Allstate Kaggle Machine Learning Project

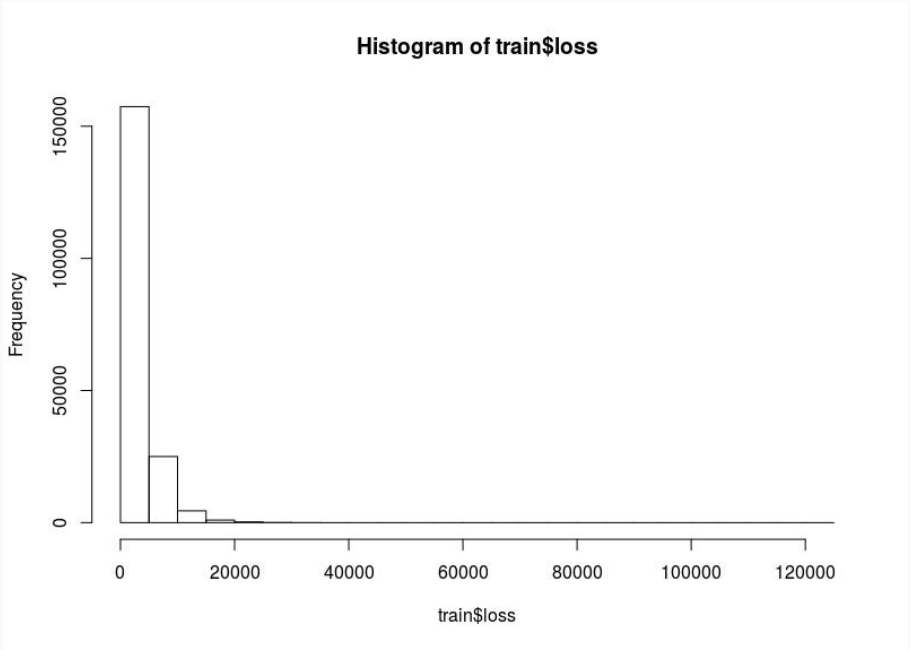
Team JoYFre

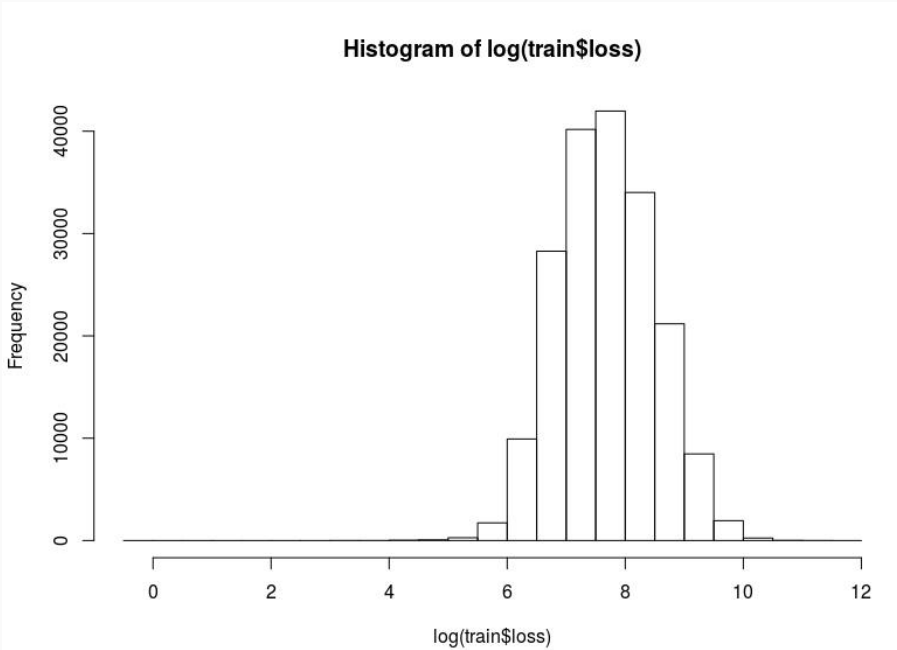
Motivation

Allstate is currently developing automated methods of predicting the cost, and hence severity, of claims. In this recruitment challenge, Kagglers are invited to show off their creativity and flex their technical chops by creating an algorithm which accurately predicts claims severity. Aspiring competitors will demonstrate insight into better ways to predict claims severity for the chance to be part of Allstate's efforts to ensure a worry-free customer experience.

