UTILIZING MACHINE LEARNING & REGRESSION ANALYSIS TO ANALYZE FLOOD INSURANCE DATA

MGT 6203 Proposal Presentation

By (order of presentation):

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

April 2022

HOSSEIN

OUTLINE

- Problem Statement
- Dataset(s)
- Data engineering
  - Data exploration through visualization
  - Data cleaning
  - Imputation
  - Feature selection and feature engineering
- Modeling
  - Linear regression
  - Stepwise method
  - Penalized Linear Regression
  - Principal Component Analysis (PCA)
  - Regression Tree
- Conclusions

PROBLEM STATEMENT

- Damage caused by floods is covered under the policy issued by National Flood Insurance Program (NFIP), overseen by FEMA (Federal Emergency Management Agency).
- Flooding is the primary vector of economic damages inflicted on local communities. There is also a projected increase in destructive flooding due to climate change; therefore, there is an enormous need to efficiently distribute financial risk.
- Our target variable is the insurance premium. We are trying to answer the following questions:
  - How do flood zone, elevation, property state, no. of floors and other features affect insurance premium? Can we predict premium based on the variables available?
  - What is the correlation between different variables?
  - What are the most important variables in premium estimation?

DATA

- The main dataset: FEMA's National Flood Insurance Policy Database (https://www.kaggle.com/lynma01/femas-national-flood-insurance-policy-database)
- More than 50,000,000 rows
- Limited the analysis to Houston, TX data (2,029,540 rows)
- Supplementary data: 2,000,000 rows of Claims (https://nfipservices.floodsmart.gov/reports-flood-insurance-data)

