The Dendrotrons
Kaggle: Allstate

Kawtar Belmkaddem
Jhonasttan Regalado
Jason Sippie
Nathan Stevens
Chris Valle
Conred Wang

Outline

- Business Problem
- EDA
- Unsupervised ML
- Supervised ML
- Takeaways

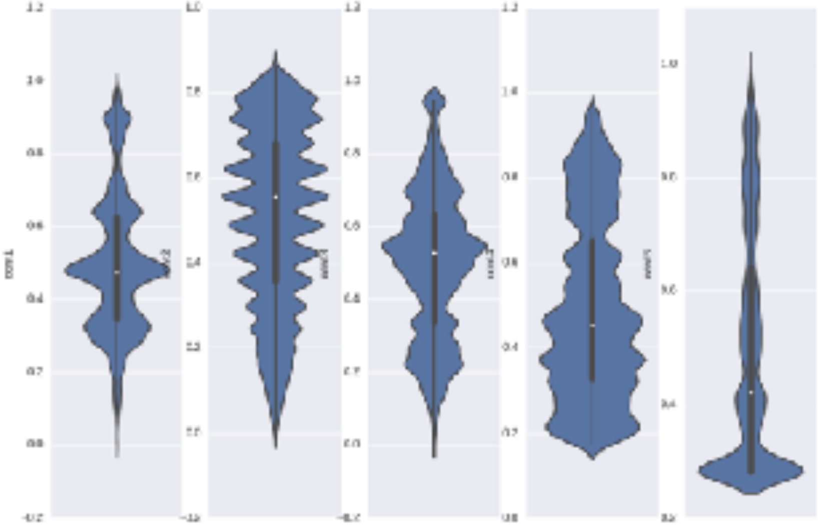
How Severe is the Insurance Claim?

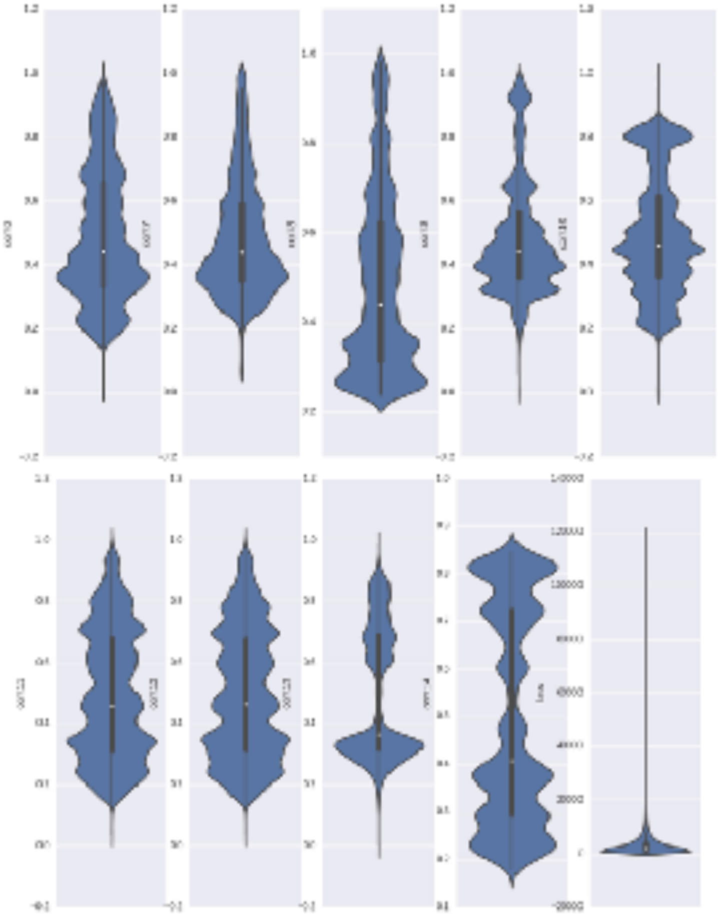
When you've been devastated by a serious car accident, your focus is on the things that matter the most: family, friends, and other loved ones. Pushing paper with your insurance agent is the last place you want your time or mental energy spent. This is why Allstate, a personal insurer in the United States, is continually seeking fresh ideas to improve their claims service for the over **16 million households** they protect.

