Allstate Insurance Claims Severity

Rajeeva Gaur, Jeff Pickelman, Hongyi Wang

Problem & Motivation

The goal of this project is to predict the cost of insurance claims in order to help improve insurance claim severity analysis and provide better targeted assistance to customers.

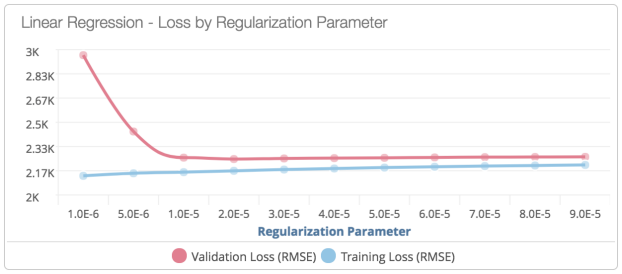
The dataset was provided by a Kaggle challenge. There are 188,318 training examples in the dataset, each having 116 categorical features and 14 continuous features. The categorical features are processed using one-hot encoding for linear regression, and are processed using label encoding for the decision tree based models.

We divided the provided dataset into train, validation, and test sets with a ratio of 80%, 10%, 10%. Each example has a label which is the cost associated with the claim, and is what aim to predict via the learning methods tried in this project. Since the label is a continuous real number, we decided to use RMSE as our evaluation metric.

Linear Regression

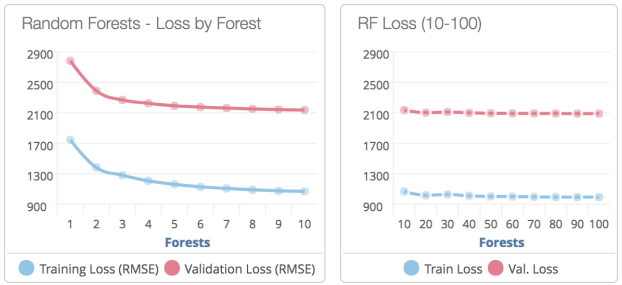
The cost function is linear least squares, with L_1 norm normalization to prevent overfitting. RMSE has been used as the error metric for the following graph and for choosing the parameter.

The graph shows the loss achieved for different values of the regularization parameter. Note that the train and validation errors are close together, indicating low variance, and thus good generalization performance.



Random Forests

Increasing the number of random forests from one to ten greatly reduced the RMSE. Surprisingly, increasing the number of tress beyond twenty yielded almost no further gains.

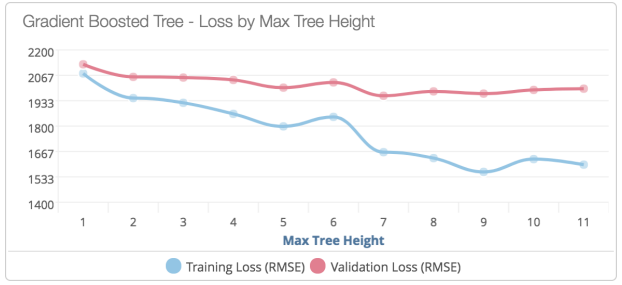


One explanation for this is that the dataset - while large - was not particularly varied, and thus the random subset of the data chosen by bootstrap aggregation didn't have enough variability to provide deeper insight for the cause of the loss value.

Gradient Boosted Tree

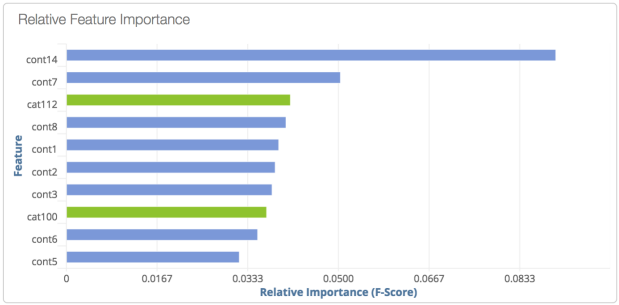
Given enough depth, the gradient boosted tree model fit the training set very well, but performed poorly on the dev/test set. Thus tuning the model parameters was very important for training this learning model.

We used k-fold cross validation at each step, which when combined with grid search to find the parameters produced the lowest cross validation RMSE. Early stopping was also used to detect model convergence.



Feature Analysis

Unsurprisingly the continuous features yielded the most information. Eight of the top 10 features by F-score were continuous (seen in blue).
Since the features are anonymized, we can't determine what real-world measures they map to.



Results

We tried a variety of different regression models to see how they would perform on estimating the claim loss. We chose a mixture of linear and non-linear models. Based on the result we picked the three best performing models and worked to further optimize them.

The Gradient Boosted Tree model produced the best result. Early works to combine 3 of the top performing models to produce a single prediction did not yield good results, as the 3 models likely capture similar underlying trends in the data.