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.

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 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.

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 may result in overfitting, so we may have to reduce the the features by PCA or some other method. We also have to find the relations between the features for that matter 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) and if required, deep learning (which is achieving state of the art in almost everything).

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). If the model trained comes below 1150 error, the model can be deemed useful and ready. The test set for this model is provided by Kaggle as a dataset. The testing will be done on this set and the benchmark model is hosted by Kaggle with which we can compare our model performance. 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