Investigation of key factors to earthquake insurance take-up rates in Quebec and British Columbia households and prediction model building

Yongcheng Jiang

Western University

USRI Program

Supervised by: Katsuichiro Goda, Jiandong Ren

A B S T R A C T

Maintaining an adequate level of earthquake take-up rate could protect the insurance industry from systemic failure. Past research has shown that British Columbia and Quebec have significant differences in earthquake insurance take-up rate. This report investigates key factors from the structure (default options and various types) of the insurance plan and personal characteristics along with socioeconomic/demographic profiles that affect the demand for earthquake protection in the form of insurance. The report also provides a prediction model for earthquake insurance take-up rate. The results show an importance ranking of key factors of earthquake insurance take up, the most important three are "annual expected loss ratio", “age” and “average household size”. An optimal prediction model constructed by random forest with 15 predictors provides 69.4% testing accuracy. An important finding is that there exist cognitive biases among participants. Possible explanations of this finding are discussed.

1. Introduction

According to Insurance Bureau of Canada, Québec City, Montreal, and Ottawa are exposed to a high-risk area in which there is at least a 5-15% chance that a strong earthquake will strike in the next 50 years. The PACICC (Property and Casualty Insurance Compensation Corporation) points out a catastrophic event with insured losses greater than $35 billion would overwhelm Canada’s insurance industry (PACICC,2021). However, the results from Goda et al. (2020) show that British Columbia and Quebec have significant difference in earthquake insurance take-up rate. The take-up rate is significantly higher in British Columbia (0.399) while the take-up rates in Quebec (0.034) are uniformly low. Some assumptions could be made: 1. In real life most Quebec residents underestimate the risk of earthquake. 2. Quebec households do not have enough understanding of the earthquake insurance provided (Lamontagne, 2016).

This report analyzes a survey dataset and selected features from the 2016 FSA dataset. The earthquake demand and personal characteristic data are based on a web-based earthquake protection choice survey that was conducted by the Institute of Catastrophic Loss Reduction (Kunreuther et al., 2021). In addition, regional demographic profiles based on the 2016 Canadian Census's Forward Sortation Areas (FSA) are combined with the survey dataset. By combining these two datasets, the demographic and seismic hazard profiles such as PGA corresponding to the region (FSA) of each participant could be evaluated. Logistic regression and decision tree methods are then applied to investigate key features to the take-up rate and finally an optimal prediction model is built.

Maintaining an adequate level of earthquake take-up rate could protect the insurance industry from systemic failure. Hence, it is important to investigate the key factors to earthquake insurance take-up rates. A prediction model might be useful for the government to estimate the take-up rate gap in regions at risk.

2. Data and Methods

To analyze the key factors to earthquake insurance take-up rates in Quebec and British Columbia households, a web-based earthquake protection choice survey by the ICLR (Kunreuther et al., 2021) is merged with some regional demographic profiles and physical hazard/risk indicators, such as PGA and annual expected loss ratio. The PGA (a seismic hazard indicator) is extracted from the Geological Survey of Canada, while the annual expected loss ratio (a quantitative seismic risk indicator) is calculated by adopting an earthquake catastrophe model. In addition, an estimation of earthquake insurance take-up rates for FSA locations is applied and it is served as a benchmark to compare with the estimation from the web survey. The approaches for the estimation of these data are illustrated in the paper of Goda et al. (2020). To achieve more accurate prediction, a subset of variables is selected by recursive feature elimination (RFE). Following that, logistic regression and classification tree methods (CART and random forest) are applied to obtain a variable importance ranking and an optimal prediction model.

2.1 Construction of Dataset

The individual survey and the regional demographic profiles from the Census data are combined by referring to the "key", which is forward sortation area (FSA), FSA. A FSA is a way to designate a geographical unit based on the first three postal code. To combine the survey data and census FSA data, an FSA column is added to the survey data by extracting the first three postal code of each participant. Each participant has their survey responses and corresponding region-wide demographic profiles. This step is important because a typical household response could be a good indicator of the regional household response.