Allstate is currently developing automated methods of **predicting the cost**, **and hence severity, of claims.**

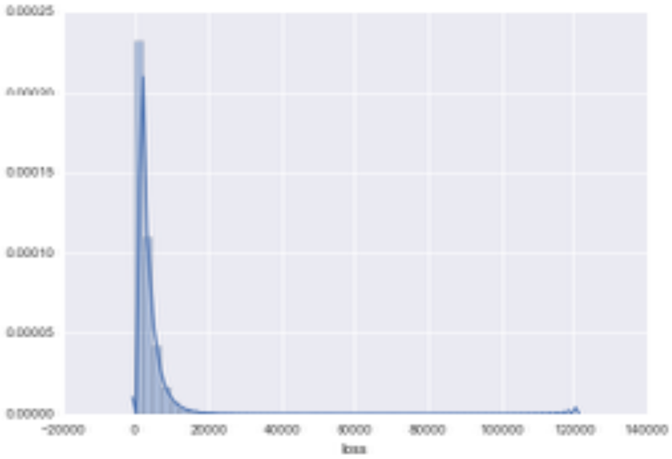
EDA - Statistical Description

rows: 188318 columns : 132	cont: 14 cat: 116
missingness: none	Remove ID column

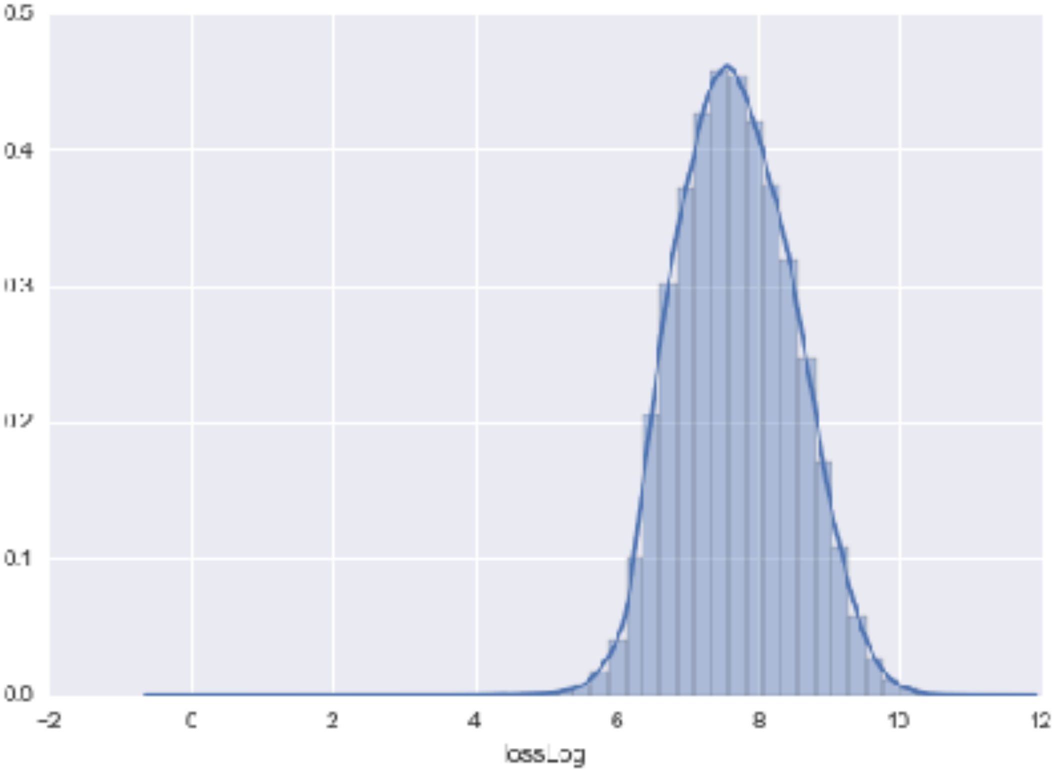




EDA - Transforming Response Variable

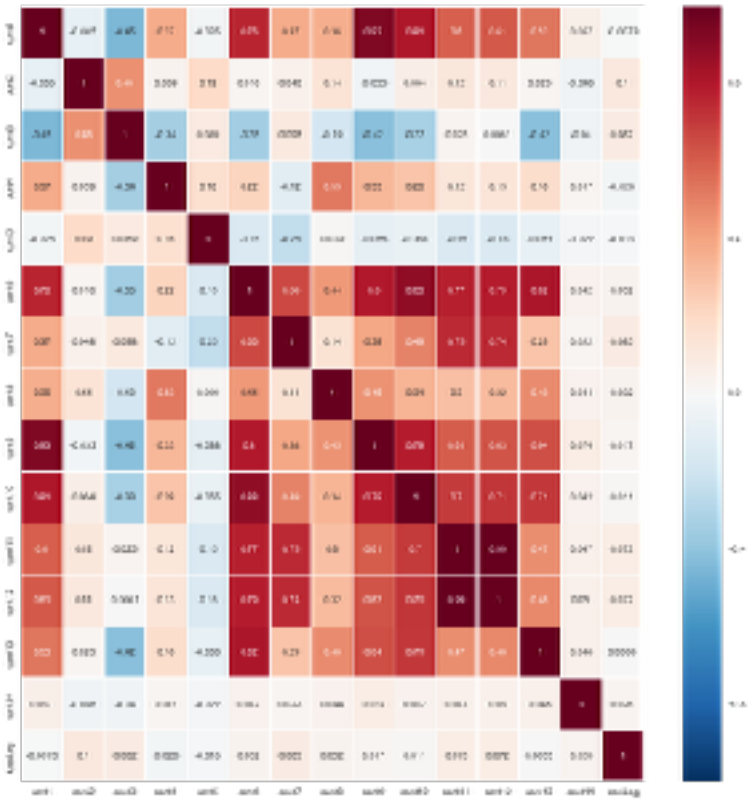



```
In [22]: print "skew:",train.loss.skew()
         train.loss.describe()

skew: 3.79495037754

Out[22]: count    188318.000000
         mean       3037.337686
