Machine Learning Project

# Project Report

Nitin Pathania

# Definition

## Project Overview

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.

Supervised learning is the task of inferring a function from labeled data.it has two types of data: training data and testing data. The training data consist of a set of *training examples*. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. This inferred function is then used to predict the output value of the testing data. We already have the results of the test data. The difference between the predicted data and the results are used as a performance metric to ensure the accuracy of the inferred model.

My personal motivation is this fact only that the very first use of machine learning in my mind is to reduce the frequency of redundant tasks, So that we can invest our times in better work and this problem addresses exactly that.

The link to my data source is: <https://www.kaggle.com/c/allstate-claims-severity/data>

## Datasets and Inputs

The dataset contains 2 .csv files with information necessary to make a prediction. They are:

* 1. train.csv and test.csv - contains
     + id - the id of a training set question pair
     + cat1 to cat116 - category variables (the range of values is not provided, neither the column names).
     + cont 1 to cont14 - continuous variables (the range of values is not provided, neither the column names).
     + loss - the amount which the company has to pay for a particular claim. This is the target variable. In test.csv,loss is not present since we are going to predict that.
  2. In train.csv -
     + Number of rows = 188318
     + Number of columns = 132
     + Highly relevant as this is the data we will train on.
  3. In test.csv -
     + Number of rows = 125546
     + Number of columns = 131
     + Highly relevant as this is the data we will test on.

As this was a Kaggle competition. The dataset is provided by Kaggle and Quora. They can be obtained [here](https://www.kaggle.com/c/allstate-claims-severity/data).

## Problem Statement

The **Allstate Corporation** is the largest personal lines insurer in the United States and the largest that is publicly held. Due to its large size, they have tackle a large number of claims which takes time done by a human.

Allstate is currently developing automated methods of predicting the cost, and hence severity, of claims. The problem is to create an algorithm which accurately predicts claims severity. Basically, given the data , predict the amount to be paid by Allstate.

We want to understand the relationship between the 130(116+14) features and the loss. Although it looks straightforward but it isn’t so. We have to find the relations between the features and 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 accuracy. The models to be used are: linear regression (as base model) and XGBoost (as trusted algorithm). This solution is a good fit for the problem as first gets a base score using naive model and later on uses complex models to attain better scores.

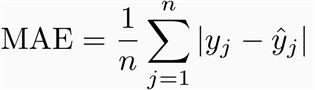
The end result we are looking for is the predicted amount that will be claimed.

## Metrics

The model prediction for this problem can be evaluated in several ways. Firstly, our problem is regression type, so we can’t use the metrics for classification or clustering like precision, accuracy, recall. We have at our disposal, mean absolute error, mean squared error, mean squared log error and r2 score. Now the r2 square checks how much variation is described by the fitted line and usually used for linear cases. Since we are more concerned as to how much do the company has to pay extra (currency wise) we should use a MAE or MSE. MSE penalizes large errors more than the small error, however, MAE gives equal weightage to every error. We will be using mean absolute error (Lower it is, better the model).

