VARSHA VISHWAKARMA
Tel: +1 (716)-533-3541 | Email: vvishwak1883@gmail.com
LinkedIN: https://www.linkedin.com/in/varsha-vishwakarma-9284958a/

EXPERIENCE SUMMARY

Data Scientist and **Machine Learning Engineer** with over **7 years** of professional experience, specializing in high-impact projects across various sectors, including retail, recommendation systems, consumer banking, revenue generation, and image data analysis. Leading impactful research, focusing on **Large Language Models**, and bringing a keen understanding of generative techniques for alignment and evaluation methods to enhance customer-facing experiences.

- Experience in working in various domains such as **Fashion, E-commerce, Retail, Finance**, facilitating the entire lifecycle of a data science project: Data Extraction, Data Pre-Processing, Feature Engineering, Dimensionality Reduction, Algorithm Implementation, Back-Testing, and Validation.
- Extensive experience in **prompt engineering**, optimizing prompts for various applications to enhance model accuracy and performance, including **healthcare** and **patient** data.
- Experience with large language models (**LLMs**) such as **LLaMA2**, **GPT-3**, and **Transformer** models, gaining a deep understanding of their architectures and functionalities and their application in the healthcare domain
- Experience in data-wrangling, loading in **Big Data** platforms such as Apache Spark, working efficiently through **SQL server** after doing enough Data Extracting process from several sources, transforming in transit and loading into the relevant platform to perform actions.
- Expertise in using Machine Learning Techniques in **Python & R (R Studio)** such as Linear Models, Polynomial, Support Vector; classification models such as Logistic Regression, Decision Trees, Support Vector Machine and K-NN (K Nearest Neighbor) also in clustering like K-means.
- Experienced in Building Recommendation engines using Association rule, collaborative filtering, and segmentation
- Proficient in Statistical Modeling and Machine Learning techniques (Linear, Logistics, Decision Trees, **Random Forest**, SVM, K-nearest neighbors, Bayesian, XGBoost) in Forecasting/ Predictive Analytics and segmentation methodologies.
- Hands-on experience in implementing LDA, and Naive Bayes, and skilled in Random Forests, Decision Trees, Linear and Logistic Regression, SVM, Clustering, neural networks, Principal Component Analysis, and good knowledge of Recommender Systems.
- Knowledge of interactive ML tools such as TensorFlow, Keras, SciKit-Learn, and PyTorch, and expertise in using strong Coding Platforms such as Spyder, Jupyter Notebook, and R Studio offered by Anaconda Navigator as well as **Google Colab**.
- Proficient in utilizing **GCP** , **AWS** and **Azure** cloud services, with hands-on experience in deploying and managing virtual resources on these platforms.
- Experience in the Hadoop ecosystem and Apache Spark frameworks such as **HDFS, MapReduce, HiveQL, and Pyspark.**
- Experience in text mining and topic modeling using NLP & Neural Networks, tokenizing, stemming, and lemmatizing, tagging part of speech using TextBlob, **Natural Language Toolkit (NLTK),** and Spacy while building Text Segmentation and Sentiment Analysis.
- Knowledge of AI & Deep Learning techniques such as **Convolutional Neural Networks (CNN)** for Computer Vision, **Recurrent Neural Networks (RNN)**, Deep Neural Networks with applications of Backpropagation, Stochastic Gradient Descent (SGD), **Long Short-Term Memory (LSTM)** and Continuous Bag of words, Text Analytics, etc.
- Hands-on experience on Apache Hive, and Apache Spark using Python for **Big Data**. Collect insights from data using Hive queries to make business decisions.
- Skilled in creating executive Tableau Dashboards for Data visualization and deploying them to the servers.
- Proficient in Data Visualization tools such as **Tableau** and **PowerBI** and Big Data tools such as **Hadoop HDFS, Spark (PySpark), and MapReduce.**
- Experience in implementing a comprehensive monitoring system to track model performance, including real-time streaming data and alert mechanisms to identify and address any anomalies or drifts in the model's behavior.

