**HIGH-FLIERS**

**Project Plan**

Project Object Objectives

To design an anomaly detection system capable of automatically catching

fraudulent transactions.

Project Scope

We will use a couple of approaches to evaluate. Since we have labeled data, we will design a classification approach. Additionally, given that fraudsters change their methods and tactics very quickly, we will also use an unsupervised approach. We

will begin with an unsupervised model and move into a classified approach.

Project Deliverables

We will deliver two approaches, one being unsupervised and the other supervised.

In the Unsupervised model, we will use the following data breakdown:

training = 60%

validation = 20%

prediction = 20%

Additionally, in the unsupervised algorithm we’ll be exploring is Random Cut Forest.

Intuitively, the model samples portions of our data and builds a small forest of

trees. After training, inference can be performed on a single data element. If the tree

changes beyond some learnable threshold, the data element is considered an

anomaly.

In the Supervised model, we will use the following data breakdown:

training = 60%

validation = 20%

prediction = 20%

Additionally, in the supervised method we will train a Linear-Learner model. Run

hyper-parameter tuning to understand the efficacy of those models for this scenario.

Project Constraint/Assumptions

* Using a dataset containing 6million records with 11 columns
* Does the order of the records matter?
* How much of the data will be used for training, validation and prediction?
* Do we need to aggregate the data time-wise (hourly, monthly, quarterly)?
* How many total models do we need?
* We will not be using “isFlaggedFraud” to compare the results of our models with the rules- based engine.