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

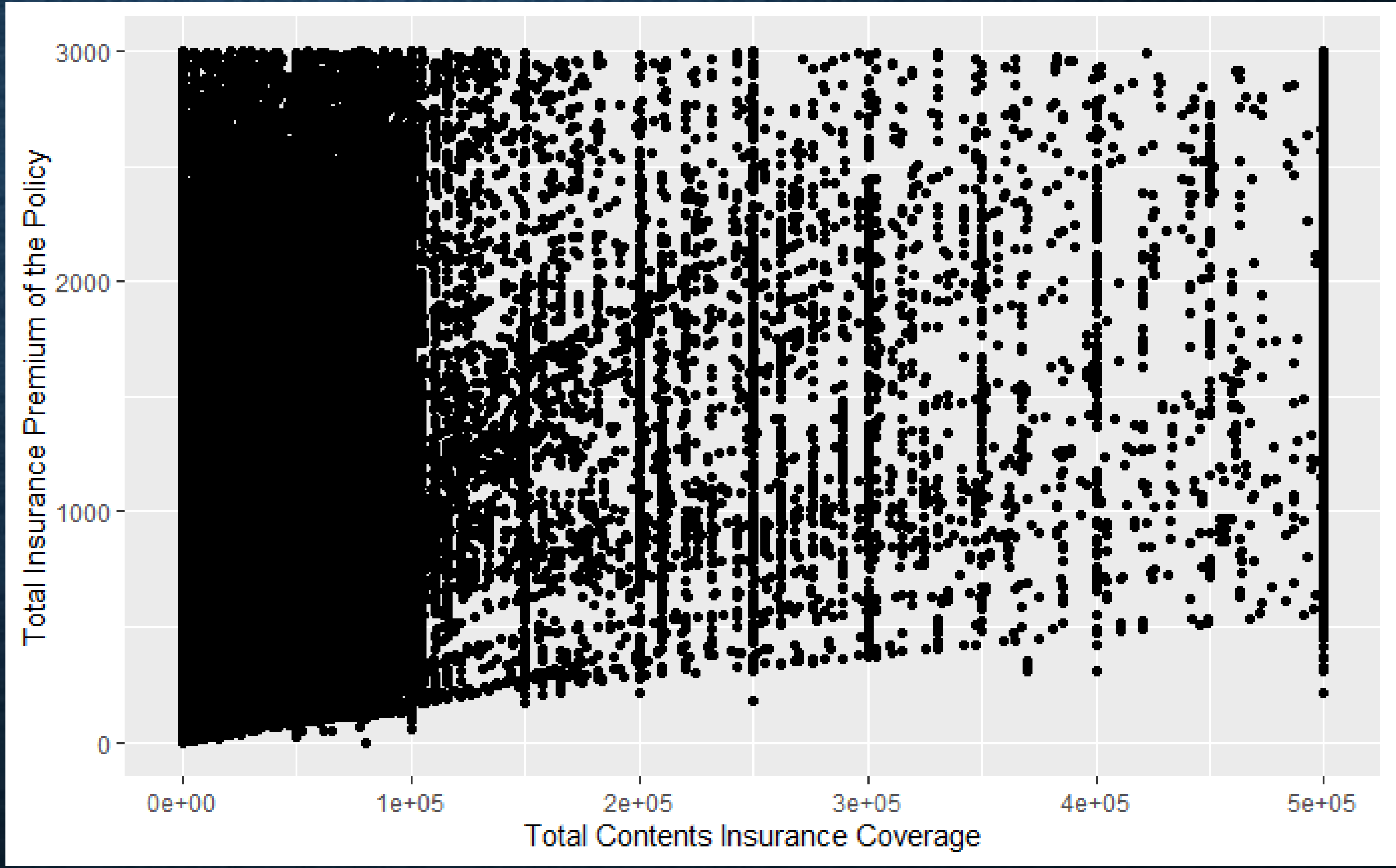
WARDA

STEPS IN DATA ANALYSIS

- Visualization and understanding data
- Data Filtering/Cleaning
- Feature Selection & Feature Engineering
- Modeling

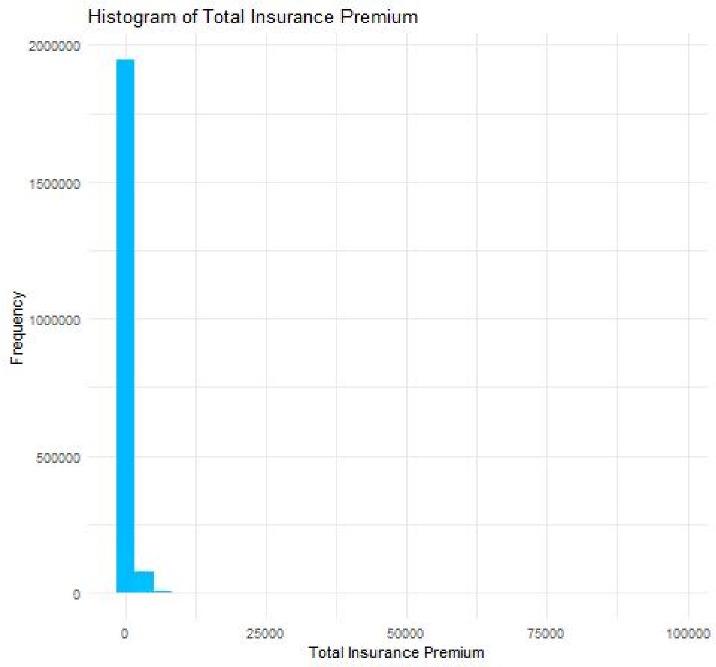
VISUALIZATION AND UNDERSTANDING DATA

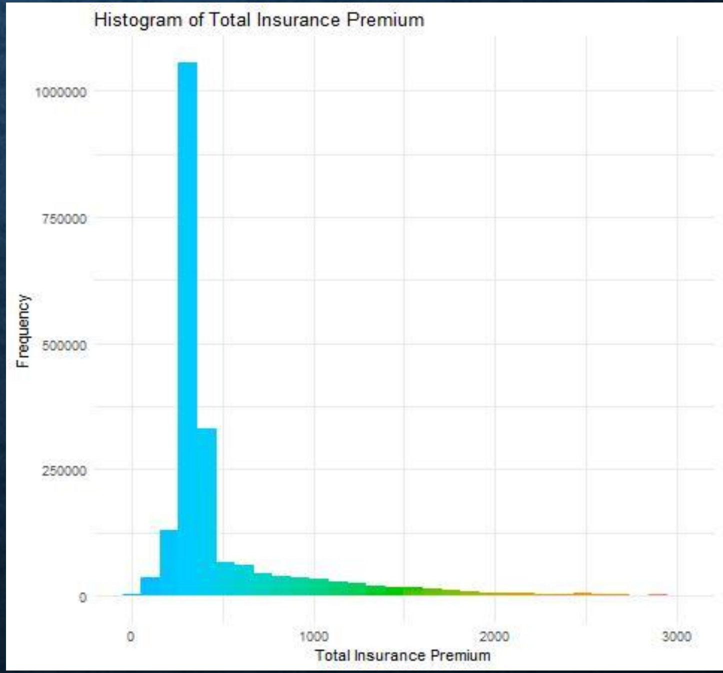
- Visualization to understand the data



DATA FILTERING/CLEANING

- Data Filtering/Cleaning
  - Right data formatting
  - Outlier removal
  - Removed columns with more than 60% data missing
  - Imputing missing data with mean, 0 and mode
  - Created dummy variables for categorical features





