Who’s Likely to sue?

Machine Learning Models for Predicting Litigated Claims

Alexander Wu, Camilla Calmasini, Helen Xu
Advisors: Ian Duncan, Janet Duncan, Xiyue Liao
Department of Statistics and Applied Probability, University of California, Santa Barbara
March 2018

Abstract

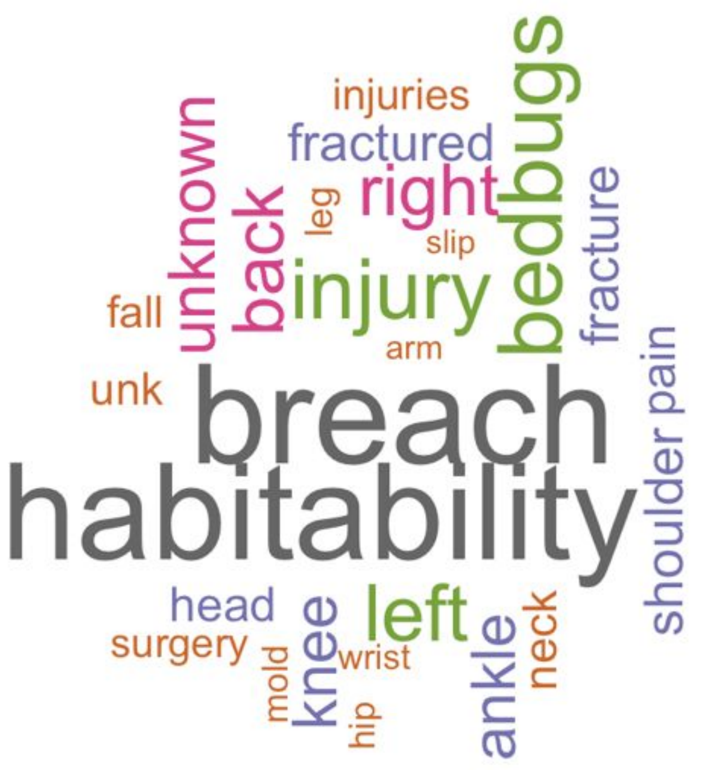
We identify factors that are predictive of insured small business claims lawsuits and we compare different models to predict whether claimants will litigate. Our model allows an insurer to direct resources to high-risk claimants. We highlight the performance of the Random Forest and the Logistic Regression models. We find the lag between claim incurral and report date to be the most significant indicator of whether a claimant will litigate. Policy state, injury severity, and claimant age were also important predictors.

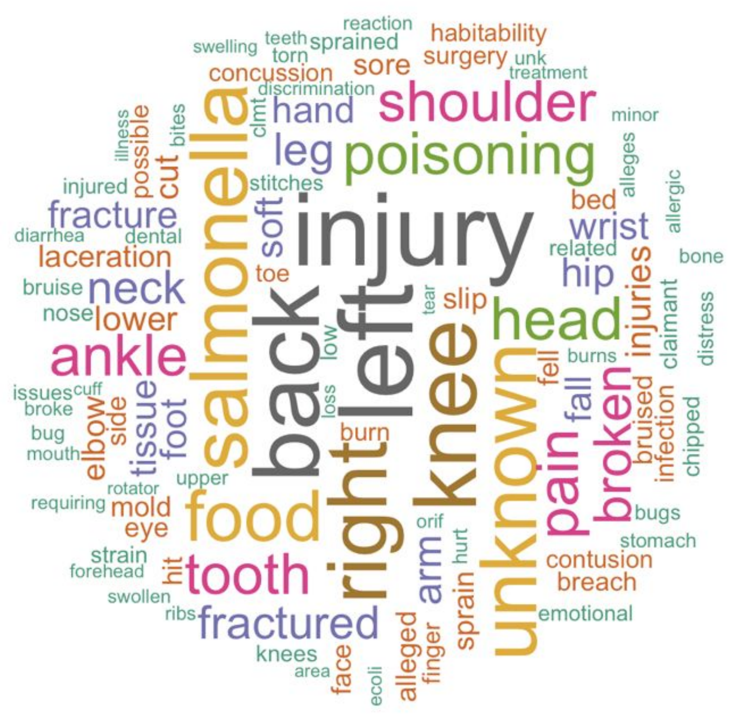
Introduction

In this project, we develop a model using claims data to predict whether a claimant will pursue attorney representation. This is important because attorney representation may delay the resolution of claims and yield no ultimate benefit for the customer. The purpose of this model is to further support the proper assignment and handling of claims, which will mitigate attorney involvement and potentially allow for timelier resolutions while maintaining a fair claims process. Our goal is to predict which claims are likely to be converted to litigated status within two years since claims that stay unlitigated for the first two years are unlikely to be litigated. Furthermore, we are using data at the 21 day mark to identify potentially litigious claims quickly so that the insurance company can triage them appropriately.