         std        2904.086186
         min           0.670000
         25%        1204.460000
         50%        2115.570000
         75%        3864.045000
         max      121012.250000
         Name: loss, dtype: float64
```



EDA - Correlation between continuous features

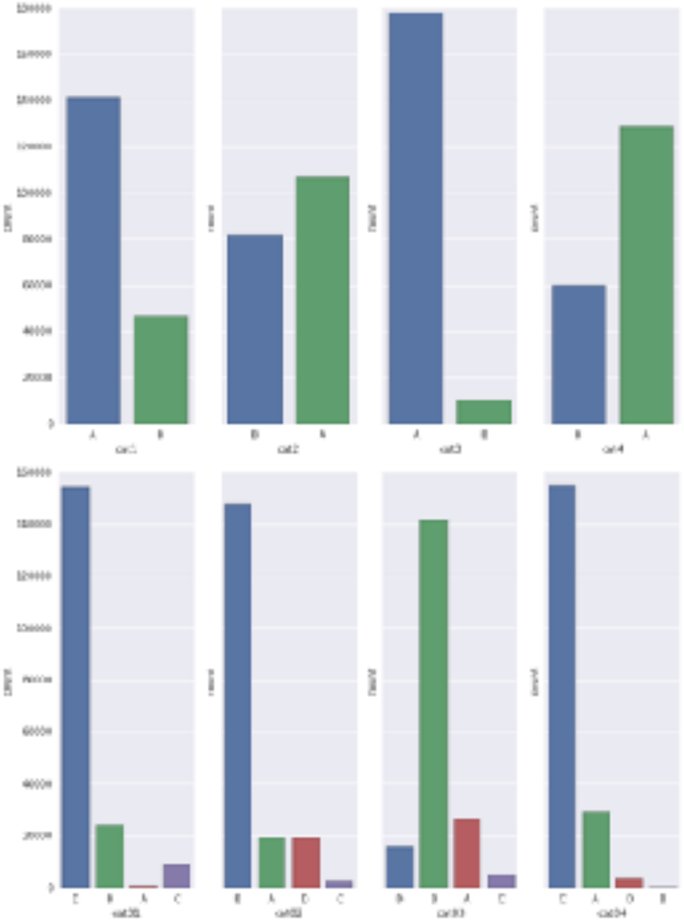
Continuous variables



Variables	Correlation
Cont 11 & Cont 12	0.994384
Cont 1 & Cont 9	0.929912
Cont 6 & Cont 10	0.883351
Cont 6 & Cont 13	0.815091
Cont 1 & Cont10	0.808551
Cont 9 & Cont 6	0.797544
Cont 9 & Cont 10	0.785697
Cont 6 & Cont12	0.785144