2.2 Variable Selection Methods

The merged survey dataset has 83 columns in total, which may be too large for prediction. Some variables are not applicable because combining individual data with aggregated data might cause extrapolation issue. On the other hand, aggregated variables PGA and annual expected loss are applicable because these variables do not change within FSA. Hence, only PGA and annual expected loss are included. Then, some "dirty" variables are deleted manually. For example, columns(variables) with insufficient responses, columns with non-numeric and non-categorical datatype (such as comments), and redundant columns. Finally, a total of 14 variables (1 response and 13 predictors) are selected (Table 1). The first row of Table 1 shows each variable has 2396 values, indicating that there are no missing values.

A variable selection method called Recursive Feature Elimination with Cross-Validation (RFECV) is applied to select the best variables. It is similar with Backward Stepwise Selection, which begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor (the predictor with highest RSS), one-at-a-time. First, the estimator is trained on the initial set of features and the importance [1]of each feature is obtained either through any specific attribute or callable. Then, the least important features are deleted from the current set of features. A subset of n-1

[1] The importance calculations can be model based (e.g., the random forest importance criterion/RSS for logistic regression) or using a more general approach that is independent of the full model (Kuhn & Johnson, 2013).

features is obtained, along with its cross-validated test score. This procedure is repeated recursively on the subset until the number of features with the highest cross-validated score is obtained (Figure1).

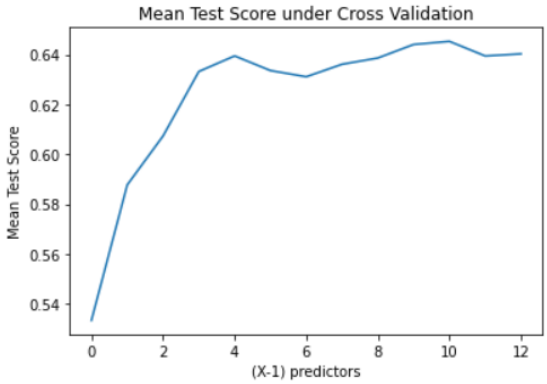

Figure 1 Mean Test Score under Cross Validation for different number of predictors

Table 1 A brief summary of each variable

	eq_choice	optout	riskpool	private2	BC	age	male	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	annual expected loss
count	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2396.000000	2.396000e+03
mean	0.525459	0.498331	0.337229	0.328047	0.500000	57.272538	0.469115	0.489983	0.532554	4.553422	0.036728	0.344324	0.315484	1.513912e-04
std	0.499456	0.500102	0.472862	0.469600	0.500104	15.015024	0.499149	0.500004	0.499043	1.952052	0.188132	0.475246	0.128903	1.814028e-04
min	0.000000	0.000000	0.000000	0.000000	0.000000	19.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	7.000000e-08
25%	0.000000	0.000000	0.000000	0.000000	0.000000	46.000000	0.000000	0.000000	0.000000	3.000000	0.000000	0.000000	0.300000	1.905000e-05
50%	1.000000	0.000000	0.000000	0.000000	0.500000	60.000000	0.000000	0.000000	1.000000	5.000000	0.000000	0.000000	0.300000	5.853000e-05
75%	1.000000	1.000000	1.000000	1.000000	1.000000	69.000000	1.000000	1.000000	1.000000	6.000000	0.000000	1.000000	0.400000	2.726800e-04
max	1.000000	1.000000	1.000000	1.000000	1.000000	90.000000	1.000000	1.000000	1.000000	10.000000	1.000000	1.000000	0.700000	8.030900e-04

2.3 Logistic Regression

To implement the RFECV method, a supervised learning estimator with a fit method that provides information about feature importance either through a coefficient attribute or through a feature importance attribute is needed (scikit-learn.org, 2022). Hence, the logistic regression is applied.