STEPS IN DATA ANALYSIS

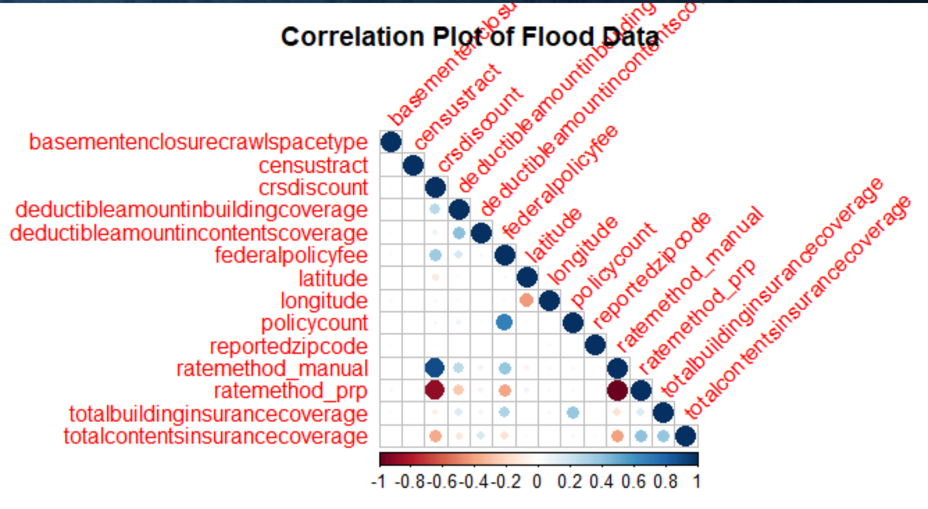
- Visualization and understanding data
- Data Filtering/Cleaning
    - Outlier removal
    - Removed columns with more than 60% data missing
    - Imputing missing data with mean, 0 and mode
    - Right data formatting
    - Created dummy variables for categorical features
- Feature Selection & Feature Engineering
    - Correlation matrix
    - VIF
    - Stepwise regression
    - LASSO regression
    - PCA
    - Feature transformation (log and cubic root)
- Modeling
    - Linear model
    - More advanced models such as Random Forest and SVM

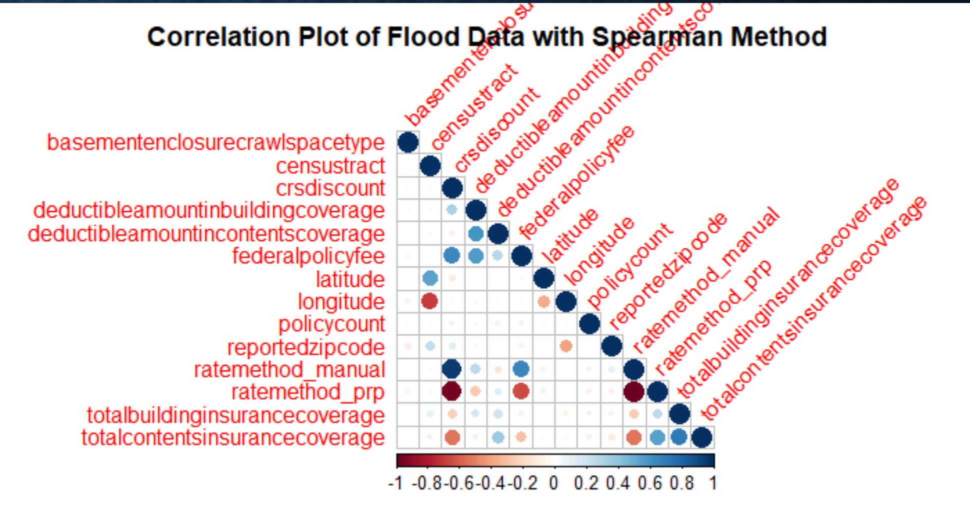
MAHSHID

FEATURE SELECTION & FEATURE ENGINEERING

After data cleaning and visualization:

- Explored both linear and Spearman relationship between features and omitted the features that are highly correlated (cut-off > 0.6)
- Explored non-linear transformation of data: logarithmic and cubic root transformation
- Omitted featured with VIF above 10.
- Stepwise regression (AIC as metric) and LASSO regression (5-fold cv) were performed for further feature selection.
- Reduced 44 variables, many of which in categorical format, to 22 through the above steps



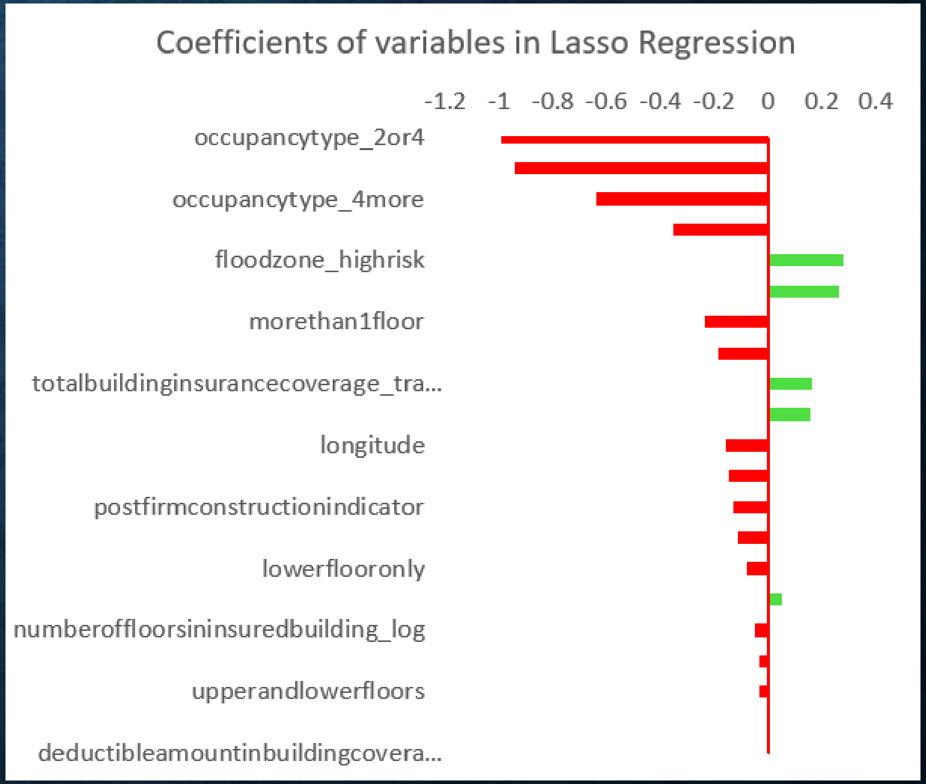


FEATURE SELECTION & FEATURE ENGINEERING

- Feature selection and transformation improved R^2 of linear regression from 0.35 to 0.44.
- All the features are significant based on linear regression
- The final dataset: still large; ~2 million rows with 21 features.
- Due to computational demand, data were filtered for the top most frequent 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) .

