Allstate Claims Severity Kaggle Competition

Dina, Josh & Nick

Presented on or after November 28, 2016

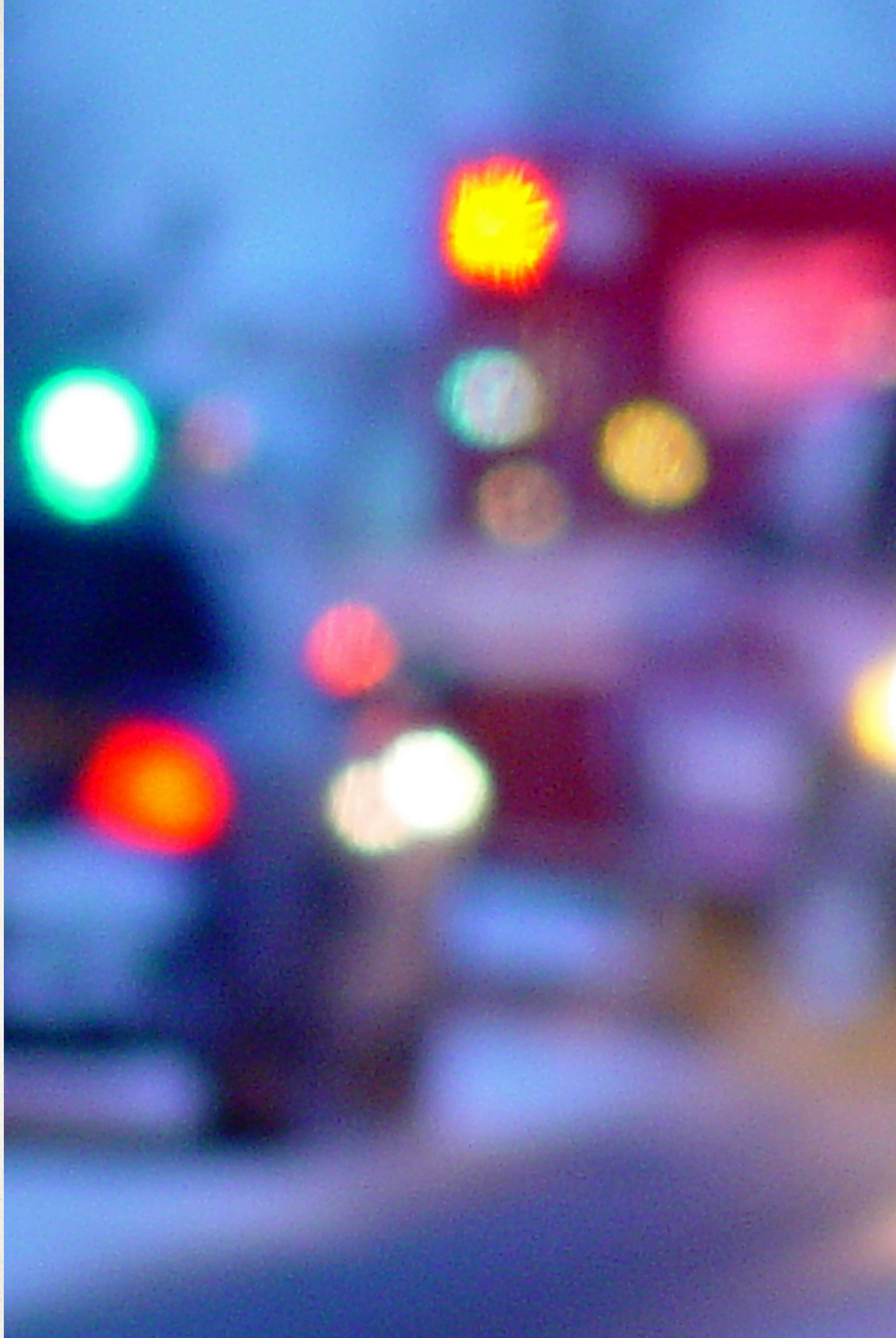
The Goals

Allstate's Goal:

To predict the cost of car accident claims

Our Goal:

To understand how to build appropriate models using different machine learning algorithms



Exploratory Data Analysis

The Data

- ID column represents anonymized customers
- Loss is the response variable to predict
- 116 categorical variables and 14 continuous variables
- The test data has 125,546 rows and the training data has 188,318 rows