TECHNICAL SKILLS

Machine Learning	Linear Regression, Logistic Regression, Gradient boosting, Random Forest, GLM, Maximum likelihood estimation, Clustering, Classification & Association Rules, Bayesian, SVM, K-Nearest Neighbors (KNN),K-Means Clustering, Decision Tree, XG Boost, Neural Networks, LSTM, RNN, CNN, Principal Component Analysis, Sampling Design, Time Series Analysis, Reinforcement Learning, Anomaly Detection, ARIMA, Basket Analysis, Text mining, Computer Vision, Word2Vec, K-Means, Model Monitoring, Generative AI, Encoder and Decoder, Open AI GPT, LLM
Data Visualization Tool	Tableau, ggplot, Plotly, PowerBI, Python matplotlib, seaborn, IBM Analytics.
Programming & Tools	Python, R, SQL, SAS, Big Query, Tableau, Power Bi, PyCharm, Hadoop, HIVE, VS Code, Excel, PowerPoint
Environment & Libraries	Linux, Ubuntu, Windows, GitLab, scikit, Pandas, NumPy, Matplotlib, Dask, NLTK, spaCy, Streamlit, Flask, Gensim, Glove, Tesseract, PySpark, OpenCV, TensorFlow, PyTorch, Keras, PIL, Global Surrogate, LIME, Dask, JAX, RestAPI, YOLO, Rasa-X, SciPy, Databricks
Statistical	Data Mining, Data Wrangling, Probability, Hypothesis Testing, Correlations, Association Rule, Decile, Principal Component Analysis, A/B Testing, ETL, Descriptive Statistics
Cloud Services	Google Cloud Platform (GCP), VPC, IAM, AWS (S3, EC2, Sage Maker), Mocrosoft Azure, Docker, Kubernetes

EDUCATION

- Master of Professional Studies: Data Science, University at Buffalo, Buffalo, NY, Jan 2024
- Bachelor of Technology: Electronics & Instrumentations, KIIT University, India, May 2016

PROFESSIONAL EXPERIENCE

University at Buffalo – Buffalo,NY, USA **Jan 2023 – Present**
Role: Research Intern

Working on two generative AI projects, the objectives were to enhance the accuracy of differentiating authentic and synthetic facial images and to improve patient-provider interactions in healthcare. The first project involved generating and analyzing realistic morphed images using advanced generative models, developing robust statistical measures. The second project focused on automating responses to common inquiries through an advanced healthcare conversational system, aiming to improve user experience, support strategic decision-making, and ultimately enhance healthcare outcomes by leveraging AI technologies.

Roles & Responsibilities:

- Generated realistic morphed facial images using **StyleGAN** and **MIPGAN**, and **Diffusion** model, which impact the ability to differentiate them from bogus images.
- Conducted **human annotation** to score image quality, including both bogus and morphed images for evaluating the accuracy and effectiveness of **FIQA(Face Image Quality Assessment)**.
- Built a user interface to provide a seamless and user-friendly platform for **annotators** to evaluate and assign quality scores to patches of facial images.
- Performed **comprehensive comparative analysis** of bonafide, StyleGAN, and MIPGAN-generated images, employing advanced metrics to assess and contrast their quality and authenticity. This analysis facilitated a deeper understanding of generative model capabilities and the characteristics distinguishing real from synthetic facial imagery.
- Pioneered the implementation of statistical tests on a facial image quality dataset, annotated by five different annotators, featuring 20% overlapping and 10% duplicate images. Strategically measured annotator consistency and **inter-annotator correlation**, enhancing the reliability and accuracy of image quality evaluations.
- Engineered a cutting-edge image quality prediction model by integrating **ResNet** for spatial feature extraction and a Transformer for capturing long-range dependencies. Enhanced prediction accuracy by 8% compared to conventional models through innovative algorithm optimization.
- Designed an advanced healthcare conversational system using the **Llama2** model, **RAG** Framework, **HuggingFace transformers**, and **LangChain**, enhancing patient-provider interactions, automating responses to common

inquiries, improving the overall user experience in the **healthcare sector**, and creating a chatbot to provide real-time responses to user queries, enhancing customer support and user engagement.
- Implemented **Rasa Framework with GPT-3** for text generation, creating a prompt workflow to enhance conversational AI capabilities.
- Analyzed **Patient Health** Information, including electronic records, patient details, and diverse **healthcare data**, encompassing patients' history, and literature on healthcare. Skilled in extracting insights and identifying trends to inform strategic decision-making and improve healthcare outcomes while adhering to **HIPAA** regulations.
- Evaluated the model's performance using various metrics like Precision, Recall, **F-Score, ROC, and AUC and** Cross-Validation to test the models with different batches of data to optimize the models.