The variable selection process with RFECV method in logistic regression is illustrated below. Firstly, the dataset (with 13 variables) is fitted with logistic regression, the variable “earthquake choice” is the response variable and others are predictors.

$$ln\frac{p}{1-p} = \beta 0 + \beta 1 \cdot \text{optout} + \beta 2 \cdot \text{riskpool} + \beta 3 \cdot \text{hdprivate} + \beta 4 \cdot \text{BC} + \cdots$$

Next the RFECV function is applied with the fitted logistic regression as an estimator. The optimal features could be accessed by the ranking attribute (Table 2). Ranking_[i] corresponds to the ranking position of the i-th feature Only variables with ranking "1" will be selected. Annual expected loss

(ranking 3) is dropped first because it has the highest RSS, then age (ranking 2) is dropped. Then the best model with the 11 predictors, are selected. Then, its test accuracy is estimated. Instead of calculate test accuracy simply by the testing set spitted from the original data is not an accurate estimator for the real test accuracy. Hence, the 10-fold cross-validation method is applied to obtain a better estimate. The selected predictors for each model as shown in Table 3 and the optimal model prediction accuracy along with the prediction accuracy of the M1, M2, M3 [2]models from (Kunreuther et al., 2021) are compared in Table 4.

After getting the optimal model for prediction, it is of interest to ask what are the most important variables contributing to the result? In logistic regression, the odds ratios of predictors could be used to interpret their effect on the response. Odds ratios that are greater than 1 indicate that the event is more likely to occur as the predictor increases. Odds ratios that are less than 1 indicate that the event is less likely to occur as the predictor increases. For example, in Table 5, the predictor "eq_exper" (which represents earthquake experience) has an odds ratio of 2.9 with a confidence interval (1.7 – 4.9), which indicates participants who has earthquake experience are 190% more likely to purchase earthquake protection comparing with participants who does not have earthquake experience.

However, the odds ratio does not show which predictor is more important for prediction. The odds ratio compares the odds [3]of two events, for example, participants living in BC/Quebec. In other words, the odds ratio compares the odds of the event (purchasing earthquake protection) occurring at 2 different levels of the predictor, it is not able to compare the effect across different predictors. Hence, a more intuitive way to compare the key factors is using the importance plot in tree classification methods.

[2] M1 has 4 predictors: optout, risk pool, high deductible private and BC; M2 is identical to M1 with additional variables based on socio-economic characteristics of the respondents; M3 is identical to M2 with one additional variable PGA.

[3] The odds of an event are the probability that the event occurs divided by the probability that the event does not occur.

Table 2 the ranking of features

	Feature	Rank
0	optout	1
1	riskpool	1
2	private2	1
3	BC	1
4	age	2
5	male	1
6	uni	1
7	inclevel	1
8	risktaking	1
9	eq_exper	1
10	insur_eq_exper	1
11	PGA	1
12	annual expected loss	3

Table 3 The selected features for the optimal model, 4-feature model, 11-feature model and 12-feature model

	0	1	2	3	4	5	6	7	8	9	10	11
0	optout	riskpool	private2	BC	male	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	None
1	riskpool	eq_exper	insur_eq_exper	PGA	None	None	None	None	None	None	None	None
2	optout	riskpool	private2	BC	male	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	None
3	optout	riskpool	private2	BC	age	male	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA

Table 4 A summary of cross-validation score of logistic models

	4 predictors	11 predictors	12 predictors	optimal logistic
my model	0.633156	0.645244	0.639393	0.645244
original	0.601336	0.638561	0.641733	-

Table 5 Odds ratios of optimal features

	OR	z-value	2.5%	97.5%
Intercept	0.476973	4.931400e-05	0.333613	0.681938
C(optout)[T.1]	1.584651	1.719433e-07	1.333429	1.883204
C(riskpool)[T.1]	2.603110	7.958735e-19	2.106617	3.216618
C(private2)[T.1]	1.125612	2.653140e-01	0.914048	1.386143
C(BC)[T.1]	1.262569	1.271078e-02	1.051019	1.516700
C(male)[T.1]	0.901686	2.487354e-01	0.756280	1.075048
C(uni)[T.1]	1.260834	9.793801e-03	1.057499	1.503266
C(inclevel)[T.1]	1.112242	2.358631e-01	0.932850	1.326131
C(eq_exper)[T.1]	2.895295	8.125385e-05	1.706324	4.912744
C(insur_eq_exper)[T.1]	2.231302	2.643996e-15	1.828752	2.722461
risktaking	0.874114	7.111343e-09	0.835185	0.914858
PGA	2.900819	2.426164e-03	1.457387	5.773862

2.4 Classification Tree and Random Forest

The main reason of using tree classification methods is that the basic idea of tree is intuitive. Tree methods have a nice graphical representation and importance plot that can be determined which predictor is more important. In this section, a basic classification tree will be introduced first to explain the algorithm of tree and then an ensemble tree method, random forest, will be introduced.

