Kaggle Competition:

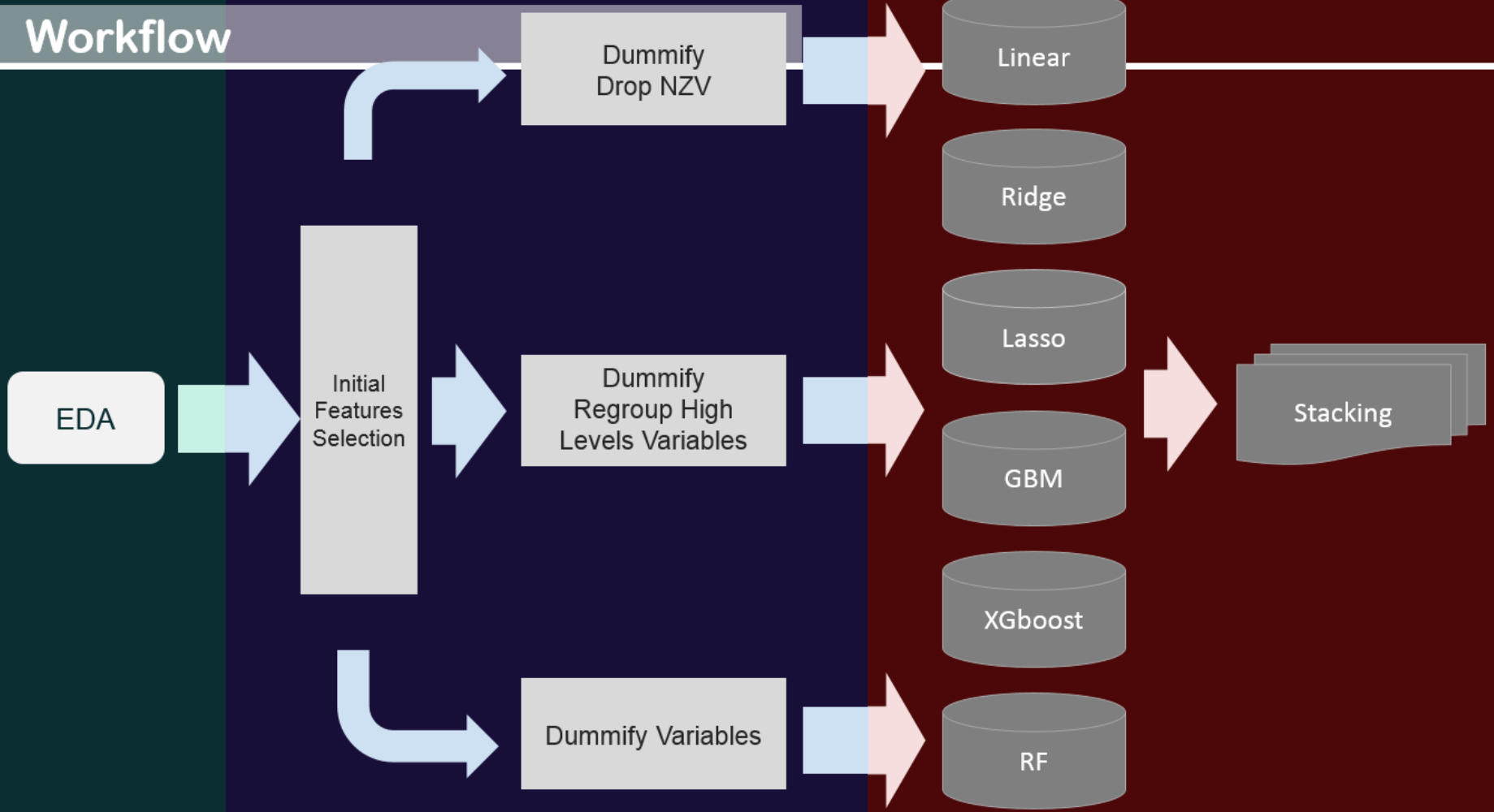
Allstate Claims Severity

Team KGW : Wen Li · Lei Zhang · Chuan Hong · Lydia Kan

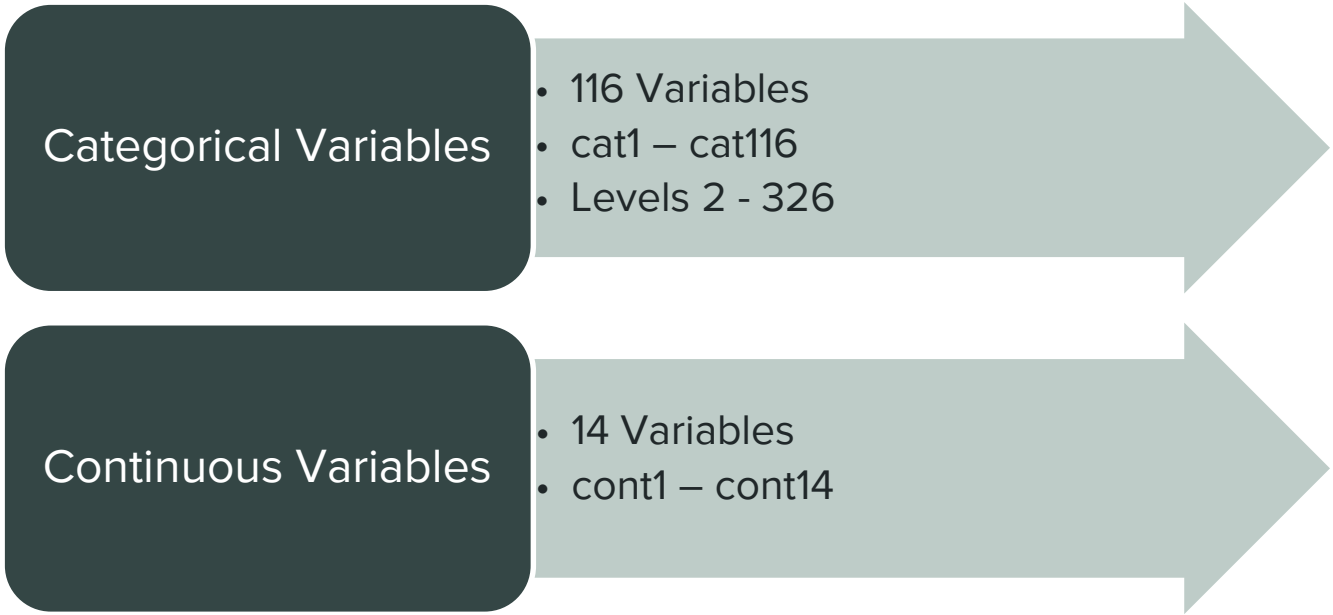
Content

- Workflow
- EDA
- Initial Features Selection
- Feature Engineering
- Supervised Learning
- Results and Finding
- Future Works

Workflow


EDA
Initial Features Selection
Dummify Drop NZV
Dummify Regroup High Levels Variables
Dummify Variables
Linear
Ridge
Lasso
GBM
XGboost
RF
Stacking