Threshold = 0.78

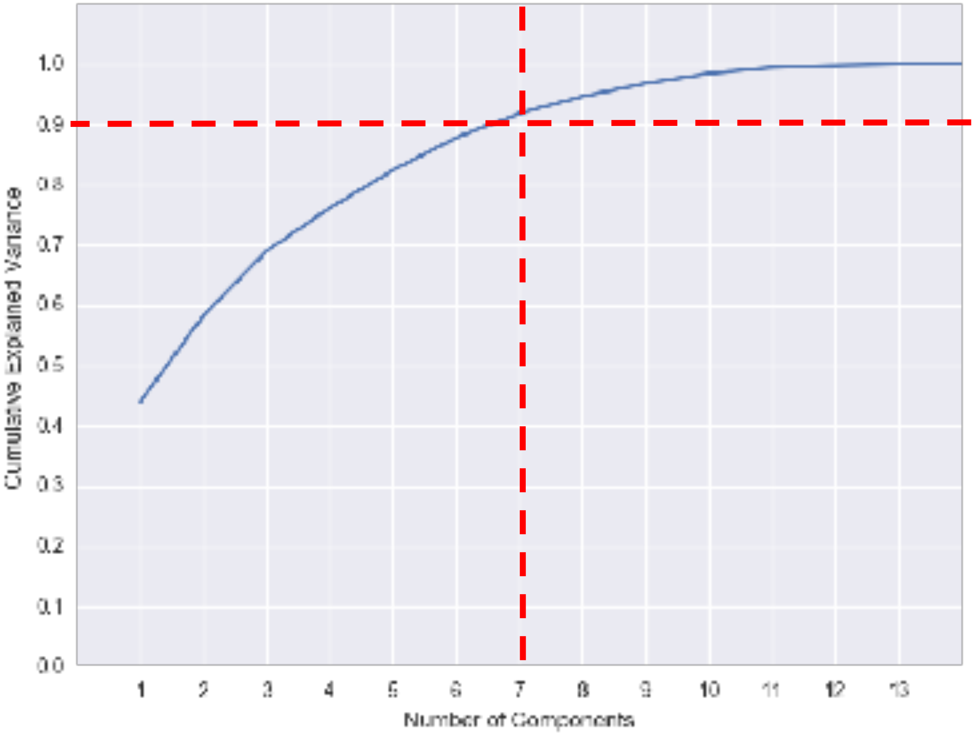
EDA - Categorical Features Frequency



Category 1 - 72	A, B
Category 73-76	A,B,C
Category 77- 88	A,B,C,D
cat112	51

Unsupervised: PCA/SVD - Dimensionality reduction

PCA for Continuous Features



- Continuous variables reduced from 14 to 7
- Binary categorical variables reduced from 72 to 26

Machine Learning for Prediction

Models Examined:

Regression

- Linear regression -- R^2 of 50%. Good for initial analysis
- Boosted trees -- XGBoost had best performance
- Neural network -- close second to XGBoost
- XGBBoost + NN => marginal improvement MAE 1126

Classification

- Logistic regression
- SVM

Tactics to Reduce Iteration Time:

Regularization

- Near-zero variance function
- Use p values from regression
- Reduced # levels (e.g., cat116)
- Penalized MAE by $300

