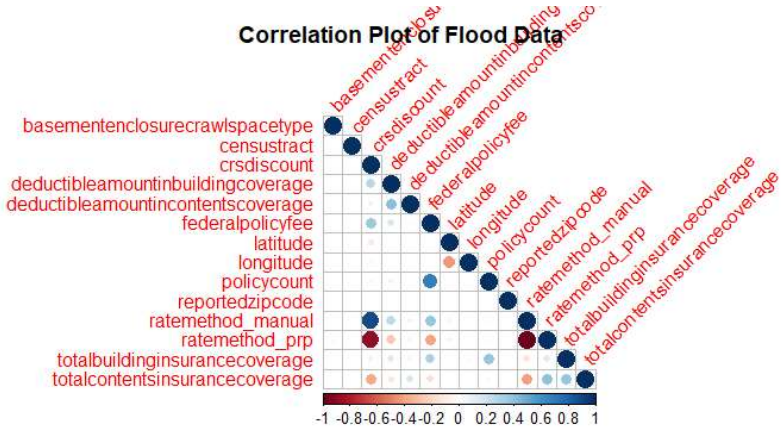

Figure 4 Linear correlation Matrix for numerical features

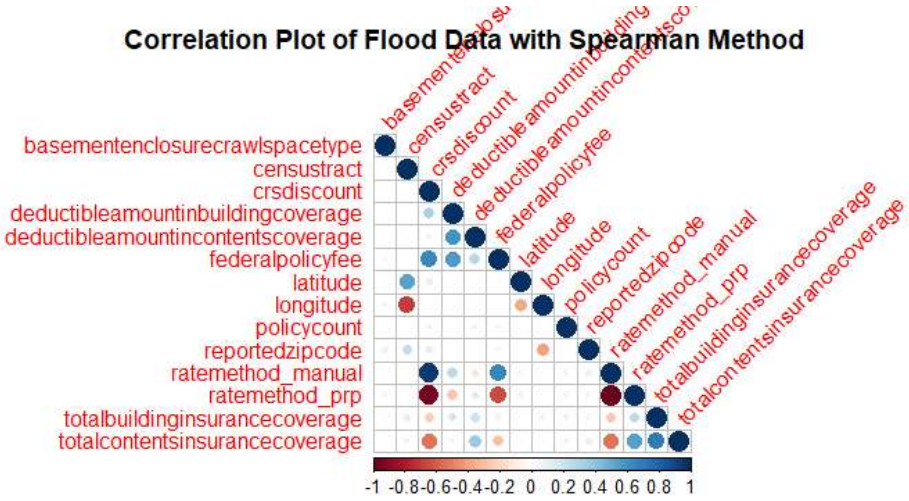

Figure 5 Spearman correlation matrix for numerical features

3.3.2 Non-linear Transformation

A linear regression on the data at this stage gives $R^2 = 0.35$ with 26 features, some of which came as insignificant. Before removing the insignificant features, we first checked the nonlinearity between target and predictor variables. An insignificant feature in linear regression may end up significant in a log-log model.

First, log-linear model was run, this improved R^2 to 0.39. Therefore, going on, all the models are made for log(insurance premium). Then, for each continuous variable two regressions were performed: first a log-linear model and then a log-log model. If R^2 is higher for log-log model, the feature is transformed to log for future models. The results of this step is logarithmic transformation of "deductibleamountinbuildingcoverage", "basementclosurecrawlspace", "numberoffloorsininsuredbuilding" and "policycount". After these logarithmic transformations, R^2 improved to 0.40.