### Mean Absolute Error (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.



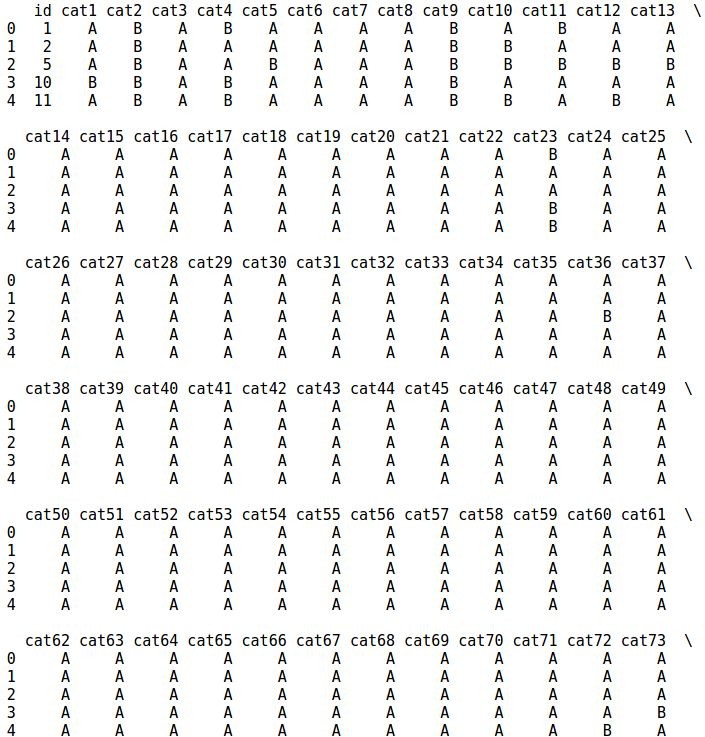
**Fig 1: Formula for calculating MAE**

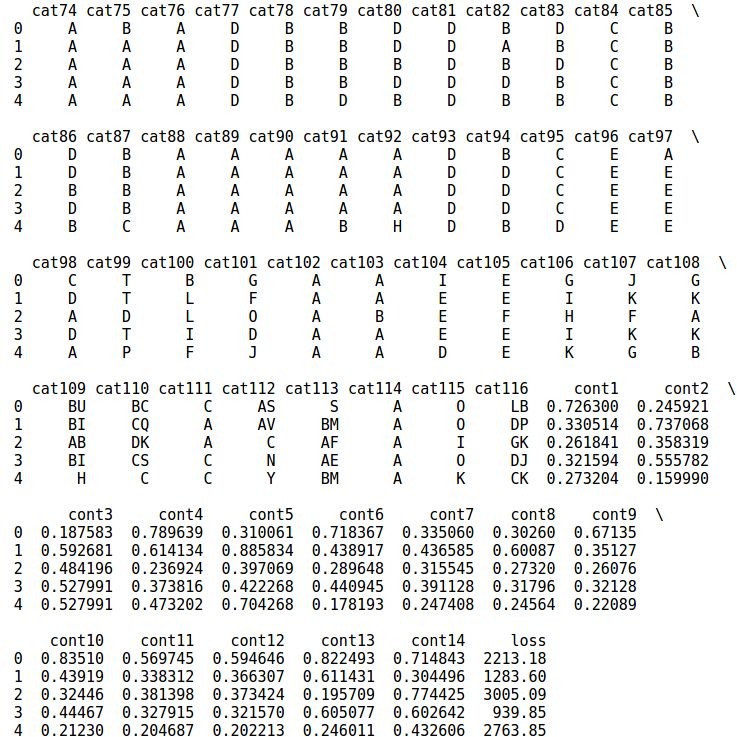
# Analysis

## Data Exploration

Data exploration is done first.

The first five columns of train data looks like this. The first thing we notice is that it contains both categorical as well as continuous features. Secondly, categorical features are alphabetical and since, the models in machine learning are designed and optimized to work well with numeric vectors, we need to convert them into numeric values.



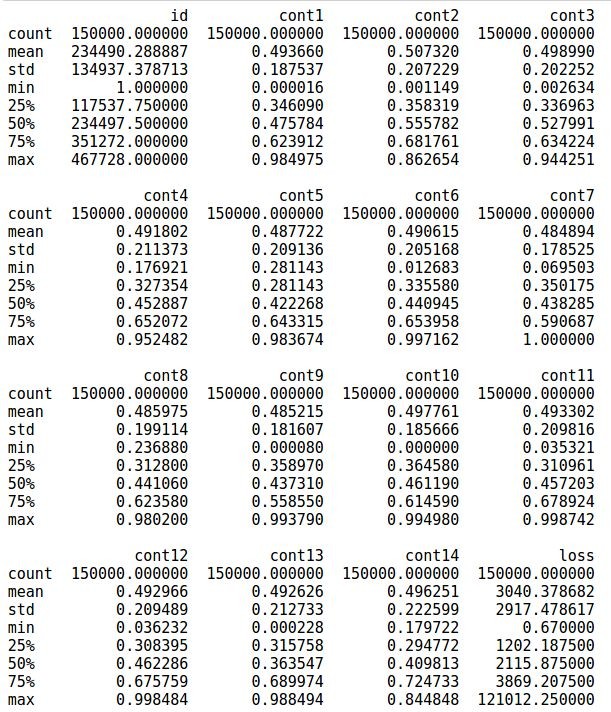


**Fig 2:Train\_data first five columns**

The test data is same with the exception of not having a loss column.