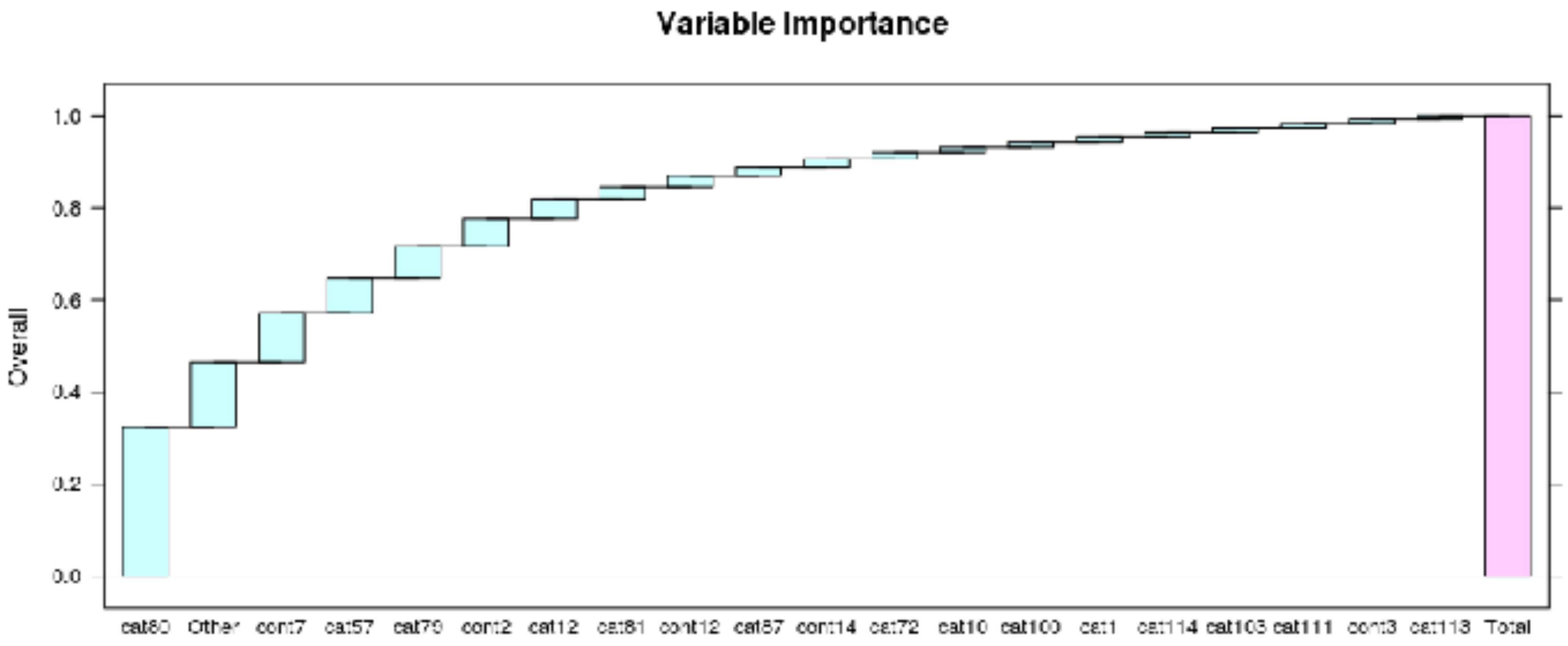
Sampling

- Random sampling
- Sampling cat80D versus B

Other

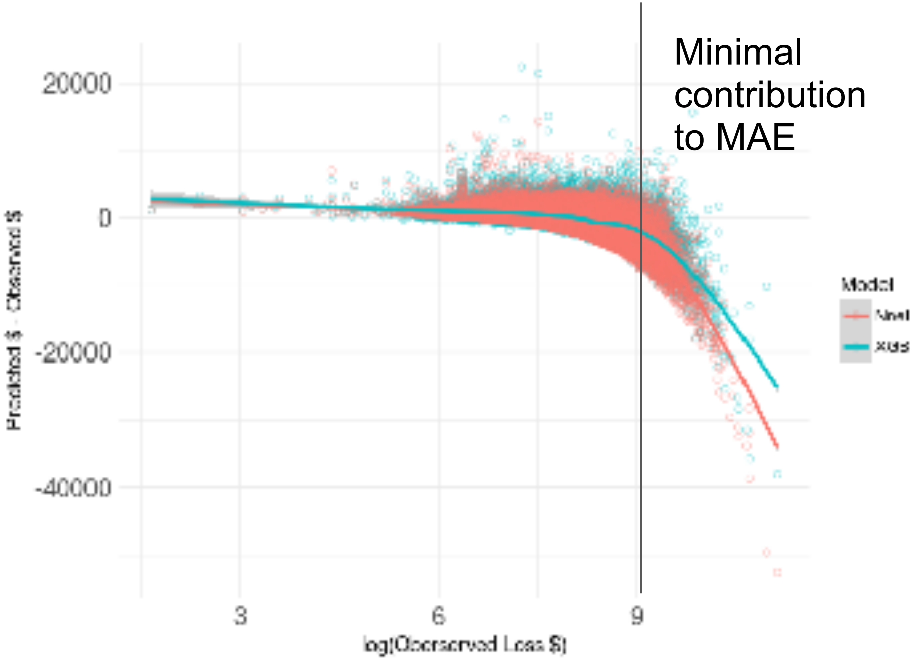
- Used AWS, but parallel processing not always a turn-key solution
- Reduce # folds in validation