In the same way, cubic root ($\sqrt[3]{feature}$) transformation is checked. It is found that cubic root transformation of "deductibleamountinbuildingcoverage", "deductibleamountincontentscoverage", "elevationdifference", and "totalbuildinginsurancecoverage" improves the R^2 to 0.45. At this stage all the features are significant:

```
Coefficients:
                                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                    9.567e+00  1.821e-01   52.531  < 2e-16 ***
log(basementenclosurecrawlspacetype + 1)       2.846e-02  2.499e-03   11.390  < 2e-16 ***
construction                                  -2.702e-01  6.567e-03  -41.139  < 2e-16 ***
deductibleamountinbuildingcoverage_trans      -2.792e-03  1.496e-04  -18.668  < 2e-16 ***
deductibleamountincontentscoverage_trans       2.345e-02  1.217e-04  192.623  < 2e-16 ***
elevatedbuildingindicator                     -1.872e-01  1.945e-03  -96.241  < 2e-16 ***
elevationdifference_trans                     -4.619e-01  1.209e-03 -382.039  < 2e-16 ***
federalpolicyfee                               4.147e-03  1.851e-05  223.997  < 2e-16 ***
latitude                                       3.611e-02  2.898e-03   12.460  < 2e-16 ***
longitude                                      1.220e-02  2.090e-03    5.835 5.37e-09 ***
log(numberoffloorsininsuredbuilding + const + 1) -4.259e-01  3.291e-03 -129.408  < 2e-16 ***
log(policycount + 1)                          -1.641e+00  1.259e-02 -130.342  < 2e-16 ***
postfirmconstructionindicator                 -1.272e-01  6.305e-04 -201.745  < 2e-16 ***
primaryresidenceindicator                     -7.217e-02  1.004e-03  -71.894  < 2e-16 ***
totalbuildinginsurancecoverage_trans           1.481e-02  2.758e-05  536.887  < 2e-16 ***
not_condo                                     -1.450e+00  1.828e-02  -79.305  < 2e-16 ***
u_condo                                       -1.607e+00  1.838e-02  -87.441  < 2e-16 ***
lowerflooronly                                -6.711e-02  1.050e-03  -63.932  < 2e-16 ***
upperandlowerfloors                            1.840e-02  1.113e-03   16.528  < 2e-16 ***
basementandabove                               8.895e-02  4.051e-03   21.960  < 2e-16 ***
basementonly                                   4.357e-01  1.351e-01    3.226  0.00126 **
morethan1floor                                -2.903e-01  6.742e-03  -43.059  < 2e-16 ***
floodzone_highrisk                             4.086e-01  8.570e-04  476.755  < 2e-16 ***
occupancytype_single                          -9.766e-01  2.820e-03 -346.350  < 2e-16 ***
occupancytype_2or4                            -9.764e-01  4.125e-03 -236.727  < 2e-16 ***
occupancytype_4more                           -5.058e-01  2.944e-03 -171.794  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4052 on 1967190 degrees of freedom
  (20670 observations deleted due to missingness)
Multiple R-squared:  0.4523,    Adjusted R-squared:  0.4523
F-statistic: 6.498e+04 on 25 and 1967190 DF,  p-value: < 2.2e-16
```

3.3.3 Variance Inflation Factor (VIF)

VIf of the above model was checked. We should choose a VIF cut-off under which a variable is retained [13]. Any VIF above 10 shows multi-collinearity and the corresponding features are removed. This only decreased R^2 to 0.44:

```
Call:
lm(formula = totalinsurancepremiumofthepolicy_log ~ ., data = temp0)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7552 -0.1801 -0.0555  0.0626  4.7847

Coefficients:
                                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)                               7.239e+00  1.813e-01   39.932  < 2e-16 ***
construction                             -2.627e-01  6.616e-03  -39.713  < 2e-16 ***
deductibleamountinbuildingcoverage_trans -8.079e-04  1.502e-04   -5.379 7.51e-08 ***
deductibleamountincontentscoverage_trans  2.393e-02  1.224e-04  195.462  < 2e-16 ***
elevatedbuildingindicator                -1.846e-01  1.925e-03  -95.909  < 2e-16 ***
elevationdifference_trans                -4.691e-01  1.217e-03 -385.369  < 2e-16 ***
federalpolicyfee                          1.899e-03  1.203e-05  157.827  < 2e-16 ***
latitude                                  4.649e-02  2.918e-03   15.931  < 2e-16 ***
longitude                                 1.734e-02  2.105e-03    8.241  < 2e-16 ***
numberoffloorsininsuredbuilding_log      -4.474e-01  3.301e-03 -135.521  < 2e-16 ***
postfirmconstructionindicator            -1.253e-01  6.347e-04 -197.425  < 2e-16 ***
primaryresidenceindicator                -8.086e-02  1.009e-03  -80.173  < 2e-16 ***
totalbuildinginsurancecoverage_trans      1.470e-02  2.690e-05  546.554  < 2e-16 ***
lowerflooronly                           -6.764e-02  1.058e-03  -63.950  < 2e-16 ***
upperandlowerfloors                       1.900e-02  1.119e-03   16.979  < 2e-16 ***
basementandabove                          1.250e-01  3.289e-03   38.012  < 2e-16 ***
basementonly                              5.130e-01  1.361e-01    3.769 0.000164 ***
morethan1floor                           -2.989e-01  6.790e-03  -44.018  < 2e-16 ***
floodzone_highrisk                        4.418e-01  8.398e-04  526.032  < 2e-16 ***
occupancytype_single                     -9.825e-01  2.840e-03 -345.945  < 2e-16 ***
occupancytype_2or4                       -1.002e+00  4.131e-03 -242.511  < 2e-16 ***
occupancytype_4more                      -5.464e-01  2.956e-03 -184.865  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4084 on 1967194 degrees of freedom
  (20670 observations deleted due to missingness)
Multiple R-squared:  0.4436,    Adjusted R-squared:  0.4436
F-statistic: 7.47e+04 on 21 and 1967194 DF,  p-value: < 2.2e-16
```

3.3.4 Stepwise Regression

For the next step of feature selection, stepwise regression is performed with AIC as the main criterion. AIC is a good metric since it is balance between likelihood and the number of predictors. The model with the lowest AIC still has the same number of features. Therefore, no feature is dropped at this stage.