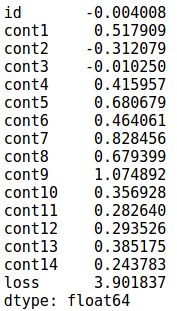
In case of continuous data, we can find the mean and standard deviation to check the outliers.

Seeing it, it doesn’t look like we need to scale it, it looks quite balanced. Although , loss column needs further inspection. Also ,the values are already between 0 and 1 so , we are not going to scale it.



**Fig 3: Mean, std, min, max, etc. of continuous features**

We will check for skewness now.



**Fig 4: Skewness of continuous variables**

This clearly shows that loss is positively skewed and needs to be rectified. Next, we look at the violin plots of continuous features.

The results drawn were -

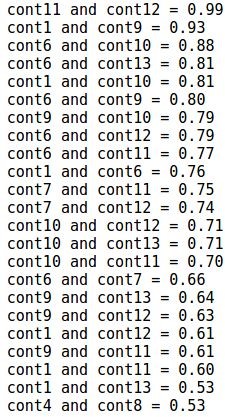
Cont1 has many values close to 0.5.

Cont2 has a pattern where there a several spikes at specific points. Cont5 has many values near 0.3.

Cont14 has a distinct pattern. 0.22 and 0.82 have a lot of concentration. Loss distribution must be converted to normal.

The necessary step to transform loss.

Next, we see the correlation between different features and print out the top entries.



**Fig 5: Top correlated features**

We see that top 5-6 entries are highly correlates which presents the opportunity to apply dimensionality reduction method like PCA.

Now, we count the number of different labels in each categorical feature.

Results are Below -

Cat1 to Cat72 have only two labels A and B. In most of the cases, B has very few entries. Cat73 to cat 108 have more than two labels.

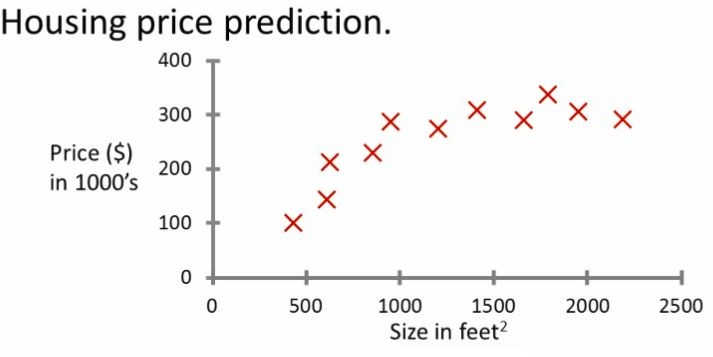
Cat109 to cat116 have many labels.

## Algorithms and Techniques

The very first algorithm to be used is linear regression.

### Linear regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. For example, consider that we have data about prices of houses. We take their sizes and plot it on x axis and their price on y axis. Then we plot our data. It looks somewhat as depicted in the image below.



**Fig 6: Graph between price and size of houses**

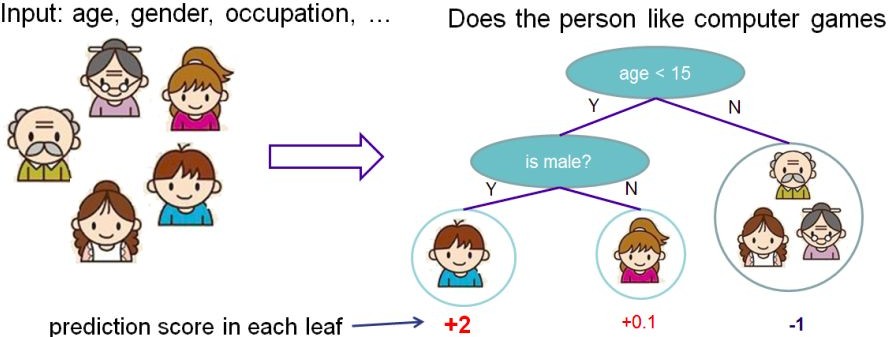
Now we can fit a curved line passing through these points. And use that line to predict the house given its size in the future.

