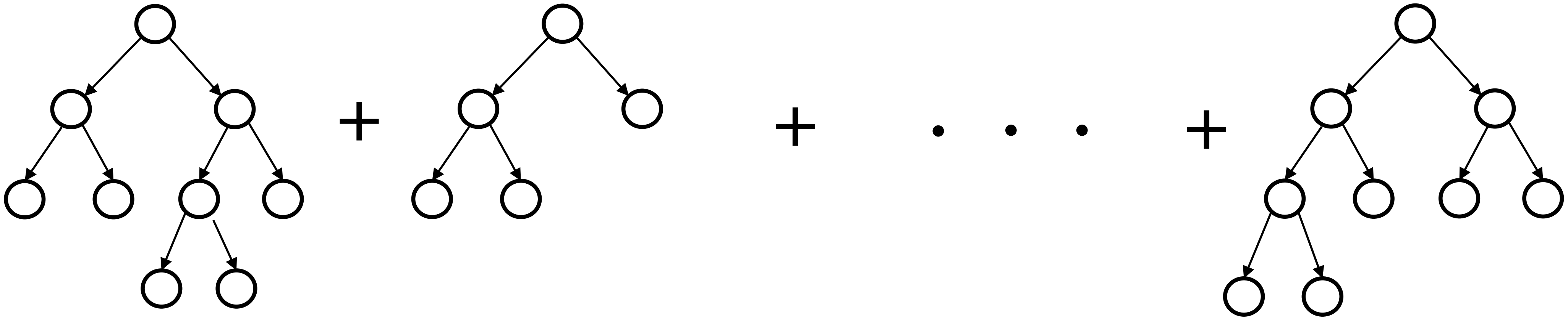
Model Automation

Gradient Boosting Machine (GBM)



Iteratively:

1. Train a decision tree
2. Add learning rate * decision tree to current model
3. Reweight records by current model residual
4. Repeat until specified max number of trees is reached

GBM Hyperparameters

1. Number of Trees
2. Learning Rate: constant weight placed on each tree in the model
3. Bag Fraction: fraction of rows to randomly select to train each tree
4. Interaction Depth: number of splits allowed in each tree
5. Minimum Observations in Terminal Nodes

Underfitting and Overfitting

Source: https://hackernoon.com/memorizing-is-not-learning-6-tricks-to-prevent-overfitting-in-machine-learning-820b091dc42

Cross-Validation and Hyperparameter Tuning

Train	
	Predict
Train	
Train	
Train	
Train	
Train	
Train	
Train	
Train	

For each fold, train model on other n-1 folds to make predictions on that fold

<u>Tuning Hyperparameters:</u>

Loop over reasonable values of hyperparameters

Optimize favorite error metric (e.g. RMSE, MAE, or Lift) on training set using cross validation

Evaluate model on holdout set

learning_rates <- c(0.01, 0.05, 0.1)
interaction_depths <- c(1,2,3)
rmses <- c()