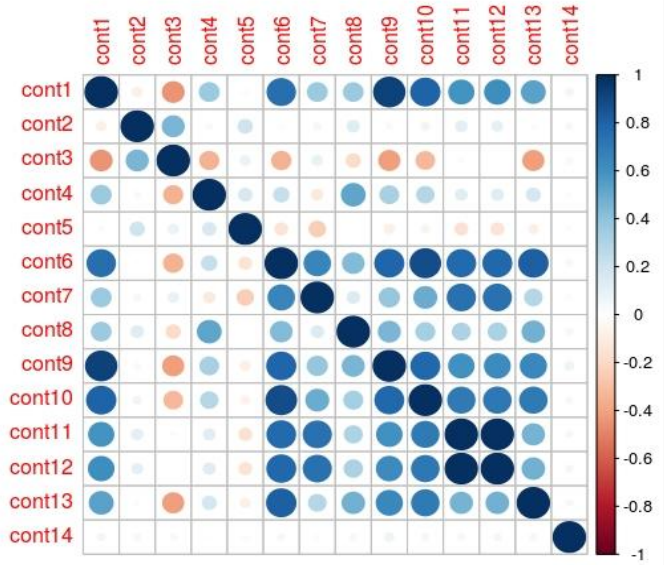
EDA Loss Prediction

Initial transformation to unskew data

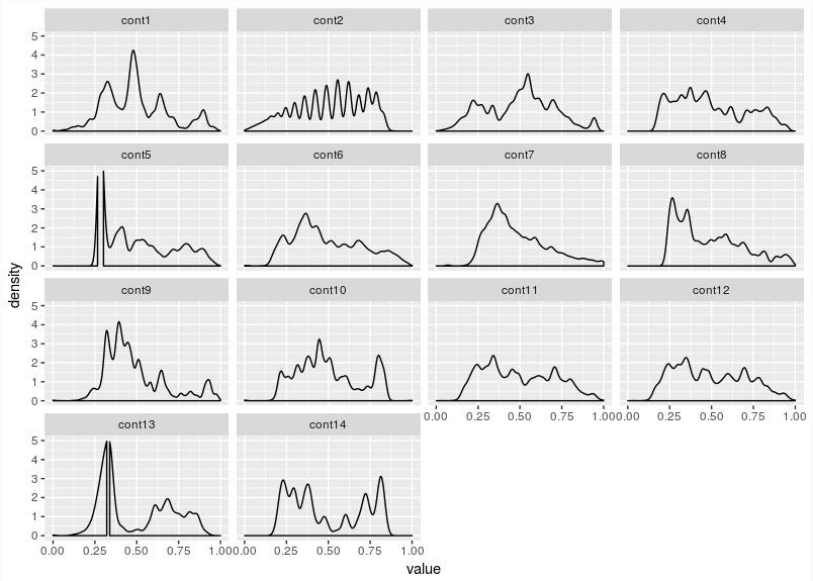




Correlation Variables EDA

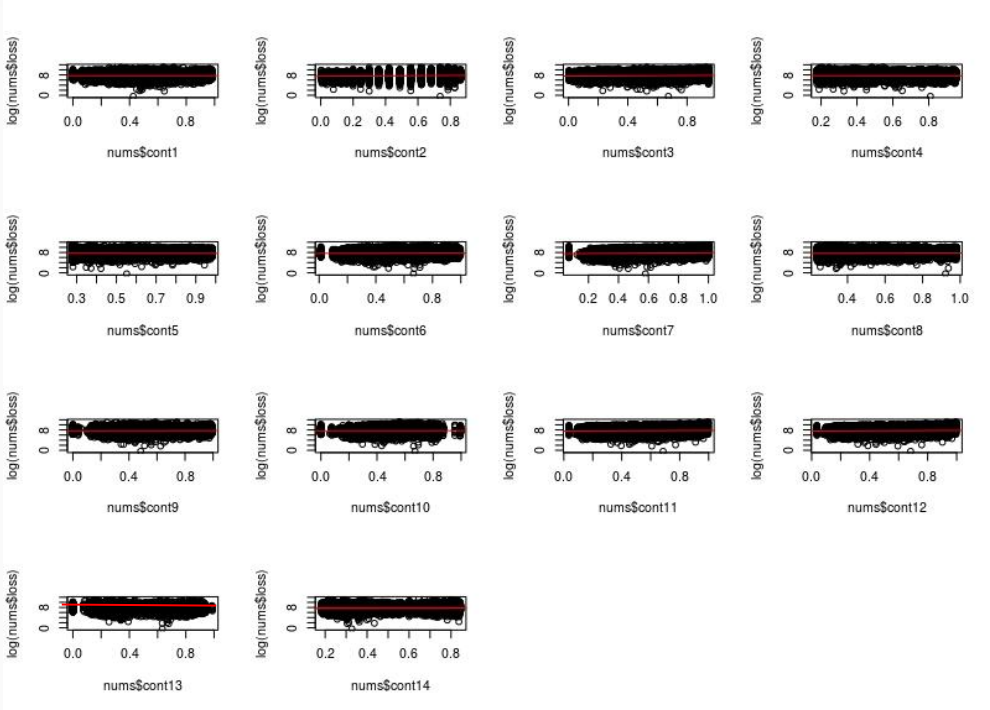


	Variables	VIF
1	cont1	12.581557
2	cont2	1.518707
3	cont3	2.565368
4	cont4	1.941685
5	cont5	1.205744
6	cont6	22.824854
7	cont7	4.878287