3.3.5 Lasso Regression

For the last step of feature selection, Lasso regression was performed with five-fold cross-validation. Before running the lasso, the data are scaled and centered, as part of the preprocessing. We can discard the variables with coefficients close to zero, in this case we will remove basementonly.

3.4 Final Dataset

The final dataset is still large: just below 2 million rows with 22 features. Some bigger models such as random forest could not run on our personal computers for this size of data. For this reason, the top most 5 frequent zipcodes are filtered. This corresponds to the following zipcodes: 77096, 77089, 77084, 77024 and 77062. More complicated models are run only on this subset of data which is still large (~293,794 rows).

4 Linear Regression Modeling

The final dataset after cleaning and feature selections includes 293,794 observations with 22 predicted and response variables. The dataset is randomly split into training set and testing set with the ratio 7:3. For a direct comparison between models, the data is scaled. The models are then built on the training set and tested against the testing set. The model accuracy is evaluated using r-squared (R^2) which represents for the variance in total insurance premium explained by the regression model. A higher r-squared, a better fitting model.

Several linear regression models are used to examine the relation of total insurance premium and predicted variables including Stepwise, Ridge, Lasso, Regression Tree, and Random Forest. Principal Component Analysis is also used to examine if we can condense the number of variables without losing accuracy but in significantly less time. All the linear regression models in this part based on the dataset with either imputation by zero or mode. However, due to the page limitation, we will report the results with imputation by zero in detail. The results from imputation by mode will be used for direct comparison by the end of this section.

4.1 Stepwise Method

Step() function is used to perform stepwise selection. The model will run through different eliminations of variables. From the output, the one with smallest statistics AIC will be selected. In this work, direction of “forward” and “both” are examined for direct comparison. The model will start with a full bag of variables. Then it starts removing the least significant variables one after the other until no variable is left in the model. Both directions combine both forward and backward direction. As the variables have been carefully selected from above process, all these variables are significant, and thus similar R^2 of 0.498

is achieved for both backward and both directions. Note that R^2 is different from the previous section since we modeled on a subset of the data.

4.2 Penalized Linear Regression

Backward and both direction models might not explain the effect of removing or adding a variable on the response. For instance, in backward direction, removing predictors from the model can be seen as settings their coefficients to zero. To minimize this effect, penalty lambda has been introduced to maintain the variables in the model as well as decrease model complexity. In this work, both Ridge (l2 norm) and Lasso (l1 norm) are considered.

The R^2 of both models are 0.498 which is similar to forward or both directions. The coefficients of each variable in Ridge, Lasso, and Forward direction modelsare not much different in value as well as its positive or negative effect on the total insurance premium. A typical values of variable coefficient from Lasso model is displayed in Figure 7. It is observed that the occupancy type has the strongest negative effect on the insurance, following by "elevatedbuildingindicator". Risk zone and construction with basement have positive coefficient, resulting in an increase in the total insurance premium as showing in Figure 6. The "basementenclosurecrawlspacetype_log" and "deductibleamountinbuildingcoveragetrans" variables have very small coefficients but not reducing to zero in both Ridge and Lasso, indicating that these variables still have effect on the total insurance calculation.

