Project Description

Nowadays, Virtual Learning Environments (VLEs)represent mature online education platforms allowing teachers to provide courses, which contain thoroughly managed resources and different engaging activities. VLEs allow creating various educational games, quizzes, e-learning tests, or prepare multimedia materials to attract students. Considering technological progress and increasing demand for online education, VLEs have perceived fast development over the past few years. VLE developers and providers have fully taken advantage of the internet and current web technologies and operate modern education systems to improve students' knowledge and skills. Educational approaches and methodologies tailored to the functionalities of the virtual platforms get rid of the limitations of traditional courses taken in classrooms and come up with higher flexibility in terms of where and when to take the courses online.

The goal of this project is to find a model, comparing several machine learning techniques and different kind of datasets, to identify students at risk in dropping out a class. As a result, suitable form of intervention at the individual e-learning course level can be applied in time.

Datasets:

1. MOOC dataset of Academic data

This dataset contains data of years 2012 and 2013 of 13 massively open online courses (MOOCs)offered by MIT and Harvard universities. One of the major recurring issues raised in academic literature is the consistently high dropout rate of MOOC learners. Although 335700 participants enrolled on these courses, the completion rate is 4.4% .

2. Demographic – Social- Economic Factors Dataset

This dataset contains data from a higher education institution on various variables related to 4424 undergraduate students, including demographics, social-economic factors, and academic performance.

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Machine Learning Methods used:

- Decision Tree Classifier
- Random Forest Classifier
- Logistic Regression
- SVM

Outcome:

I adjusted the two datasets to be composed of the same proportion of target data, and more specifically around 2200 graduate students and 1400 students who drop out. Next, I implemented the machine learning models mentioned above in both datasets. I compared the evaluation metrics of the two classifiers that had the best performance in each of the two datasets. As a result, academic data is far more capable of predicting students' dropout than the social- economic data of the students.

Deviations from your initial project plan:

Due to some google cloud issues I did not have the opportunity to use Dataproc to run Apache Spark.