It's a naive classifier and won’t work in our dataset since it has too much dimensions. So, it will be used as base classifier and we will try to minimize the loss it gives using a better model.

The second model we will be using is XGBoost.

### XGBoost

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. XGBoost stands for e**X**treme **G**radient **B**oosting. The term “Gradient Boosting” is proposed in the paper Greedy Function Approximation: A Gradient Boosting Machine, by Friedman.



**Fig 7: Depiction of trees**

A tree is simply a series of decisions. For example, as seen in the diagram, to check if a person like computer games, given features like age, gender,occupation etc. we construct a tree as if age is less then 15 then furthermore check if the person is male and outputs the probability. In gradient boosting,we take several weak learners(models whose performance is just better than chance) and combine them to get a strong model. To take our previous example again, let’s say that the decision tree shown in diagram have a performance just better than chance(>50% accurate). Let’s say we have constructed another decision tree model which asks whether the person is student or employed and its performance alone is also just better than chance. Then, we use both the trees and to combine their result, we use voting(give a weightage to their individual decision). This is tree boosting and taking many trees and combining them to give a greater performance than any tree individually.

Xgboost just uses the full extent of the machine it is running on.

The two reasons to use XGBoost are also the two goals of the project:

* 1. Execution Speed.
  2. Model Performance.

XGBoost is openly available to download and in sklearn.

Before implementing both the algorithms, we also have to preprocess the data. This will be discussed later.

## Benchmark

Due to hassle and limitations (like big dataset), 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 model to get a base MAE. 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.

First we run the linear regression. Train it on 150000 entries and then find the mae for the rest around 30000 entries. The result comes to be **1267.48**

The final score obtained using fine tuned XGBoost model is **1132.54788745**. This is better than both the base classifier score and the untuned XGboost score.

This is our base score and our aim is to beat this score, hopefully by a good margin.

# Methodology

## Data Preprocessing

First, we will normalize the loss column since during our exploratory analysis, we found that it is positively skewed. Apart from this, we perhaps does not require anything else to be done.

Next, we store the loss column in target variable and drop the id and loss column. We will take log of the target + shift.

Shift is another hyper-parameter. It sometimes gives better performance. I have taken it from the discussion linked here -

<https://www.kaggle.com/c/allstate-claims-severity/discussion/24611#141269>

After taking log, our loss is normalized.

Since we can’t use alphabetical categories, we factorize the columns cat1 to cat114.

We are going to use a part of the train data as test data. So we take 150000 entries as train data and the rest 38,318 entries as test data. In other words, roughly 20 percent of training data is taken as test set.

Since, there are entries which are in train data and not in test data and vice versa, we will first join them and then factorize and then divide them again into train and test data.

It doesn’t look promising to remove outliers because there aren’t much of them and those which are present, are not affecting much.

## Implementation

Although, the implementation is pretty straightforward, I would still walk through it. We have already preprocessed the data as described above. Now that we have our training and testing data, next step is to implement our base classifier (linear regression). We simply import linear regressor from sklearn and also mean absolute error. Then we train the model. Then, we predict the results of the test set. One thing to note here is that, since the target variable was transformed as log(loss+shift), we have convert the predicted result back. This could be done using the exponential. So we predict the loss and take their exponential and then subtract shift from it. Then we calculate the MAE between these values and the actual values.

Next we have to implement the XGBoost model.

First we does what we did for linear regression model, ie. Preprocess the data and divide it into training and testing dataset which we already did for linear regression.

Then, a simple XGBoost classifier is built using usual parameters without any tuning and compare the results.

There was no major part in the code which needed to be heavily documented. I have already documented bits and pieces almost everywhere in the notebook.

Since, it requires a lot of computing power , I ran it on google cloud, but for some reason , the server stopped before the code could fully run several times while doing tuning on the XGboost classifier, when I ran it on full dataset on google cloud, it always gave an error of bad\_alloc memory and even after hours.

