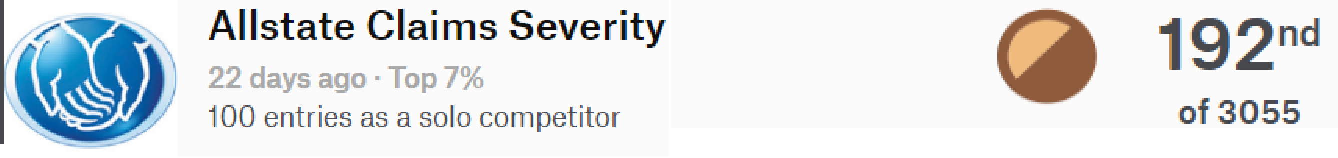
kaggle



Allstate Claims Severity

22 days ago · Top 7%

100 entries as a solo competitor

192nd

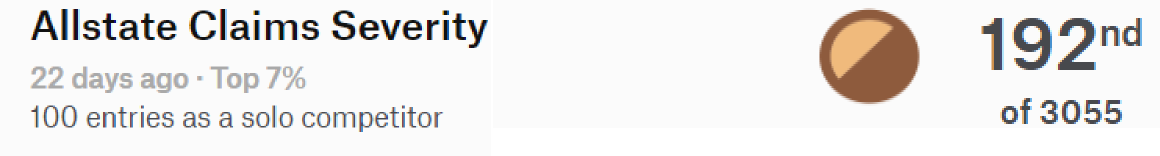
of 3055

Allstate Claims Severity Dataset:

- 130 anonymous features (116 categorical features + 14 continuous features)
    - Training data: 188,325 observations
    - Test data: 125,547 observations
    - No missing data
    - Dependant variable: amount of loss in $
- 72 categorical features have 2 levels, four features have 50-320 levels, others have 3-20 levels
- 14 continuous features are moderately skewed, strong correlation is observed among some of the continuous features. All features have been rescaled in the range of [0,1]
- The dependant variable is substantially skewed, hence it is log transformed
- Metric: MAE

Platform and Packages:

- AWS Elastic Cluster 2: m4.4xlarge (16 cpus+64g), sometimes m4.10xlarge (64 cpus)
- Python 2.7 + Jupyter notebook
- XGBoost, LightGBM, Sci-Kit Learn, Keras



Allstate Claims Severity

22 days ago · Top 7%

100 entries as a solo competitor

192nd

of 3055

	Solution 1	Solution 2
Rank – Private LB	Top 7%	Top 10%
1st Level Models	1 XGBoost model (unskewed, combined categorical features) 1 LightGBM model (unskewed, combined categorical features) 1 Keras model (standardized, one-hot-encoded)	1 XGBoost model 1 LightGBM model 1 Keras model 5 SciKit-learn models (unskewed, combined categorical features): RandomForest, ExtraTree, Adaboost (linear), Adaboost (decision tree), MLP
2nd Level Model	Markov-Chain-Monte-Carlo optimized weights (Metropolis-Hasting)	XGBoost
PCA on 1st level predictions	No	Yes, top 6 components
Bootstrap Aggregation	All models, both levels	All models, both levels
Pros	Good complexity-performance balance MCMC runs in minutes	Fine tuning is not necessary for every model
Cons	All models must be tuned to the top 20%	Long running time (5-10 hours for each model)