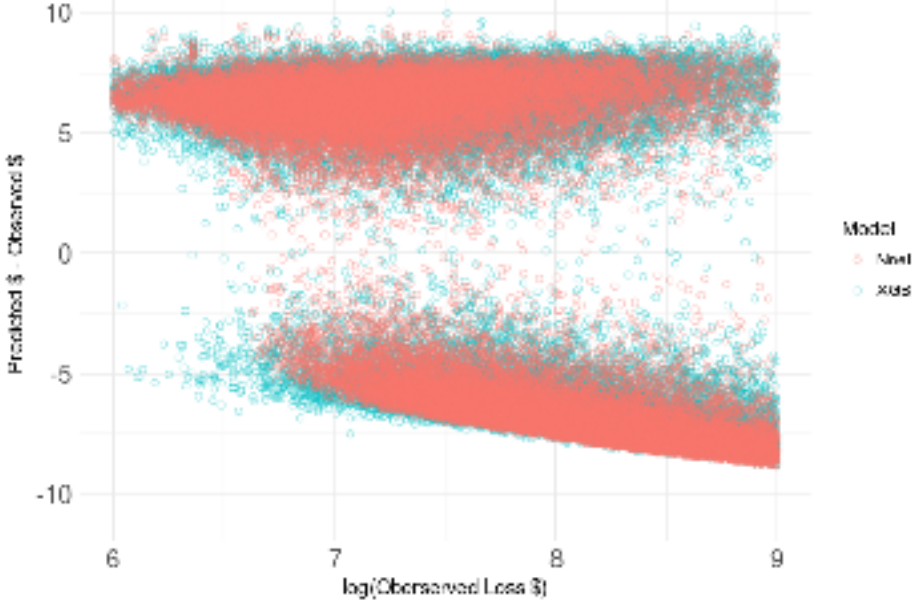
Machine Learning for Prediction -- Model Assessment



Cat80 largest single predictor

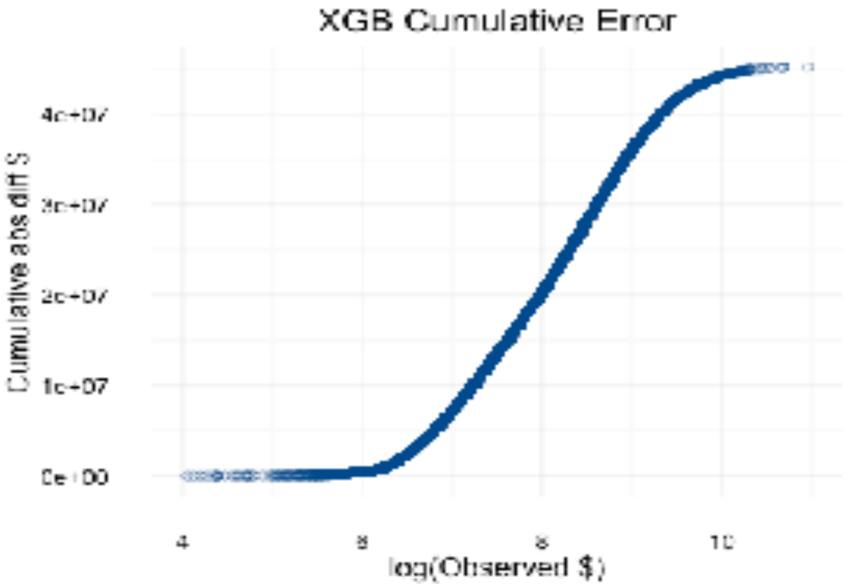
Machine Learning for Prediction -- Model Assessment