8	cont8	2.011909
9	cont9	10.987792
10	cont10	6.779808
11	cont11	93.680761
12	cont12	97.561770
13	cont13	7.160353
14	cont14	1.015809



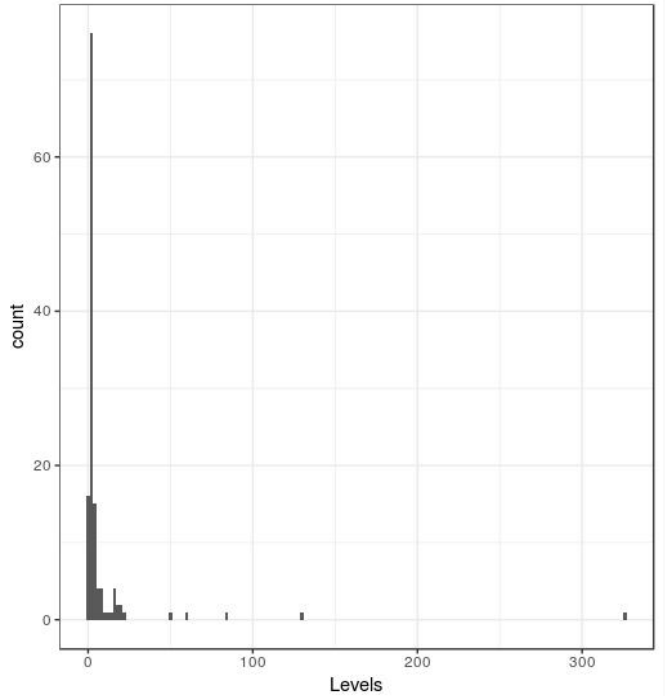
Continuous Variables

little to no slope in linear models predicting loss from individual variables indicates the models account for very little of the data's variability (R^2)