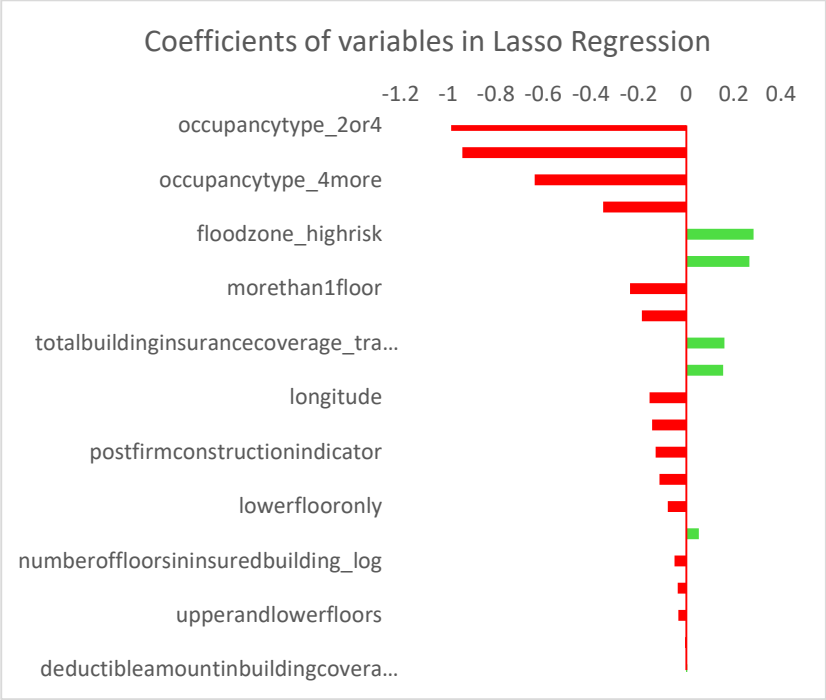

Figure 6: Variable's Coefficient in Lasso Regression Model

4.3 Principal Component Analysis (PCA)

In this part, PCA is used to examine whether the number of variables can be condensed to speed up the machine learning models of such a large dataset without loosing much information. Function prcomp() with scaled is used for that purpose.

PCA output return 21 principal components, which equals to number of variables in the data set, as rotation. The summary function on the result object gives us standard deviation, proportion of variance explained by each principal component, and the cumulative proportion of variance explained. For example, the first three principal component, explains 0.4991 or 49.91% variability (based on "Cumulative proportion") of the variability.

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12
Standard deviation     1.4030 1.3032 1.1651 1.0760 1.0046 0.93751 0.90091 0.69031 0.47982 0.33506 0.32501 0.28129
Proportion of Variance 0.1955 0.1687 0.1349 0.1150 0.1003 0.08731 0.08063 0.04734 0.02287 0.01115 0.01049 0.00786
Cumulative Proportion  0.1955 0.3642 0.4991 0.6141 0.7144 0.80167 0.88230 0.92964 0.95251 0.96366 0.97416 0.98202
                          PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.25920 0.22529 0.19163 0.11376 0.08174 0.05250 0.03841 0.03732 0.03309
Proportion of Variance 0.00667 0.00504 0.00365 0.00129 0.00066 0.00027 0.00015 0.00014 0.00011
Cumulative Proportion  0.98869 0.99373 0.99738 0.99867 0.99933 0.99961 0.99975 0.99989 1.00000
```

To determine the number of principal components required to represent most of the predictors, screeplot is used. From the screen plot (Figure 7) which represents the variation of each PC captured from the data, we look for the "elbow point", where the amount of Variance significantly drops off. With a good selection of the number of PCs, we can ignore the reset without loosing any important information. It was observed from Figure 8 that the elbow started at 7 PCs and then at 10 PCs where the "Cumulative Proportion" increases from around 88% to 96%. It was also observed from ggbiplot graph that predictors distributed on both positive and negative sides of the most two PCs: PC1 and PC2.

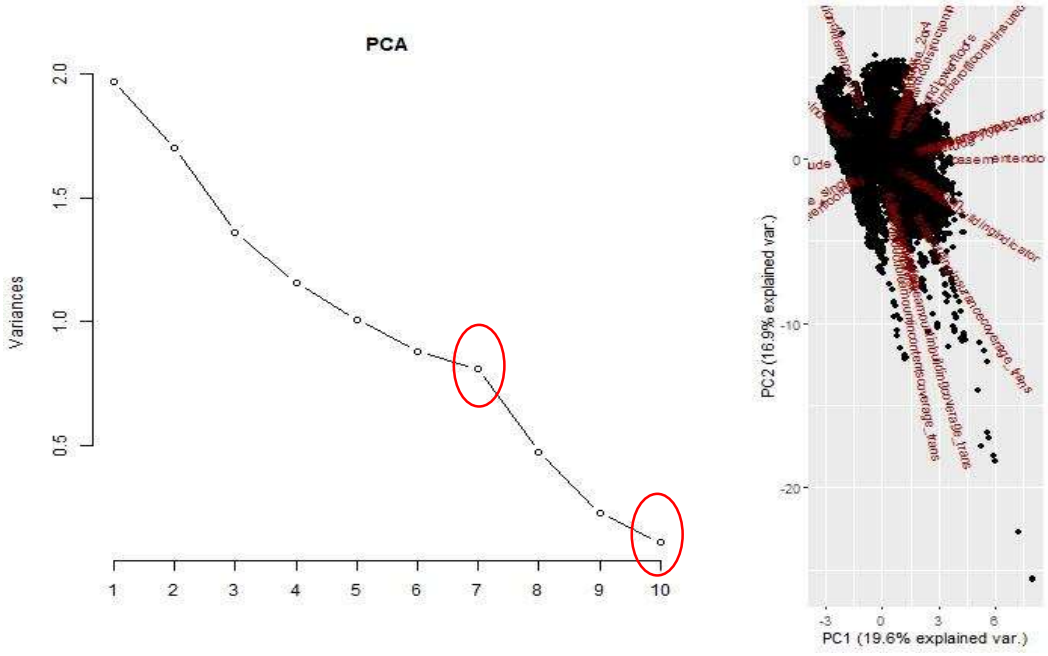

Figure 7: Left: Screeplot, Right: ggpiplot of PCA

With the PCA results, we now tested the validation of the model against Lasso. We run both the linear regression and cross validation with the number of principal components (PC) either 7 or 10. The R^2 values with 7 PCs and 10 PCs are 0.397 and 0.463 respectively, which is not much different to 0.498 from Lasso with all 21 variables. This promising result suggest that PCA can be used to reduce the number of variables for linear regression models in our study.

4.4 Regression Tree

Further to stepwise and penalized regression models, regression tree is used to evaluate the linear regression. Generally, trees split observations into several partitions using binary recursive partitioning algorithms. At each split point, variables are selected in such a way that maximizes the closeness of the response variable within partitions and minimizes the similarity of the response between the two partitions. This splitting process is continued until each partition reaches a minimum size or the sum of the squared deviance from the mean in a partition becomes zero. When the splitting process is complete, these partitions become terminal nodes. In this part, single tree and random forest are used.

4.4.1 Decision tree

The single regression tree was built considering all 21 variables. However, it is observed from the output, only 4 variables ("federalpolicyfee", "floodzone_highrisk", "elevationdifference_trans" and "totalbuildinginsurancecoverage_trans") were included in the model, as shown in the Figure 8. The split initially starts with "federalpolicyfee", followed by "floodzone_highrisk" and "elevationdifference". There are 12 terminal nodes. The R^2 of the decision tree is 0.784 that can explain 78.4% of the dataset.

