Udacity MLND Capstone Proposal
Manish Thakrani

Domain Background
The problem arises in the domain of particle physics. Physicists collide particles at high energy and observe the reactions that follow. These observations have led to deeper understanding of matter, and also to the discovery of previously unknown particles that exist only at very high energy.

In order to be certain of the fact, that new particles resulted from a given collision, the scientists collect a set of measurements from the collision and use these as the feature set. These features are then fed to a set of known physics functions to augment the feature space. The total set of features is then analysed by a pre trained model to see whether the collision resulted in a new particle of not. Since training data is very expensive to generate (large particle colliders take billions to construct and operate)[1], the models are pre trained on simulated data.

Good quality models can alleviate the need of augmenting the data sets (theoretically the model should be able to self learn the set of nonlinear functions) and directly affect the economics of running these experiments.

Problem Statement
The problem statement is to find a model that accurately classifies particle collisions that result in the generation of a new particle versus those that do not. We take a set of labelled particle collision data and train a classifier on this data for binary classification (new particle discovered vs not discovered). The problem is described in the paper 'Searching for Exotic Particles in High-Energy Physics with Deep Learning'

The paper proposes some solutions to the problem and also provides metrics to qualify those solutions. These will be used to benchmark the results obtained in this capstone.

Datasets and Input
The dataset for model training and validation is provided at
https://archive.ics.uci.edu/ml/datasets/HIGGS

The data has been produced using Monte Carlo simulations. The first 21 features (columns 2-22) are kinematic properties measured by the particle detectors in the accelerator. The last seven features are functions of the first 21 features; these are high-level features derived by physicists to help discriminate between the two classes.

[1] https://www.forbes.com/sites/alexknapp/2012/07/05/how-much-does-it-cost-to-find-a-higgs-boson/#514a36173948

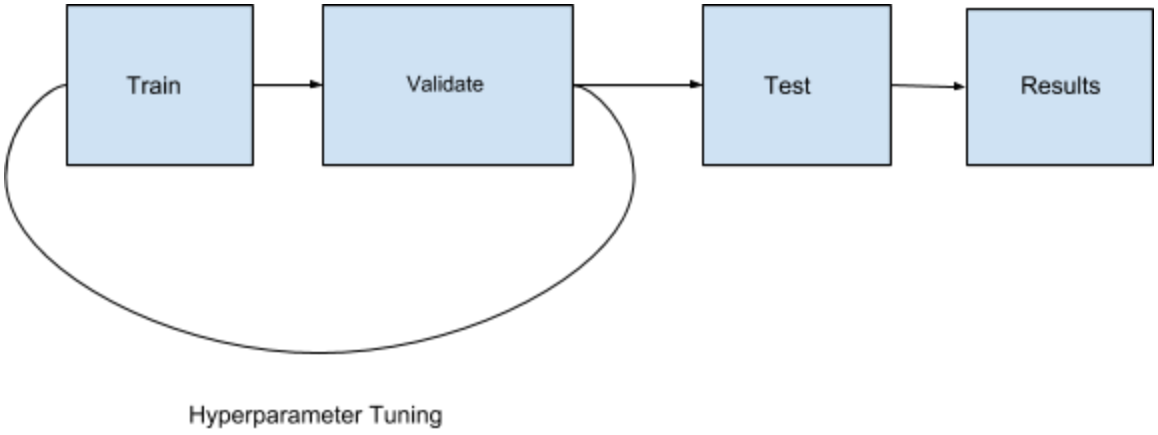
Solution Statement
I want to propose a solution that uses boosted trees (XGBoost), to construct a classifier that can discriminate between collisions that result in the generation of new particles versus those that do not. The expectation is that since ensembles can learn hidden non linear relationships in the data, we might be able to get a good performing classifier without having to augment the data with known nonlinear data mappings.

Since the original paper contains classification results obtained using a Bayes Decision Tree classifier and varying depth deep neural networks, these can be used to compare the quality of the results obtained using the boosted trees approach.

Benchmark Model
The paper contains results of running a Bayes Decision Tree classifier and varying depth deep neural networks for classifying the data. These results are used as a benchmark to evaluate the results obtained in this capstone.

Evaluation Metrics
I plan on using area under a ROC curve to evaluate the performance of my models. Since this is the same metric used in the paper, it will make it easier to compare the performance of ensembles to the models described in the paper.

Project Design
I expect the sample workflow for this task to follow the process as shown below:



The main algorithm of focus in training is boosted trees.
The main evaluation criterion is area under the ROC curve.
The main method for hyper param tuning is expected to be grid search.

I plan on using the raw dataset both with and without the augmented features to evaluate performance. My expectation is that, with right set of hyper params, the learning algorithm should be able to perform well even without the use of the augmented feature space.