PAVANI MARAGOWNI

Contact no: +1 (240) 584-1232; e mail: pavanimaragowni@gmail.com

SUMMARY

- **Experienced in R and Python programming languages.**
- **Proficient in planning and architecting end to end data science projects.**
- **Expertise in engineering and building models based on Regression, segmentations, and clustering.**
- **Developed scalable Machine Learning and AI platform projects.**
- **Worked with Text Mining, and Natural Language Processing (NLP)**
- **Used Statistics and Probability in decision modelling of data projects.**
- **Working Knowledge of Machine Learning including Supervised, Unsupervised and Deep Learning algorithms**
- **Worked on both Deep learning and LLM models such as GPT-3.5, GPT-4 and Langchain models.**
- **Worked on SQL and NO-SQL databases like MongoDB, Redis**
- **Expertise using AWS, MS Azure environments, RabbitMQ, Flask**
- **Strong knowledge in Intelligent Document Processing (IDP) and AI Platforms**
- **Working experience in domains such as Healthcare and Market Research**
- **Exposure working cross-continents across US, Europe, and India work settings.**
- **Experience in coordinating across business and technical teams and shipping high quality products/services.**
- **Strong Analytical, organizational and Leadership skills**

TECHNICAL SKILLS

Programming Languages Python, R, SQL	Python Libraries NumPy, Pandas, Scikit-Learn, Matplotlib, OpenCV, SciPy, TensorFlow, Keras, Pymongo, Minio, ggplot2, Seaborn, NLTK, Pika, BeautifulSoup4, Flask, Fastapi
Data Science Predictive Analytics, Descriptive Statistics, Data Mining, Data Wrangling, Data Visualization, Feature Extraction, Pattern Discovery, Topic Modelling and Sentiment Analysis, NLP, Computer Vision, Deep Learning and MLOps	Machine Learning Regression, Classification, Clustering, Causal Discovery, Neural Networks and Deep Learning, Linear Regression (Rigid, Lasso) & Logistic Regression, Decision Trees, KNN, SVM, Time series analysis, K-Means Clustering, Ensemble methods (Random Forest, Ada Boost, Gradient Boost, XG Boost), Predictive Modeling with Time Series (AR, MA, and ARIMA), Association rules, CNN and RNN
NLP & GENERATIVE AI TensorFlow, spaCy, PyTorch, NLTK, Regex, CoreNLP, GPT, LangChain, BERT models	Computer Vision OpenCV, Scikit-Image, OCR (Microsoft, ABBY, Marx V3, Google Vision), ICR
Data Visualization Tableau, PowerBI, Matplotlib, Seaborn, ggplot2, Plotly	Storage S3, EFS, Minio
Version Control GitHub, Git, Gitlab	IDE RStudio, Jupyter Notebook, PyCharm, VS
Databases MongoDB, Redis, MySQL, Oracle SQL, MS Access	Technologies & Tools RabbitMQ, Postman, JIRA, Zendesk
Cloud Technologies AWS, Azure, Amazon S3, EC2, RDS, GCP	Operating Systems Linux, Mac OS, Windows

