DOPPELGÄNGER
SYNTHETIC DATA GENERATOR

INDIVIDUAL REPORT FOR **PRASHANT CHAUDHARY (A0213485E)**

GRADUATE CERTIFICATE IN PATTERN RECOGNITION SYSTEMS – PRS GROUP 19

Your personal contribution to project?

Project Journey:

A little bit of history before I share about my personal contribution and learnings – after our first class in the semester, I thought of forming a team at an early stage and so I first asked Anirban, Ankeit (Taksh) and then Nirav, and soon we all formed a team.

Initially, I was thinking of implementing some vision-based intelligent system as I already had some basic framework and ideas ready but after few rounds of discussions, we finalized to proceed on Synthetic Data Generator. Initial idea was floated by Nirav with a good business case and concrete knowledge on how to implement it.

What I personally liked about the idea of Synthetic Data Generation was the flexibility to not only use different types of input data formats (Numeric, Image or Time-series etc.) but also there are different types of problem categories (Regression/Classification) and implementations using various Machine and Deep Learning approaches available (covering supervised, unsupervised and semi-supervised learning methods). Depending upon individual interests, team members could choose different areas. And it truly turned out to be the case. As everyone got something important to work on (both as individual or as a sub-group) to contribute in this project. Anirban took charge from ML expertise perspective helping to resolve algo bottlenecks and Taksh actively took care of deployment concerns because of his DevOps expertise. Nirav provided good guidance and solution insights during the entire project journey. Overall, it was a good learning and working experience with everyone.

Project Work:

1. **Module 1** (Numeric Data Regression) –
    a. I researched on various numeric and timeseries datasets for POC (later I dropped timeseries implementation to keep overall scope reasonable).
    b. Once we finalized on life-expectancy dataset, I worked on data pre-processing steps as it required a lot of data cleaning before we apply PCA.

2. **Module 2** (Image Data Classification)– After we zeroed down on MNIST image data, I helped Nirav on designing / training of VAE model and worked on CNN Evaluator/Validator parts with Anirban.

3. **Entire Application/UI** – I designed and developed the Application & UI using various technologies – Python, Flask, Dash & React.JS etc. Its intuitive tabbed interface to depict workflows and show dashboard, charts and visualizations adds to the user experience. App is fully productionize to execute same Python functions used in ipynb and display outcome on the Dashboard in realtime in the form of charts and images.

4. **Application Deployment / Testing** – Paired with Ankeit to deploy the application package in production env and its testing.

5. **Installation and User Guides –** I created the installation guide to depict application structure and user guide for a simple functionality walk-through.

6. **Marketing Video** – I conceptualized and prepared 2 min marketing video advertisement (comes in the beginning of our video presentation). I also paired with Nirav to customize the look and feel of the Application from demo perspective.

7. **Team Presentation -** We worked as team to complete, review and deliver the mid-term presentation.

8. **Team Project Report –** I contributed to few sections of final project report and did peer reviews.

Project Coordination -

1. Apart from the team formation, I helped with project management aspects as well like doing regular catch-ups and progress tracking. Also -
2. Setup & managed team repository, GitHub organization and Google-drive locations for easy collaboration.
3. Defined project implementation roadmap into three phases (architecture goals) -
    a. Phase 1 for completing different case studies on Google Colab notebooks.
    b. Phase 2 for completing App/UI focussing on user experience as a goal.
    c. Phase 3 to deploy/productionize the final product along with various documentations and guides.
4. Resolved various bottlenecks by engaging with teammates to solve some problems together.
5. Took ownership to pick-up and finish the critical items that were on back-burner like App/UI, video ad or guides.

What you have learnt from the project?

1. I believe the business case of Synthetic data is quite strong but in practical it is equally hard and challenging. This is because of inherent associativity of domain expertise with data as well as technology challenges in pre-processing, generation and quality of synthetic data. But this project also gave us a good insight on how to apply an interface driven approach where various

low level implementation strategies can be applied as "a detail" like data cleaning or different ML/DL models under a common framework/architecture.

2. Simple concepts like PCA or auto-encoders are quite versatile. While doing PCA in the lectures I did not imagine it could also be used to generate good quality synthetic data.
3. Also, auto-encoders can do so many things like data generation, compression, filtering/de-noising data or anomaly detection. I specifically got mesmerized with its generative aspects that have immense possibilities like music or video generation.
4. With project I got a lot of hands on experience while training CNN models or ML techniques like XGBoost, Linear Regression, Random forests or data analysis, visualization and pre-processing aspects as well.
5. Developing a complete Python based ML solution product– from backend data processing to show visualizations at the UI. It was a great addition in my tech arsenal.

How you can apply this in future work-related projects?

1. I work in a Global bank's Data warehouse system where almost all trades happened during the day are captured. Although there is humongous amount of data but due to data sensitivity concerns, it cannot be freely shared even within the Bank's various sub-divisions, downstream systems, or testers/developers for running test cases in the lower environments. Here, Synthetic data Generator can solve these data privacy concerns by generating sensitive columns data and make it available for further usage.
2. I have also developed (and now maintaining) Data Quality framework for our data warehouse. It hosts thousands of data quality checks, that are created by domain experts to find errors while data is ingested. But these hand crafted rules can only go so far due to huge number of dimensions in the data and its volume. Here, we can utilize auto-encoders to detect anomalies in data by training auto-encoders only on the actual valid data and let it then transform incoming data where it can help finding the human errors or even fraud detection making Data Quality engine much more effective.
3. Auto-encoders and PCA both are also useful in dimensionality reduction and hence it can help to reduce the dataset dimensions for further analysis/usage by rule authors.
4. There is huge potential for peta-bytes of data stored in our data warehouse for doing machine and deep learning. A full-fledged Python/Flask solution similar to what I developed in our project can be very useful for our end-users, senior managers, business-analysts, to let them visualize underlying data and help data-scientist & developers to build ML/DL models for further improving our data engineering & analytics solutions.

--- END ----