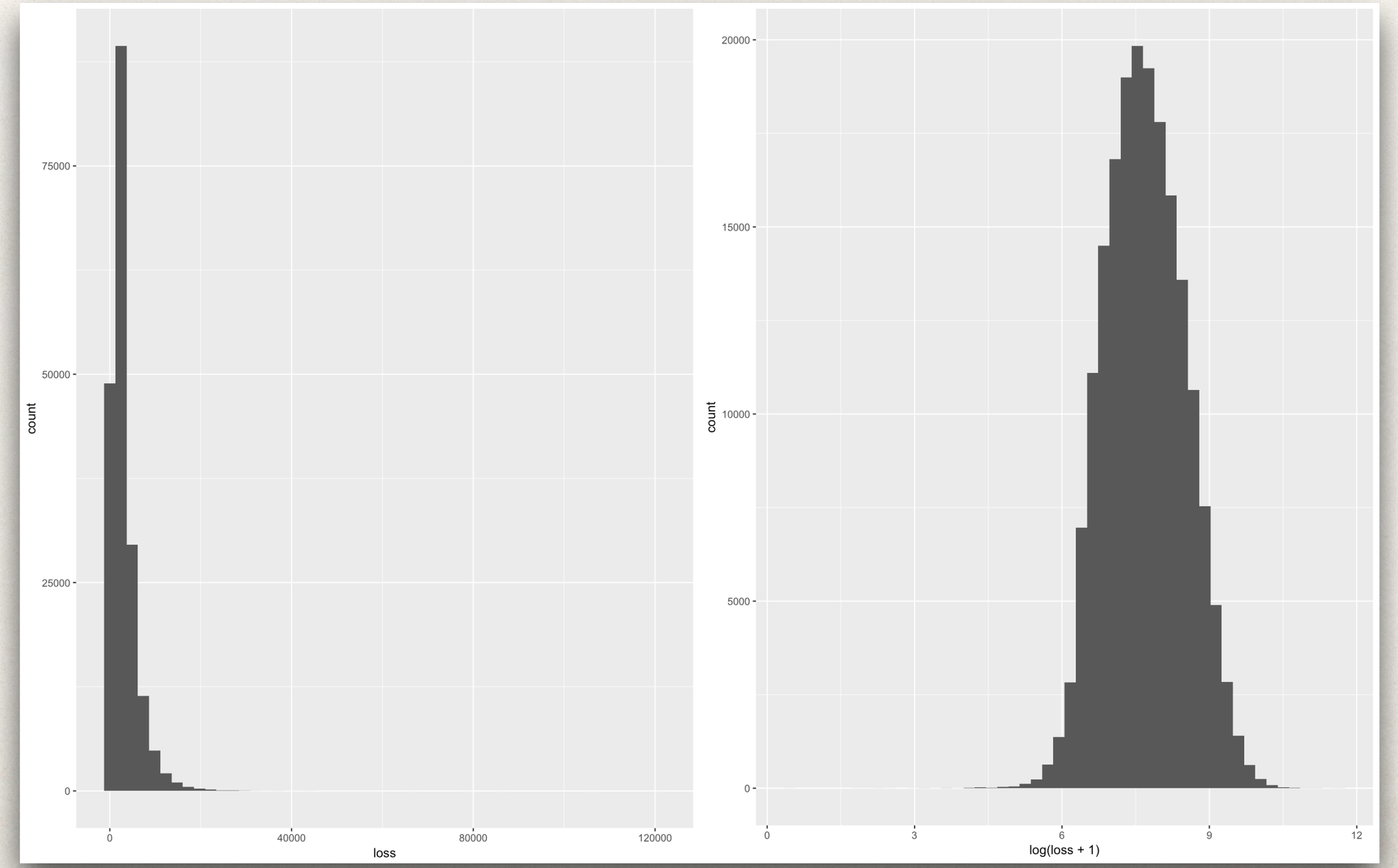
Lack of Domain Knowledge

- Categorical data was only in the form A,B,C,D,E…
- Continuous data ranged from 0 to 1

