Question week5

1/ How does the Gradient-Boosted Tree Algorithm work in Classification? How does Gradient Boost differ from AdaBoost and Logistic Regression?

Gradient boost Tree chains the creation of a first tree with the creation of a second tree on the residual of the first one. This continues iteratively. The trees are combined at the end.

Gradient boost Tree uses gradient descent costs method, Adaptive Boosting assign a weight to a result. GBT is more complex than AdaBoost

Linear regression only provide linear relationships, Gradient Boost can capture non linear relationships. The results of a Linear regression are easier to understand than gradient boost

2/ What is a Delta Lake and how does it offer a solution to building reliable data pipelines?

Delta lake is a storage layer situated on top of your data lake file storage. It allows to do ACID transactions (atomicity, consistency,isolation, durability) which allow to resolve conflicting accesses which help securing the pipeline

3/ When working with Pandas, we use the class pandas.core.frame.DataFrame and when working with the pandas API in Spark, we use the class pyspark.pandas.frame.DataFrame, are these the same, explain why or why not?

The pandas.core… is a pandas library running under spark. It uses one thread

The pyspark.panda is a spark library that distributes the work

4/ What is a Machine Learning Pipeline is and why it’s important? What are the steps in a Machine Learning workflow?

A pipeline is a sequence of steps to build a model. Typically, data collection, EDA, Feature creations, model research, model build, model deploy. It is important as it allows to standardize how to do the build.