PROFESSIONAL EXPERIENCE

Architect (AI/ML & NLP)
American Express, Phoenix, AZ ***June 2024 - Present***

Project: Gen AI Marketing Content Generator
Role: Developing a Marketing content generator (MCG) for the Amex Marketing team enabling Generative AI processor. Working with GPT 4o model to build Brand copies that align with the company brand guidelines and marketing dam requirements. Conducted Prompt engineering to use internal guidelines and user input to generate the best output from the LLM models. Researched and implemented fine-tuning techniques for GPT-4 models to align outputs with business compliance and operational standards. Built Python-based testing scripts to validate GPT-4 model performance, focusing on accuracy, compliance, and user-specific scenarios. Implemented OCR technologies to extract data from scanned documents, images, and PDFs, achieving high accuracy across diverse document types as part of native text extraction. Integrated AI-powered image recognition tools to generate metadata and improve content searchability within systems. Managed and optimized Amazon PostgreSQL databases for efficient querying and storage of extracted data. Leveraged advanced SQL queries for analytical operations and data performance optimization. Developed custom solutions for extracting structured data from PDFs and PowerPoint files using Python and AI libraries. Automated document parsing workflows, integrating seamlessly with downstream data processing systems.
Utilized AWS cloud services for scalable data pipelines and ELK monitoring dashboards. Automated CI/CD pipelines using GitHub Actions, Jenkins, and XLR templates to streamline deployment processes. Monitored GenAI pipelines post-deployment, resolving bottlenecks and improving model performance based on business feedback.
Conducted compliance reviews for LLM outputs in adherence to Model Risk Management Guidelines, focusing on data privacy, PII detection, and relevance.
Collaborated with teams to gather requirements for GenAI solutions, aligning technical designs with organizational objectives. Created comprehensive documentation for model architectures, training datasets, and compliance protocols to promote transparency.

Environment/Tools: Google Cloud databases for images and AWS S3 for pdf and doc file storages. Python- Pycharm for development, LLModels (GPT-4, LangChain wrapper), OCR Models (Teseract, Microsoft OCR)

Solution Engineer (AI/ML & NLP)
Instabase, San Francisco, CA ***May 2022 – May 2024***

Project: Instabase AIHub App.
Role: Developed AIhub Aplications for Loss Runs, Bank statements and PayStub documents. AI hub uses LLM models to recognize, classify and extract important fields from the documents provided. Used GPT 3.5 and 4 to extract and build solutions from Loss Run documents for **Nationwide**. Have used Form Recognizer to recognize and extract table information from Bank statements.

Project: E2E solution development for **Cass Information systems**.
Role: Handled Classification of documents for Cass Waste, Utilities, Telecom, and freight services and developed a solution that enables split classification. The E2E development included annotating the documents, model training, processing the invoice files using OCR, refining the extractions, and sending it to the human review page using validations checkpoints. The manually reviewed or straight through processed docs are then sent to downstream API consumption.

Project: POV solutions for **Sonic Automobile**, **Nationwide**, and **ColPal**.
Role: Worked on multiple POVs for building deep learning and LLM solutions for Sonic Automobile invoice processing, Nationwide Loss runs data extraction and ColPal Remittance advisory documents. Improved LLM solutions with continuous prompt engineering. Built solutions and tested solution accuracies to yield the best automation and extraction accuracies.

Project: Instabase Marketplace Apps.
Role: Handle Marketplace App development on the platform for Tax documents, Checks and Invoice processing. Work with Business leaders to provide end to end solutions using the IB Image and Document processing platform. Work with data

engineers to annotate the data for model training. Training OCR models for better accuracy on a variety of Identity and photo-proof documents such as Passports, Drivers Licenses, and W2. Have worked on ABBY, Microsoft OCR, Google Cloud and MarxV3 OCRs to develop solutions for Bank statement processing for a large Bank in the US.

Project: Classifier model retraining pipeline
Role: Created a package that provides a pipeline which automates the classifier retraining task based on the human review results. It takes the output from flow review API & builds an annotation set to train an existing classifier model. A post-flow trigger schedules the classifier training job also provides email updates on the status of the pipeline. This package helped provide value to the customer for real time feedback during development. This package is also used for retraining a classifier depending on flow results, auto-improvement of classifier with CI-CD in production environments, scaling up the classifier with new classes, and improving the classifier based on validation results & edge cases.

Developed machine learning models using TensorFlow and scikit-learn for predictive analytics and pattern recognition. Designed Python scripts for advanced data transformations and API integrations. Ensured secure data handling by implementing IAM roles, VPCs, and encryption mechanisms within cloud environments.