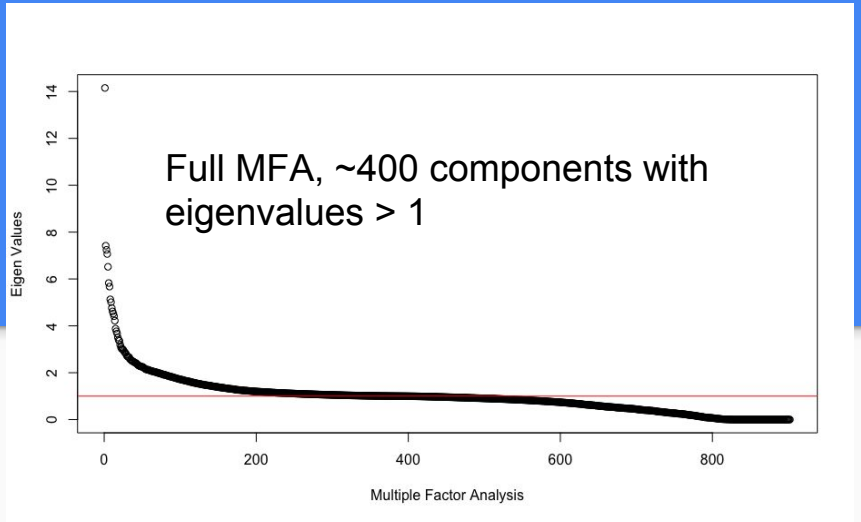
Categorical Variable EDA

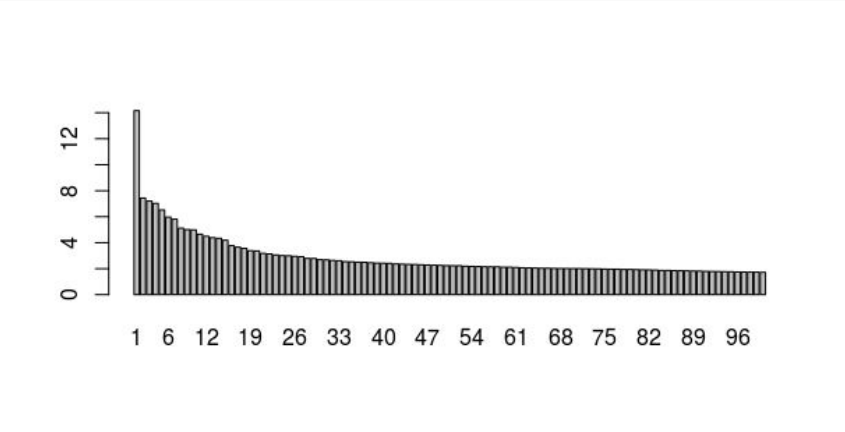
- 116 categorical variables
  - 72 have 2 levels
  - Cat116 has 326 levels
  - Cat112 has 51 levels and is speculated to be States +D.C.
    - Seems to be irrelevant as variance in loss by state(cat112) is very similar
  - 27 of the 2 level variables have <1% in one of the levels



MCA (Fred)

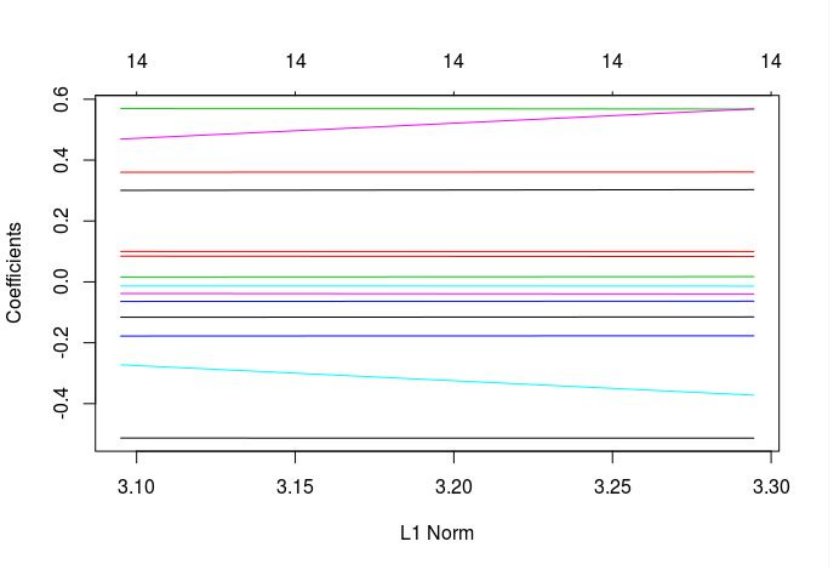
- Ran MCA, no eigenvalues dropped beneath 1 even with 300 PCA