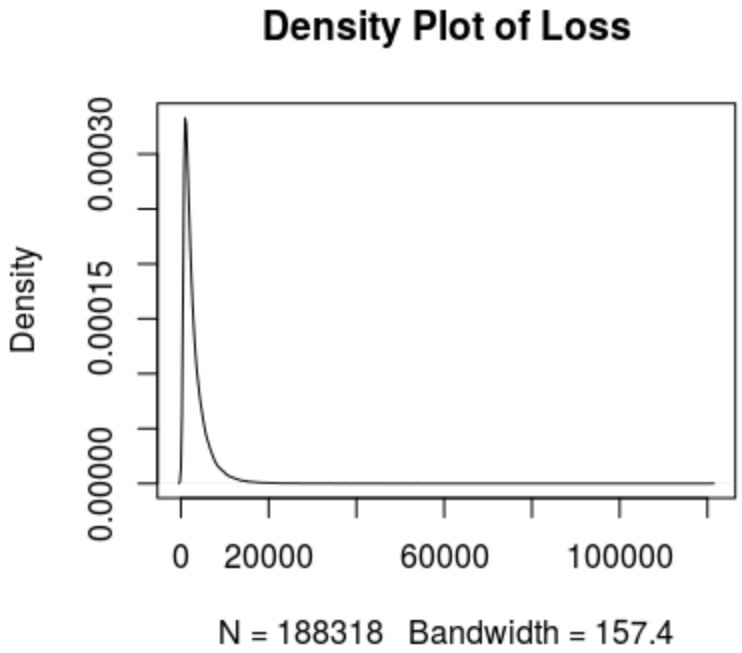
Numeric Graphic: Dataset

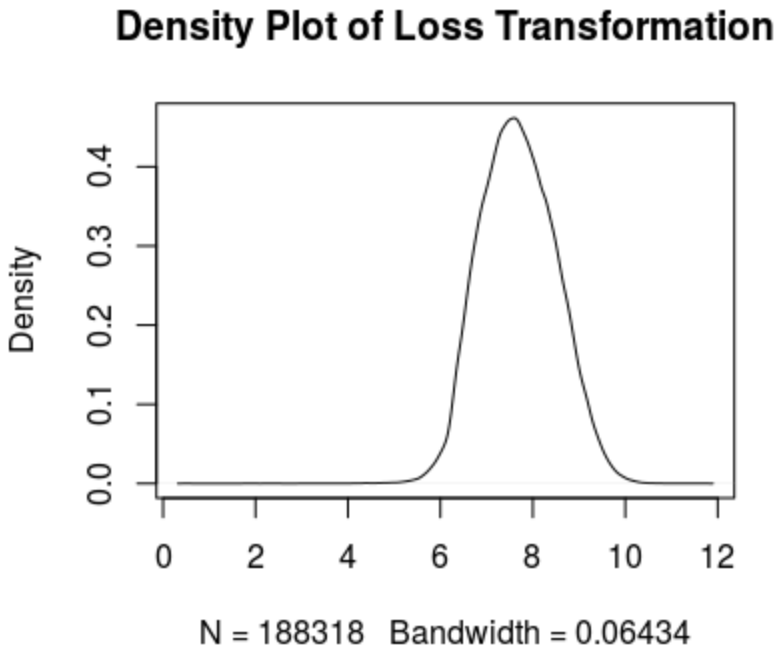


Numeric Graphic: The Categorical Variables

Variable	Train	Test	Variable	Train	Test
cat89	I	F	cat105	R S	
cat90	G		cat106		Q
cat92	F	G E	cat109	BM CJ BV BY BT B BF BP J AG AK	AD
cat96		H	cat110	BK H BN DV EI BD BI AN AF CB EH	BH CA EN
cat99		U	cat111	D	L
cat101	N U		cat113	BE T AC	AA R
cat102	H J		cat114	X	
cat103		M	cat116	BI V BL X FS P GQ AY MF JD AH EV CC AB W AM IK AT JO AS JN BF DY IB EQ JT AP MB C IO DQ HO MT FO JI FN HU IX	AQ EM FY AI N ET KO BJ IW DB LP MX BR BH JS ER A BN BE IS LS HS EX