LINEAR REGRESSION MODEL

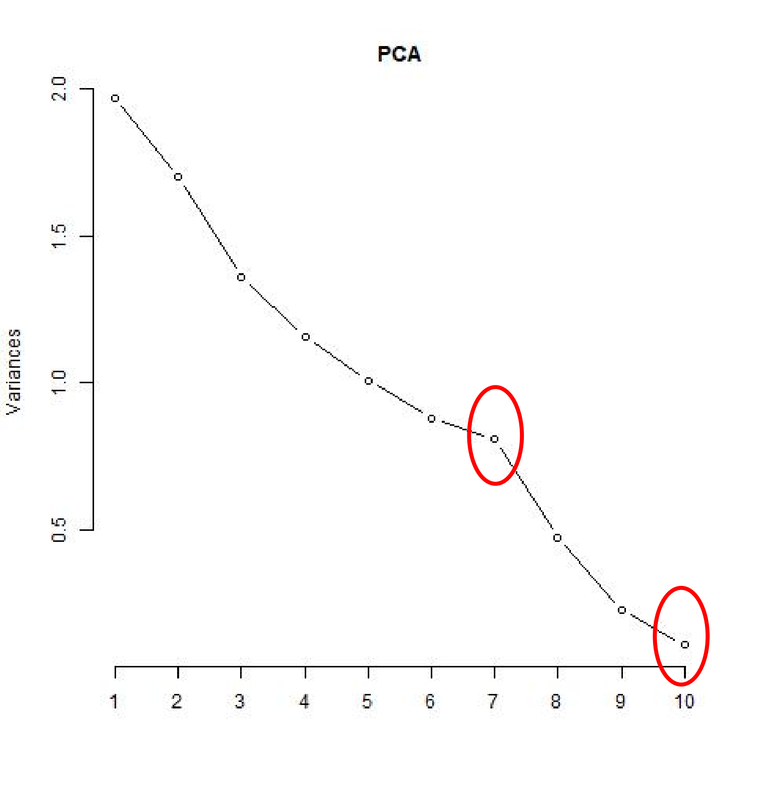
- Train-test split: 70% - 30%
- Stepwise: forward and both direction; AIC as a metric, all variables significant; $R^2 = 0.498$
- Penalized Linear Regression (Ridge & Lasso): occupancy type has the strongest negative effect on the target variable, following by "elevatedbuildingindicator".
- Risk zone and construction with basement have positive coefficient, resulting in an increase in the total insurance premium

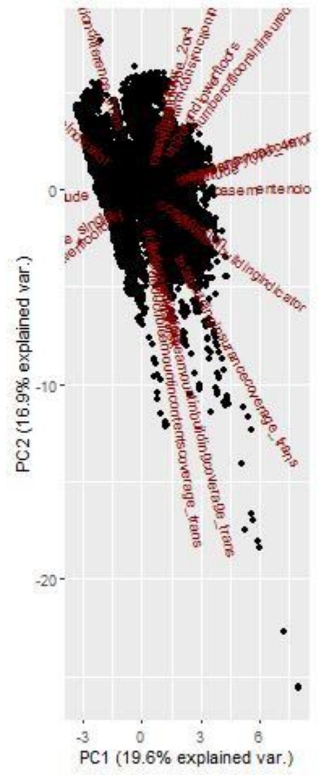


ANH

PRINCIPAL COMPONENT ANALYSIS (PCA)

- PCA was run on the remaining 21 variables. From standard deviations of each PC and screeplot it was concluded that 7 to 10 PCs are sufficient to account for 88% to 96% of variability respectively.
- R^2 with 7 PCs and 10 PCs is 0.397 and 0.463 respectively.