Environment: Python 3.x (Scikit-Learn/Scipy/Numpy/Pandas/Matplotlib/Seaborn), Machine Learning (StyleGAN, MIPGAN, Diffusion, Transformer, Llama2, GPT-3, OpenCV, ResNet, VGGNet), LangChain, HugginFace, Rasa Framework, RAG Framework, Git 2.x, HTML, PHP, VS Code, Deepbull Server, Flask, GPU, PowerPoint, Ms Excel, TMUX.

Zensar – Bengaluru, India **Mar 2021 – Aug 2022**

Role: Technical Specialist/Data Science/Machine Learning

The project aimed to develop and optimize advanced chatbot workflows for IT support and HR functions, utilizing NLP techniques to enhance accuracy and performance. Through integration of CI/CD pipelines and Docker containers, the objective was to establish a streamlined deployment process and foster continuous improvement of chatbot functionalities within team workflows.

Roles & Responsibilities:
- Led the development of advanced **AI agents** and **chatbot** workflows using Rasa Framework for IT and HR, employing NLP techniques like intent recognition and entity extraction to refine model accuracy and performance.
- Effectively handled a variety of **structured and unstructured data** sources, performing **annotation** tasks to prepare training data for chatbot training and development.
- Integrated **GitLab** for **CI/CD pipelines** and **Docker** containers to streamline the deployment process into team workflows to facilitate continuous improvement and rapid iteration of chatbot functionalities.
- Collaborated closely with **cross-functional teams** to ensure chatbot solutions were user-centric, aligning with specific requirements and enhancing overall user interaction and engagement experiences.
- Architected and deployed a production-ready end-to-end AutoML system on AWS Cloud utilizing **Dataiku** to fully automate the **data visualization** process using **Tableau**, achieving a 75% reduction in time and effort for producing interactive visualizations. Utilized **Streamlit** to develop interactive web applications, enabling dynamic data visualization and user interaction.
- Built the Vinci Machine Learning framework for Engineers and Data Scientists to build and deploy models as managed services on **Kubernetes**. Orchestrated the integration of real-time streaming data sources into **Azure Databricks** pipelines, ensuring exceptional performance and reliability, and enabling prompt insights for clients.
- Performed **Data Collection, Data Cleaning, Data Visualization,** and Text Feature Extraction, and performed key statistical findings to develop business strategies.
- Employed NLP to classify text within the dataset. Categorization involves labeling natural language texts with relevant categories from a predefined set.
- Developed Sentiment Analysis using Machine Learning & NLP by training historical Data provided by organizations to understand the sentiment of end-users.
- Initiated various pre-processing phases of text like Tokenizing, Stemming & Lemmatization, Stop Words, Vocabulary Phrase Matching, POS Tagging using **NLTK**, and Spacy libraries on **Python,** and converting the raw text to structured data.
- Trained model in Python to predict the sentiments on word embeddings of the reviews using **Word2Vec**.
- Build classification models based on Logistic Regression, Decision Trees, and Random Forest to classify the texts through labels
- Closely monitored the performance of the model by using confusion matrix, classification report, Accuracy, Recall, Precision, and F1-score, additionally performed A/B testing to see which of the proposed solutions worked better.

Environment: AWS RedShift, EC2, S3, Azure, Databricks Hadoop Framework, HDFS, Hive, PySpark, Spark (Pyspark, MLlib, Spark SQL), Python 3.x (Scikit-Learn/Scipy/Numpy/Pandas/Matplotlib/Seaborn), Tableau Desktop

(9.x/10.x), NLTK, Spacy, Streamlit, VADER, NLP, Bag of words, TF -IDF, Kubernetes, Docker, Rasa Framework, BERT, Count Vectorization, Word Embeddings, Word2Vec.

