# Machine Learning Engineer Nanodegree

Capstone Proposal

*Bryan Luke Lathrop Nov 1st, 2016*

## Domain Background

This Project will be based around the Kaggle competition detailed at: https://www.kaggle.com/c/allstate-claims-severity

This competion centers on the idea that the severity(or cost) of a claim may be predicted based on the several factors in the data set. Much of the work in statistics to date has been used by the insurance industry in pursuit of this goal, and this particular challenge is aimed at recruiting the particpants for work in an already tested field.

I choose this competition as it dataset and goals allow us to explore various machine learning techniques without focusing on data collection. I also believe that the techiques and goal used are very close in style to those used in industry currently, and so more applicable to future projects.

## Problem Statement

Using the provided dataset we will use various machine learning techniques to predict the claims severity. Claim severity is expressed as a cost, where a higher cost is a higher severity. The predictions will be made on a per claim basis, and are intended to be applied to future claims as an indicator for customers/agents.

## Datasets and Inputs

The dataset is provided by the competition organizer, and is anonymized, including removing labels from from each data point. We are to assume that this data was gathered in the normal course of the business of prior insurance claims, and will be continue to be gathered so that new predictions may be made.This means that we may not use intuition to provide new features. We are to assume that the data is relevant to the problem and accurate. We may test this relevance, or use methods such as PCA to examine the most relevant labels.

We are provided training and test data set, where the training set includes the "loss" field that we are attempting to predict, and test does not. When looking at the common features, we see 116 categorical and 14 continuous features. The features seem well matched between train and test, with similar mean/standard deviation/min/max.

## Solution Statement

The model will output a predicted 'loss' for each claim in the test data. This data will then be submitted to the Kaggle competition, where a score will be assigned to the model. The solutions score will be evaluated using the mean absolute error(MAE) between the actual and predicted loss.

## Benchmark Model

The base model for this project is planned as a simple linear regression based on the data, with minimal pre-proccessing only, and submitted for scoring according to the previously mentioned method. This will provide a definitive measurement of the improvement we see in the final model.

## Evaluation Metrics

The project success may be evaluated on the improvement in score over the benchmark model, as returned from the competetion. Both models will be trained using the same data and submitted for the same test data. As we are using MAE for scoring, we will be looking for the lowest score as the winner.

Additionaly we will track prediction time for the scores acheived, as well as training time, in an effort to quantify the effort needed to use the score in a production environment. These times will be used with the final scores to determine viability of the model

## Project Design

The project may be broken into several categories:

- Data Analysis - looking for details of the data, such as size, layout, usefulness of features.
- Pre-processing data - transform/scale data as appropriate for the chosen uses.
- Benchmark model - run the benchmark and evaluate the scores
- Final modeling and prediction - use our chosen model and generate a score
- Submission for scoring - submit both Benchmark model and our model for a final score. record score as well as computation times for each model
- Evaluation of final model vs benchmark

### Data Analysis

We will load and examine the dataset as a panda dataframe and ensure that the test and train dataset's are similar enough for use. We will take a look at the features and make predictions about feature correlation, scaling, etc. We will also review outliers in each feature. For categorical data, we will make sure that test data is included in the train data.

### Pre-processing data

Here several tasks will take place:

- categorical data will be transformed to numerical
- all data will be scaled 0-1
- new features will created (clustering, etc)
- unneeded features will be removed(ex:id)

- data will be transformed to final state for handoff to the model (numpy array)

**Benchmark model**

The benchmark will be run on a minimally scaled and prepared data set, without new features, etc. The planned benchmark model will be a linear regression, from the scikitlearn tool chain.

**Final modeling and prediction**

The final model will be trained and validated on the train data. As part of the training, we will use various cross-validation and gridsearch techniques to tune the hyper-params.

The planned final model is expected to be a stacked model using various regressors at each layer. Each layer will contain several standard models, such as random forest/KNN/XGB, run using cross validation to make a prediction . The output of each model for each prediction will then be used as an input in the next model for a prediction. Predicted MAE and train time will be tracked for each, as well as for the final model. Each model in each layer will have hyperparameters tuned for best MAE using gridsearch, or other appropriate tuning.

**Submission for scoring**

The Benchmark and final model will be submitted and the relative score of each recorded. For this submission, the final runtime of training and prediction will be recorded, as well as the run time of individual segments.

**Evaluation of final model vs benchmark**

The model score will be compared to the benchmark, along with the run time. The train time of each will be compared and evaluated in the context of use for updating the model as new data comes in. The predict time will be evaluated in the context of providing agents/customers with results for for use in claims.