2023 Fall Session Seven: Intro to Machine Learning using Python in Databricks

1. **Are there challenges or opportunities when dealing with rare events/outcomes? Do you have recommendations for different evaluation metrics like F1-scores, Jaccard, etc.?**
   A - That is a rare event that can be addressed with either over sampling or under sampling the minority dataset or data sampling. It can also be framed as an anomaly detection problem.

   There are various modeling techniques for detecting anomalies. One mentioned during the demo was autoencoders, where we train deep neural network ware on normal data, and then it produces reconstruction error based on normal data.

   What our encoder does is takes an input and predicts, the same input, it reduces the dimensionality and then it brings it back to the same dimension. In the process it gets their sense of normal data, then it reconstructs the same input data. Then you compute the different bit of actual input and the predicted input which gives the reconstruction error. That is a measure of whether we can detect abnormal or rare events or behavior. Then we apply that auto encoder model to a new data. If that reconstruction error is much higher, then that is an indication that there is something anomalous or out of the ordinary.

2. **Will we have access to Databricks machine learning (ML) runtime that includes libraries such as Torch, TensorFlow? Also, will we have access to a graphics processing unit (GPU)?**
   A - Yes, we will eventually have this capability. Currently, there are limitations due to Ambari and Hive. Once we are off both of those, AutoML and other ML runtime clusters will be accessible. CMS will need to produce specifics around GPU access with ML clusters, but they would be available.

3. **Will OpenAI API be accessible to us?**
   A - This feature is still under consideration, we will inform the community when available for use.

4. **Do you use XGBoost Library also for these models ?**
   A - XGBoost can be used. It is very popular and has several models that can be used for classification or regression type problems.