STAT-542 Project proposal

Claims cost Prediction (Allstate Kaggle competition)

Team details:

Name	Net ID
Jayachandu Bandlamudi	bandlmd2
Wenke Huang	whuang67
Yubai Yuan	yubaiy2

Motivation: Insurance domain is one of the prominent areas 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 the severity. This automated process ensures a worry-free customer experience, offers insight into better ways to predict claims severity.

Description: As part of the final project, we would like to work on a "kaggle" competition by Allstate related to claim cost prediction. Our task is to predict the target variable "loss" (numeric quantity) based on the several anonymous predictor variables which include 116 categorical variables,14 continuous variables. And the error metric 'MAE' (mean absolute error) will be used to assess model performance.

Scope: For the learning process, we have two datasets one is for training (188K observations), the other is for testing (125k observations). We train several machine learning algorithms from the class such as KNN, GLM, Ridge, LASSO, CART on the training dataset using 10-fold CV and make predictions on the testing set to evaluate model performance. Also, we will try more complex models such as Random Forests, Boosting, SVM, Neural Net in the regression setting. Since we have several variables it is possible to reduce the number of variables by feature selection through exploring the data and feature selection, feature engineering will be the most challenging part of the project. And at the final stage we will try ensemble of different models to achieve better model performance.

Note: If time permits we will implement a new feature selection method based on importance sampling and incorporate the new methodology into one of ensemble tree method.

Pipeline for the project: