ALLSTATE CLAIMS SEVERITY

CREATE AN ALGORITHM ACCURATELY PREDICTING CLAIMS SEVERITY

Claudia Wang

ABOUT THIS PROJECT

- This is a previous recruitment prediction competition held by Kaggle (an online community of data scientists) and Allstate Insurance.
- Allstate invited machine learning practitioners to show off their creativity and flex their technical chops by creating an algorithm which accurately predicts claims severity.
- Submissions are evaluated on the mean absolute error (MAE) between the predicted loss and the actual loss.

DATA DESCRIPTION

- Each row in this dataset represents an insurance claim. Competitors must predict the value for the 'loss' column. Variables prefaced with 'cat' are categorical, while those prefaced with 'cont' are continuous.
- File descriptions
  - **train.csv -** the training set
  - **test.csv -** the test set. You must predict the loss value for the ids in this file.
  - **sample_submission.csv -** a sample submission file in the correct format

EXPLORATORY DATA ANALYSIS

- Train Dataset: 188,318 rows, 132 columns

	id	cat1	cat2	cat3	cat4	...	cont11	cont12	cont13	cont14	loss
0	1	A	B	A	B	...	0.569745	0.594646	0.822493	0.714843	2213.18
1	2	A	B	A	A	...	0.338312	0.366307	0.611431	0.304496	1283.60
2	5	A	B	A	A	...	0.381398	0.373424	0.195709	0.774425	3005.09
3	10	B	B	A	B	...	0.327915	0.321570	0.605077	0.602642	939.85
4	11	A	B	A	B	...	0.204687	0.202213	0.246011	0.432606	2763.85

- Test Dataset: 125,546 rows, 131 columns

	id	cat1	cat2	cat3	cat4	...	cont10	cont11	cont12	cont13	cont14
0	4	A	B	A	A	...	0.38016	0.377724	0.369858	0.704052	0.392562
1	6	A	B	A	B	...	0.60401	0.689039	0.675759	0.453468	0.208045
2	9	A	B	A	B	...	0.30529	0.245410	0.241676	0.258586	0.297232
3	12	A	A	A	A	...	0.31480	0.348867	0.341872	0.592264	0.555955
4	15	B	A	A	A	...	0.50556	0.359572	0.352251	0.301535	0.825823

EXPLORATORY DATA ANALYSIS

- **Train Dataset:**

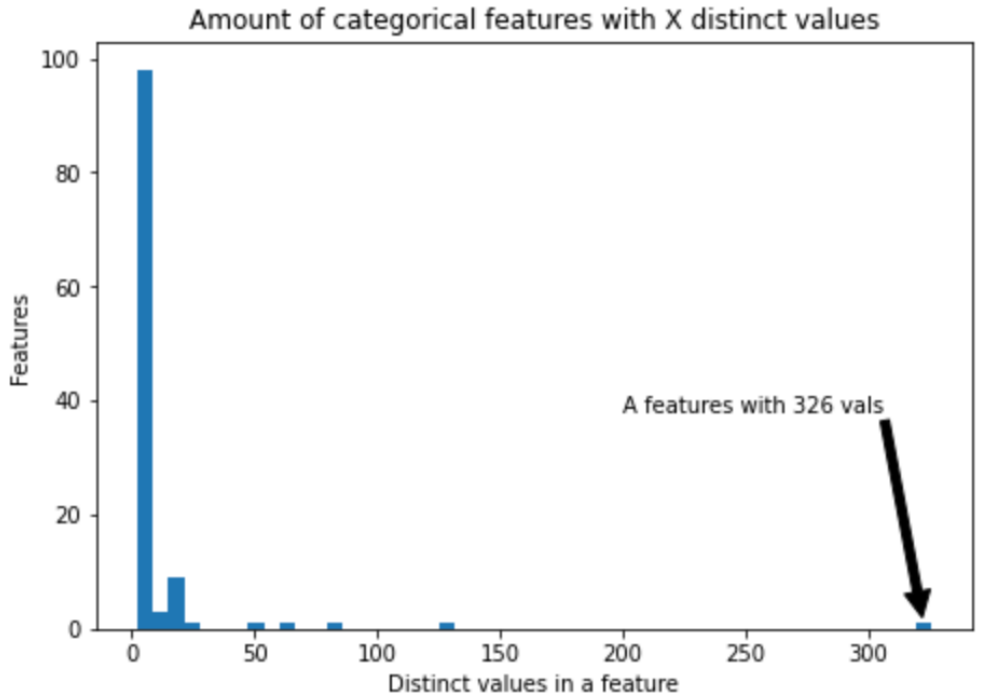
```
First 20 columns: ['id', 'cat1', 'cat2', 'cat3', 'cat4', 'cat5', 'cat6', 'cat7', 'cat8', 'cat9', 'cat10', 'cat11', 'cat12', 'cat13', 'cat14', 'cat15', 'cat16', 'cat17', 'cat18', 'ca
Last 20 columns: ['cat112', 'cat113', 'cat114', 'cat115', 'cat116', 'cont1', 'cont2', 'cont3', 'cont4', 'cont5', 'cont6', 'cont7', 'cont8', 'cont9', 'cont10', 'cont11', 'cont12', 'c
```

