German International University of Applied Sciences
Informatics and Computer Science
Dr. Nada Sharaf
Eng. Aya Abdallah
Eng. Omaima Ahmed

Big Data & NoSQL Databases, Spring 2024
Project
Due date is Thursday, May 23rd, 2024 at 11:59 PM
Submitted in groups of maximum 4 (can be cross tutorial)
NO exceptions

You are employed as a data analyst by a company that specializes in leveraging data to predict and enhance the success of crowdfunding campaigns. A dataset has been sourced from Kickstarter, containing detailed information about various crowdfunding projects, including their financial goals, pledge amounts, project categories, and outcomes. Kickstarter is a popular online resource used by ambitious entrepreneurs to pitch their products to the wide audience of the internet in an attempt to obtain funding. When someone launches a project on Kickstarter, the most important question for them is, "Will my project reach its goal before the deadline I've set? And also, "What can I do to make my project succeed?"

Data Description

- ID : Unique Kickstarter ID
- name : Name of proposal
- category : Specific category of project
- main_category : Parent Categories
- currency : Original currency
- deadline : Project expiry date
- goal : Goal of product in respective currency
- launched : Project launch date
- pledged : Amount of pledged before deadline
- state : State of project -- fail, success, ...
- backers : Number of backers for a project
- country : Country of project
- usd pledged : Pledged amount converted to USD (by Kickstarter)
- usd_pledged_real : Pledged amount converted to USD (by Fixer.io)
- usd_goal_real : Goal amount converted to USD (by Fixer.io)

Requirements

- Perform any necessary data cleaning & engineering that renders your data useable (i.e. handling missing values, duplicates, classification, transformation...etc.)
- Take note of the multiple attributes and perform 3 queries that provide different insights about the data (Please ensure to run these queries using one of your preferred tools: Spark (SQL or Dataframe), Cassandra, or MongoDB).
- Use the results for each of your queries to produce 3 different visualizations.
  *Hint: The more creative and insightful your queries and findings are the better! (higher marks for more fruitful results)
- Using SparkML, prepare data for Machine Learning by combining all the feature columns into a single vector column as input and produce at least 3 models to predict whether a project will be successful or not. Hint: To prepare Target Column: First, examine the project states and convert the "state" column into something usable as targets in a model. Drop projects that are "live" and count "successful" states as outcome = 1, while combining every other state as outcome = 0.

You are allowed to use any of the tools, frameworks and technologies covered throughout the course.

Deliverables

- Your Source Code: Your program should implement the specifications indicated above.
- Part of the grade will be on how readable your code is. Use explanatory comments whenever possible
- Please submit your work by compressing it into a zip file and sending it to **bigdata602.s24@gmail.com** with subject: “Project” and IDs of the team members in the body of the email).

PLAGIARISM IS NOT TOLERATED AND COPIED WORK WILL BE AWARDED 0 POINTS FOR BOTH TEAMS INVOLVED or IF YOU COPIED IT FROM THE INTERNET OR ELSEWHERE (NO. EXCEPTIONS.)!

Note: You will be asked for the reasoning of any actions, queries or decisions you have taken in the implementation of this project during the evaluations that have led to your answers/results