Environment/Tools: S3, EMR, and python- Anaconda: Jupyter notebook, Libraries (Deep Learning: Tensorflow, PyTorch, Bigdata: Pyspark), LLModels (GPT-4, LangChain), OCR Models (Teseract, ABBY, Microsoft OCR, Google Cloud)

Datascience-AI-ML Consultant

Echotech Software Solutions, Jacksonville, FL ***March 2019 – May 2022***

Project: PHealth Assist App
Role: PHealth Assist is a medical application which features personalized discharge summary plans for patients. Physical Patient discharge summaries are digitized, OCRed and extracted to be sent to downstream user interface screens. Has a unique gamification with scores for doctor appointments and Physician assistant interactions. Built a Question Answering model using Deep learning for symptoms, and their medical treatment. The model answers questions based on the Discharge summary of the patient.

Project: Oversee product roadmaps
Role: Oversee analytics related to engagement of app features and new feature roll outs. Used experimentation to drive product engagement and improve customer experience. Was responsible for suggesting and creating different UI screens for healthcare users and patient users. Contributed to enhancing PHI data masking for HIPAA compliance. Analyze and size opportunities to drive prioritization and strategy. Hypothesize and recommend product enhancements by leveraging the data, engagement, and business. Built dashboards to track the progress of feature rollouts and the long-term success of the products.

Environment/Tools: AWS - EC2, EMR, and python- Anaconda: Jupyter notebook, SpaCy, pronto ontology libraries, Libraries (Regression/ Classification: scikit, Deep Learning: tensorflow).

Machine learning Engineer

Exponential AI, Hyderabad ***August 2016 – Jan 2019***

Project: Developing an AI platform-as-a-service with Machine learning and Analytics capabilities.

Role 1: Build a microservice for running R based predictive models through the cloud-based platform.
Approach: Scripted and built classification and regression models for supervised and unsupervised learning using R language. Implemented logic to execute models scripted in R language on the platform using python program to execute from the cloud.

Role 2: Risk score measurement depending on patient engagement.
Approach: Written python programs to implement matching of user profiles with the risk based on engagement and historical data. Risk scoring, matching, and sending risk scores to the outcomes microservice. Published to the microservice APIs using Python stacks. Used GitLab to push changes. Unit testing and SonarQube code testing for code coverage and quality.

Role 3: Worked as an Agile Scrum Master for product development.
Approach: Used *Atlassian Jira* as an issue management platform that allowed team to easily manage their tasks throughout the lifecycle. Identifying key items to be delivered and planned sprints, writing user stories, assignment, and discussion with team on tasks and sub-tasks for the user stories, participated in daily stand ups, tracking sprint tickets, and clearing Jira tickets with proper comments and videos for stories before completing the sprint tasks.

Role 4: Process documentation
Approach: As part of GXP compliance, helped in product and process documentation. Used *Atlassian confluence* as a wiki page for product documentations including 1. Document platform features that are being built 2. A walk-through document for developing solutions on the platform. 3. Document the capabilities and requirements of each micro service. 4. Documentation for a solution engineer and business user guides.

Role 5: Working on a POC to deploy the platform learning service for external solutions.
Approach: Worked on integrating the platform learning service to pick external pickled model binaries coming from an external source/solution. Run the model through custom model jupyter extension of the platform learning service. The output and insight generated are available for review, feedback and download to the business end user through the business user UI screens.

Environment/Tools: AWS, Azure ML, IBM Watson, Teserract, python- (openCV, tensorflow, NLTK, SPaCy, Flask, Scikit, Numpy, Pandas), R Studio, Atlassian JIRA, confluence.

Data scientist

Nielsen India, Bangalore ***April 2015- April 2016***

Project: Sample design for Cash slip store intercept of retailers.
Role: Determine Basket penetration of category and brands for country by retailers. Slot weights calculation, slot designing and session designing for data collection as part of Nielsen Retail Measurement Services (RMS).

Approach: Calculate Basket penetration at category and brand, category and pooled category levels from customer data and calculate target number of slips, slot weights based on universe data for sampling based on micro-representativity and session design. Presentations of the analysis to business leaders and retail clients to give insights.