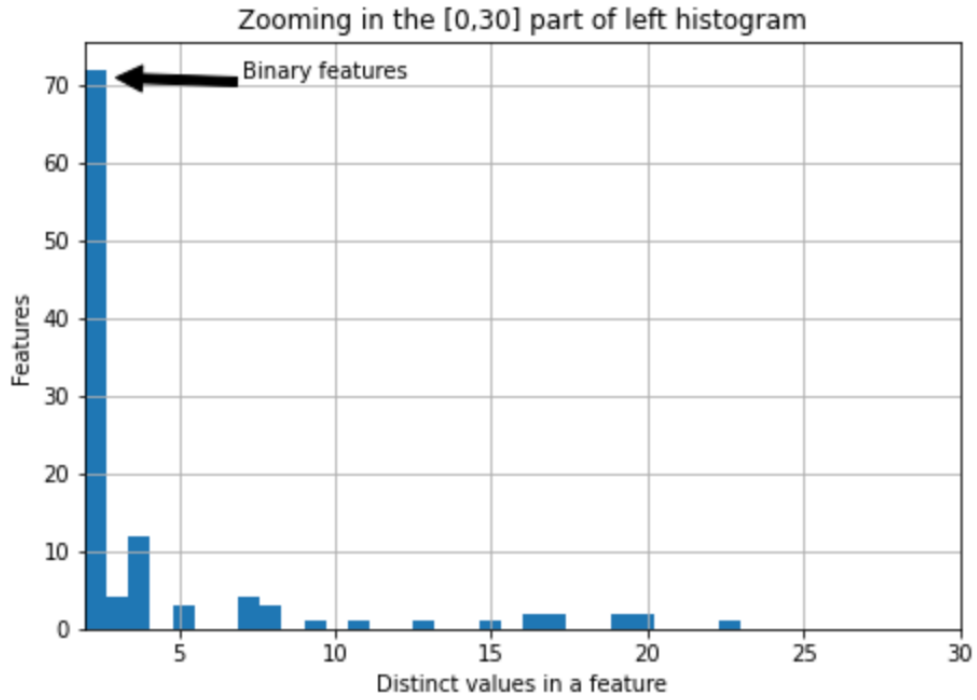
	id	cont1	cont2	cont3	cont4	cont5	cont6	cont7	cont8	cont9	cont10	cont11	cont12
count	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000	188318.000000
mean	294135.982561	0.493861	0.507188	0.498918	0.491812	0.487428	0.490945	0.484970	0.486437	0.485506	0.498066	0.493511	0.493150
std	169336.084867	0.187640	0.207202	0.202105	0.211292	0.209027	0.205273	0.178450	0.199370	0.181660	0.185877	0.209737	0.209427
min	1.000000	0.000016	0.001149	0.002634	0.176921	0.281143	0.012683	0.069503	0.236880	0.000080	0.000000	0.035321	0.036232
25%	147748.250000	0.346090	0.358319	0.336963	0.327354	0.281143	0.336105	0.350175	0.312800	0.358970	0.364580	0.310961	0.311661
50%	294539.500000	0.475784	0.555782	0.527991	0.452887	0.422268	0.440945	0.438285	0.441060	0.441450	0.461190	0.457203	0.462286
75%	440680.500000	0.623912	0.681761	0.634224	0.652072	0.643315	0.655021	0.591045	0.623580	0.566820	0.614590	0.678924	0.675759
max	587633.000000	0.984975	0.862654	0.944251	0.954297	0.983674	0.997162	1.000000	0.980200	0.995400	0.994980	0.998742	0.998484

  - Datasets have been preprocessed since all the continues variables have been reduced to [0,1] and means of them are around 0.5. Therefore, this is a feature dataset.
  - No missing values.

