Who is Dhruv Shah?

Dhruv Shah is an aspiring Data Scientist and Data wizard who is in his final semester pursuing a master's degree in applied data science, from Syracuse University, New York. He has completed coursework's like Natural Language Processing, Applied Machine Learning, Quantative reasoning for data science, Building Human Centered AI Application, Responsible AI, Database Management systems and Deep Learning in Practice. He also done a Bachelor's Degree in Computer Engineering from NMIMS University in Mumbai, India.

What is his work Experience?

1. His Most recent work experience is at *Center for Computational and Data Science* where he worked as a Data Scientist for a period of approximately 9 months starting from October 2023 to May 2024. Along with his teammates he helped building an NLP pipeline which helps in classify whether a news is fake or genuine. They implemented LSTM model on top of BERT transformers achieving an accuracy of 88.12%. why was this approach taken because a key feature of LSTM is that it has a larger context window compared to other models and this can be beneficial understanding longer phrases capturing the overall sentiment of a phrase, he also implemented various NLP techniques to analyze syntactic, grammatical, and semantic features of the dataset. He also helped in conceptualizing a neo4j graph to explain a widespread of misinformation to the non-technical stake holders of the projects.
2. He has also worked as Machine learning Intern at Talakunchi Networks Pvt Ltd, Mumbai where he Formulated and deployed an **enterprise-wide risk assessment model** that quantifies **asset risk scores** and identifies **critical vulnerabilities** impacting overall **security posture**. He built the model using **random forest classifier** utilizing various parameters to categorize vulnerabilities, compute risk scores on a 10-point scale, and predict post-remediation impact, achieving **77.63%** accuracy while reducing **manual assessment** time by **40%**. He also Innovated an **automated Python-regex** script for a premier Indian bank to extract critical data from mobile and web application dumps, improving data integrity and reducing manual processing time.
3. He has also worked as a ***Security Intern*** at Talakunchi Networks Pvt Ltd, India during Dec 2020 – July 2021, where he Orchestrated vulnerability assessments across **1000+ network endpoints** using **Tenable Security Center** and **Nessus Professional**, identifying and categorizing critical security gaps he has also Executed comprehensive **black box** penetration testing on **15** web applications, discovering and documenting **30+** high-severity vulnerabilities including **SQL injection** and **XSS vulnerabilities**. He also Formulated a custom regex-based compliance policy files, streamlining security validation processes aligned with client specifications.

What are his projects?

1. ***Predictive Solution for Enhancing Preventive Care Engagement*****,** Humana Sept 2024 – Oct 2024 Pioneered a team effort to tackle Humana's 55% preventive care non-

engagement among LPPO members, a critical factor in the decline of **CMS Star ratings** and **quality bonus payments**, Our breakthrough approach propelled us into the **top 50** of over **400+** competing teams across the nation. Designed and implemented a pipeline integrating **14 datasets and 250+ features** for over **1.5 million members**, achieving **69% accuracy** in predicting disengaged members through a high-performance stacking model (XGBoost, LightGBM, CatBoost). Created a multi-component strategy to encourage regular check-ups and health tracking. Projected to **increase engagement by 10%**, directly improving CMS Star ratings by 0.5 and Medicare Advantage bonus payments.

2. ***Context-Based risk Prioritization of Vulnerabilities*****,** Talakunchi Networks Pvt Ltd Aug 2022 – May 2023, He Identified a critical gap in existing vulnerability prioritization models and developed a machine learning model in Python that assesses **32 distinct vulnerability factors** using the Random Forest algorithm, achieving **71.38% accuracy in threat identification**. **He also** Utilized **NLP techniques and Data analysis techniques** to handle and analyze **over 100 million vulnerability data points**, ensuring a comprehensive and data-driven approach to vulnerability risk. **I also** Presented my work at ICICC 2022 in the Springer LNNS publication; introduced techniques utilizing machine learning which improved threat identification accuracy.

3. I developed a medical chatbot using Retrieval-Augmented Generation (RAG) architecture to assist users in identifying potential medical conditions based on their reported symptoms. The chatbot not only diagnoses the condition but also recommends an appropriate medication to manage or alleviate the symptoms.

   To ensure accuracy and relevance, the chatbot was trained on two extensive datasets:

   - **Gale Encyclopedia of Medicine**: A comprehensive resource covering a wide array of diseases and their associated symptoms, spanning six volumes with over 700 pages per volume.
   - **Medicinal Data**: A dataset containing information on over 2 million medicines, aligned with the conditions outlined in the Gale Encyclopedia.

   The model was built using **BIOMISTRAL 7B**, a powerful language model tailored for medical applications, and further enhanced with **MedPub Embeddings**, a specialized set of medical embeddings optimized for chatbot use.

   Given the sensitive nature of healthcare information, the chatbot integrates robust **user authentication** features to ensure privacy and data security. Additionally, users can view their previous conversations, offering a personalized and continuous support experience.

   This solution combines advanced natural language processing with domain-specific medical knowledge, providing users with accurate and reliable health guidance while maintaining the highest standards of privacy and security.

4. This project focused on developing an anomaly detection system for IoT devices using sensor data. The system analyzed data from multiple sensors measuring environmental parameters like temperature, humidity, gas levels, light intensity, and pressure across eight IoT endpoints. The team employed sophisticated data processing techniques, including timestamp transformation and cyclical encoding, to handle approximately 9.9 million data points.
   We utilized the Isolation Forest algorithm for anomaly detection, successfully identifying abnormal patterns in gas and temperature readings. For predictive modeling, we implemented both ARIMA and SARIMA models for forecasting, along with Bisecting K-Means clustering to classify sensor behaviors into normal and anomalous patterns.
   We successfully demonstrated the effectiveness of their approach in detecting device anomalies and predicting future sensor values. The system's practical applications include early detection of malfunctions and potential security risks in technological environments, enabling proactive maintenance strategies and enhanced operational efficiency.

My Skills:

TECHNICAL SKILLS

- **Programming Languages**: Python and R
- **Data Analysis**: Pandas, NumPy, SciPy, Matplotlib, NLP and Seaborn
- **Data Visualization**: Tableau, Power BI, ggplot2 and Seaborn
- **Machine Learning & AI**: Supervised/Unsupervised Learning, Deep Learning (CNNs, RNNs) and NLP (BERT, GPT)
- **Data Engineering**: Spark, Hadoop and SQL
- **Tools:** MS Excel, MS PowerPoint, MS Word, and Postman.
- **framework:** Lang chain and Hugging Face