The basic idea of a classification tree is that the predictor space is divided into some distinct and non-overlapping regions. For every observation falls into a specific region, the same prediction is given by the majority vote: most commonly occurring class of training observations in that region (James, 2021). A recursive binary splitting method is applied to grow a classification tree. At each step of the tree-growing process, the best split is made based on the smallest entropy on that particular step. Entropy is referred to as a measure of node purity—a small value indicates that a node contains predominantly observations from a single class (James, 2021).

The classification tree itself has the function of variable selection by controlling the maximum depth (i.e., number of splits) a tree could reach. Figure2 illustrates a three-depth tree example. Take the first node as example: first decision boundary is "earthquake insurance experience <= 2.5", which is decided by testing all the possible decision boundaries splitting the dataset and choosing the one that minimizes the Gini index of the two splits. At this node, the smallest entropy is 0.999. Samples indicates the number of observations fall into this class. Here "samples = 1605" simply represents the number of samples in the training set since the split has not been made at this node. Value = [773,832] describes the repartition of these participants among the tree possible classes of choice, i.e., 773 participants chose not to purchase and 832 chose to purchase the earthquake protection. By the rule of majority vote, this class is predicted to be “1”, which is purchasing the earthquake protection. Then, this node is spited by “earthquake insurance experience”, samples with earthquake insurance experience <= 0.5 will fall into the left branch while the others fall into the right branch. At each branch, the best split is decided again by the minimum entropy. This process continues until the maximum depth is reached. One problem with the

decision tree is that it suffer from high variance (James, 2021). To reduce the variance, random forest method is applied.