EXPLORATORY DATA ANALYSIS

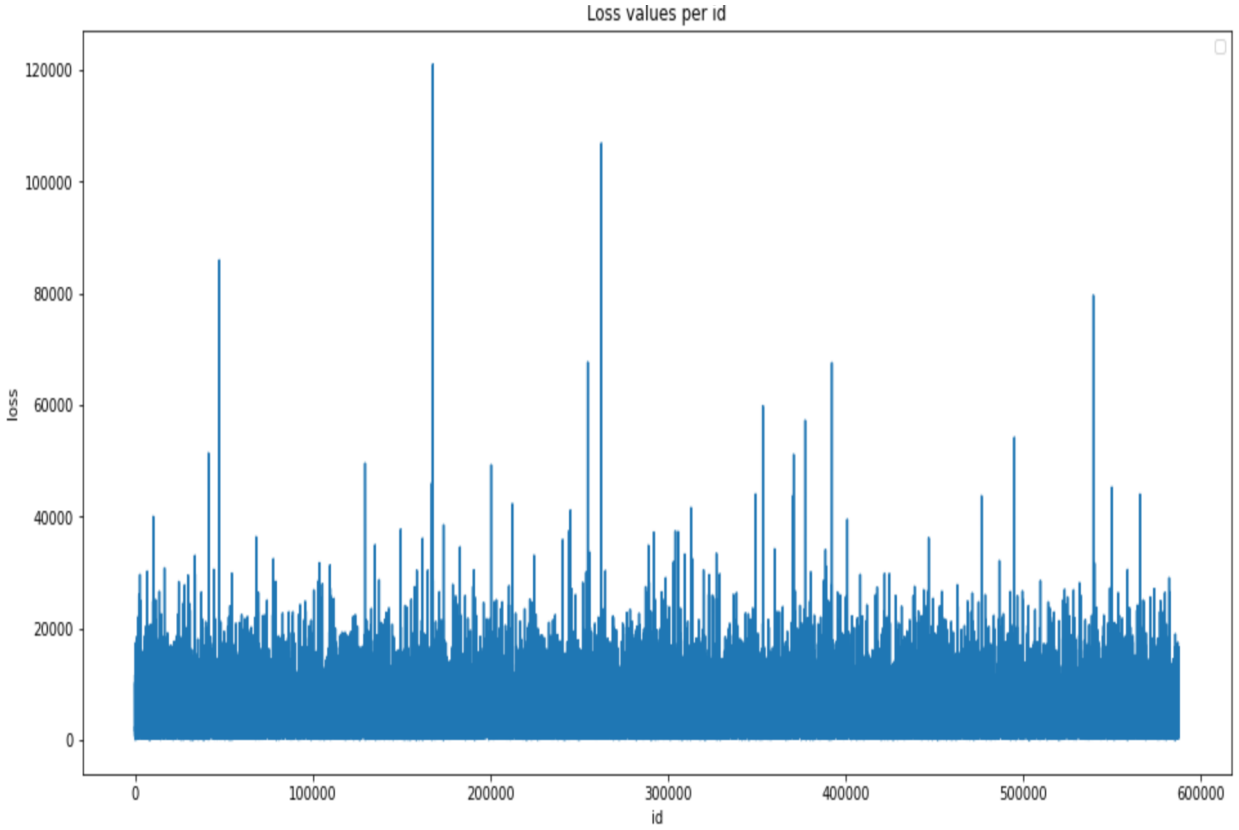
- Categorical:116 features
- Continuous:14 features
- A column of int64: ['id']