Other projects: Universe Estimates, Sample Design, and causal / promotional data estimations for RMS projects. Consumer Panel studies for store visit behavior (Nielsen consumer Panel services).
Environment/Tools: SAS environment, Statistical Standard error and Basket Penetration calculation, Segmentation, regression analysis, R-RStudio, sqldf, CART, Python- Numpy, pandas, scikit. Analysis using MS-Excel, **R, SAS, MATLAB**, and python tools like matplotlib, seaborn.

Programmer/Data Analyst

Objects consulting solutions, Hyderabad **January 2010 – March 2015**

Project: Milward Brown Market Research Analysis
Role: Statistical analysis on Brand and Ad market survey data.

- **Coding and Optimizing R-code** for Statistical tests on Brand and Ad market survey data.

- Involved in importing, cleaning, transforming, validating, and modeling data using **R** as a tool.
- Also used Advanced Microsoft Excel and other statistical tools and was responsible for understanding and making conclusions from the data for decision making purposes.
- Worked on data input screens and data collection screens.
- Made presentation of data in charts, graphs, tables, and knowledge of using data visualization tools like Tableau.

Research Technician II

Fred Hutchinson Cancer Research center, Seattle, WA **February 2007- October 2008**

Project: Determining the polymorphic coding sequences in genes giving rise to minor Histocompatibility antigen (Very important factors in Graft versus Host Disease developed because of donor blood rejection during blood transfusions in leukemia patients). We were interested in looking at the polymorphic coding sequences using **bioinformatics** tools using **text mining algorithms** for gene sequence locations.

Assistant Scientist

Celera Genomics, Rockville, MD **April 2006-January 2007**

Project: Interacting with clinical sites to facilitate sample procurement, assist with the QC for tissue processing, tissue processing of samples, protein assays, Isolation of Nucleic Acids from clinical samples (RNA, DNA)

INTERNSHIP PROJECT EXPERIENCE

Internship

International School of Engineering (INSOFE), Hyderabad **July 2014-December 2014**

Project: Predictive Analysis on Patient Non-Adherence

Aim: In the health care industry non-compliance with medication regimens is a huge problem with socio-economic consequences. Not just losses to pharmacies but the patients also fall sick and there is a healthcare expense which amounts to 100 to 200 billion USD annually. My project helps in targeting those members who are at risk and predict their non-adherence patterns.

Approach: Build a model that can account for the risk of non-adherence per patient and predict if the patient is likely to non-adhere to the medication

Techniques:

1. **Logistic Regression**: Classification of patients into members who adhere and who do not adhere.
2. **Time series analysis**: To make a prediction of when a particular patient is likely to move from adherence to non-adherence.
3. **Decision trees**: Build decision trees based on all the factors responsible for a patient to become non-adherent to give business insights.

Project on Big Data: Counting Frequency of words using N-gram.

Aim: N-gram models are used widely in computational linguistics and computational biological especially in protein and DNA sequence analysis. My project deals with counting the frequency of individual words from a million-line tri-gram text to see the top 20 highly occurring words in the text corpus.

Approach: First uploaded the text file onto **Hadoop HDFS** and then used a higher-level language; wrote a program to COUNT the frequency of words.

Techniques: Using **PIG**, loaded the data and tokenized it into word strings with corresponding frequencies and COUNT the frequencies of each of the words and OUTPUT the top 20.

EDUCATION & Certifications

- Master of Science, University of Minnesota, Minnesota, USA
- Master of Science, University of Hyderabad, Hyderabad, Andhra Pradesh, India
- Bachelor of Science, Nizam College, Osmania University, Hyderabad, Andhra Pradesh, India

- Certificate Program in Bigdata Analytics and Optimization, INSOFE, Hyderabad, Telangana, India
- Diploma in Computer Applications from Global Infotech, Hyderabad, India
- SAFe 6.0 Certified Professional