VINAYAK KHARE

+1-878-834-9240 vinayakk@andrew.cmu.edu https://linkedin.com/in/vinayak-khare/ Pittsburgh, PA

EDUCATION

Carnegie Mellon University (CMU), Heinz College **August 2025**
Master of Information Systems Management.
- **Coursework:** Introduction to Deep Learning, Advanced Natural Language Processing, Machine Learning in Production, Data Science for Product Managers, Statistics for IT Manager, Database Management, Agile Methods

Manipal Institute of Technology, Manipal (MIT), India **Jul 2016 - Jun 2020**
Bachelor of Technology, Electrical & Electronics Engineering

WORK EXPERIENCE

PricewaterhouseCoopers (PwC) AC, India | *Senior Data Analyst* **Jul 2021 – Apr 2024**

- Led a team of 4 to design scalable ETL pipelines with PySpark and Databricks on Hadoop, reducing annual reporting efforts by over **500 hours and operational costs by 25%,** enabling efficient data processing and real-time analytics for **PowerBI reporting dashboards.**
- Collaborated with cross-functional teams to develop a near-real-time **PowerBI dashboar**d for SOC analyst metrics (ticket volume, schedules, hours), using **Spark on Databricks to process large-scale data in real-time,** delivering actionable insights that improved operational efficiency and client budgeting decisions.
- Reduced pipeline failures by **50% by developing CI/CD workflows in Azure DevOps**, ensuring robust version control of Databricks notebooks and scripts for big data analytics, supporting accurate and timely reporting for stakeholders.
- Streamlined data ingestion from diverse sources like APIs, Azure Blob Storage, and Azure SQL Database, using PySpark on Databricks to automate workflows, unifying large-scale data and cutting manual processing by **18 hours weekly** to support downstream analysis and reporting.
- Optimized Azure Data Lake Storage for analytics by implementing a **Medallion architecture with Databricks Delta Live Tables and Spark,** enabling efficient batch and streaming processing of big data to generate insights through advanced reporting and visualization in **PowerBI.**

Accenture, India | *Data Analyst* **Jan 2021 - Jun 2021**

- Integrated data from 10+ sources, including APIs and relational databases, using **Python (Pandas, requests)** and **SQL** to create unified datasets for analytics, reducing manual processing time by over 600 hours annually.
- Created **Python** scripts using **Pandas** and **Matplotlib** to perform exploratory data analysis and visualize trends, generating actionable insights that enhanced decision-making for analytics projects.

ACADEMIC PROJECTS

RAG-Based Q&A ChatBot (Python, LangChain, LlamaIndex, Pinecone) | **Feb 2025–Mar 2025**
- Developed a Retrieval-Augmented Generation (RAG) chatbot using LangChain and LlamaIndex, fine-tuning the Llama 3.1 model with to answer Pittsburgh and CMU queries, achieving an F1 score of 0.7 and exact match rate of 0.6.
- Built a web scraping pipeline to process 500+ pages, creating a Pinecone vector database with MPNet embeddings, improving retrieval accuracy by 80% and reducing query latency by 40%.

Automatic Speech Recognition Transformer (Python, PyTorch, Torchaudio) | **Sep 2024–Dec 2024**
- Designed an ASR Transformer model using PyTorch and Torchaudio, processing 100+ hours of Librispeech audio to transcribe 10K+ utterances, achieving a 6.9% Character Error Rate.
- Improved model accuracy using SpecAugment, mixed-precision training, and beam search decoding on 5K validation samples.

Movie Recommendation System (Python, Surprise, Flask, MLflow, Docker ,Kubernetes) | **Jan 2025–Mar 2025**
- Built a collaborative filtering recommendation system using Surprise and Python, tuning SVD with GridSearchCV and conducting A/B testing to achieve 0.68 precision and 0.63 recall, improving suggestion accuracy by 30%
- Deployed a Flask API with MLflow for model tracking, using Docker and Kubernetes to containerize the system, ensuring scalable delivery of recommendations for simulated user testing.

NLP Analysis of Diabetes CGM Consumer Posts (Python, NLTK, RoBERTa, K-means) | **Jan 2025–Feb 2025**
- Analyzed 37K+ consumer posts using NLTK and RoBERTa with Python, achieving 88% sentiment accuracy and identifying key attributes like accuracy and cost for CGM brands.
- Applied K-means clustering to segment users into four groups, boosting product insight relevance by 30% for brands like Dexcom.

Customer Churn Prediction & Segmentation (Python, Gradient Boosting, SHAP) | **Nov 2024–Dec 2024**
- Developed a Gradient Boosting model using Python and SHAP, achieving 88% accuracy and identifying key churn drivers like tenure, enabling up to 50% churn reduction potential.
- Segmented customers into high-risk groups using K-means clustering, improving retention strategies by 35%.

SKILLS

- **Machine Learning**: Linear Regression, Logistic Regression, Gradient Boosted Machines, Support Vector Machines, Decision Trees, Principal Component Analysis (PCA), K-Means Clustering, Random Forest, Hypothesis Testing, A/B Testing
- **NLP**: Text Processing, LSTM, Transformers, Word Embeddings (Word2Vec, BERT), Sentiment Analysis, Topic Modeling (LDA)
- **Tools/Frameworks**: NumPy, Pandas, Pytorch, HuggingFace, Git, PowerBI, Docker, Jenkins (CI/CD), Kafka, Azure, MLflow, Databricks, Azure SQL Database, Hadoop
- **Programming Languages**: Python, Java, SQL, PySpark