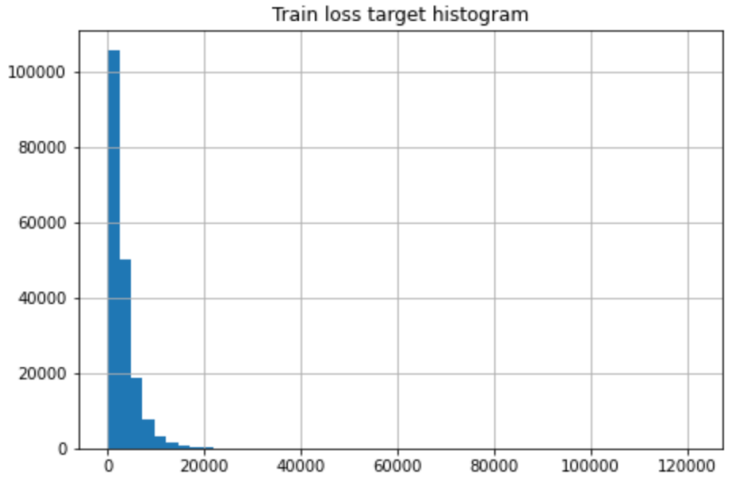


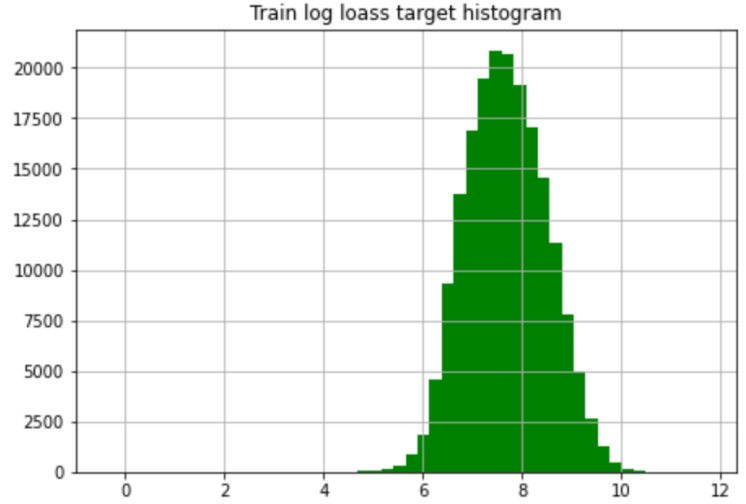


EXPLORATORY DATA ANALYSIS

- Target Feature: Loss Value
- Since the skewness of the loss value is high (3.67), use log transform