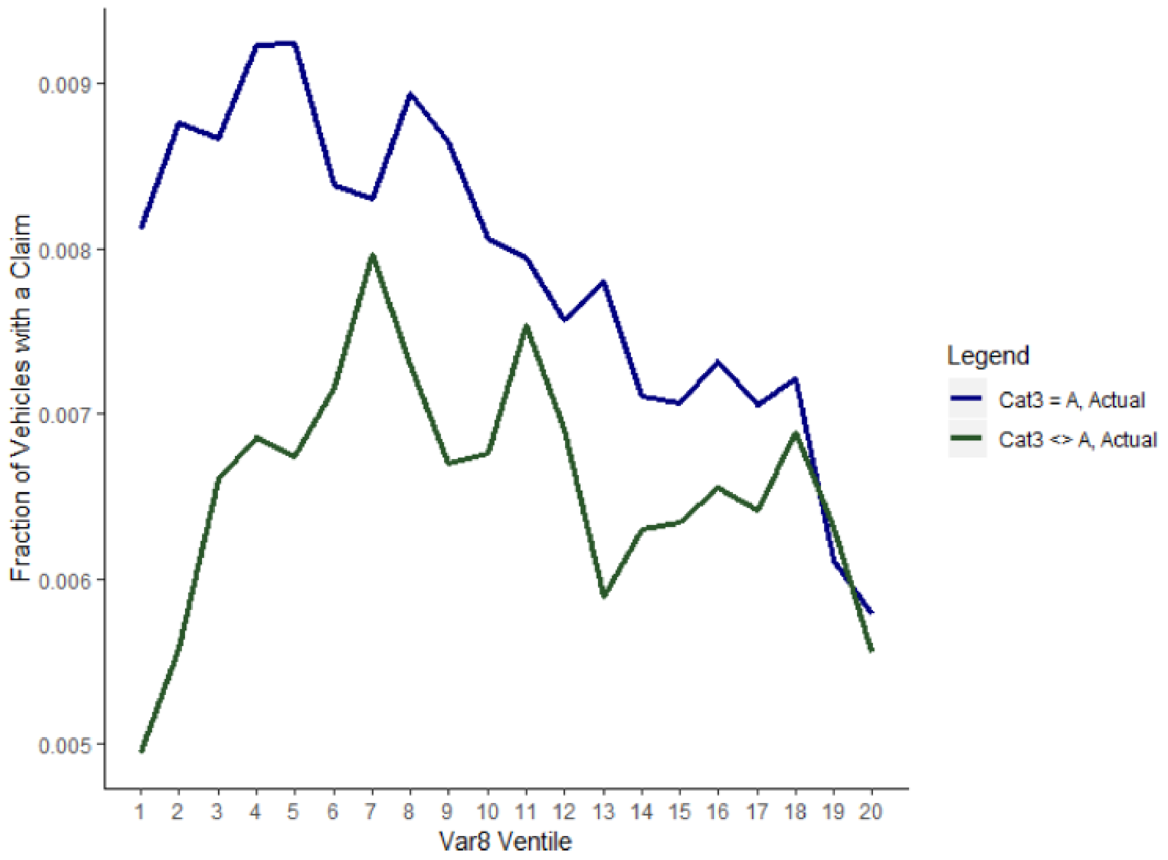
for (learning_rate in learning rates){
for (interaction_depth in interaction_depths){

- train the model
- on each subset of n-1 folds
- using learning_rate and interaction_depth for those hyperparameters
- append rmse on cross-validated holdout set to rmses

}}

Using GBM for Enhanced Pricing Accuracy

An Interaction



Data from 2011 Allstate Kaggle Competition

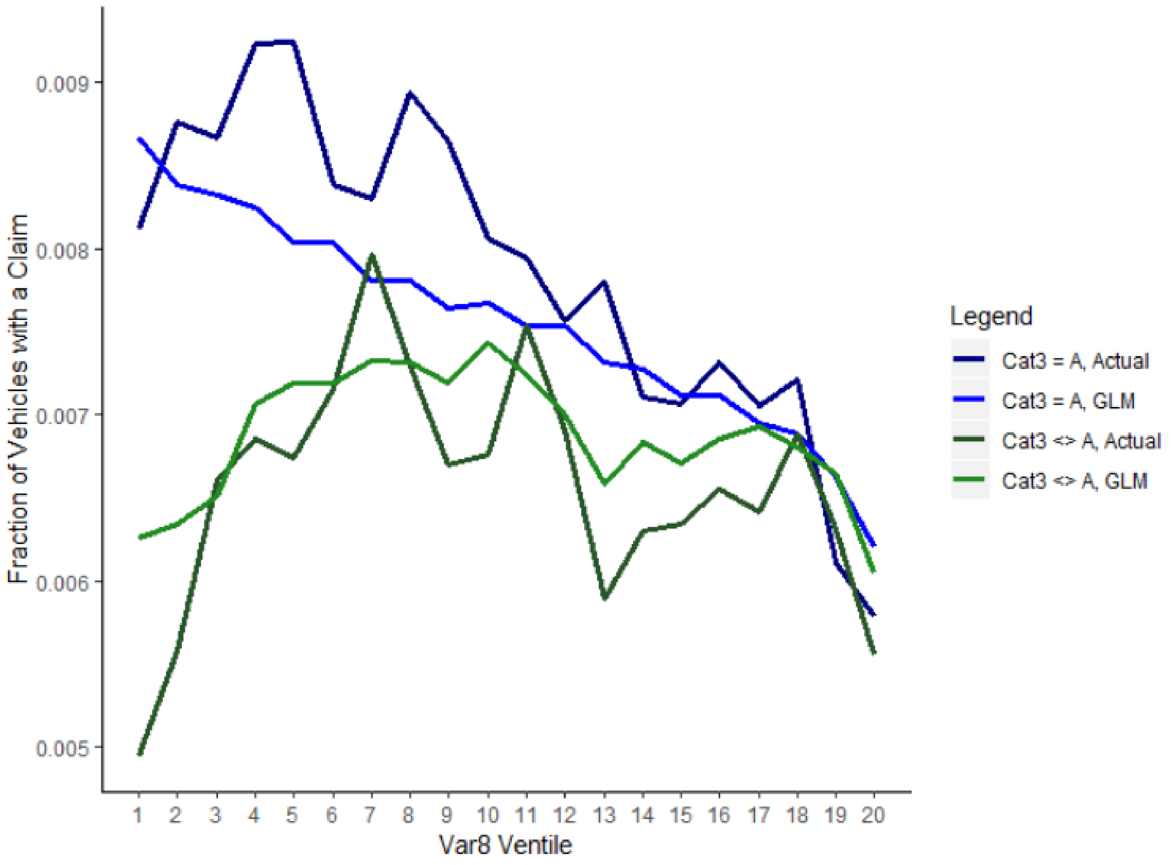
Target variable: bodily injury liability claim indicator

