# Machine Learning Engineer Nanodegree

## Capstone Proposal

**Anshuman Patel**

October 02, 2017

## Domain Background

Supervised learning is one of the most promising field in machine learning where already much development has taken place and is currently used in real world application. The customer is the focal point of any business and is directly proportional to growth of the business. One area which affect the customers the most is the services provided and the delay in them. In order to reduce the waiting period, the business can use the machine learning algorithms. It will not only help the customers but also let businesses to focus on other things which require regular human intervention.

Allstate, the second largest personal lines insurer in the United States, hosted a contest on kaggle, about predicting how severe is the claim of an insurance. According to them, when an accident occurs,your focus should be on the things that matter the most: family, friends, and other loved ones. Pushing paper with your insurance agent should be the last place you should spend your time or mental energy. And according to me we should automate every task that is redundant and can be done by the enormous power of computation, So that we can do new discovery that will push mankind to a new height.

The dataset used in this project can be found at :-
https://www.kaggle.com/c/allstate-claims-severity/data

Some of the works that use similar approach to solve real world problems and influenced this work are :-

1. Author Age Prediction from Text using Linear Regression [ Link ]
2. Correlation Study and Regression Analysis of Water Quality Assessment of Nagpur City, India [ Link ]
3. The Prediction of Indian Monsoon Rainfall: A Regression Approach [ Link ]

## Problem Statement

The Allstate Corporation is the second largest personal lines insurer in the United States and the largest that is publicly held. Due to its large size, they have to tackle a large number of claims which takes time when done by a human. Allstate is currently developing automated methods of predicting the cost, and hence severity of the claims. The problem is to create an algorithm which accurately predicts claims severity. As input, we are provided with different variables which the agents look at in or order to decide the status of the claims. They can be both continuous or discrete.
Since the target variable is a continuous quantity(the amount to be paid to client), it is essentially a regression task.

## Datasets and Inputs

The dataset contains 2 ".csv" files with information necessary to make a prediction.

1. They both contain the follwong data fields :-
    - id :- Probably the unique id of the insurance holder
    - cat1 to cat116 :- Category variables (the range or the column names of the values are not provided).
    - cont 1 to cont14 :- Continuous variables (the range or the column names of the values are not provided).
    - loss :- The amount of compensation that the Allstate insurance has to pay to the insurance holder. This is the target variable and is not present in the testing dataset(test.csv).

2. Training dataset (train.csv) -
    - Number of rows = 188318
    - Number of columns = 132
    - Relevance :- Highly relevant as this is the data we will train on.

3. Testing dataset(test.csv) -
    - Number of rows = 125546
    - Number of columns = 131 (excluding target variable)
    - Relevance :- Highly relevant as this is the data we will test on.

The link to the dataset can be found above in the domain background.

## Solution Statement

We want to understand the relationship between the 130(116+14) features and the target variable loss. Although it looks straightforward but, due to the large number of features, and in turn curse of dimensionality, may result in overfitting, so we may have to reduce the the features by using PCA, t-sne or some other method. We also have to find the relations between the features for that matter and if they are highly related, it would make sense to use PCA to reduce the dimensionality. We also have to convert categorical values from alphabets to numbers which can be used in models. Then we would test a few models to check which performs best using Kfold splitting and finally get the mean absolute error. The models to be used are: linear regression (as base model) and XGBoost (as trusted algorithm) and if required, deep learning (which is achieving state of the art in almost everything).Currently, its not decided what kind of neural network it would be.To tune parameters in XGBoost, we will use Grid Search.

## Benchmark Model

As this is a Kaggle competition a benchmark model would be the best Kaggle score for the test set, which comes in at 1109.70772 mean absolute error(lower is better).
However, due to hassle and limitations (like 5 entries per day), for academic purposes, we would use a part of training data as testing data. More precisely, we would use the last 38,318 entries of training set as testing set.
We will run the linear regression classifier to get a base MSE. Then, we can compare our next model with it to see of it can beat it and by how much extent.
We will take the best model and for satisfying the curiosity, run it on test set provided by
Kaggle as test dataset. The submission file will be uploaded on kaggle website to check the score. Then, we can also compare our model with the benchmark model hosted by Kaggle. A personal goal would be to be in the top 20% ie. less than 1121.21401 error of the Kaggle Private Leaderboard.

## Evaluation Metrics

The model prediction for this problem can be evaluated in several ways.Since the official evaluation of this project is done by Kaggle using mean absolute error (Lower it is, better the model), same will be used for evaluation of models.
And among our internal models, linear regression model will seve as the benchmark model.

## Project Design

We will proceed into this projects by 5 steps.

1. The first step would be to setup the environment and prepropcessing the data. During preprocessing we would check the structure of data, check for skewness and if any irregularity is found then related hyperparameters will be tuned. Also we would check for correlation between features, if any, so that it will be easy to implement PCA. Also we would assign each categorical variable a numerical value.

2. Then we would first implement Linear Regression and test the predicted data's mean absolute error to the competitions best answer, also this would serve as our local benchmark score.

3. Then we would employ XGBoost model without tuning and compare its result to the linear regression model's. And also we would try to tune the XGBoost model.

4. And then we would go for our last model, Keras model, with deep learning.

5. In our last step we would try to tune the keras and XGBoost model. And whichever method gives lowest mean absolute error will be the final model to submit.

Tools and Libraries used:
- Python
- Jupyter Notebook-
- Pandas
- Scikit learn
- Seaborn
- Matplotlib
- Tensorflow
- Keras
- XGBoost

## References


[1] Kaggle, ”Allstate Claims Severity” (2017). [ Link ]
https://www.kaggle.com/c/allstate-claims-severity

[2] Allstate wikipedia page. [ Link ]
https://en.wikipedia.org/wiki/Allstate

[3] Supervised Learning Wikipedia page. [ Link ]
https://en.wikipedia.org/wiki/Supervised_learning

[4] Author Age Prediction from Text using Linear Regression.
https://homes.cs.washington.edu/~nasmith/papers/nguyen+smith+rose.latech11.pdf

[5] Correlation Study and Regression Analysis of Water Quality Assessment of Nagpur City, India
http://www.ijsrp.org/research-paper-1115/ijsrp-p47110.pdf

[6] The Prediction of Indian Monsoon Rainfall: A Regression Approach
http://www-personal.umich.edu/~copyrght/image/solstice/sum07/Solstice_GoutamiED.pdf