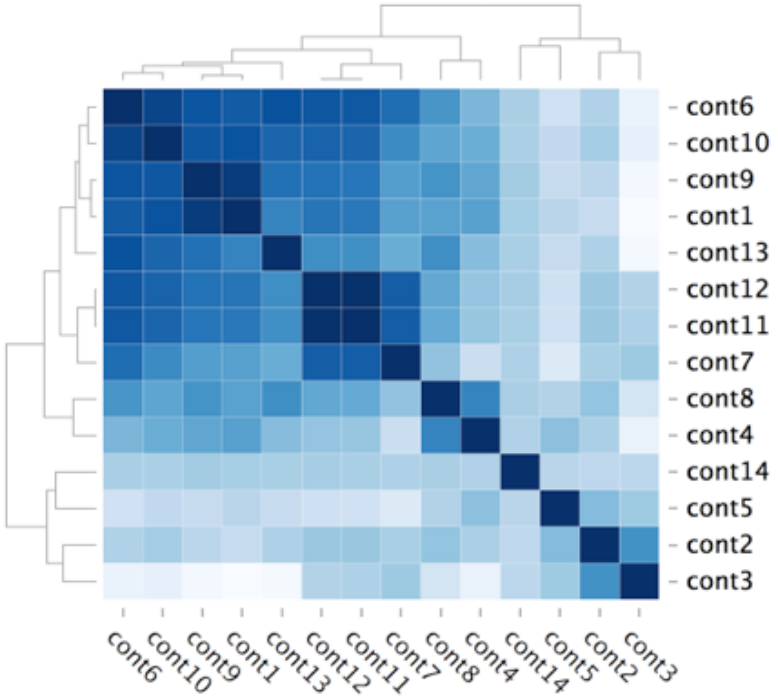
Graphic EDA: Output Variable



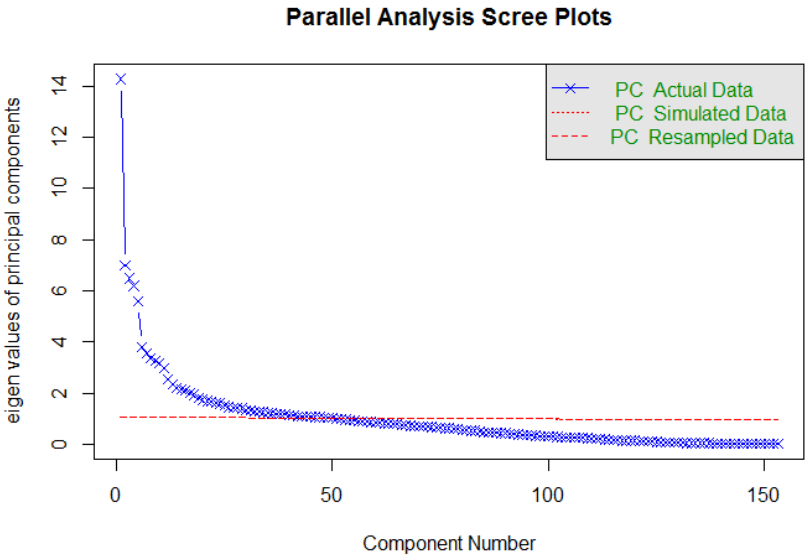


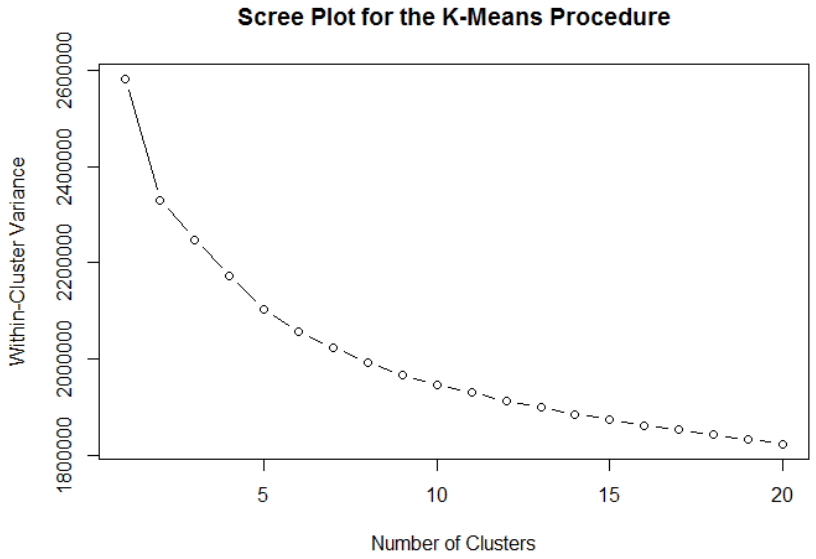
Graphic EDA: Input Variable

Correlations of all continuous variables



Initial Features Selection: Unsupervised





- **Goal: Check if the models are able to simplify the dimensions**
- **Result: There is no significant classification**

Features Engineering

Dummify Categorical Variables

(Keep All Features)

Pro:

Keep all information

Con:

Take too much time / Not suitable for MLR / Overfitting

Dummify Categorical Variables

>> Drop Near Zero Variance

Pro:

Time saving/Required by some models

(e.g. MLR)

Con:

High error and may lose some information

Select the Variables Have > = 15L

>> Group the Levels (variables) by Count and Avg of Loss

>> Dummify Categorical Variables

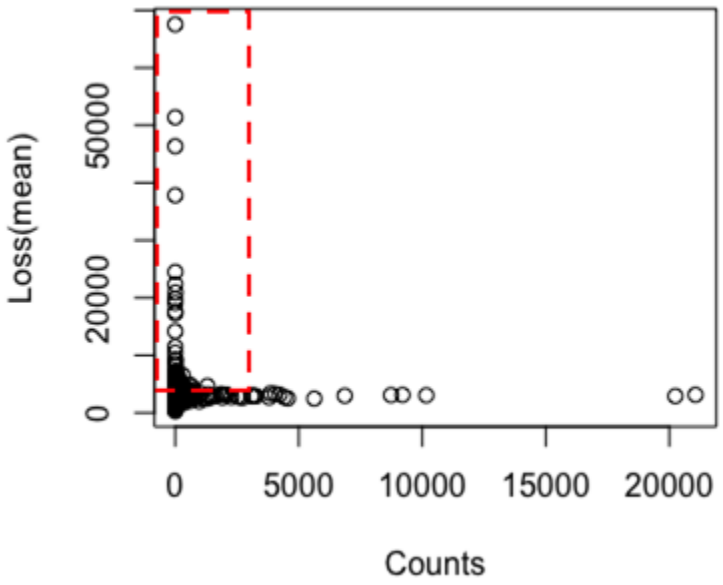
Pro:

Do Not throw away useful features

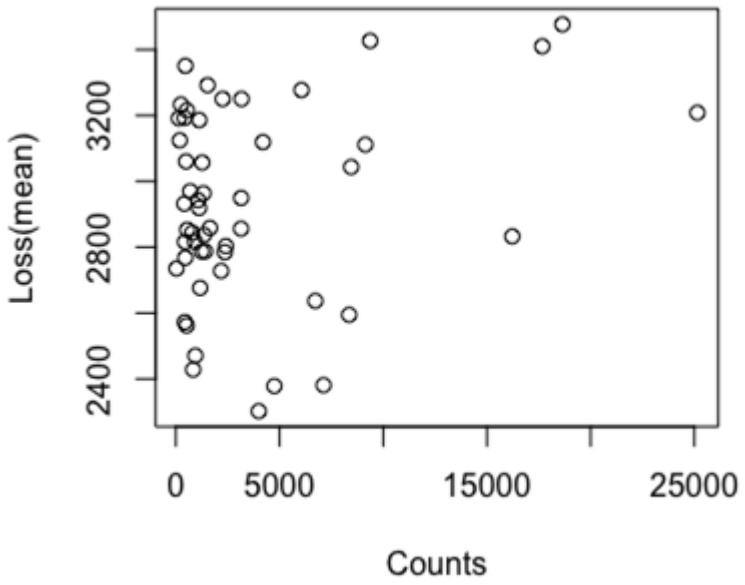
Con:

Multiple ways to group variables

Features Engineering



cat116
326 levels → 10 new groups



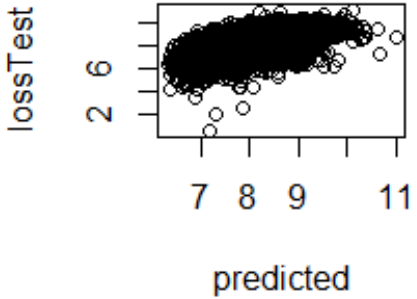
cat112
51 levels → 11 new groups

Supervised Models

Multiple Linear Regression

Features Engineering: Drop NZV

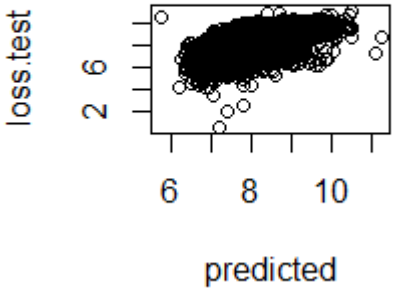
RMSE: 0.57659



Multiple Linear Regression

Features Engineering: Drop Correlated V. + New Group

RMSE: 0.56557



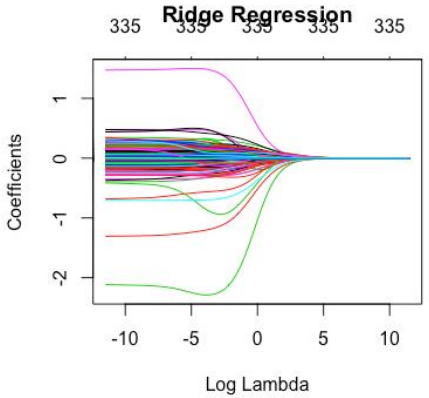
Supervised Models

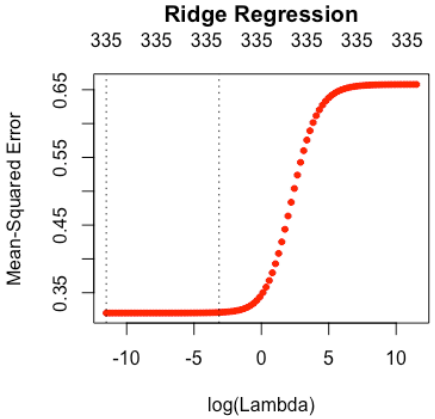
Ridge Regression

Features Engineering: New Group

Parameter: Lambda 1e-05

RMSE: 0.56414



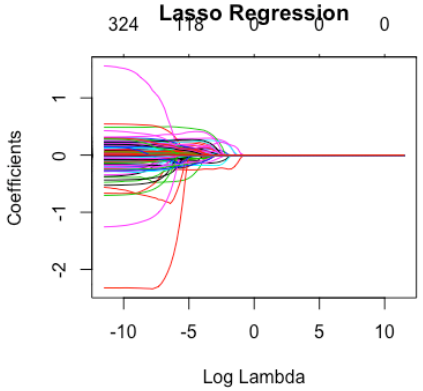


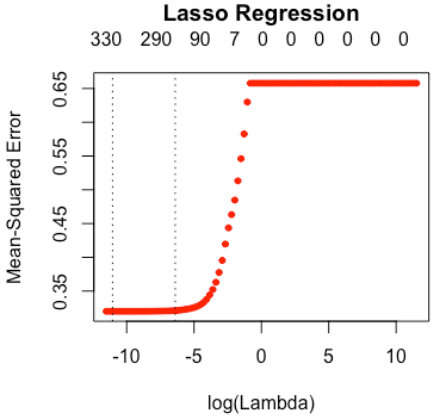
Lasso Regression