Word Clouds

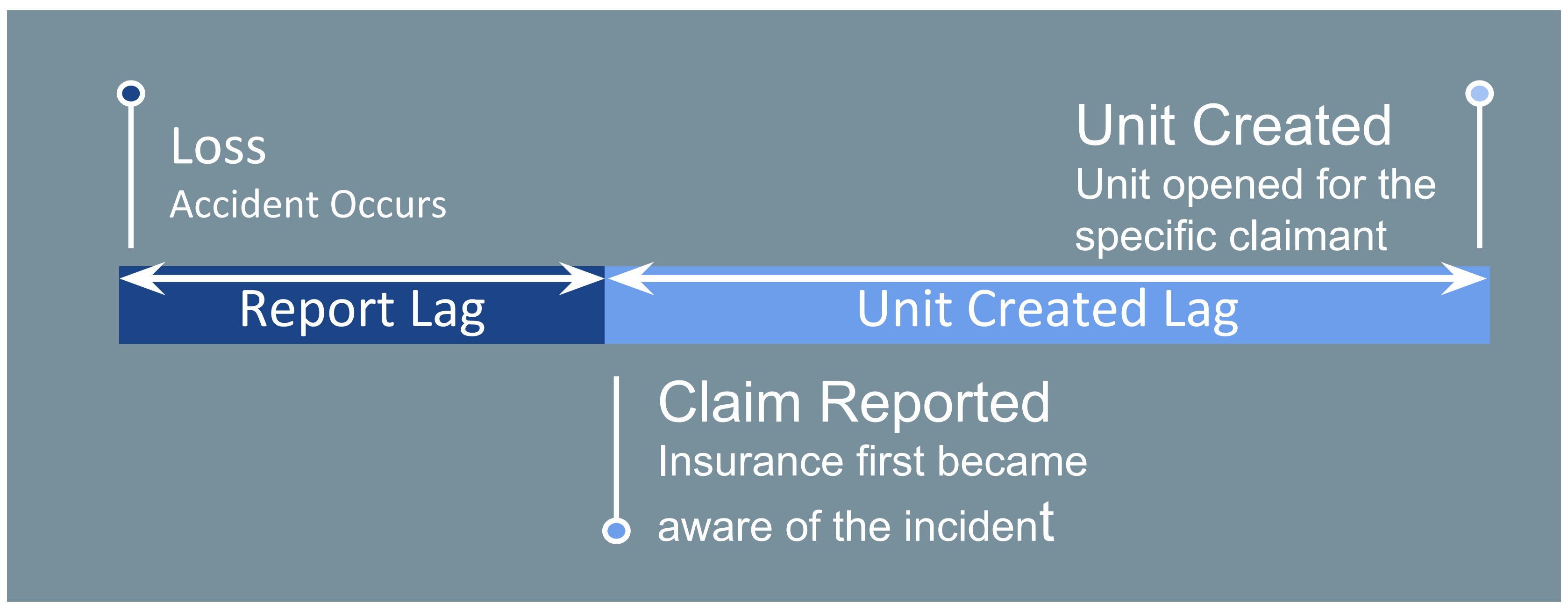
- Bigger font size indicates higher word frequency
- Same color indicates similar word frequency





Methods

- Data preprocessing
  - Created “Report Lag” and “Unit Created Lag”
  - Grouped variables with too many levels
  - Merged datasets
- Sampling techniques
  - K-fold cross validation
  - Balancing the data
- Predictive models
  - Random Forest
  - Logistic Regression
  - Decision Trees
- Model selection criteria
  - AUC
  - Specificity
  - Sensitivity



