Predicting Insurance Claim Litigation

Syen Yang Lu, Mingxi Chen, Aaron Barel
Advisors: Professors Xiyue Liao, Janet Duncan
Department of Statistics and Applied Probability

Background

Litigated claims are the most costly claims for insurance companies. Predicting which claimants are likely to litigate enables proper handling of claims prior to attorney involvement. Using small business claims data from an insurance company, we develop predictive models that will indicate whether or not a claimant will litigate. Our models utilize machine learning algorithms and natural language processing to perform prediction, using quantitative data and text data such as claim adjuster's notes. By evaluating each model's performance, an insurance company can choose which model to use in order to decide which claims and claimants need proper assignment.

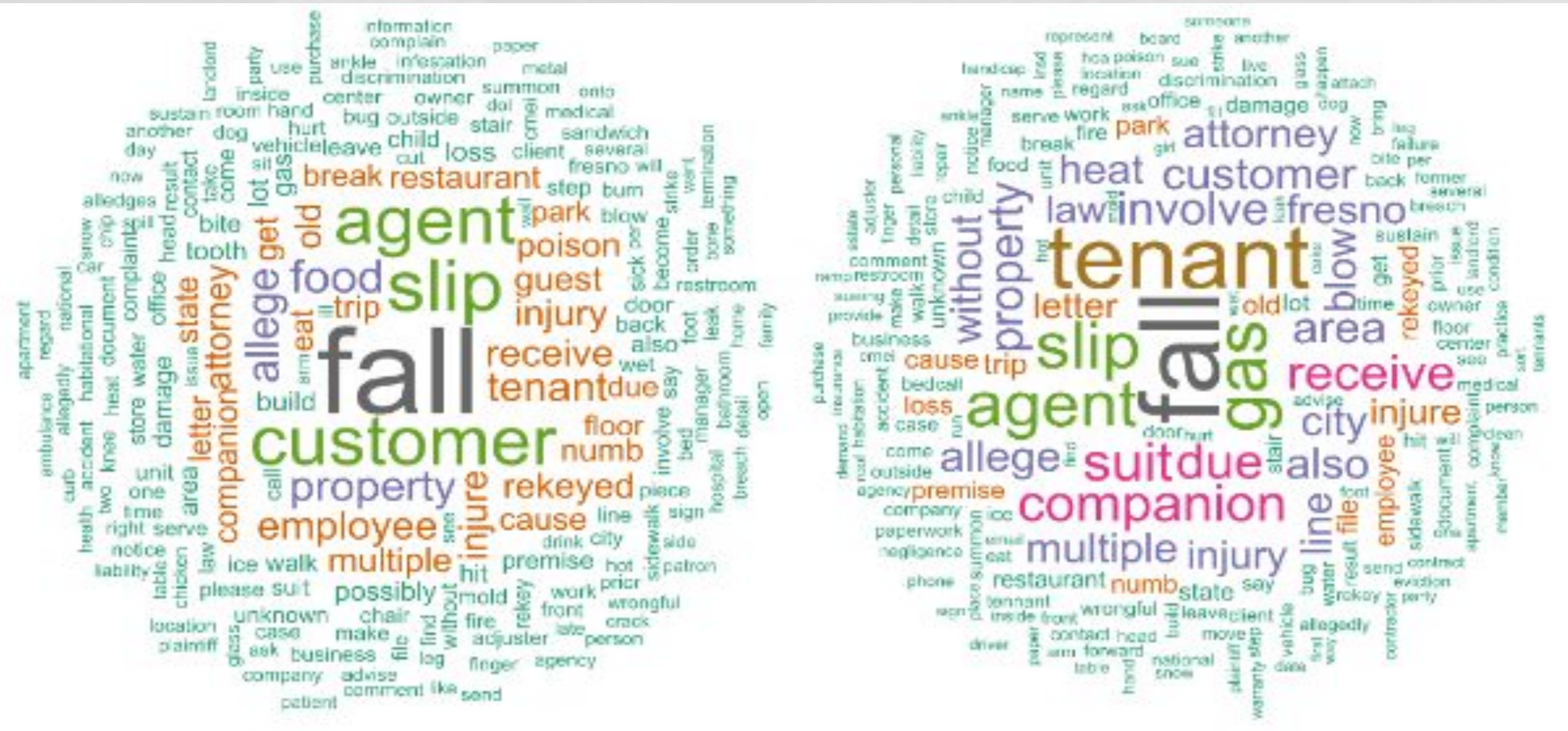
Methods

Data Preprocessing:

- Clean text fields with stopword removal and stemming
- Fix data errors such as negative Age
- Reduce high dimensional categorical fields
- Creation of new variables such as Report Lag and Unit Creation Lag

Data Exploration:

- Common words for unlitigated claims (left) and litigated claims (right) in Loss Description variable shown in word cloud below



Predictive Models:

- Support Vector Machines (SVM)
- Naive Bayes
- Logistic Regression
- Random Forest

Model Selection and Evaluation:

- 5-Fold Cross Validation
- Class balancing
- Threshold adjustments
- Recall, Accuracy, Area Under Curve (AUC), Precision