Features Engineering: New Group

Parameter: Lambda 1.592283e-05

RMSE: 0.56415





Supervised Models

Random Forest

Features Engineering: NZV

Parameter: Number of trees = 500 , No. of Variables tried at each split = 51

RMSE: 2014.217

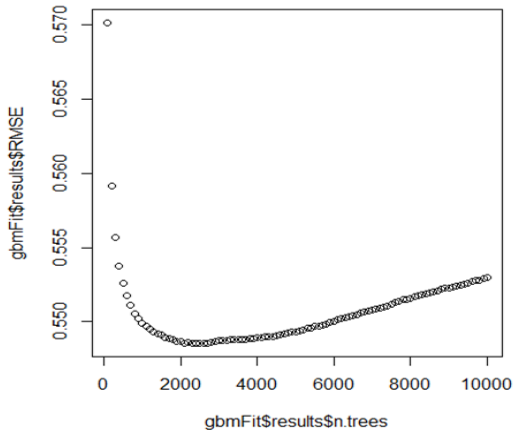
Supervised Models

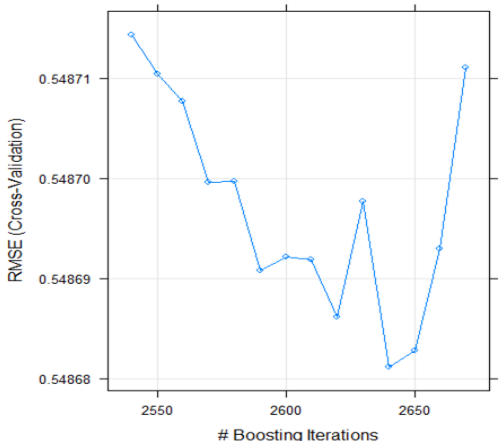
Gradient Boost

Features Engineering: NZV

Parameter: ntree=2640
n.minobsev = 20
interation.depth = 5
shinkage = 0.1

RMSE: 0.51

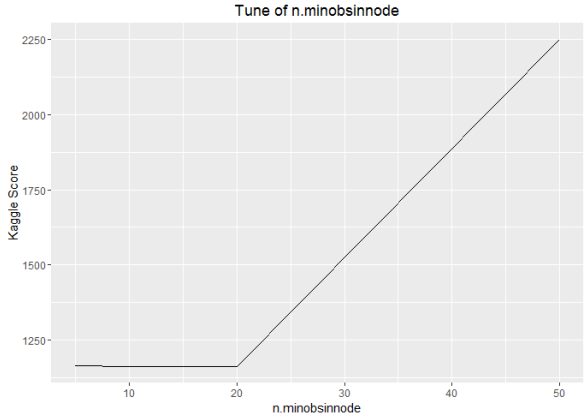




Increase Kaggle score by tuning parameters

ntree
2600: 1163.89861
2640: 1162.56392

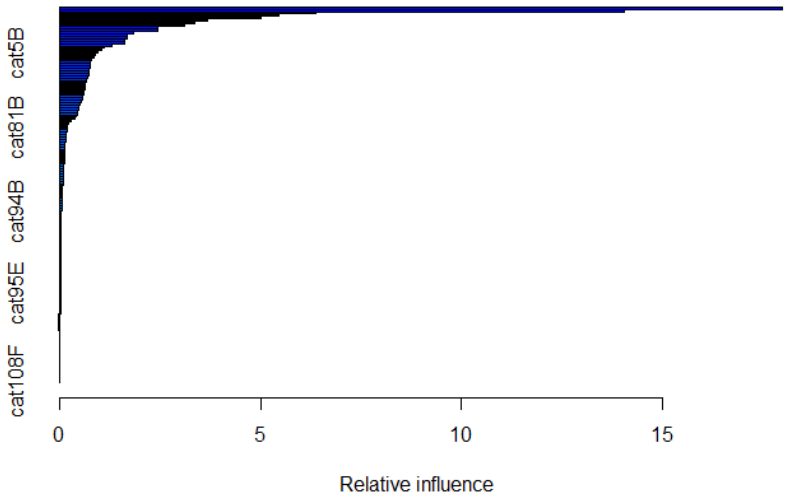
n.minobsev
50: 2251.57822
5: 1165.24778
10: 1162.56392
20: 1162.22589