Loss
Accident Occurs

Unit Created
Unit opened for the specific claimant

Report Lag

Unit Created Lag

Claim Reported
Insurance first became aware of the incident

Results

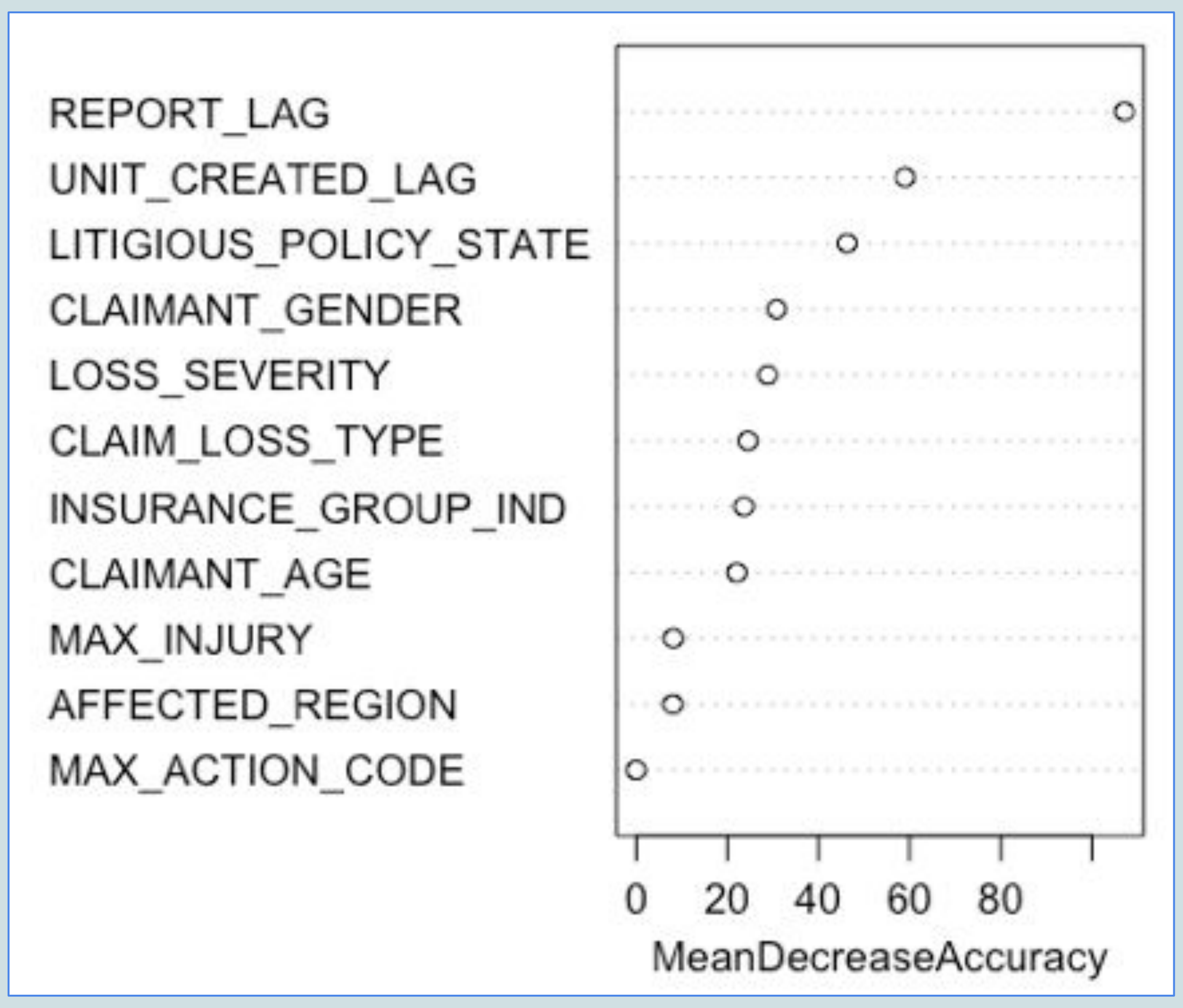
Random Forest is selected as best model, followed by Logistic Regression.

Model	Threshold	AUC	Sensitivity	Specificity
Random Forests	0.50	0.70	0.09	0.98
Random Forests (Undersampled)	0.50	0.72	0.67	0.65
Random Forests (Oversampled)	0.50	0.7q	0.74	0.60
Random Forests (ROSE)	0.50	0.67	0.93	0.19
Logistic Regression	0.50	0.65	0.00	1.00
Logistic Regression (Undersampled)	0.50	0.64	0.66	0.55
Logistic Regression (Oversampled)	0.50	0.64	0.64	0.58
Logistic Regression (ROSE)	0.50	0.63	0.64	0.54

A Lift Chart allows us to identify top claims likely to be litigated.

Quantile	Amount	Probability Range		Predicted #	Actual #
1	166	0	0.114	11.2	2
2	165	0.114	0.198	25.6	10
3	165	0.198	0.272	38.7	10
4	165	0.272	0.336	50.4	16
5	165	0.336	0.410	61.9	14
6	166	0.410	0.490	74.1	20
7	165	0.490	0.566	87.3	26
8	165	0.566	0.664	99.5	32
9	165	0.664	0.740	113.1	47
10	165	0.740	1	134.8	53

Most important variables:



Conclusion

The Random Forest and Logistics Regression models give us similar important variables and we conclude that the top three reasons why claimants litigated are: high report lag, low unit created lag and policy state. Furthermore, based on the AUC value, Random Forest gives us the best classifier.