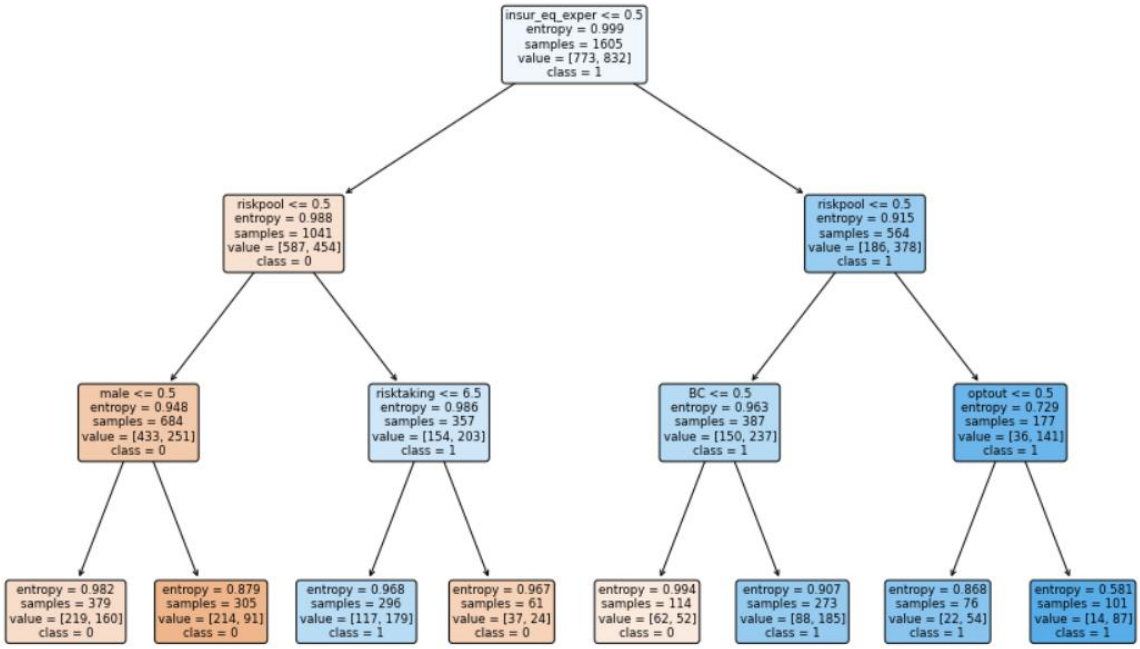

Figure 2 A classification tree with 3 depths

An ensemble method is an approach that combines many simple "building ensemble block" models in order to obtain a single and potentially more accurate model (James, 2021). The random forest algorithm is an ensemble of tree method that consists of a large number of decision trees (controlled by n estimators in the code). Each individual tree spits out a class prediction and the class with the most votes will be the model's prediction (the class can be visualized with importance plot). Moreover, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. This step is same as the bagging method, however, when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split use only one of those m predictors and m is usually the square root of the number of total predictors. The purpose of this process is to decorrelating the trees. Because averaging

many uncorrelated quantities leads to a large reduction in variance. And uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions.

The feature randomness is the key of random forest. In a normal decision tree, when splitting a node, the predictor produces the most separation between the observations is selected. However, in a random forest each tree can choose only from a random subset of features, usually the squared root of total number of features. This leads to more variation of the trees in the model and lead to a lower correlation across trees and more diversification. Hence, it is impossible to visualize the whole forest. But the selected feature ranking could be visualized with the importance plot (Figure3).

One crucial problem with both tree and forest methods is that as the maximum depth increases, the testing accuracy might not be increasing because of overfitting. Without controlling the depth of tree, the tree will be split until all leaves are pure. Hence, a tree pruning process should be performed. This is very similar with the variable selection. A simple and direct method is comparing the testing accuracy for each depth. According to Figure 4 and Figure 5, the maximum depth with the highest cross-validation score is selected, which are 5 depths for tree and 9 depths for forest.

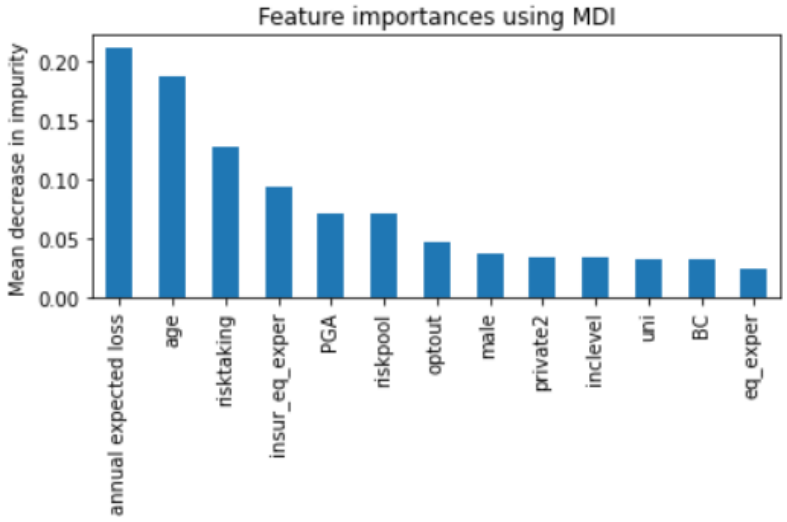

Figure 3 Feature importance plot for random forest

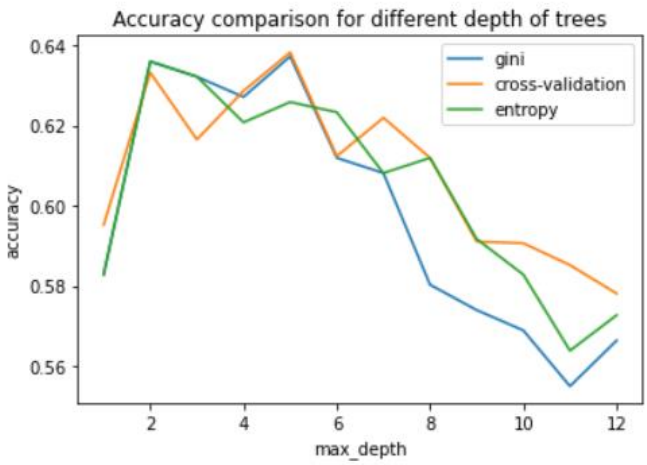

Figure 4 Accuracy comparison for different depth of trees for the basic tree model