The Search For Redundant Variables

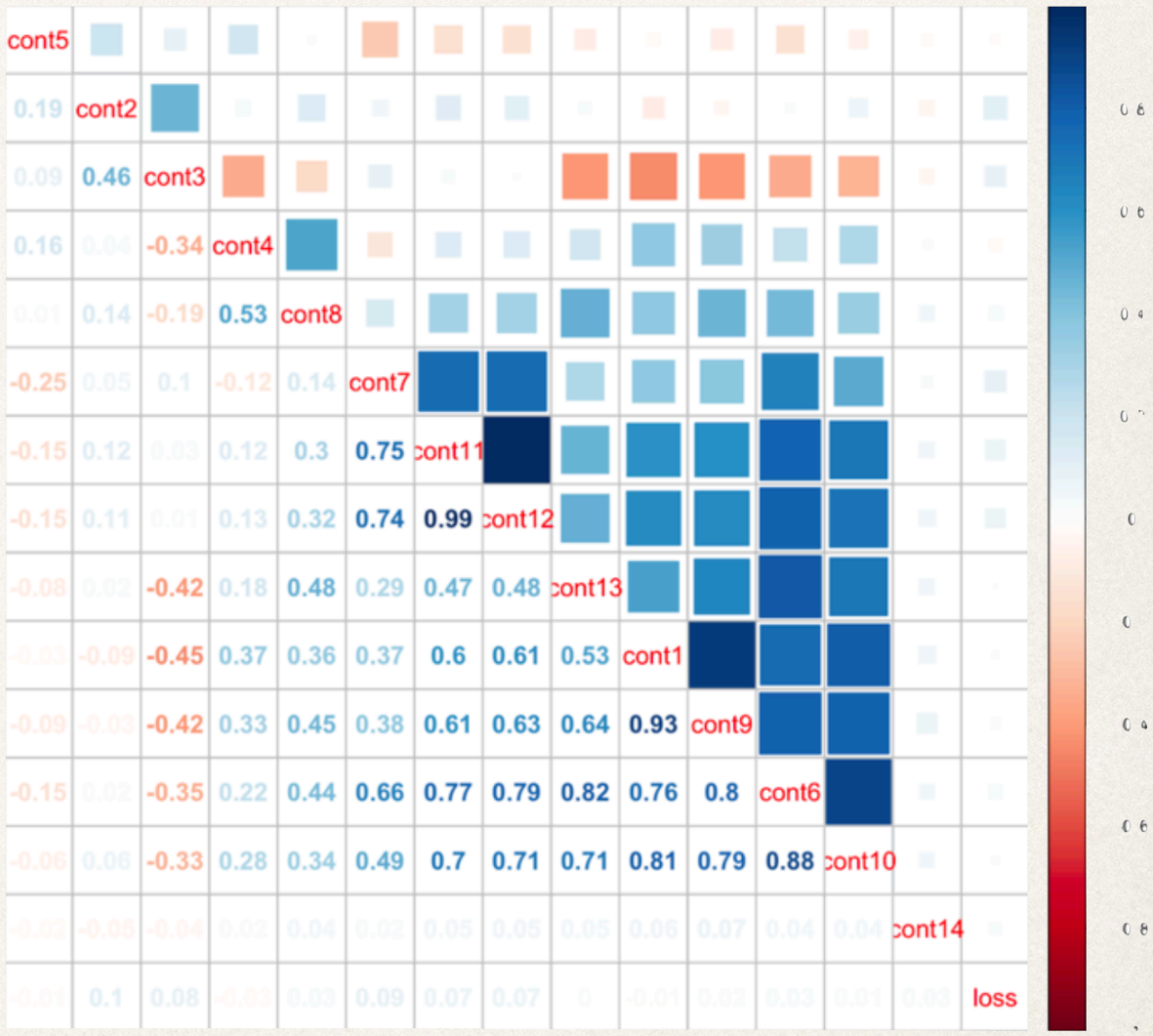
1. Created dummy variables
2. Tested near zero variance on the data
3. Found that 54 variables had all factors eliminated