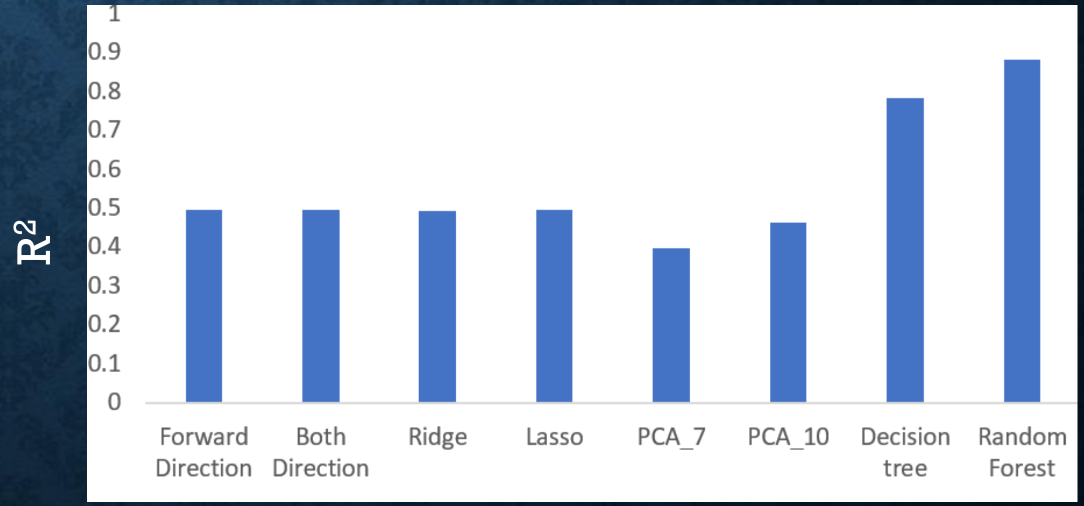




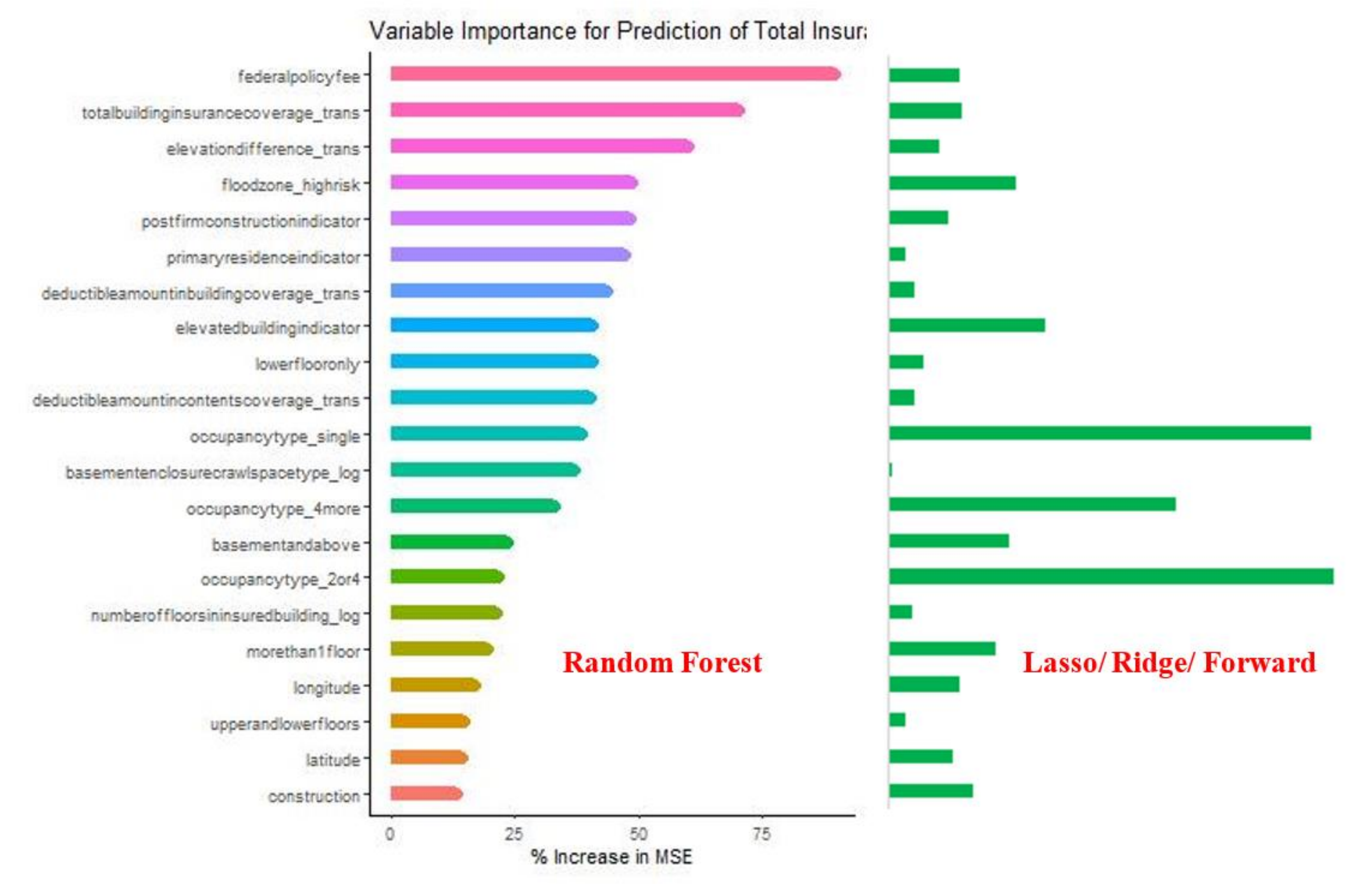
TREE-BASED REGRESSION

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

- Decision Tree: input: 21 variables;
  - only 4 variables ("federalpolicyfee", "floodzone_highrisk", "elevationdifference_trans" and "totalbuildinginsurancecoverage_trans") were included in the model
  - 12 terminal nodes.
  - $R2 = 0.784$
- Random Forest:
  - 500 weak regression trees
  - $R2 = 0.884$

