All-State Severity

PROPOSAL

Abdoulaye Diallo
224apps@gmail.com

Domain Background:

Because of new innovations in computing power, machine can predict certain output with models that can be trained using training data generated by users. There are three ways machine learning are used today: supervised, unsupervised and Reinforcement learning.In this paper, we will focus on supervised learning which is using the training data to get a hypothetical function. By minimizing the cost function we can accurately predict future results.

As a car owner, the last thing I want to have when involved in an accident is pushing paperwork with my insurer. Symmetrically, **AllState**, one of the biggest car insurers in the United State wants to predict the costs of an accident, hence the severity of claims.

Goals

1. Show creativity and flexibility by using technical algorithm which accurately predicts the severity of claims.
2. Dig out insights that can help AllState to give great customer service experience to their valuable customers.

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 the 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 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 link of this project taken from **kaggle** can be found **here**

File Description:

1) Variables noted with 'cat' are categorical, while those noted with 'cont' are continuous
2) **train.csv** - the training set:
    - Id - the id for each training set
    - Category variables cat1 to cat116
    - Continuous variables cont1 to cat14
    - The loss variables for each training set. And, these will be our target variables. They are only present in the train.csv since we will use the test.csv to predict them
3) **test.csv** - the test set which has
4) Sample_submission.csv - a sample submission file in the correct format.

Solution Statement:

We want to predict the relationship between the features and the loss. Due to the curse of dimensionality, a lot of the features might be overfitting and we need to cram our data into a lower dimensional space by using the principal component analysis(PCA) or other methods. Also, we can use kfold to test which model perform better for our case and then use the mean square error. We will use a linear regression and XGBoost.

Benchmark model:

Since this is a kaggle competition, the benchmark model is the best score provided for the testing data which is at 1109.70772 mean absolute error(lower is better)

So, let's run a linear regression classifier to get the base MSE. then, we can compare our model with the kaggle benchmark.

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), the same will be used for evaluation of models.

Project Design:

We will use a linear regression classifier by converting the categorical features from alphabet to numbers and run a cross validation set to test it. Also, we'll do the same for XGBoost. Finally, we will use keras to do a deep learning. The XGBoost and the DL will be optimized and whichever model gives us the lowest mean absolute error will be our final model. We will use python and a jupyter notebook which is hosted in AWSagemaker and the libraries will be pandas, scikit learn, seaborn, matplotlib, tensor flow, keras and XGBoost and others

References:

- Kaggle - https://www.kaggle.com/c/allstate-claims-severity/
- AllState - https://www.allstate.com/auto-insurance.aspx
- Machine learning - https://en.wikipedia.org/wiki/Machine_learning