# Machine Learning Engineer Nanodegree

## Capstone Proposal

## Allstate Claims Severity Project

Alok Biswas on April 2nd 2019

### Domain Background

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

This competition asked to develop models that can automatically predict the cost, and hence severity, of claims. It may be predicted based on the several factors in the data set. Much of the work in the field of statistics has been used by the insurance industry in pursuit of this goal, and this particular challenge is aimed to create an algorithm which accurately predicts claims severity by using machine learning.

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

### Problem Statement

The Kaggle competition provider Allstate is searching for an automated way to predict the amount of money claimed by their insurance policyholder. For this purpose a model using machine learning shall be developed.

### Datasets and Inputs

The dataset is provided by the competition organizer Allstate. We are provided training and test data set, where the training set includes the "loss" field that we are attempting to predict, and test set 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. The train set has 188318 rows, and the test set has 125546.

As we can't verify the test set directly, we will further break out a validation set from the train data, for use as our own test set for the purpose of validating the models before use with the provided test set. This validation set will be sized to about 25% of the train data.

## Solution Statement

We want to understand the relationship between all 130 features and loss, the target feature. There are many features which, due to curse of dimensionality, may result in overfitting, so we may have to reduce the features by using PCA or some other method. We also have to find the relations between the features and if they are highly related, it would make sense to use PCA to reduce the dimensionality. We will also convert categorical values from alphabets to numbers that can be used in models. Then we would test a few models to check which performs best using k-fold splitting and finally get the mean squared error. The models to be used are: linear regression (as base model) and XGBoost (as trusted algorithm) and if required, deep learning. Currently, it's not decided what kind of neural network it would be. To tune parameters in XGBoost, we will use Grid Search.

## Benchmark Model

The base model for this project is planned as simple linear regression based on the data and get MAE, with minimal pre-processing.  We would use a part of training data as testing data and more precisely, we would use the about 25% entries of training set as testing set for validation. Then, we can compare our next model with it to see if it can beat it and by how much extent. This will provide a definitive measurement of the improvement we see in the final model. We will take the best model and for satisfying the curiosity, run it on test set provided by Kaggle as test dataset. A personal goal would be to be in the top 25% of the Kaggle Private Leaderboard.

## Evaluation Metrics

The project success may be evaluated on the improvement in score over the benchmark model, as returned from the competition. 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.

Additionally, we will track prediction time for the scores achieved, 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

Citation:

http://xgboost.readthedocs.io/
http://scikit-learn.org/
https://www.kaggle.com/c/allstate-claims-severity/discussion