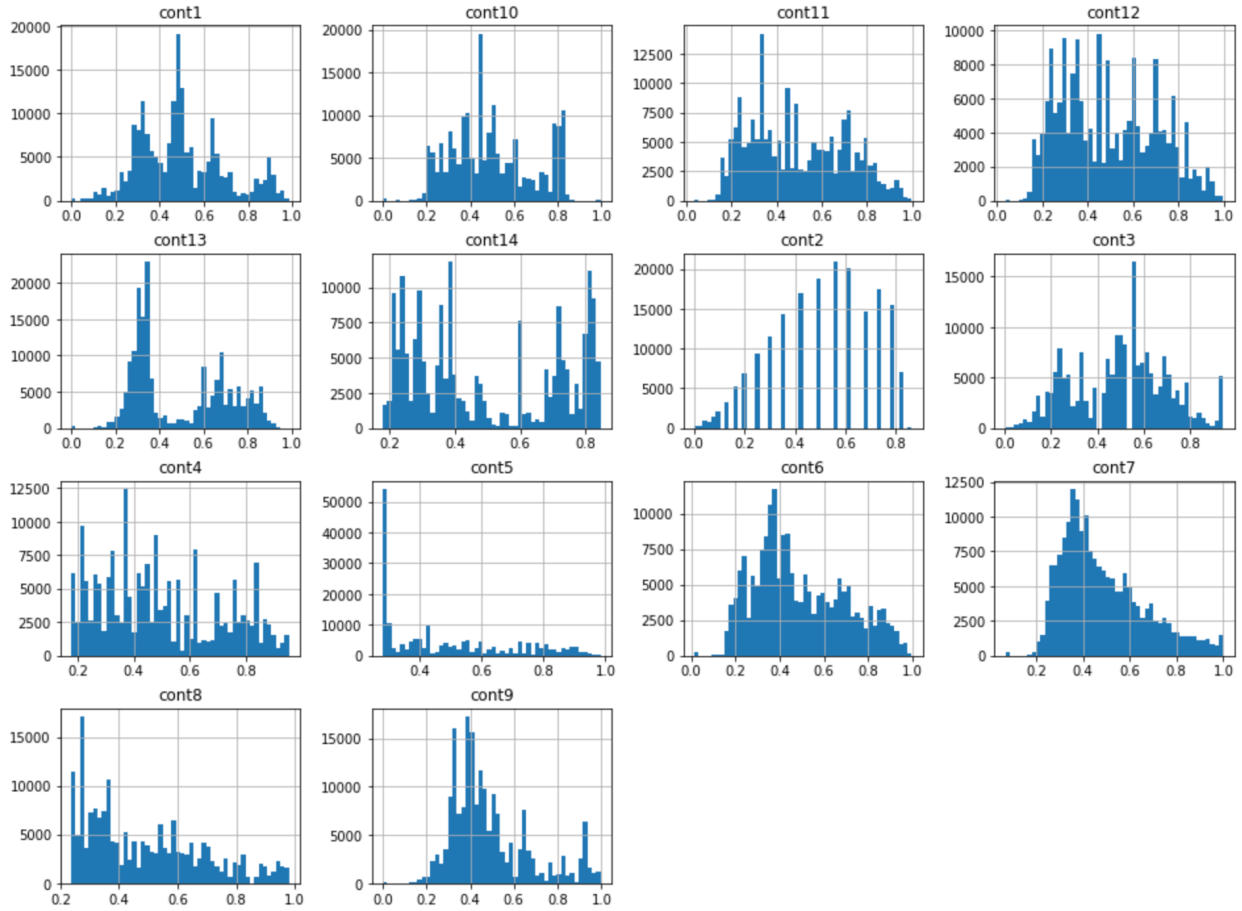






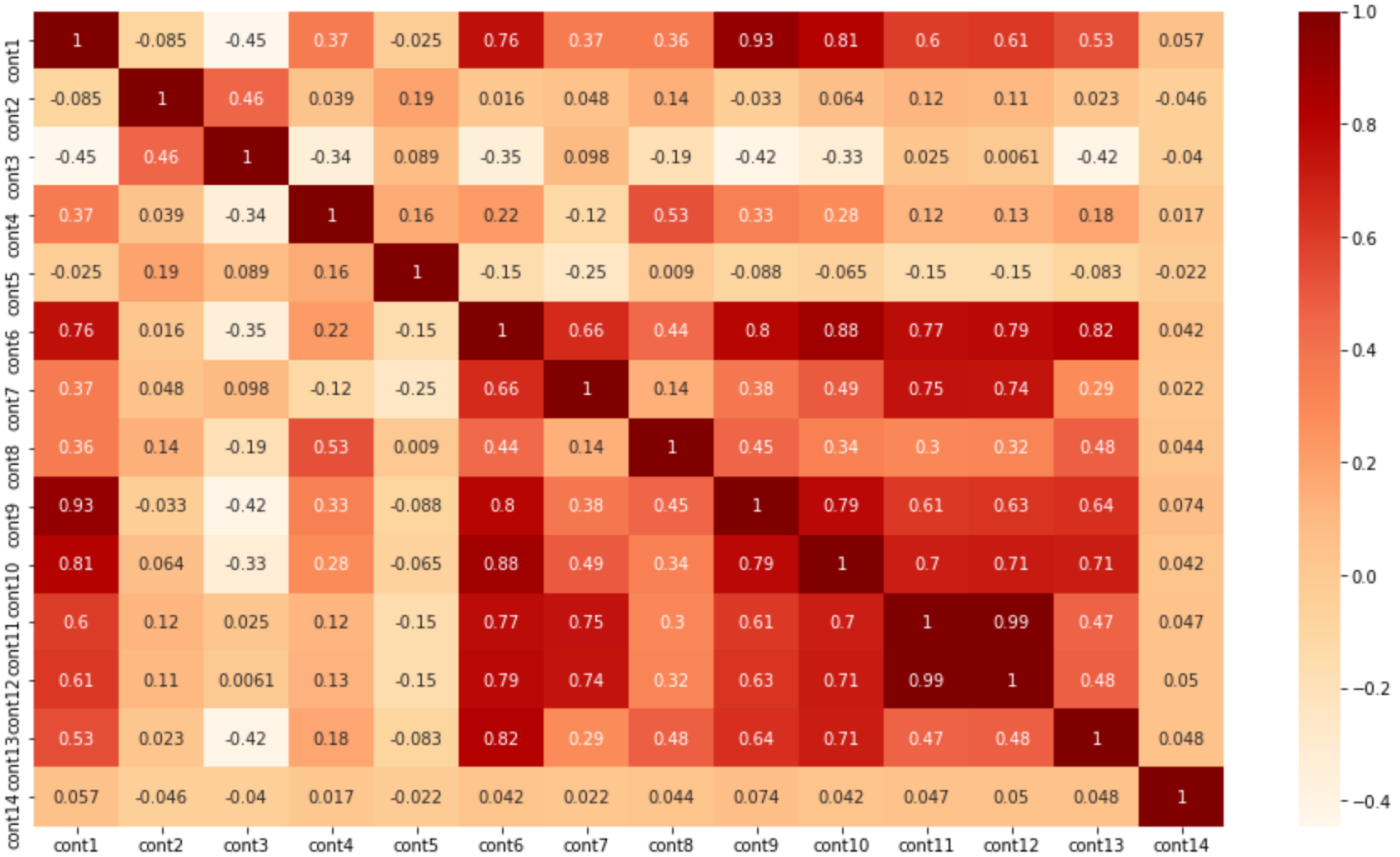
EXPLORATORY DATA ANALYSIS

- Continuous Features



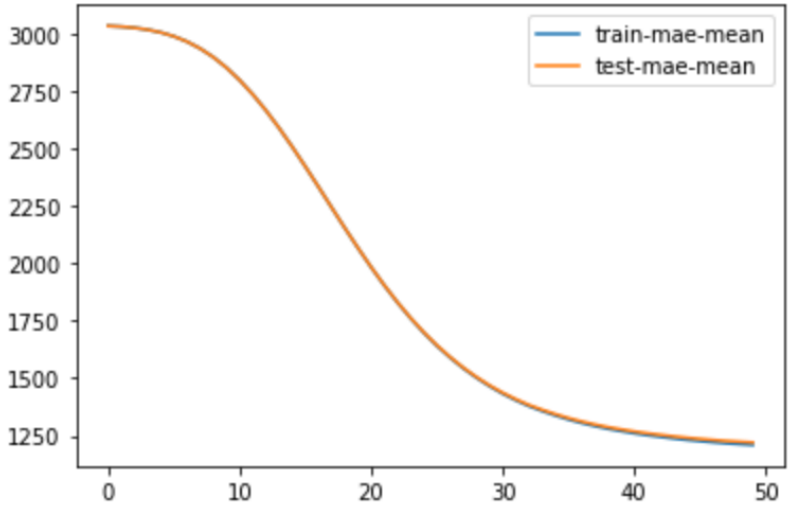
EXPLORATORY DATA ANALYSIS

- Correlations among features

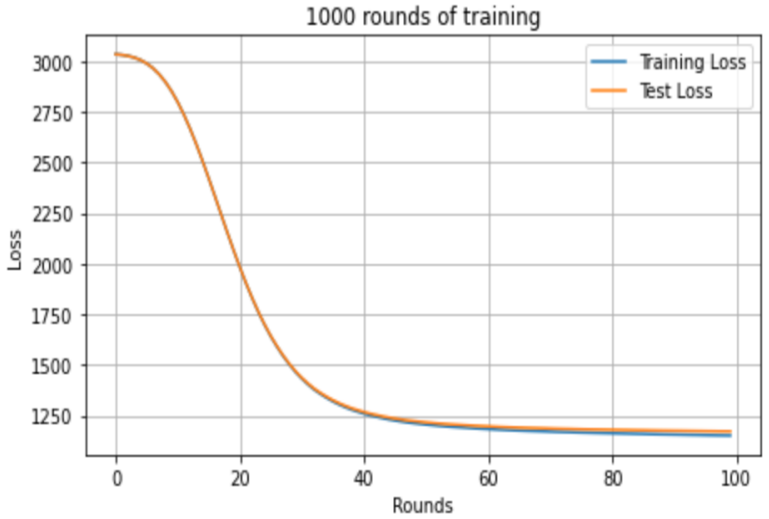


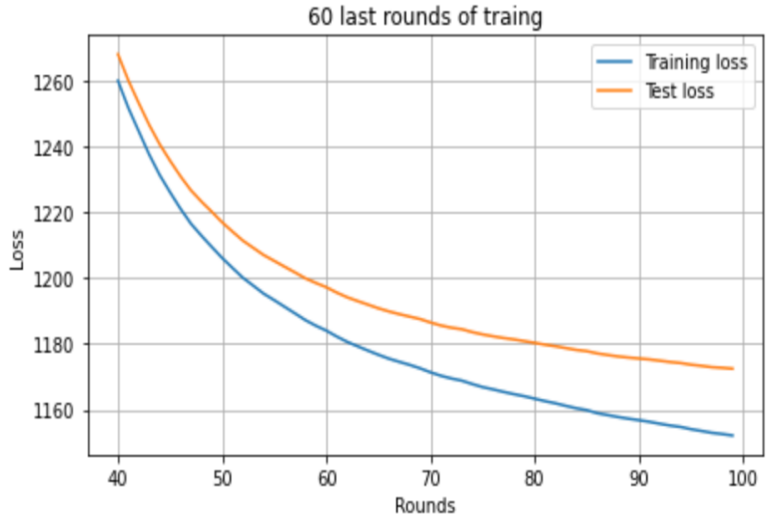
MODEL