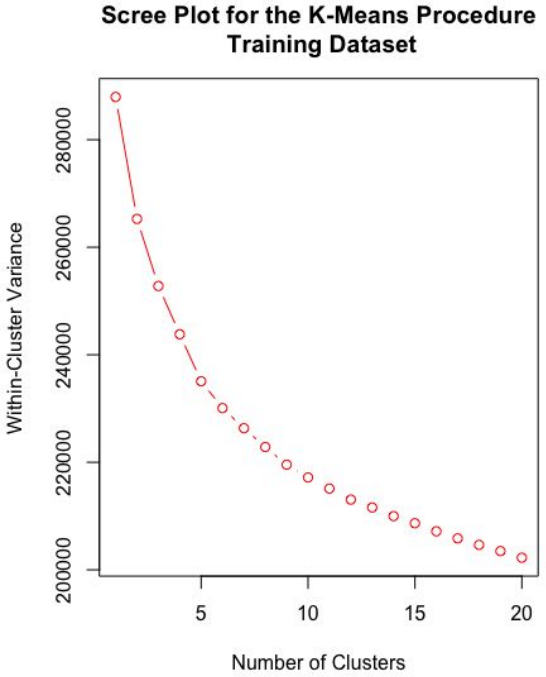


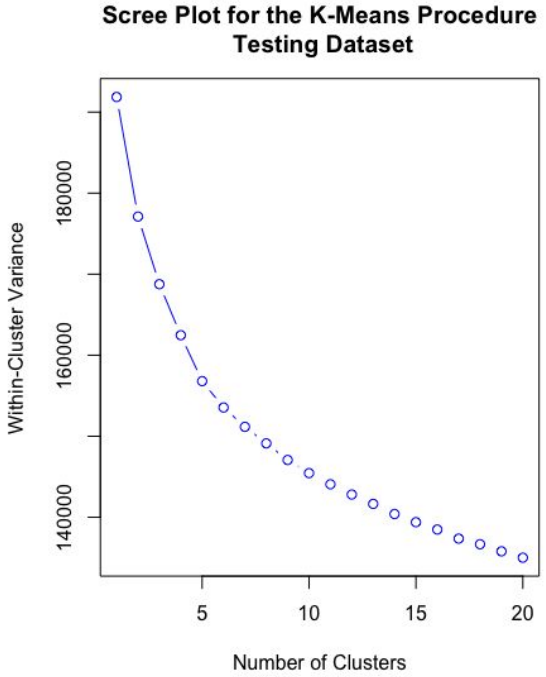
Ridge and Lasso (Fred)

- Ran Lasso and Ridge (pictured) with no convergence, only using the continuous variables
- Next Steps: Boosting