```
Regression tree:
tree(formula = totalinsurancepremiumofthepolicy_log ~ ., data = train)
Variables actually used in tree construction:
[1] "federalpolicyfee"                    "floodzone_highrisk"
[3] "totalbuildinginsurancecoverage_trans" "elevationdifference_trans"
```

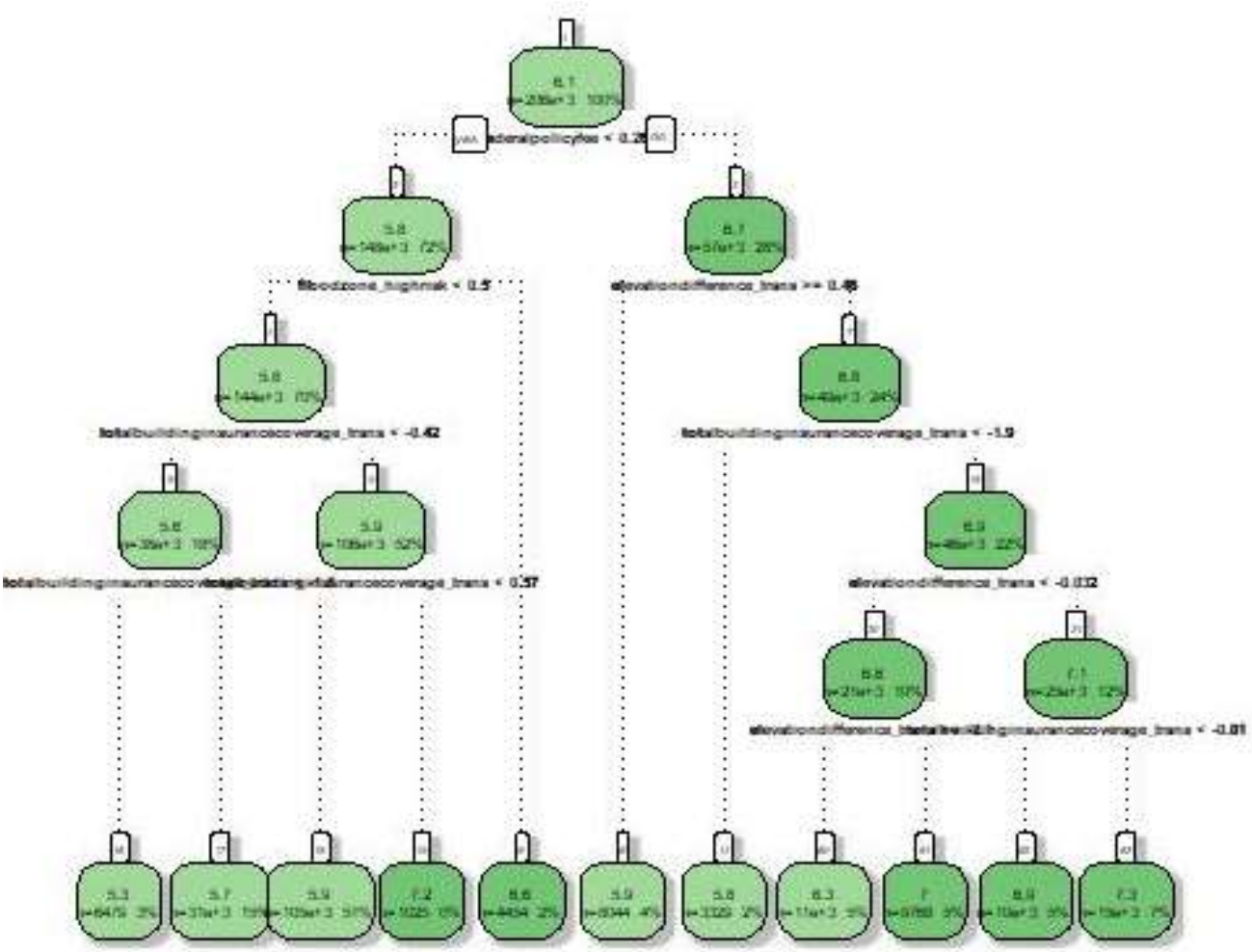

Figure 8: Decision tree of 4 variables and 12 terminals

4.4.2 Random Forest

Finally, Random Forest is used for our model. It is noted that RF is very time consuming for a large dataset. In the RF, the variables is randomly selected by the total number of predictor variables divided by 3. There are 21 variables in given dataset, and thus 7 will be randomly selected for each weak tree. 500 weak regression trees are generated by randomly selecting and considering 7 variables for each tree and the average across the models is computed for prediction. The R^2 value for RF is observed at 0.88%

The importance of each variable is examined using importance(). The "%IncMSE" is the amount that the MES of predictions increases if a variable is chosen from the model. The latter is a measure of the total decrease in node impurity that results from splits over that variable, averaged over all trees. These values are used to identify the importance of variables in the dataset and plotted as below using varImpPlot(). From the plot in Figure 9, the top 4 variables are well-aligned with decision tree. However, random forest includes more variables than decision trees and thus a higher R^2 is expected. Since random forests tendency to overfit decrease with the number of trees, and 500 trees is quite large, this improvement in the R^2 could

