AI Skunkworks Project

AutoKaggler in Spark and TF - Statistical Models

Project Manager: Karan Barai (barai.k@husky.neu.edu)

Expected Start Date: 18th March 2019

Expected Project Duration: TBD (For Spring 2019 students who need to submit a project, the deadline for tasks will be course project deadline)

Abstract:

To replicate H2O.ai automl function using Apache Spark and TensorFlow. These functions will include models such as Support vector Machines, Gradient Boosting Machine, Factorization machines and ensemble techniques etc. H2O.ai is gaining popularity but still has some limitations and lacks popular machine learning models, hence to overcome these drawbacks we can come up with our own automl function built on spark clusters using TensorFlow.

Deliverables:

1. A running implementation of statistical model assigned to an individual. Includes any one or multiple models like SVM, GBM, Factorization machines etc. Not limited to models listed, contributor/student can come up with other models as well.
2. Test cases to prove implemented model runs and gives desirable output.

Prerequisites:

Able to conduct basic machine learning tasks

Milestones:

No.	Task	Due Date
1.	Setting up the development environment and getting comfortable with spark and tensor flows.	03/25 – 04/01
2.	Develop data preprocessing function suitable to the statistical model you are implementing.	04/02- 04/08
3.	Developing an automl function for any one statistical model using tensorflow on spark cluster.	04/09 – 04/22
4.	Design test cases and create a documentation of the work.	04/22- submission date