Machine Learning Engineer Nanodegree

Capstone Proposal

Shikhar Bansal

(shikharbansal111@gmail.com)

August 18th, 2017

Proposal

Abstract

The project proposal is to create an ML model which accurately predicts claims severity.In order to do that , we have to compare the actual amount company had to pay and the amount predicted by our algorithm using the 116 categorical features and 32 continuous features given to us.this information is crucial as it automates the process of claim checking and is advantageous to both the customers and the company.Keywords: Allstate,Supervised learning, scikit learn, keras, XGBoost.

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.

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 datasource is:https://www.kaggle.com/c/allstate-claims-severity/data

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. 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.

Solution Statement

We want to understand the relationship between the 130(116+14) features and the loss. Although it looks straightforward but it isn't so. There are many features which, due to 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 squared 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.

Project Design

First method would be using classic linear regression. To do that , we will read the data, then convert the categorical features from alphabets to numbers. Then , we will make a cross validation set to test our model on. Second method would be to use XGBoost. Finally , third would use deep learning using Keras. Second and third method will be optimized as necessary. 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,tensor flow,Keras,XGBoost. Other libraries will be added if necessary.

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