Data Mining and Visualization (83676)

Final Project

Part 1

Due: May 1st 2023

General

This project will have two submissions ,during which you will apply in Python the concepts you have learned in the course. The submission is in pairs and should include a notebook file (.ipynb) with the full code and a report (word or PDF) with the project description and explanations. The first submission constitutes 40% of the project final grade. You will be given a dataset, and in this part, you will investigate the dataset and apply preprocess techniques. The output of this part will be an input for the next part, which will be a classification task.

The grade evaluation takes into consideration:

- Techniques used: Did you select appropriate and diverse techniques and justify why?
- The process you followed: Is it correct (given the techniques you used), did you describe it well?
- Interpretation of results: Did you correctly understand and interpret the results you obtained?
- Quality of writeup: Did you present your work well, in an understandable and usable manner?

Problem Description

Startups are new businesses that aim to develop a scalable economic model. Although many fail, some become successful and influential, known as unicorns. Predicting the success of a startup is crucial for investors seeking rapid growth opportunities. In this project, students will analyze a data set of startup companies to accurately classify which ones worked and which ones didn't.

The Goal

The aim of this project is to employ data analysis techniques to forecast the probability of a startup transforming into a successful enterprise or failing to remain operational. Success is defined as an event in which the company's founders receive substantial monetary compensation through either M&A. (Merger and Acquisition) or IPO (Initial Public Offering). Conversely, the failure of a company will be deemed if it ceases operations.

Data Attribute's:

The data contains industry trends, investment insights and individual company information.

1. ID- id of the startup
2. state code: what state the startup is located in (from the USA).
3. latitude: latitude of startup.
4. longitude: longitude of startup.
5. zip code: zip code of the startup.
6. city: where the startup is located in the state.
7. Name: name of the company.
8. foundation_date : date of when the startup was founded.
9. First funding date: the date when the company received its first funding.
10. Last funding date: date when the company received its last funding.
11. First funding age : when was the first funding in units of years
12. Last funding age : when was the last funding in units of years
13. first milestone age : information on when milestones were first performed in units of the year.
14. Last milestone age: information when the last milestone was done in units of years.
15. Connections : how many relationships does a startup have. For example, a startup can have relationships with accountants, investors, vendors, mentors, etc.
16. funding rounds: how many funding rounds the company had (for further understanding of funding rounds :" https://www.forbes.com/sites/alejandrocremades/2018/12/26/how-funding-rounds-work-for-startups/?sh=48f708117386 ")
17. total Funding: how much money did the company get in fundings.
18. Milestones: number of milestones the company passed so far, for startups milestone is a tracking mark. Just like a milestone on the side of a road marks how far you've gone, a milestone in startups tracks progress as an startup grow and implement their plan.
19. In CA: is the startup in California.
20. in NY: is the startup in New York.
21. in MA: is the startup in Massachusetts.
22. in TX: is the startup in Texas.
23. in other state: is the startup in another state.
24. category: what category does the startup relates to
25. is software: binary if the startup works with software.
26. is web: binary (same as software for web).
27. is mobile: binary ..
28. is enterprise: binary.

29. is advertising: binary.
30. is games video: binary.
31. is ecommerce: binary.
32. is biotech: binary.
33. is consulting: binary.
34. is other category : if non of the other categories .
35. round-A: has gotten to round A (further read in the link of funding rounds)
36. round-B: has gotten to round B
37. round-C: has gotten to round C
38. round-D: has gotten to round D
39. Average group size: what is the average amount of workers in each group of the startup.
40. In Top500 : binary if the startup is in the top500 startups.
41. Target: did the startup succeed or not.

Instructions:

In this first part, you will initially explore and understand the given data set using statistic method and visualization tools and then apply the preprocess that will be used for the classification task in the following project part.

You should implement the following sections and add more necessary actions, analytics and visualizations to enrich your work.

- Show the data information, e.g., types of attributes, the attributes values etc.
- Show the data statistics, e.g., distribution, skewness, median and more.
- Show and explain attributes correlations.
- Show and explain visualizations that present interesting insights from the data, e.g., identify relations, trends, the effect of an attribute on the target variable etc.
- Data cleaning - check for each one of the problems and take care of them properly, e.g., missing values, inconsistent etc.
- If necessary, add and/or delete attributes.
- Data reduction - apply at least one of the methods we learned.
- Data transformation - apply the appropriate methods to the required attributes, e.g., normalization, discretization etc..

Additionally, you should use methods that have not been demonstrated in class. You should explain in the report all the steps you made.

*Students will lose points from their final mark on the assignment if they use internet published analysis of the dataset in the project.