point to nonlinearity of data that was not captured in linear regression even with logarithmic and cubic root transformation.

The variables importance of Lasso model (similar to Ridge and Forward direction) is also included in Figure 10 for direct comparison. It is observed that there is a huge difference in variables' effect on the total insurance premium.

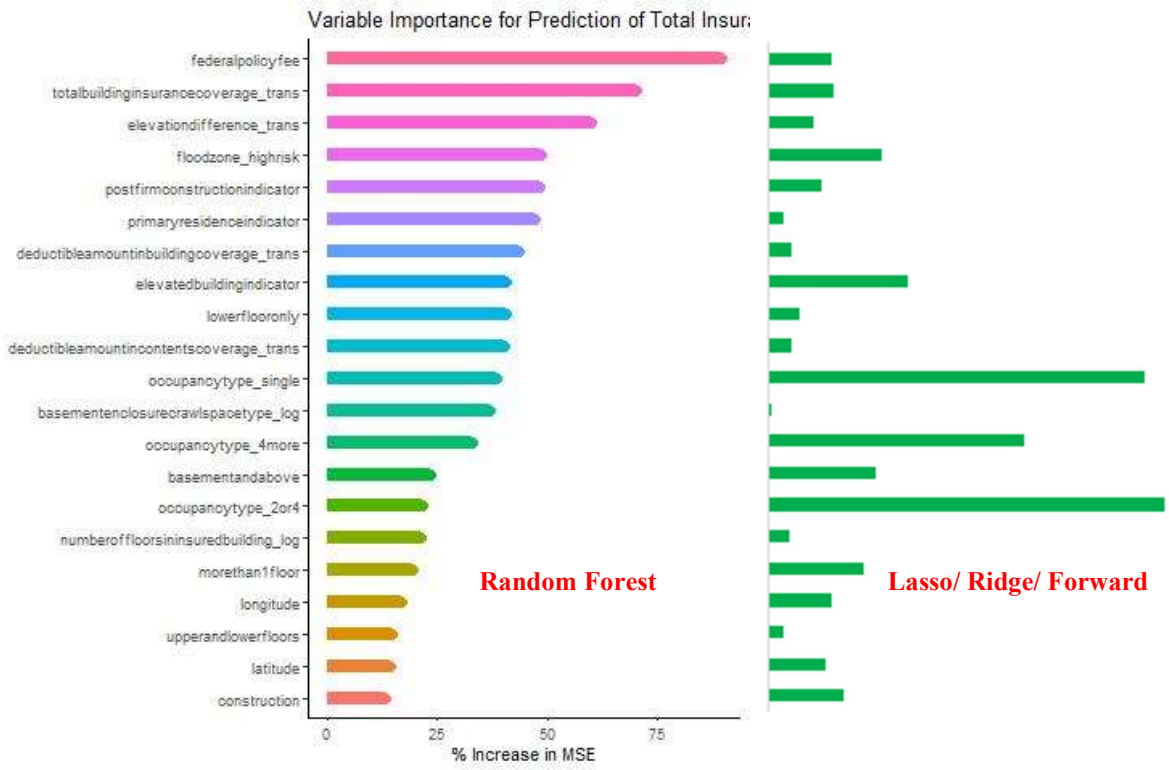

Figure 9: Relative Influence of Variables on Total Insurance Premium

The whole process is carried out for the case with imputation by mode. It is found that there is slightly improvement in the case with mode imputation. For instance, the R2 of regression (Lasso, Ridge) from the dataset with mode imputation is 51.7% in comparison with 50% from the one with zero imputation.

5. Conclusions & Interpretation of Results

In summary, different linear regression models have been carried out in this study. The values of R^2 of all models are displayed in Figure 10. Several findings have been achieved as follows:

- Successfully extract and preprocessing a large dataset, including identify negative values, outliers;
- Applying several methods for feature selection including histogram, Spearman, linear correlation, VIF, Ridge and