Predictor variables: unnamed characteristics of insured customer's vehicle

Interaction between continuous variable #8 and categorical variable #3

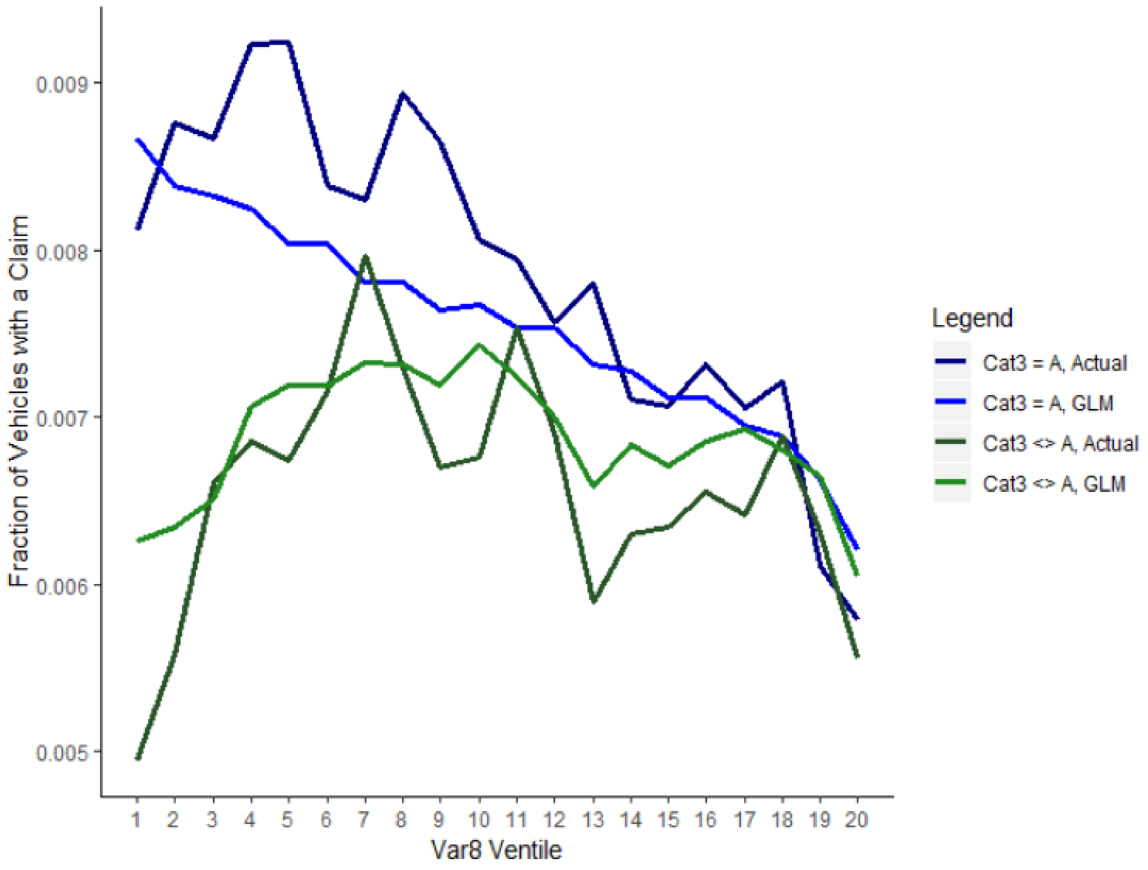
Data source: https://www.kaggle.com/c/ClaimPredictionChallenge