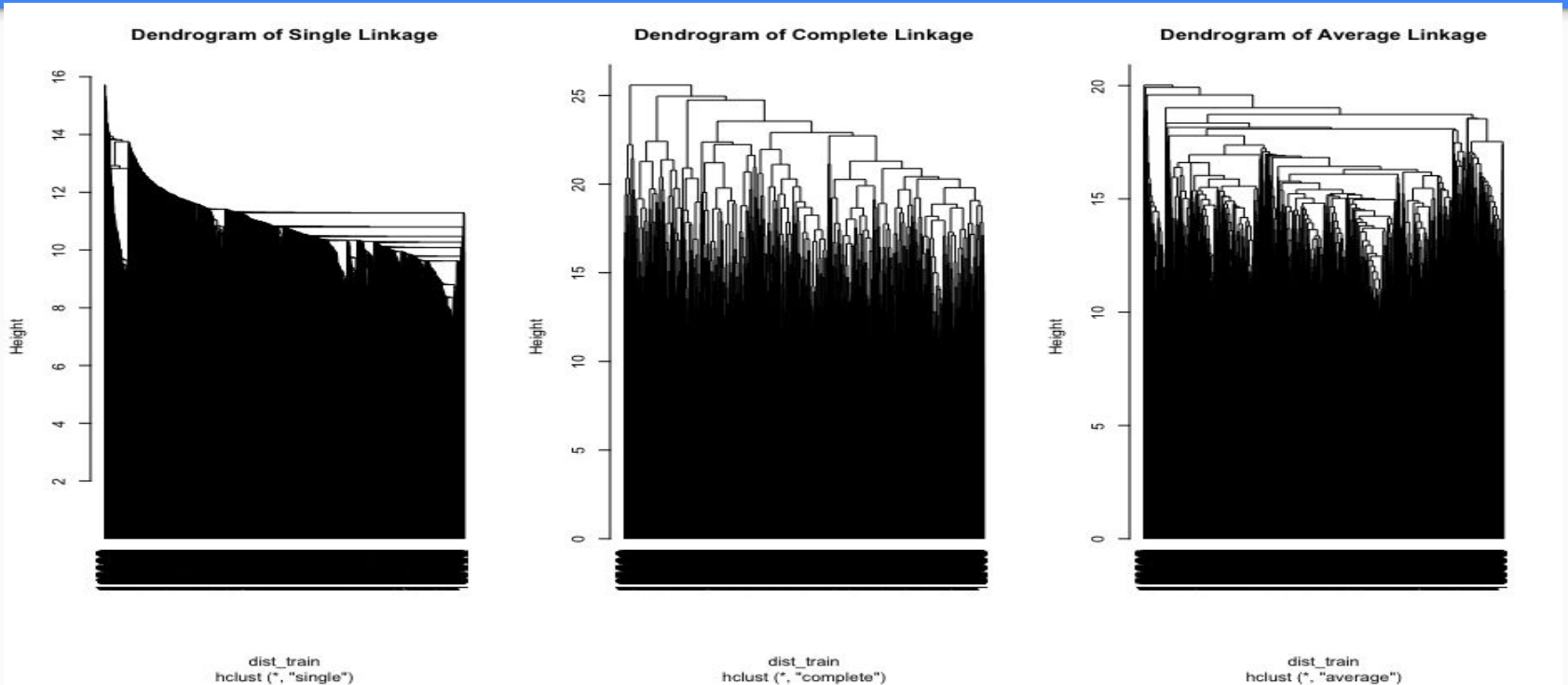


Clustering Analysis - Kmeans Clustering





Clustering Analysis - Hierarchical Clustering



Cloud Computing (AWS) (Fred)

- Had to resort to cloud computing for more calculating power
- Would have been better to use GPU for more parallelized processes
- >24hours
- Ran XGBoost and Keras/Theano (Neural Network)

AWS setup for XGBoost (Fred)

Xgboost Model

Initial xgboost run with all variables: MAE 1116.39031

Optimizing parameters with xgboost and mlr:

```
ps = makeParamSet(
  makeNumericParam("obj_par", lower = 1.5, upper = 2),
  makeNumericParam("eta", lower = 0.1, upper = 1)
)
rdesc = makeResampleDesc("CV", iters = 3L)
ctrl =  makeTuneControlRandom(maxit = 5)
```

Xgboost Model

Feature selection with xgboost and mlr:

ctrl = makeFeatSelControlRandom(maxit = 100L, prob = 0.5) # random search feature selection

ctrl = makeFeatSelControlGA(mu = 10L, lambda = 5, crossover.rate = 0.5, mutation.rate = 0.05, maxit = 100L) # feature selection with gene algorithm

1st round: random feature selection -> 113 features; genetic algorithm -> 84 features; 47 features in common

2nd round: 36 features in common

Final model: 83 features MAE: 1115.65561

AWS setup for Neural Net (Fred)

Neural Networks Model

4 hidden layers (400 nodes, 200 nodes, 50 nodes, 1 node)

Running time: 32 CPU ? Hours

Using all variables

MAE: 1111.90084

Ensemble Model

Ensemble model: xgboost + neural networks

MAE: 1105.91477

Lessons learned

Reading on forum is most helpful in EDA and getting us started with models.

Optimizing parameters cost lots of time but helped very little.

Some of the machine learning methods (clustering and kNN) needs a lot of computational resources.

Need more models for better ensembling and stacking models.