Machine Learning Engineer Nanodegree

Capstone Project Proposal

Omid Heravi
December 20th, 2017

Doman Background:

Allstate, one of the most prominent insurance companies in US has provided data on several key indicators by which data analysis could help uncover the underlying pattern in hope of detecting the severeness of an insurance claim. Data is provided in the form of categorical, continues format. This problem in my opinion, is a very great challenge to tackle as the automation process for insurance claims improves so does productivity and efficiency which inevitably leads to happier customers and overall improved process. To tackle this data analysis issues, we can use several machine learning algorithms and improve from our baseline.

References to the actual challenge on Kaggle: https://www.kaggle.com/c/allstate-claims-severity#description

Problem Statement

The goal of this project is to use any available machine learning algorithm to fully and to the best accuracy predict the 'loss' for each given row in the test dataset.

Datasets and Inputs

This project will use comma separated publicly available dataset provided by Allstate. each row represent a separate insurance claim, with a total of 136 different columns. There are four different types of columns, excluding the sub-variety of each column. The 'id' column corresponds to the claim identification, the columns prefaced with 'cat' represent the categorical data, the columns prefaced with 'cont' are continues format data, and 'loss' is the dependent variable which will be predicted.

Solution Statement

The solution consists of attempting to use several machine learning algorithms, or EDA, in an attempt to achieve the best accuracy. Will try linear regression, XGBoost, multi-layer NN, and a KNN. The libraries which will be used are primarily; SciKit, Tensorflow, and Keras.

Benchmark Model

The proposed Benchmark models that I can mentioned would only be the current existing public Kernels on this challenge can be found on Kaggle; https://www.kaggle.com/c/allstate-claims-severity/kernels. However for the purpose of being explicit, and accounting for each proposed algorithm, the best public voted kernel scores for the requested 'loss' are 1278, 1745, 1169, 1168 for Linear regression, KNN, XGBoost, and Multi-layer NN, respectively. This represents an ambitious yet attainable value for 'loss' for which I will try to improve upon.

Evaluation Metrics

The proposed and final evaluation metric that have been chosen is the MAE, mean absolute error, of the predicted 'loss' and the actual 'loss'. Since the 'loss' data are integers, no other sophisticated evaluation metric is required and MAE as the evaluation metric would simply get the job done. The lower the outputted MAE value, the better since it describes a smaller difference among the predicted and actual 'loss' value.

Project Design

The first step of the process will include all the downloading, preparing, preprocessing and transformation if necessary. Second, will include visual analysis the data to better understand the shape and correlations among the dataset. Third, will implement the various ML algorithms suggested above, record the outputted values. Finally, assess the accuracy and evaluate the accuracy of the different models.