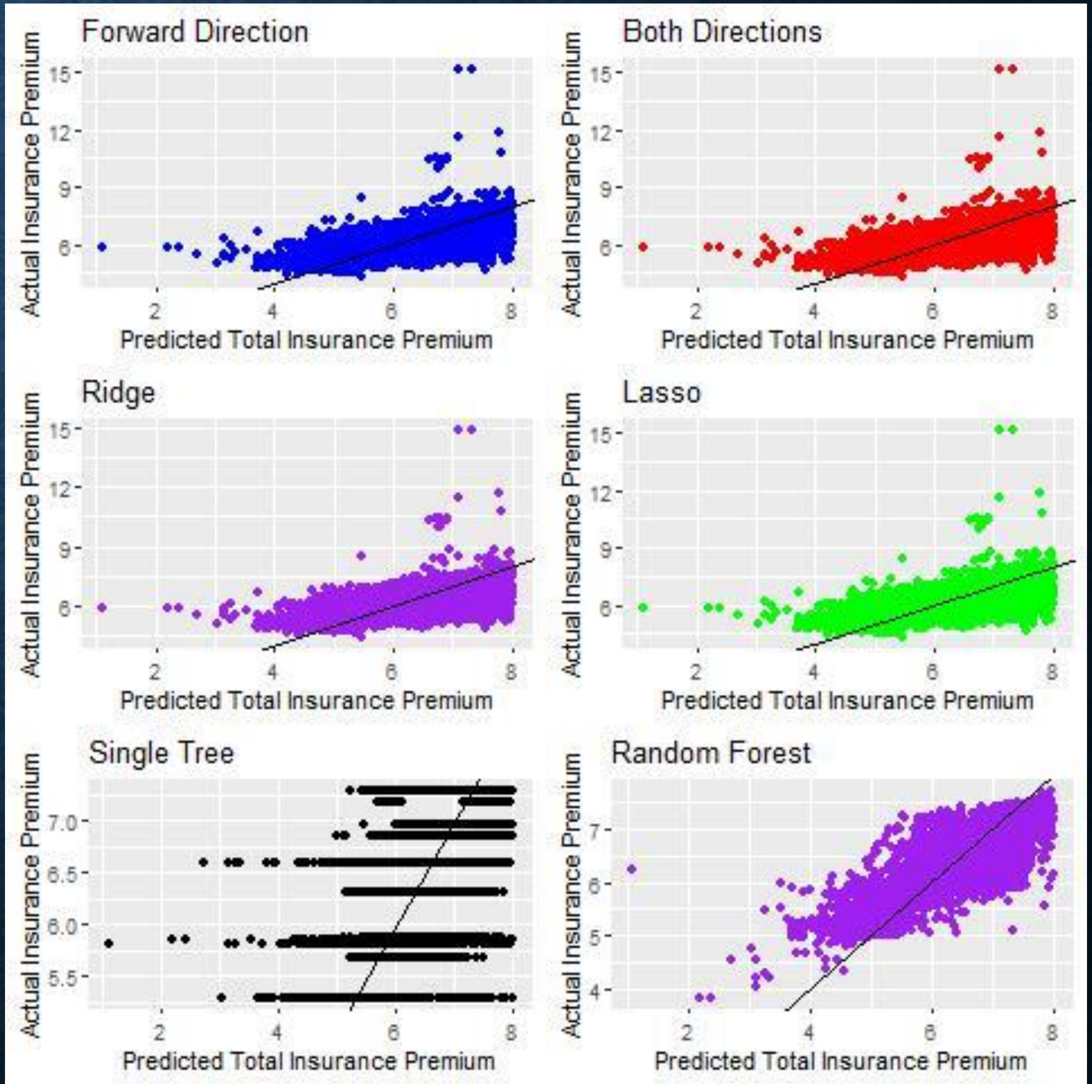


COMPARISON OF LINEAR & TREE-BASED REGRESSIONS



CONCLUSIONS

- The most important features based on linear regression are:
  - Occupancy type (3 dummy variables),
  - Elevated building indicator, and
  - Floodzone_high risk.
- Random Forest shows the highest predictive power. Based on this model the following are the most important features in predicting insurance premium:
  - Federal policy fee,
  - Total building coverage,
  - Elevated building indicator, and
  - floodzone_high risk.



REFRENECES

[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)

THANK YOU!