CIMB Bank – KL, Malaysia **Feb 2019 – Feb 2021**
Role: Data Scientist

The project involved building a bank system to leverage data science methodologies and techniques aimed at enhancing operational efficiency, reducing costs, and increasing revenue. Tasks included forecasting ATM cash demand, optimizing product sales strategies, mitigating fraud risk, gaining insights into customer behavior, and making informed strategic decisions through rigorous statistical analysis

Roles & Responsibilities:

- Performed **Data Cleaning, Data Exploration, Data Visualization**, Feature Selection, and Engineering using **Python libraries** such as Pandas, Numpy, Sklearn, Matplotlib, and Seaborn.
- Modeled a **time series** network for forecasting ATM cash demand, achieving a 28% reduction in logistic costs for cash replenishment, improved operational efficiency, and reduced logistic expenses, implemented real-time data streaming to enable the model to continuously learn and adapt.
- Increased **Mortgage/ CPL** product sales in the first quarter of model implementation by allowing the sales team to focus on high-value **customers** with the highest likelihood of **product** purchase.
- Integrated **fraud risk** scoring models mitigated the risk of fraudulent transactions, safeguarded sales integrity, minimized financial losses, targeted high-value customers, and positively impacted the company's bottom line by reducing **fraudulent activities** and increasing sales revenue.
- Applied Market Basket Analysis for detailed customer **purchase behaviour insights**, enabling the creation of tailored bank offers and **product recommendations**, subsequently enhancing cross-selling effectiveness and customer satisfaction.
- Employed matrix factorization and vector databases to efficiently handle high-dimensional user-item interaction production data and enhance the scalability and performance of the models.
- Implemented rigorous statistical tests (including t-tests, **A/B testing**, **ANOVA**, and regression analysis) on large datasets to identify key **financial trends** and **risk factors**, leading to more informed strategic decision-making and operational efficiency.
- Integrated machine learning algorithms to analyze and allocate credit across multiple touchpoints, enhancing the understanding of customer behavior and **marketing** effectiveness.
- Developed a **reinforcement learning-**based credit scoring system to optimize loan approval processes and minimize default rates. Utilized Multi-Armed Bandits algorithms to dynamically select the best loan approval strategies based on historical data and continuously updated the model.
- Implemented security best practices and compliance standards within **Data Hub** environments, ensuring data governance and regulatory adherence in sensitive banking data environments.
- Extensive hands-on experience and high proficiency with structures, semi-structured and unstructured data, using a broad range of data science programming languages and big data tools including **R**, **Python**, **PySpark**, **SQL**, **Scikit Learn**, and **Hadoop MapReduce**.
- Removed outliers using standard deviations present in the dataset to avoid the data being skewed in either direction while visualizing the data for various insights and to avoid a biased model.
- Feature Engineered raw data by doing imputation, normalization, and scaling as required on the data frame. Converting an object to Numerical Features and categorical variables to numerical values using Label Encoder.
- Used Logistic Regression, Support Vector Classifiers & ensemble learning like Random Forests & Gradient Boosting Machine and XGB to train the model & the models were optimized using Grid Search & the predictions were made on the test set using each trained model.
- Ensured personal data used in fraud risk scoring was securely processed and protected, employing encryption and access controls as per **GDPR** requirements to safeguard against unauthorized access or data breaches.

Environment: AWS RedShift, EC2, Hadoop Framework, S3, HDFS, Spark (Pyspark, MLlib, Spark SQL), Python 3.x (Scikit-Learn/Scipy/Numpy/Pandas/Matplotlib/Seaborn), R Studio, Tableau Desktop (9.x/10.x), Tableau Server (9.x/10.x), Machine Learning (Regressions, KNN, SVM, Decision Tree, Random Forest, XGBoost, LightGBM, Ensemble), Teradata, Git 2.x, Agile/SCRUM, JSON, R Studio, Microsoft Excel.