- **XGBoost Regression**
- **Cross validation**
  - 50 trees: cv score = 1220.14



100 trees: cv score = 1172.46 (overfitting risk)



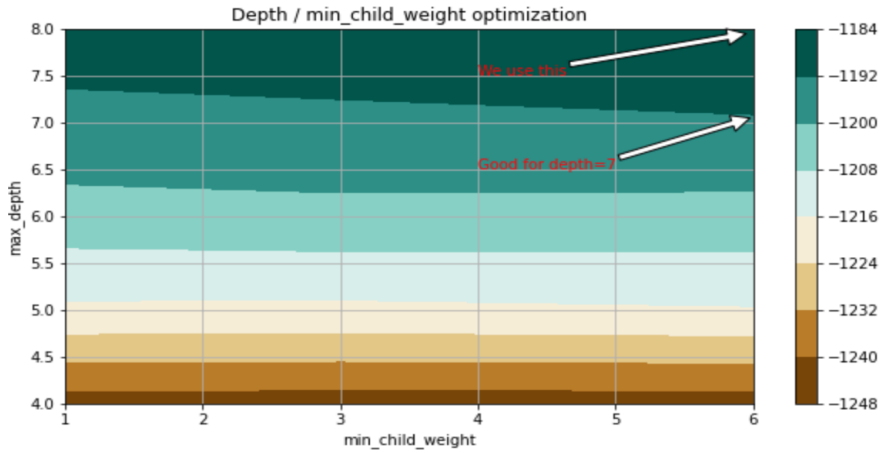


MODEL

- **Grid Search**
  - Step 1 : Build base model: eta=0.1, colsample_bytree=0.5, subsample=0.5, max_depth=5, min_child_weight=3, num_boost_round=50