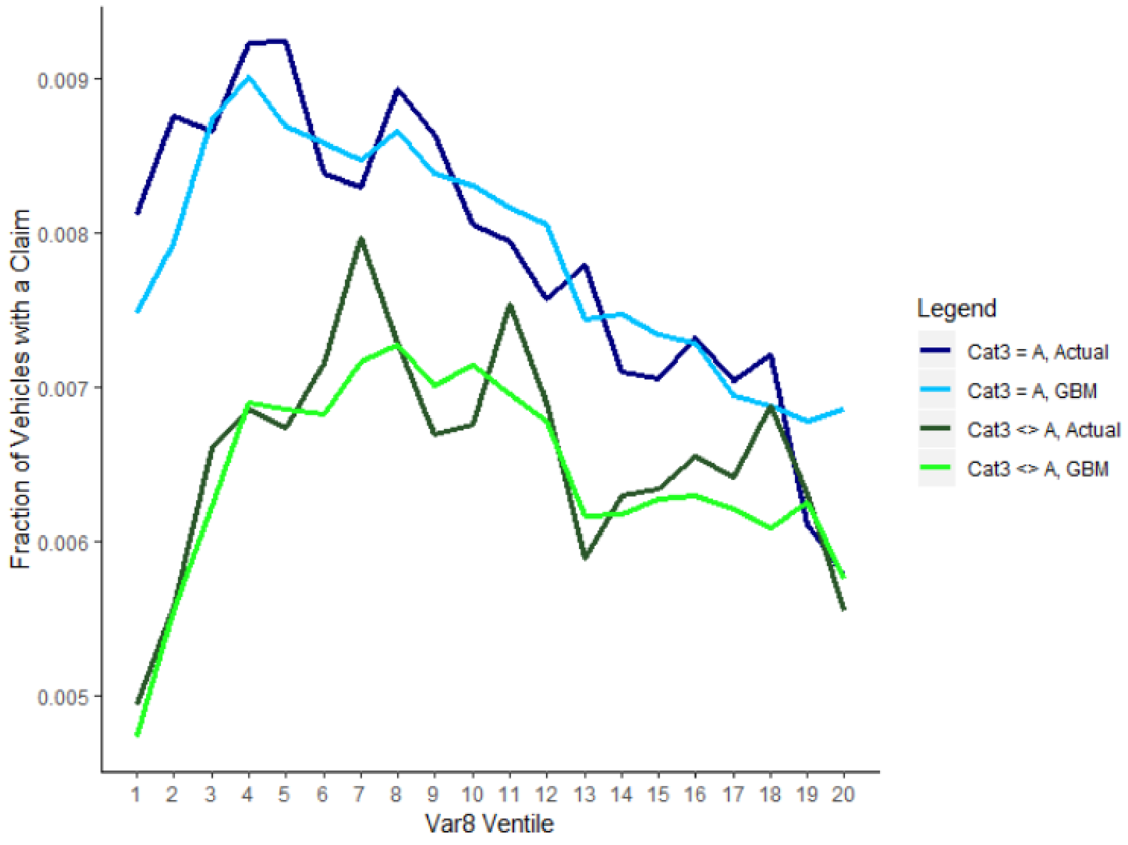
GLM – Does not fit very well



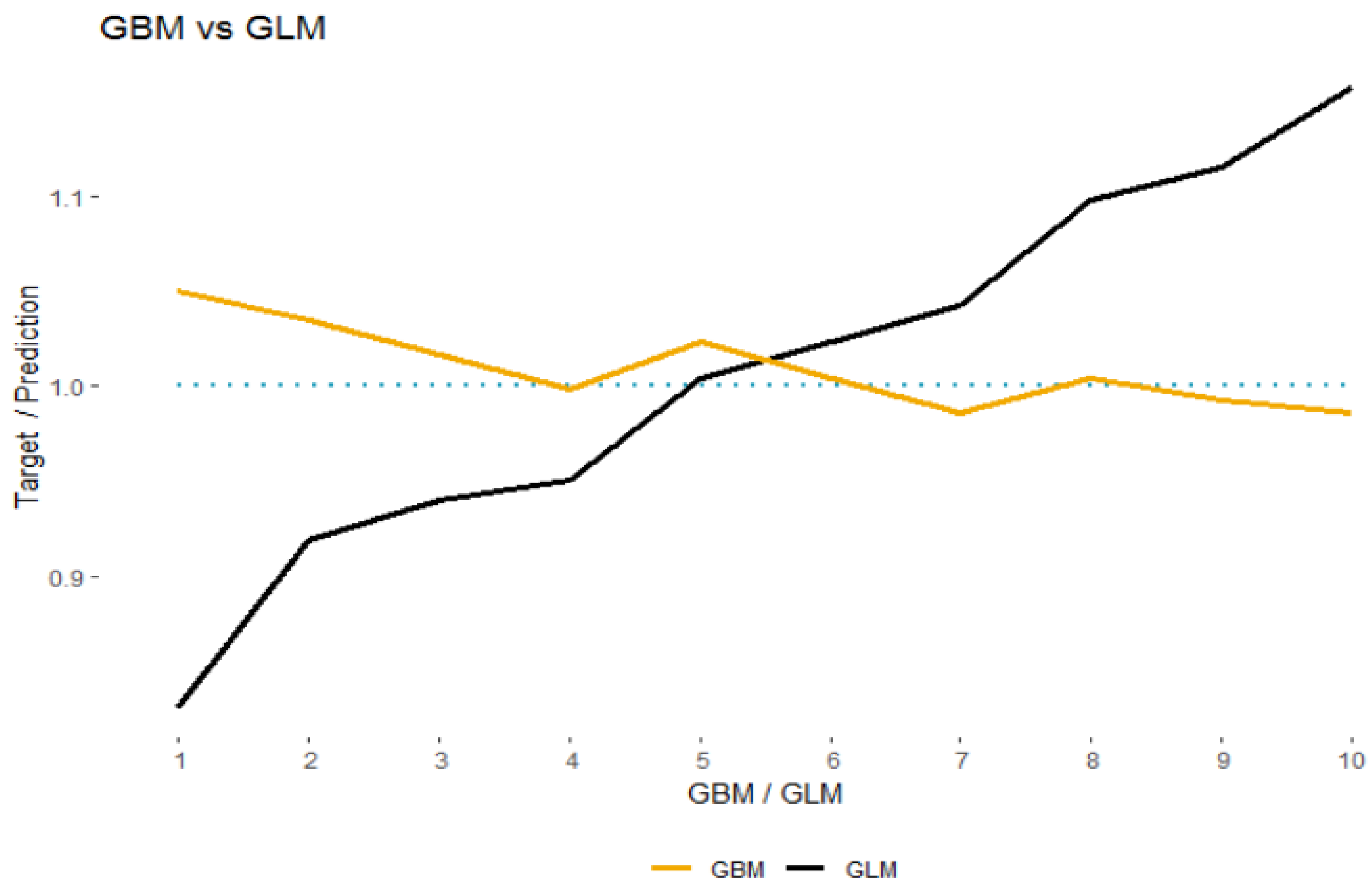
Underpredicting when Cat3 = A
Overpredicting when Cat3 <> A
Missing significant increase in Ventiles 1-4 for when Cat3 <> A

GBM – Fits better





Generate Lift



Adverse Selection

Evolution of Auto Insurance Rating Sophistication

1960's
Accidents and Violations

1990's
Prior Coverage Info
GLM's
Vehicle Characteristics

2000's
Credit Data
Vehicle History

2010's
Telematics
ADAS
Machine Learning

New data sources and modeling methods evolve with similar affect

Adverse Selection

GLM Quartile	GLM Premium	Insurer's Expected Loss	Insurer's Expected Combined Ratio	GBM > GLM	Accurate Expected Loss	GBM Expected Combined Ratio
1	$ 98	$ 49	90%	FALSE	$ 48	89.0%
1	$ 98	$ 49	90%	TRUE	$ 51	91.6%
2	$ 115	$ 58	90%	FALSE	$ 54	87.0%
2	$ 115	$ 58	90%	TRUE	$ 60	92.4%
3	$ 153	$ 76	90%	FALSE	$ 72	86.8%
3	$ 153	$ 76	90%	TRUE	$ 81	92.8%
4	$ 217	$ 109	90%	FALSE	$ 95	83.6%
4	$ 217	$ 109	90%	TRUE	$ 122	96.1%

Motivation: https://www.casact.org/newsletter/index.cfm?fa=viewart&id=5584