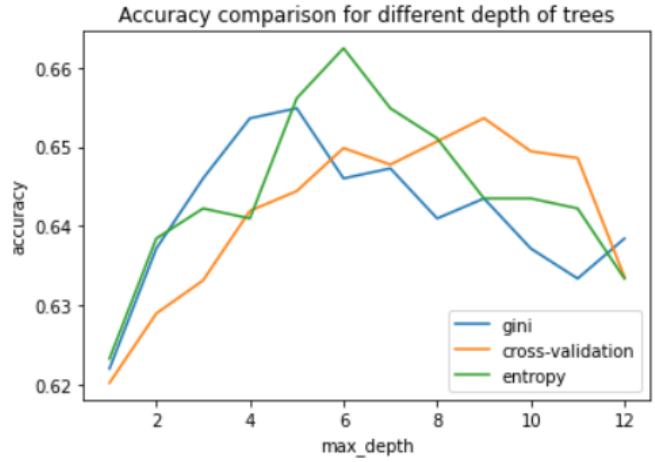

Figure 5 Accuracy comparison for different depth of trees for the random forest model

3. Results

By applying variable selection methods to logistic regression, tree classification and random forest respectively, a table containing their selected features (Table 6) and a table containing their accuracy results (Table 7) is summarized below. In Table 6, row 1-3 correspond to 4 features, 11 features and 12 features with logistic regression. These three models are compared with the three logistic models (M1, M2, M3) implemented by Kunreuther et al. (2021). As shown in Table 7, each accuracy is higher than the original one. This result is not surprising because the original models are designed to investigate the impact of default option and the structure of the insurance plan to the demand for earthquake protection. The optimal logistic model has 11 features, which is shown in row 0 of Table 6. One

important observation from Table 7 is that the accuracy of the optimal logistic model is not significantly higher than the accuracy of a 4 features model. The trend could be visualised clearly by observing Figure1. When including more predictors from 1 to 5, the test accuracy by cross-validation increases dramatically. Then it starts fluctuating and the highest score appears at 11 features. The cost of including more features is more computation time and harder to interpret the features. Hence, it is a "trade off" to decide how many features should be included.

The optimal random forest model with all variables selected gives the best predicting accuracy, which is 65.4%. Although its accuracy is very close with the optimal logistic model, it indicates the importance of variables by an importance plot (Figure3). The importance plot shows that annual expected loss ratio, age and risk taking are the three most important predictors since they have relatively high MDIs (mean decrease in impurity). Recall from section 2.4, when building a classification tree, the quality of a particular split depends on the node purity, either the Gini index or the entropy are typically used to evaluate the quality of a particular split, since these two approaches are more sensitive to node purity. The mean decrease in impurity represents the total amount that the impurity is decreased due to splits over a given predictor, averaged over all forest trees. Hence, a large value indicates an important predictor.

Table 6 Selected features for each model

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	optout	riskpool	private2	BC	male	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	None	None
1	riskpool	eq_exper	insur_eq_exper	PGA	None	None	None	None	None	None	None	None	None
2	optout	riskpool	private2	BC	male	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	None	None
3	optout	riskpool	private2	BC	age	male	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	None
4	insur_eq_exper	riskpool	age	annual expected loss	risktaking	optout	BC	private2	male	inclevel	uni	None	None
5	annual expected loss	age	risktaking	insur_eq_exper	PGA	riskpool	optout	male	private2	inclevel	uni	BC	eq_exper

Table 7 Cross-validation score for each selected model

	4 predictors	11 predictors	12 predictors	optimal logistic	optimal tree	optimal forest
my model	0.633156	0.645244	0.639393	0.645244	0.638171	0.653624
original	0.601336	0.638561	0.641733	-	-	-

4. Discussions

4.1 Weakness of Models

The variable importance ranking in random forest classification might be inaccurate because of the correlation between variables. Random Forest is a machine-learning method that is suitable for integrating complex data as it generally works well with high-dimensional variables and can determine strong predictors without the assumption of underlying models. However, a common problem of high-dimensional dataset is that the correlation between variables has effects on the variable importance ranking by decreasing the estimated importance scores of correlated variables (Darst, 2018).