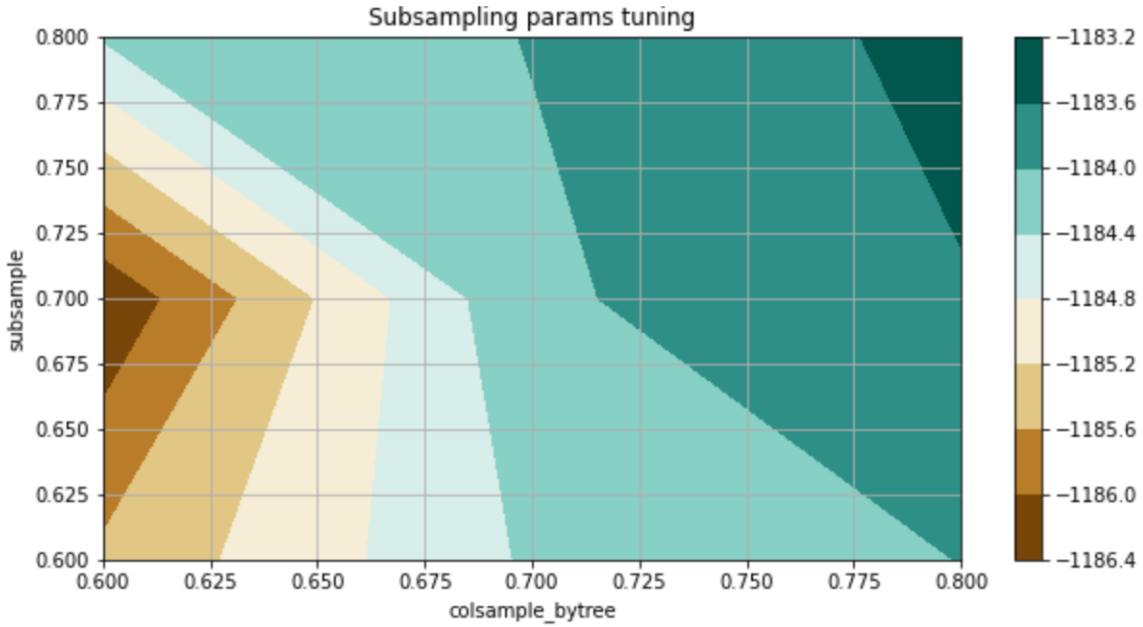
```
train-mae-mean     1208.575903
train-mae-std         2.065637
train-rmse-mean       0.558723
train-rmse-std        0.000888
test-mae-mean      1217.096240
test-mae-std         11.171228
test-rmse-mean        0.562311
test-rmse-std         0.002872
Name: 49, dtype: float64
```