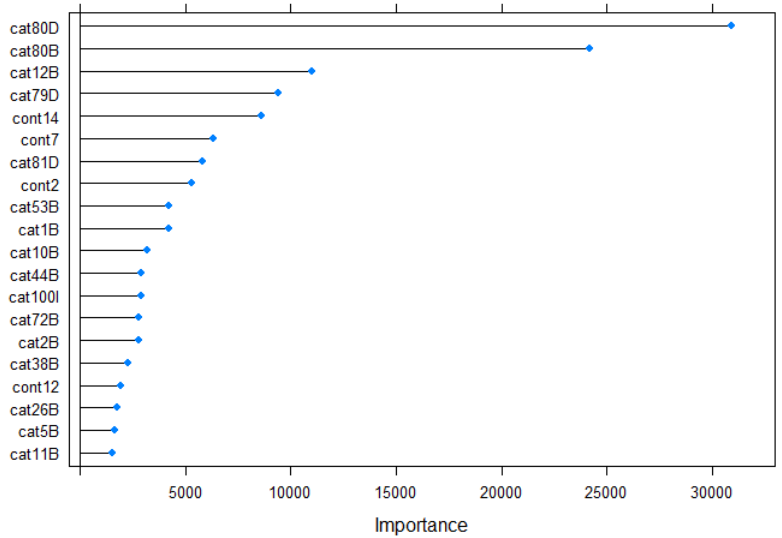


Supervised Models

Importance of Variables

- cat99R, cat99T, cat108F : 0.000000000000
- 83 out of 153 variables influence more than 0.05





Supervised Models

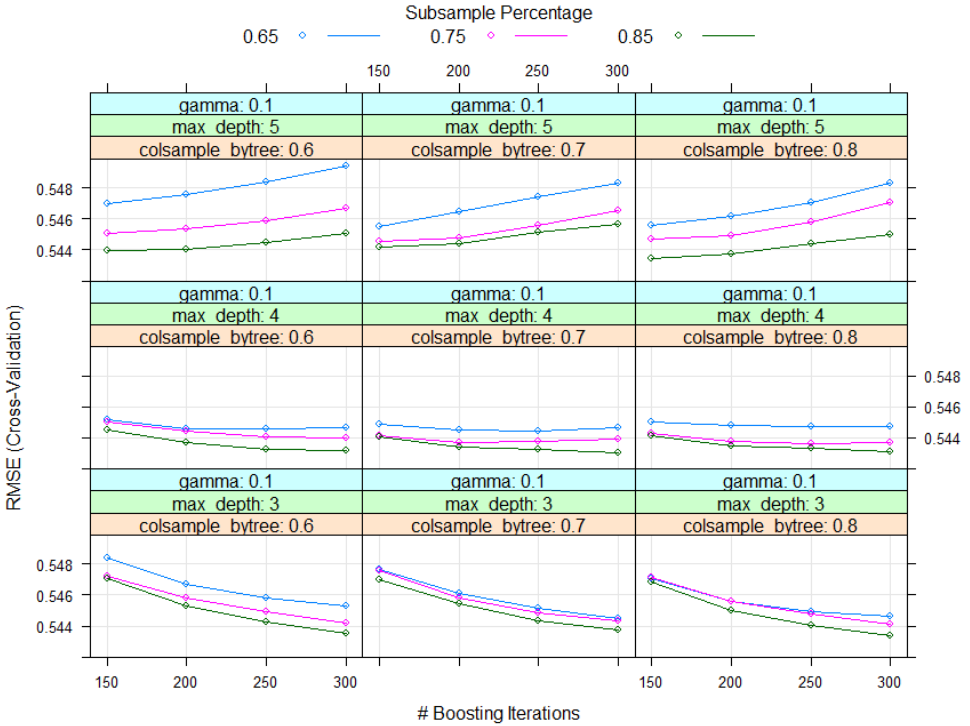
XGBoost - xgbTree

Features Engineering: New Group

Parameter:

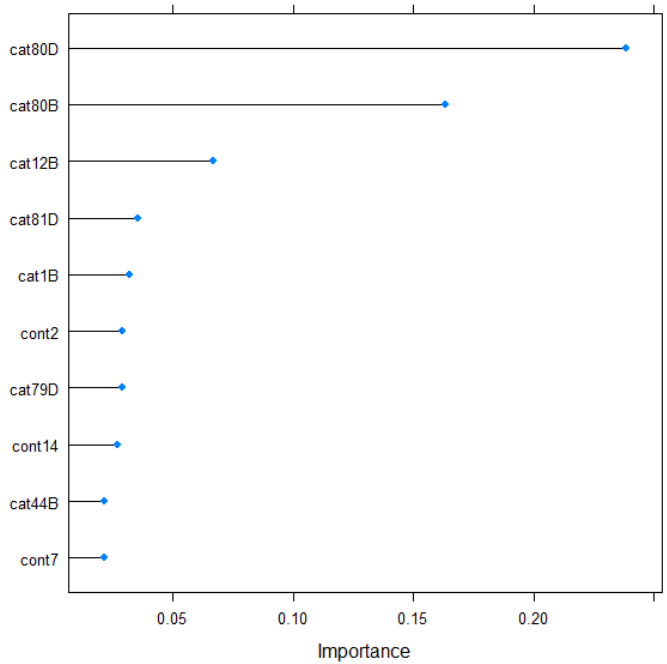
nrounds = c(150, 200, 250, **300**)
max_depth = c(3, **4**, 5)
eta = 0.3
gamma = c(0.1, **0.2**)
colsample_bytree = c(**0.6**, 0.7, 0.8)
min_child_weight = 1
subsample = c(0.65, 0.75, **0.85**)

RMSE: 0.5451

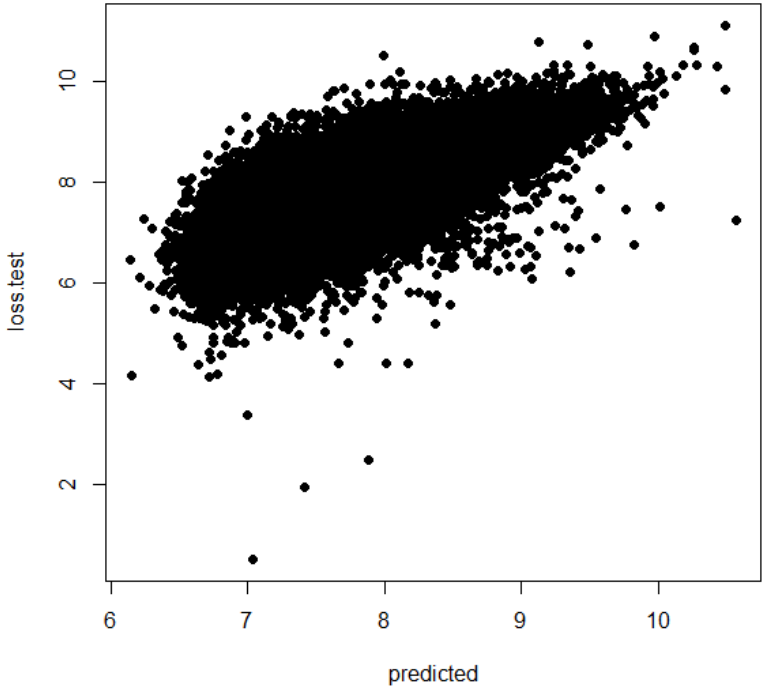


Supervised Models

Relative importance of variables (top10)



Scatter plot of loss.test vs. predicted loss



Results and Finding

Model	Features Engineering	Parameters	RMSE	Kaggle Score
MLR	Drop NZV		0.57659	
MLR	Drop High Cor + New Group		0.56557	
Ridge	New Group	Lambda: 1e-05	0.56414	
Lasso	New Group	Lmabda: 1.592283e-05	0.56415	
RandomForest	Drop NZV	Ntree: 500 mtry = 51	2014.217	
GBM	Drop NZV	Ntree: 2640 n.Minobsev: 20	0.51	**1162.22589**
XGB (xgbTree)	New Group	nrounds = 300 max_depth = 4 eta = 0.3 gamma = 0.2 colsample_bytree =0.6 min_child_weight = 1 subsample = 0.85	0.5451	

Future Works

1. **Gradient Boosting with “zv”**
   - “nzv” cut off the variables with 5% less variance
   - The kaggle score of our best boosting model is 5.6% higher than rank 1
   - With all variables, tune the parameter again
2. **Transform the y value-----reduce the variance?**
3. **Stacking different models to get higher accuracy**