**Analyzing All State claims severity (Kaggle competition)**

Insurance domain is one of the prominent domains where statistical learning/Machine learning plays a crucial role. With the use of statistical learning we can automate the process of claim-cost prediction and, also the severity. This automated process ensures a worry-free customer experience and offers insight into better ways to predict claims severity.

As part of the final project, we choose to work on a “kaggle” competition by Allstate that is related to claim cost prediction. In this project, our task is to predict the target variable “loss” (numeric quantity) based on the several predictor variables which include 116 categorical variables,14 continuous variables. And the error metric ‘MAE’ (mean absolute deviation) will be used to assess model performance.

For the learning process, we have two datasets one is for training, testing set. We implement several machine learning techniques on the training dataset using 10-fold CV and make predictions on the testing set to evaluate model performance. Since we have several variables it is possible to reduce the number of variables by feature selection and by exploring the data. Also, we try to implement several regression models such as GLM, Ridge, LASSO, CART, NNet models etc. And at the final stage we will try ensemble of different models to achieve better model performance. The pipeline is below.

Ensemble of model.

Model Evaluation

Model training

Feature selection

Data Exploration