Lasso to remove any insignificant variables;

- Random forest and decision tree have the highest R^2 value. However, decision tree have used only 4 out of 21 variables and thus it is not reliable.
- Other linear regression models including Ridge, Lasso, Forward or both directions have similar R^2.
- PCA successfully employed in this work to condense 21 variable down to 7 or 10 components. Linear regression with 10 components from PCA produce a smaller R^2 in comparison with the model with a whole dataset but with a minor difference (0.47vs. 0.498). However, modeling with PC's makes the interpretation of results difficult, therefore, the initial variables were used in decision tree and random forest.

Linear regression shows occupancy type, whether the building is elevated and whether the building is in a high-risk flood zone as the most important parameters.

Random forest has the highest predictive power among the models tried on the data. The most important features in predicting insurance premium based on this model are: federal policy fee, total building coverage, whether the building is elevated and whether the building is in a high-risk flood zone.

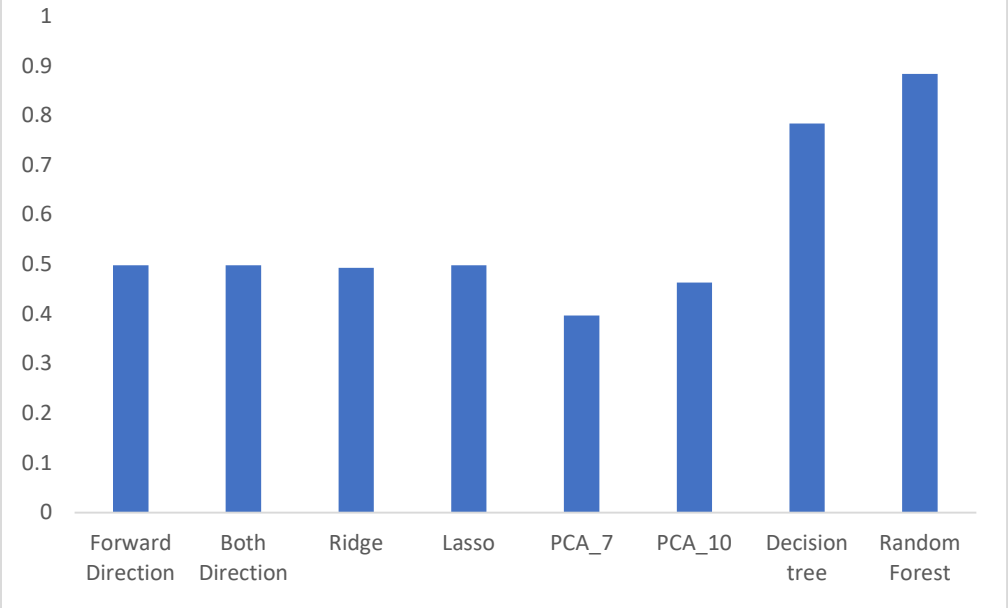

Figure 10: R^2 of different models

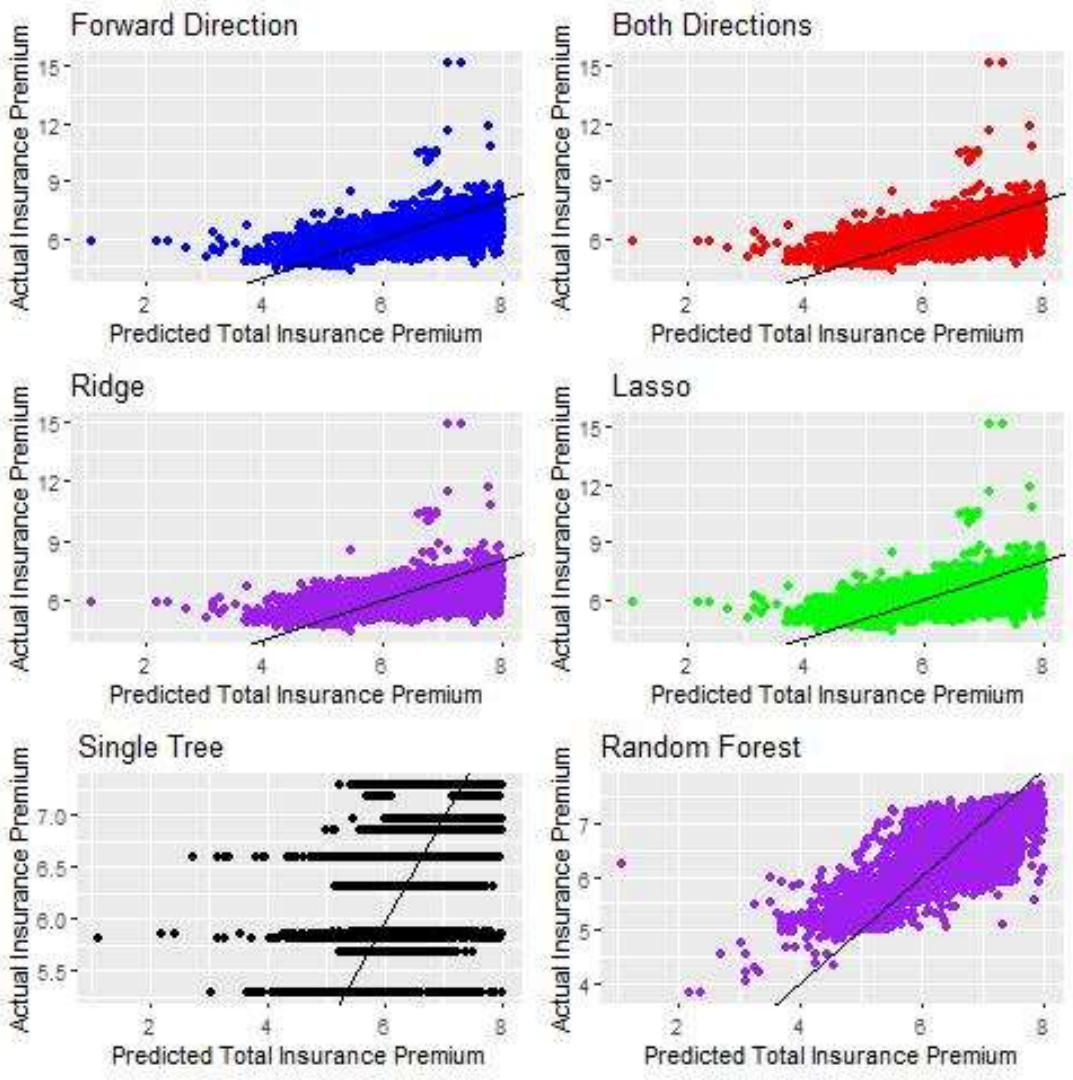

Figure 11: Results on test set of all models

6. References

[1] Munich Re. NatCatSERVICE https://natcatservice.munichre.com/ (2019).

