BACKGROUND

For my Capstone Project, I have decided to work on the All State Claims Severity competition from Kaggle.com which was first posted 2 years ago. The main reason I selected this challenge is because I am a working professional in the insurance industry and part of my job is to look for patterns in car accident data and come up with hypothesis and recommendations for managing associated risks.

OBJECTIVE

The goal of the challenge is to predict the severity or costs of each auto accident claim in the dataset given by All State as closely as possible. Conventionally in insurance, claim adjusters, human beings, have been responsible for estimating the costs of claims based on the accident/claim description and put up reserves of money to be paid to the claimant in the future. All State seems to be entertaining the feasibility of employing an algorithm that is as good as if not better than the human claims experts.

DATA

I will use the training dataset to train and refine my model, which will mostly likely be a regression model, so that I can develop a model that is expandable and works well with new data. The training dataset has roughly 188K records with 132 features including both categorical and continuous variables and the target variable which is the severity of claims. The test dataset has 126K records with 131 features without the severity variable.

<https://www.kaggle.com/c/allstate-claims-severity>

METHOD

This seems to be a typical supervised machine learning problem with a continuous y variable so I expect to start with a simple linear regression. But I as learn more technical skills as I go through this project, I may be able to try out different methods to better fit the data under the guidance of my mentor A.J.