Since there exists a potential inaccuracy of the importance ranking by random forest. The variable importance in logistic regression might be assessed to compare the difference. In general, coefficients of variables are not good indicators of relative importance since their units might be different and predictors and categorical predictors cannot be compared directly. Possible methods could be: 1. standardized coefficients 2. Comparing the drop in deviance [4]from adding each variable (James, 2021).

4.2 Cognitive biases among participants

One observation is that if worry and likely rating (i.e., participants were asked how likely they thought it was that the house in the scenario would be damaged by an earthquake in the next 50 years?) are also selected as variables, the results would be changed significantly. Worry and likely rating would become two significant predictors as their Z-value are low (see Table 8). They would be selected in the optimal model as shown in the row 0 of Table 9. One interpretation is that those who rate more likely and express worry tend to select to buy the insurance coverage. On the other hand, it is very difficult to

[4] The deviance is negative two times the maximized log-likelihood; the smaller the deviance, the better the fit (James, 2021)

measure such variables – in other words, these variables can explain how data are crated but may not be suitable as predictors. Hence, they are dropped in the variable selection process.

According to the conclusion from Goda et al. (2020), British Columbia and Quebec have significant difference in earthquake insurance take-up rate. The take-up rate is significantly higher in British Columbia (0.399) while the take-up rates in Quebec (0.034) are uniformly low. By analyzing the response of participants from the survey data, the hypothetical take-up rate of different earthquake plans and the total take-up rate of BC and Quebec participants are shown in Figure 6. The estimated take-up rates are 0.58 and 0.47 for BC and Quebec respectively. It is worth noticing that the hypothetical take-up rate could not be compared with the real one directly. However, the response could indicate the cognitive biases among participants. As discussed above, those who rate more likely and express worry tend to select to buy the insurance coverage when they were provided with the real earthquake possibility and detailed earthquake insurance plans in BC/Quebec.

Table 8 Odds ratios of the optimal logistic model after adding worry and likely rating

	OR	z-value	2.5%	97.5%
Intercept	0.146231	1.900052e-19	0.096291	0.222070
C(optout)[T.1]	1.551906	1.660636e-06	1.296527	1.857588
C(riskpool)[T.1]	2.796659	7.296466e-20	2.242262	3.488132
C(private2)[T.1]	1.090895	4.333824e-01	0.877519	1.356155
C(BC)[T.1]	0.841476	9.199971e-02	0.688412	1.028573
C(uni)[T.1]	1.143229	1.541623e-01	0.950988	1.374330
C(inclevel)[T.1]	1.131420	1.868981e-01	0.941869	1.359117
C(eq_exper)[T.1]	2.484851	1.254774e-03	1.429368	4.319729
C(insur_eq_exper)[T.1]	1.711999	4.383625e-07	1.389648	2.109124
risktaking	0.889857	1.437104e-06	0.848616	0.933102
PGA	1.596884	2.058786e-01	0.773252	3.297809
worry	1.586964	1.018234e-22	1.447091	1.740356
likely_rating1to7	1.146579	2.023271e-07	1.088934	1.207276

Table 9 Selected features for each model after adding worry and likely rating

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	optout	riskpool	private2	BC	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	worry	likely_rating1to7	None	None	None
1	riskpool	eq_exper	insur_eq_exper	worry	None	None	None	None	None	None	None	None	None	None	None
2	optout	riskpool	BC	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	worry	likely_rating1to7	None	None	None	None
3	optout	riskpool	private2	BC	uni	inclevel	risktaking	eq_exper	insur_eq_exper	PGA	worry	likely_rating1to7	None	None	None
4	worry	annual expected loss	riskpool	age	insur_eq_exper	likely_rating1to7	risktaking	optout	male	PGA	eq_exper	uni	None	None	None
5	worry	annual expected loss	age	likely_rating1to7	risktaking	riskpool	insur_eq_exper	PGA	optout	male	BC	uni	inclevel	private2	eq_exper

Table 10 Cross-validation score for each selected model after adding worry and likely rating

	4 predictors	11 predictors	12 predictors	optimal logistic	optimal tree	optimal forest
my model	0.674050	0.686567	0.688649	0.688649	0.676142	0.705352
original	0.601336	0.638561	0.641733	-	-	-

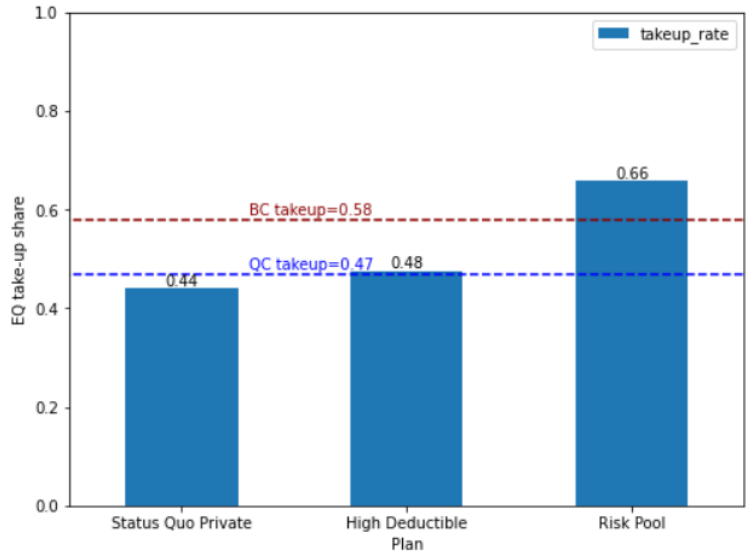

Figure 6 Comparison of earthquake insurance take-up rate

4.3 Application of Models

This prediction model could be applied to estimate the earthquake protection willingness of people in different regions. If provided a more representative and complete sample of household information, a relatively accurate estimation of the earthquake insurance take-up rate could be made by applying the prediction models illustrated in this report. With the estimated insurance take-up rates in FSAs, the government and insurance companies could make policy adjustments or some positive actions to mitigate the gap of insurance take-up rates in some regions.

5. Conclusions

In this project, we analysis on a merged dataset containing both individual and regional socioeconomic/demographic profiles on key factors that affect the demand for earthquake protection. In

addition, prediction models for earthquake insurance willingness were constructed. Though RFECV method with logistic regression and random forest as estimators, the following conclusions were drawn:

I. The importance plot of random forest shows that annual expected loss ratio, age and risk taking are the three most important predictors, followed by earthquake insurance experience, PGA ...

II. The 13 features random forest model provides 65.4% prediction accuracy of earthquake insurance take up.

References

AIR Worldwide. (2013). Study of Impact and the Insurance and Economic Cost of Major Earthquake in British Columbia and Ontario/Québec. http://assets.ibc.ca/Documents/Studies/IBC-EQ-Study-Summary.pdf

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics, 19(Suppl 1),* 65. https://doi.org/10.1186/s12863-018-0633-8

Grant Kelly. (2021). How Big is Too Big? Update to the PACICC P&C Industry Model the Tipping Point for Systemic Failure. Property and Casualty Insurance Compensation Corporation.

Goda, K., Wilhelm, K., & Ren, J. (2020). Relationships between earthquake insurance take-up rates and seismic risk indicators for Canadian households. *International Journal of Disaster Risk Reduction*, *50*, 101754. https://doi.org/10.1016/j.ijdrr.2020.101754

James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2021). *An introduction to statistical learning: With applications in R*. Springer.

Kunreuther, H. C., Conell-Price, L., Kovacs, P., & Goda, K. (2021). The impact of a government risk pool and an opt-out framing on demand for earthquake protection. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3901895

Kuhn, M., & Johnson, K. (2013). A short tour of the predictive modeling process. *Applied Predictive Modeling*, 19–26. https://doi.org/10.1007/978-1-4614-6849-3_2

Lamontagne, M., Flynn, B., & Goulet, C. (2016). Facing the communication challenges during an earthquake swarm period. *Seismological Research Letters*, *87*(6), 1373–1377. https://doi.org/10.1785/0220160036

Sklearn.feature_selection.RFECV. scikit. (n.d.). Retrieved August 9, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html