# Let the Notes Say The Price

Grant Nolasco, Isaac Golberg, Zijie Fang
Faculty Advisor: Janet Duncan, Xiyue Liao
Sponsor: CSAA
University of California, Santa Barbara - Dept. of Statistics and Applied Probability

## Abstract

The purpose of this study was to organize auto insurance claim adjuster notes into useful data for actuarial analyses and find useful variables that will improve the structured model to predict the severity of bodily injury claims. After processing adjuster notes, Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) were used to find the structure and topics of the given notes. Then, these notes were used to predict the severity of such claims using a variety of different predictive models. After comparing each model, LDA demonstrated better topics while Random Forest had better model performance.

## Introduction

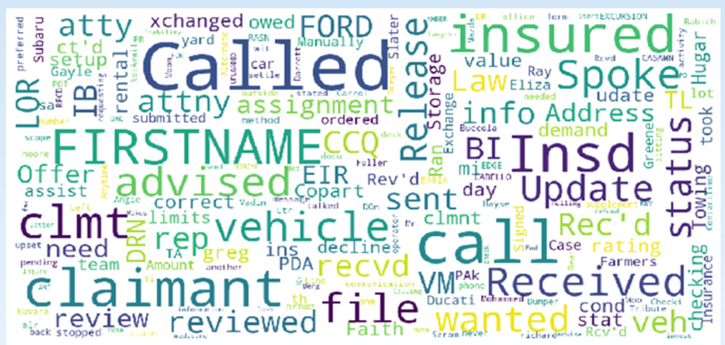
This case study will focus on automobile insurance. Two key components of automobile insurance are the property damage liability and bodily injury liability. Compared to the property damage liability, bodily injury liability is harder to predict and often more costly than the main part of the claim. Therefore, predicting the severity could help company better manage reserve. Claim notes contain a vast amount of unstructured data, so we perform topic modeling to find the important topics. We use LDA and NMF to extract the main topics, and then utilize machine learning algorithms to predict the severity of the claims.

## Data processing

Data processing will modify our merged dataset into a more usable format.

- Remove variables that do not add valuable information
- Change text into lower case and remove punctuation
- Remove stop words (offer no additional information)
- Stemming/Lemmatization

**Word cloud before data processing:**




**Word cloud after data processing:**


Word clouds present words with high frequencies in larger font and words with lower frequencies in smaller front. The word cloud before data processing is used to find high-frequency words that are not important e.g. FIRSTNAME. By deleting uninformative words during data processing, we are left with a dataset that is better suited for analysis.

## Topic Modling

**Latent Dirichlet Allocation (LDA)**

- Assumes each document is a mixture of topics and each word can be attributed to one of the topics
- Dirichlet distribution over topics

**Non-Negative Matrix Factorization (NMF)**

- A linear-algebraic optimization algorithm that is used for dimensionality reduction
- Consists of a non-negative matrix of topics and corresponding weights

**Find Topics:**

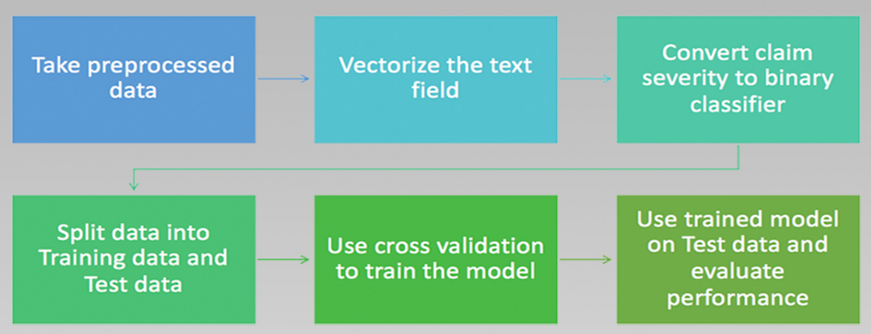
LDA:
-Split data into 10 subsets
-Cross validate each subset to get best parameters
-Find topics for each subset
-Group topics into categories

NMF:
-Tuning parameter alpha
-Find topics
-Group topics into categories

**Subset 1 (LDA)**
Topic 1:
rear driver bumper passenger hit seat veh dmg accident end
Topic 2:
pain medical treatment accident work neck medicare offer tx wage
Topic 3:
shop estimate created photo pm approved inspection final body assignment
Topic 4:
Update message alternate leave mail cwp voice city home enterprise
Topic 5:
drn offer detail status notavailable comment contacttimeanytime policereport
Topic 6:
oa vm atty provided lmtcb rep attorney discus spoke message
Topic 7:
adv pr asked liab stated spoke pending veh tow ob
Topic 8:
lane cv police iv turn report witness dac light stop
Topic 9:
pending driver listed paf range special ack prop settlement iv
Topic 10:
demand tl pd payment file total review subro activity issued

**Subset 1 (NMF)**
Topic 1:
lane hit accident turn light police stop passenger witness rear
Topic 2:
pending total detail medicare medical special verified claimed ssn wage
Topic 3:
range settlement special negotiation plan total demand visit lien attorney
Topic 4:
estimate shop photo created pm rear payment total bumper approved
Topic 5:
pain medical accident treatment work neck medicare wage doctor health
Topic 6:
driver ameriprise mercury turning st yield avoid collision decision state
Topic 7:
pd paf file ack rear driver sr prop exposure review
Topic 8:
stated adv asked oa veh spoke vm tl pr provided
Topic 9:
dac csv dcc iv state poi police witness report struck
Topic 10:
drn offer status detail notavailable contacttimeanytime comment policereport

## Predictive Modeling




- Logistic Regression
- Naïve Bayes
- Gradient Boosting
- Random Forest
- Linear SVM

## Result

| Methods | Accuracy | F1-score | Precision | Recall | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.62 | 0.63 | **0.65** | 0.62 | 0.52 |
| Naïve Bayes | 0.65 | 0.64 | 0.64 | 0.65 | **0.57** |
| Gradient Boosting | 0.65 | 0.65 | **0.65** | 0.65 | 0.55 |
| Random Forest | **0.71** | **0.66** | **0.65** | **0.71** | **0.57** |
| SVM | 0.60 | 0.61 | 0.64 | 0.60 | NA |

- Random Forest gains the highest accuracy
- Random Forest gains the highest F1-score
- Random Forest gains the highest precision
- Random Forest gains the highest recall
- Random Forest gains the highest AUC

**Random Forest classifier offers the best overall performance**