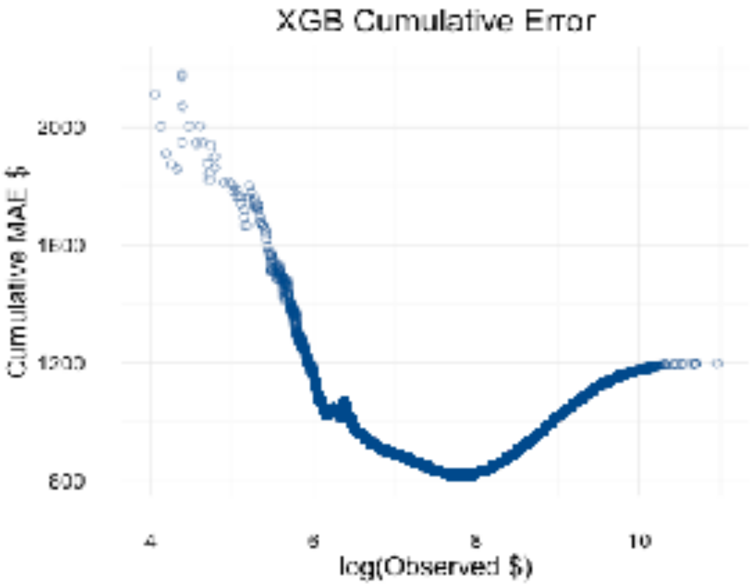




Underestimates increase with loss

Machine Learning for Prediction -- XGB Model Tuning





Most of error for claims between exp($6) and exp($9) ~($400-$8000), therefore no need to get distracted by tails

Model gets more accurate until exp($8) ~$3000, then performance degrades

What’s Salvageable?

When you’ve been devastated by a serious car accident, your focus is on the things that matter the most: family, friends, and other loved ones. Pushing paper with your insurance agent is the last place you want your time or mental energy spent.

Conclusion: claim size **can not be accurately predicted** based on provided features

Root Problem

- Doing paperwork for claim **protects insurer** against fraud
- May be able to **reduce paperwork** burden for claims if they don’t look odd
- Can features support a classification question?

New classifier: “smallClaim”

- 80% of customers account for 50% of claims by value -- all below $4500

Confusion Matrix

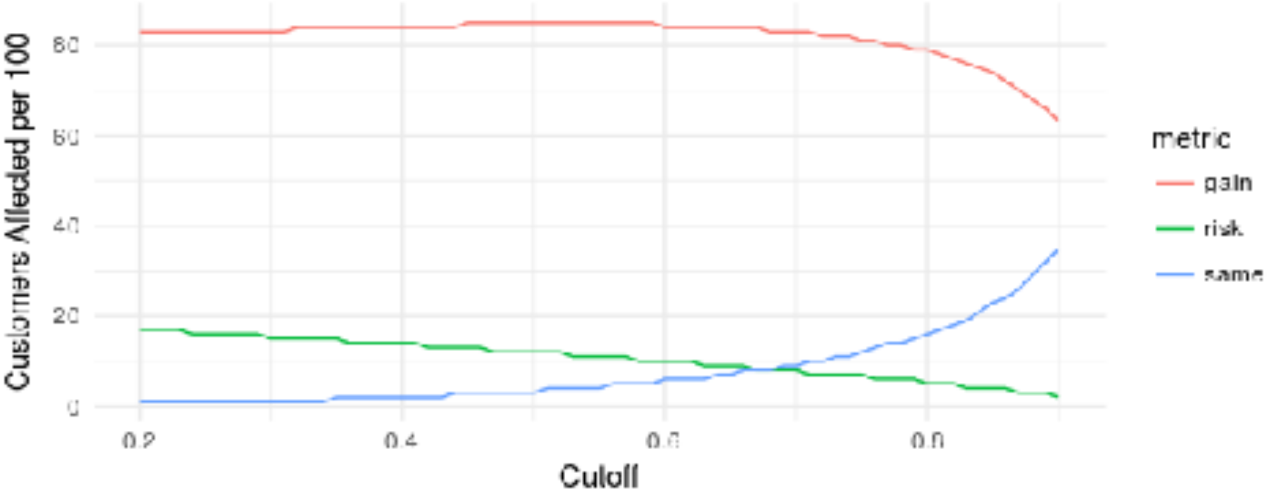
- What happens if we streamline the claims process for “normal-looking” claims?
- Confusion matrix can be recast as a business trade-off:

Truth/Predict	**Small**	**Large**
Small Claim (80%)	Gain	Same
Large (20%)	Risk	Gain

Need a cutoff that balances the gains of customers dealing with less paperwork with fraud risk

Fishing for Questionable Claims

Customer Impact of Varying Logistic Regression Cutoffs



Next Steps: quantify dollar risk of misclassification and dollar benefit to customer of reduced paperwork