Accenture - Bengaluru, India **May 2016 – Jan 2019**
Role: Machine Learning Engineer

The project was to leverage web scraping techniques to gather data from diverse e-commerce platforms and integrate it into a Retail Trend model. This model, utilizing Google Cloud Platform's natural language and vision APIs, analyzed women's garments to forecast trend projections using Recurrent Neural Network (RNN), providing valuable insights for retail clients' demand planning, investment, and warehouse management.

Roles & Responsibilities:

- Executed web scraping techniques to extract data from various **e-commerce** platforms like Amazon, eBay, and Facebook, ensuring a rich and varied dataset for analysis.
- Integrated Retail Trend model using **Google Cloud Platform's** natural language API and vision API by analyzing **women's garments** based on customer interest and forecasts **trend projections** using Recurrent Neural Network (RNN) to provide valuable insights to **retail clients** for demand planning, investment, and **warehouse management**.
- Efficiently deployed the model on Google Cloud Platform, leveraging **GCP buckets** and **BigQuery** for data storage and management, enhancing operational scalability and data accessibility.
- Launched and managed Google **Pub/Sub** for real-time messaging and event ingestion, ensuring robust data flow and integration across systems
- Developed a deep learning-based **OCR** system for bank document verification with an 82% accuracy rate, implemented in a production environment on **Azure**, ensuring robust and scalable document processing.
- Constructed a Convolution Neural Network (CNN-based) text segmentation and **BERT LLM model**, integrating advanced NLP techniques and word embeddings to effectively match resumes with job descriptions for HR recruitment processes. Achieved a 79% accuracy rate by leveraging natural language understanding and semantic analysis, and actively collaborated with HR teams to fine-tune the model for enhanced effectiveness.
- Performed data cleansing on a huge dataset that had missing data and extreme outliers from and explored data to draw relationships and correlations between variables.
- Worked on data cleaning and ensured data quality, consistency, integrity using Pandas, Numpy and participated in feature engineering such as feature intersection generating, feature normalization, and label encoding with Scikit-learn preprocessing.
- **Linear Discriminant Analysis (LDA)** is used as a dimensionality reduction technique in the pre-processing step for pattern classification and machine learning models.
- Used t-SNE to project these higher-dimensional distributions into lower-dimensional visualizations.
- Used various metrics such as F-Score, ROC, and AUC to evaluate the performance of each model and Cross-Validation to test the models with different batches of data to optimize the models.
- Created multiple custom **SQL queries** in Teradata **SQL** Workbench to prepare the right data sets for Tableau dashboards. Queries involved retrieving data from multiple tables using various join conditions that enabled to utilize of efficiently optimized data extracts for **Tableau** workbooks.
- Honored with the "**Next Gen Machine Learning**" award and received a cash prize in recognition.

Environment: GCP, Big Query, GCP vision API, GCP NLP API, Azure, ETL, Tableau, Python 2.x (Scikit-Learn/Scipy/Numpy/Pandas), Linear Discriminant Analysis (LDA), Machine Learning (Naïve Bayes, KNN, Regressions, Random Forest, SVM, XGBoost, Ensemble, Neural Network, BERT, LLM, Pyteseract), Web Scrapping, Microsoft Azure, SQL, Import.io, Agile/SCRUM.

PATENTS & PUBLICATIONS

Patent Published: An Imaging System and a Method for Image Quality Enhancement (Machine Learning) | 202221043339
Patent Published: Method and Device for Performing Data Encryption using Quantum Computing (Data Science) | 202221035778
Patent Published: Data-Driven Method and System for Data Visualization and Useful Insights (Data Science) | 202221058762
Journal Paper: ATM Cash Replenishment with Clustering Series (LSTM Network, Machine Learning) | May 20 | IJSER | ISSN2229-5518
Journal Paper: Approaches for Offline Cursive Handwritten Character Recognition (OCR, Machine Learning) | Jul 19 | IJSR | ART20199819
Journal Paper: Iris Recognition using CNN with Normalization (Machine Learning) | Nov 19 | IJRAR | IJRAR19K6911