[2] Winsemius, H. C. et al. Global drivers of future river flood risk. Nat. Clim. Change 6, 381–385 (2016).

[3] Davis, S. A. & Skaggs, L. L. Catalog of Residential Depth-Damage Functions used by the Army Corps of Engineers in Flood Damage Estimation IWR-92-R-3 (USACE Institute for Water Resources, Fort Belvoir, VA, 1992).

[4] Wing, O.E.J., Pinter, N., Bates, P.D. et al. New insights into US flood vulnerability revealed from flood insurance big data. Nat Commun 11, 1444 (2020).

[5] Bübeck, P., de Moel, H., Bouwer, L. M. & Aerts, J. C. J. H. How reliable are projections of future flood damage? Nat. Hazards Earth Syst. Sci. 11, 3293–3306 (2011).

[6] Merz, B., Kreibich, H., Schwarze, R. & Thieken, A. Review article "Assessment of economic flood damage". Nat. Hazards Earth Syst. Sci. 10, 1697–1724 (2010).

[7] McGrath, H., El Ezz, A. A. & Nastev, M. Probabilistic depth–damage curves for assessment of flood-induced building losses. Nat. Hazards 97, 1–14 (2019).

[8] Lehman, W. & Nafari, R. H. An empirical, functional approach to depth damages. E3S Web Conf. 7, 05002 (2016).

[9] 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.

[10] 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

[11] https://nfipservices.floodsmart.gov/reports-flood-insurance-data

[12] https://www.kaggle.com/lynma01/femas-national-flood-insurance-policy-database

[13] Zuur, A.F., Ieno E.N., Elphick C. S., A protocol for data exploration to avoid common statistical problems, Methods in Ecology & Evolution, (2010).