  - Step 2: Tune max_depth, min_child_weight



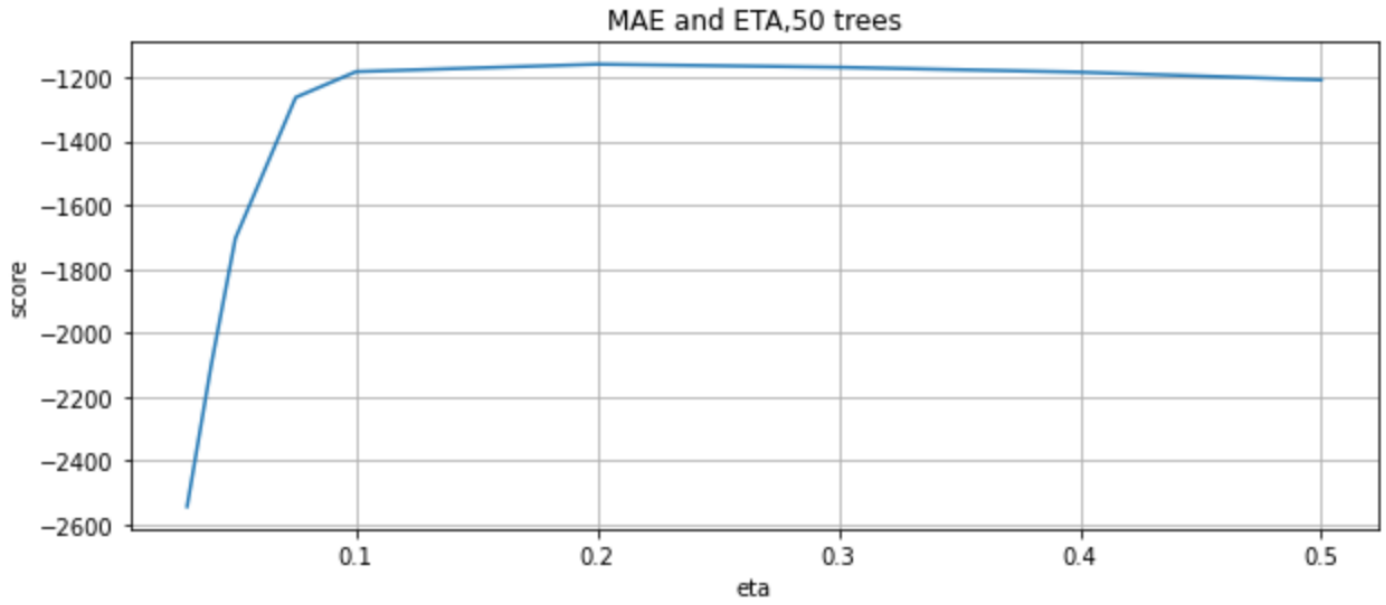
MODEL

- **Grid Search**
  - Step 3: Tune gamma to lower overfitting risk
  - Step 4: Tune subsample, colsample_bytree to change sampling



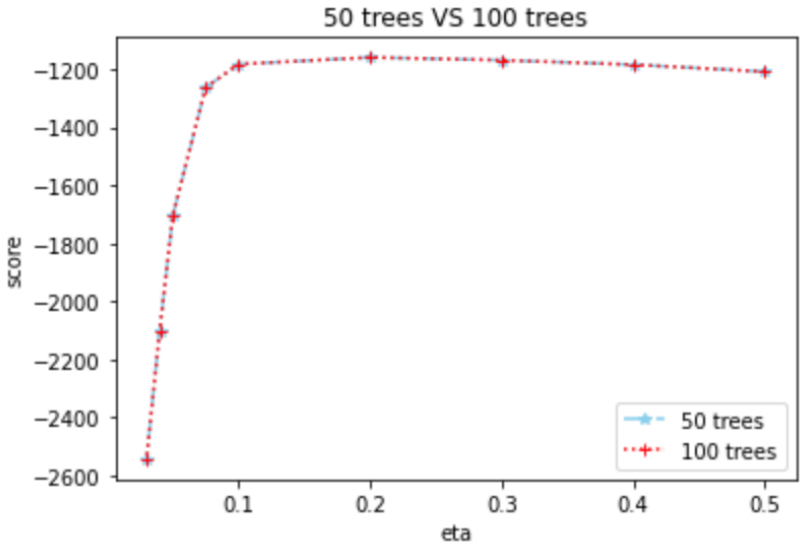
MODEL

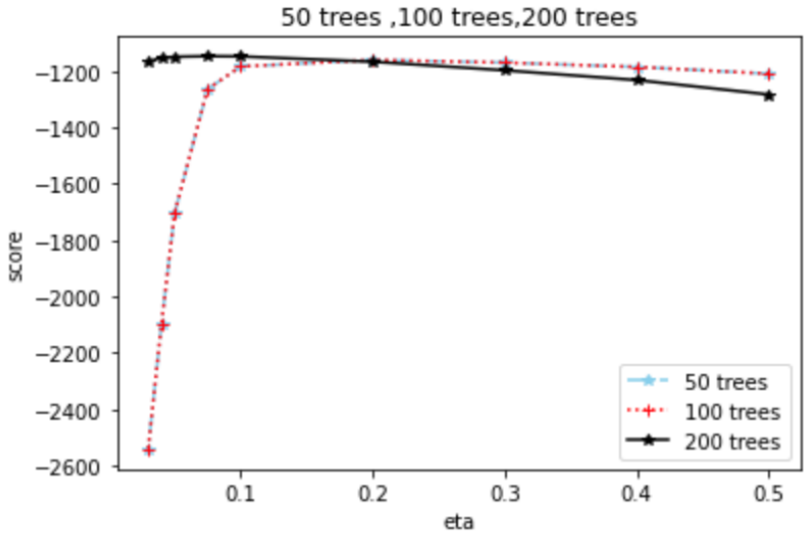
- **Grid Search**
  - Step 5: Tune eta



MODEL

- **Grid Search**





MODEL

- **Find the best combination of trees and learning rate: eta = 0.075 num_boost_round = 200**
- **Make predictions**
  - array[7.3063416, 7.6374493, 9.159702 , ..., 7.842589 , 6.9370203, 7.945583]