## Refinement

The result obtained by naive xgboost was **1150.68**, which is really good improvement over our base classifier (**1267.48**).

Since , max\_depth is lower, we should check other values also , so we ran it for the third time with max\_depth in range 5 to 9 and the best performance is for **max\_depth = 6** and **min\_child\_weight = 8**. These are the final values for them.

Next, we tune gamma. We first give it the values from [0.0,0.1,0.2,0.5,0.8,1.0].the best value comes at **gamma = 0.0**. We fix that. Next, we check for col\_sample(0.3,0.5,0.7,0.9) and subsample(0.3,0.5,0.7,0.9) and get 0.5 and 0.9. Now, check for col\_sample(0.45,0.5,0.55) and subsample(0.85,0.9,0.95,1.0). Results come as **col\_sample = 0.45** and **subsample = 1.0.**

At last, we use these parameters and reduce the **learning rate to 0.01** to find the score on the test set now.

# Results

## Model Evaluation and Validation

Our fine tuned model gives us good results. Its evident how we have chosen it. After training the model with the good parameters, we find the MAE between loss and predicted test set. Since we transformed the loss using log, we have to take the exponential of it before finding the MAE.

The final score obtained using fine tuned XGBoost model is **1132.54788745**. This is better than both the base classifier score and the untuned XGboost score.

## Justification

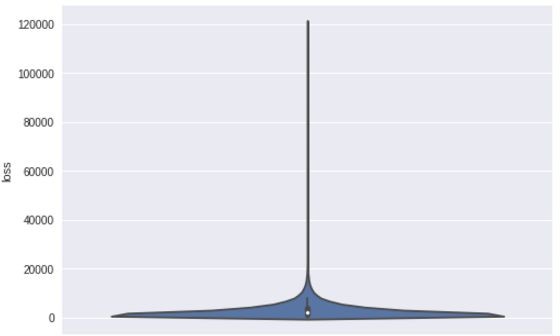
The model passes our base classifier’s performance.

As we don’t have enough knowledge about the domain the problem, we can’t say for sure if we have solved the problem.

# Conclusion

## Free-Form Visualization

The violin plot for loss was plotted and it can be clearly seen that it is positively skewed, concentrated only on the lower end of the graph and therefore, as long as we don’t transform it. We cannot have a low error.



**Fig 8: Violin plot of loss**

## Reflection

After doing the whole project, I have developed a model which is going to take about 130 features of any potential claim and gives the amount the insurance company may have to pay, although i can’t say for sure it is ready to be used due to lack of domain knowledge, however, the MAE is quite fine. It could have been better if we had more computing resources, but for the sake of academic purposes, I’d say we can stop for now.

This was the black box approach.

To go inside, i have created a simple linear regressor and used its MAE as base score, However, the aim was to beat that i deduced that XGboost is working much better than other models like SVM. Also, SVM is quite time taking so i went with XGBoost as next model. Even without parameter tuning it gave a better result. Next, we fine tuned the hyper-parameters and got the score of **1132.54788745**.

## Improvement

1. Keras also looks a potentially good model , however it would take even more resources than the current model with people on Kaggle reporting a training time of more than 10 hours and that too on good hardware. Therefore, i don’t believe we need it for academic purposes.
2. PCA could have been tried for feature reduction , however , since don’t have column names and there’s no point in visualizing the data in 2d (converting 130 features to 2 or 3 features), and it will result in loss of data, giving higher MAE, i didn’t tried it based on this intuition.

**References**

1. Kaggle, ”Allstate Claims Severity” (2017). <https://www.kaggle.com/c/allstate-claims-severity>
2. Allstate wikipedia page. <https://en.wikipedia.org/wiki/Allstate>
3. Supervised Learning Wikipedia page

<https://en.wikipedia.org/wiki/Supervised_learning>

[4] Intro to XGBoost -

[https://machinelearningmastery.com/gentle-introduction-xgboost-applied - machine-learning/](https://machinelearningmastery.com/gentle-introduction-xgboost-applied%20-%20%20machine-learning/)

[5] MSE and RMSE

<https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d>