Model Evaluation

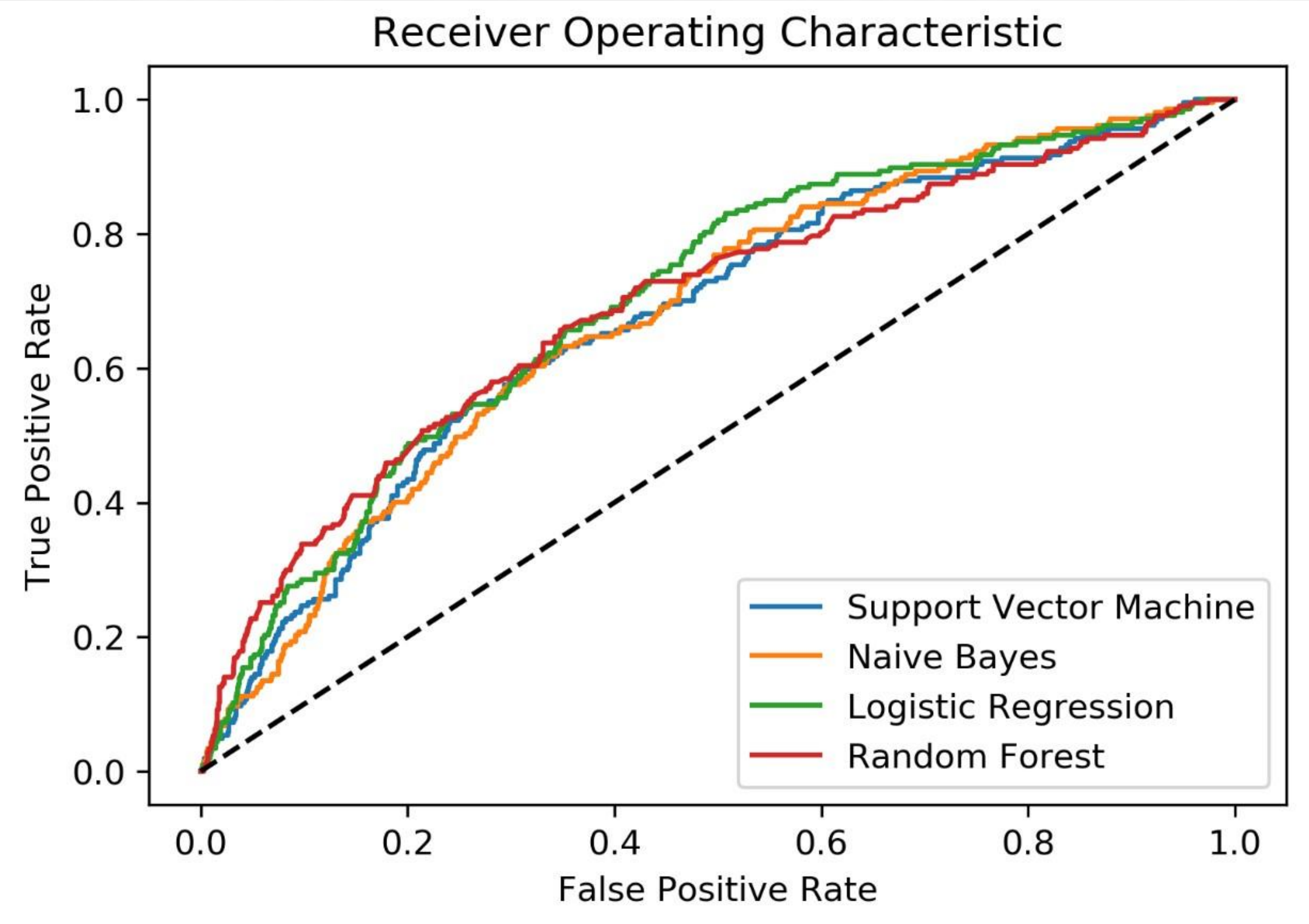
Class Imbalance:

Class	Number of Claimants	Distribution
Litigated	1,215	14.4%
Unlitigated	7,232	85.6%

Model Comparison:

	Accuracy	Precision	AUC	Recall
SVM	0.603	0.202	0.676	0.657
Naive Bayes	0.557	0.195	0.680	0.729
Logistic Regression	0.627	0.240	0.703	0.671
Random Forest	0.695	0.240	0.693	0.580

Naïve Bayes is most aggressive in classifying claims as litigated (highest Recall) while Random Forest is the least aggressive (lowest Recall).



Logistic Regression maximizes the True Positive Rate (TPR) and False Positive Rate (FPR) and thus it has the highest AUC score.

Top 3 Variables & Words:

Variable	Importance
Report Lag	0.47
Claimant Age	0.14
Unit Creation Lag	0.13

Word	Importance
fracture	0.104
attorney	0.103
eat	0.062

Conclusion:

- Using a combined dataset containing both variables (quantitative) and words (text) optimizes the modeling process.
- Based on current data available, Report Lag is the most important feature in predicting litigated claims.
- Each model's metrics can be adjusted to suit the company's needs by changing hyper-parameters and probability thresholds.