Log Transform



Correlation

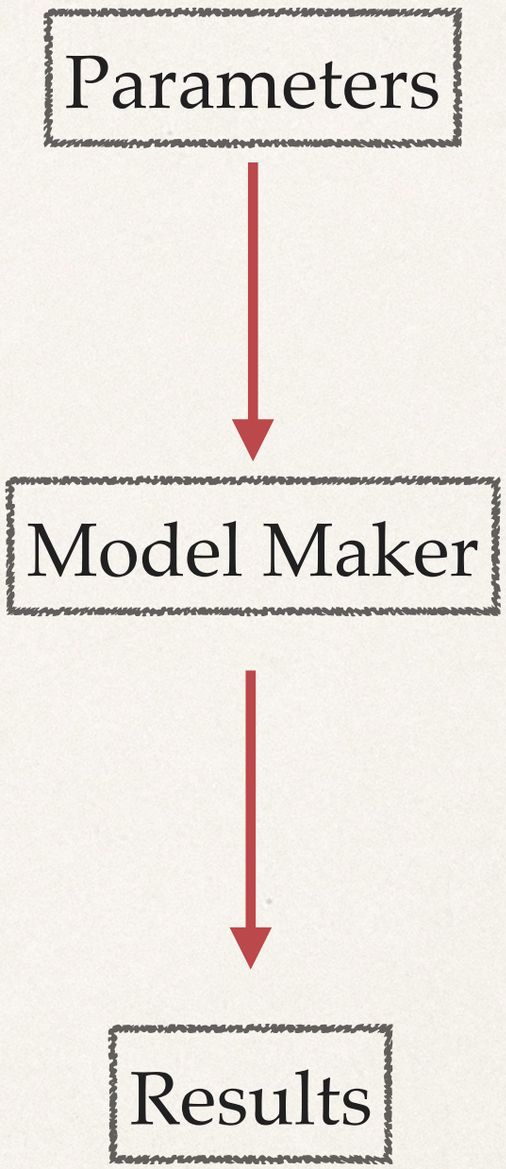
- Loss is not linearly correlated to any continuous variable
- Correlation won't affect the models we used



Loss is shown as log(Loss)

Streamlining Teamwork With Modularity

Make model testing simple to run and quick to iterate on



Notable Model Parameters

- Model method
- Grid of parameters
- Number of cross-validation folds

A function then…

- Runs the created model
- Returns metrics including:
  - estimated MAE
  - best model parameters
- Creates a Kaggle submission file

The “Model Maker”

Pre-Processing

- Removed bad categorical data found during EDA
- Created dummy variables of factors
- Near zero variance applied
- Scaled and centered continuous variables

Processing

- Perform cross validation to find the best parameters
- Train final model on all of the data with best tuning parameters

MAE Used in Cross Validation

- MAE is the Mean Absolute Error
- The Kaggle score for this competition is determined using MAE

Finding a Model Method

Methods Tried

Methods	About
KNN	Simple, easy to understand baseline
Gradient Boosting Machines	Robust, plug-and-play
XGBoost	Parallelized gradient boosting
Neural Networks	Extremely flexible, regression assumptions not required, robust to outliers

Parameter Tuning

Model Method	Approach
KNN	K = sqrt(nrows of training)
Gradient Boosting Machines	Forum suggestions
XGBoost	Forum suggestions & cross validation
Neural Networks	Experimentation

Neural Network Challenges

- Tried multiple packages
  - nnet, neuralnet
    - Chose nnet for simpler topology
- Finding initial range of parameters
- Lack of documentation on parameters
- Fine tuned parameters
  - Little return on investment

Neural Network Model

$$w_i^{new} = w_i - \eta \frac{\partial E}{\partial w_i} - \lambda \eta w_i$$

λ values

nodes	0.0	0.01	0.1	0.5
20	1216	1206	1199	1203
25	1221	1210	1197	1204
30	1230	1224	1212	1223

XGBoost Challenges

- Requires tuning many parameters
- Grid search too slow
- Used forum suggestions as starting guides

Results

Methods	Best Kaggle Score
KNN	1533.743 [Not Submitted]
Gradient Boosting Machines	1163.47311
XGBoost	1148.65697
Neural Networks	1206.69697

Future Improvements

- Ensemble methods
- Expand scope outside of Caret like H2O or Keras
- Further optimizing parameters

Insights

Insights on the Model Maker

- Model maker expedited teamwork and efficiency
- Limited our work to caret and R
- Debugging the model maker slowed progress

Comments on Servers

- High performance computing cluster allowed us to try multiple models
- Special node on the server gave ability to use massive amount of RAM for KNN
- Fully utilized multi-threading in R

Insights

- It's easy to make a model, but difficult to make a competitive model
- It takes many models to discover what is a good and bad model
- Parameter tuning can be more of an art than a science
- Don't do parameter tuning on a subset of the data
- Neural networks hard to optimize

Acknowledgements

Brandeis University HPC Cluster

Thank You