Kapil Sahu

Boston, MA | (+1) 857-867-2861

kapilsahu2102@gmail.com | www.linkedin.com/in/kapilsahu- | Portfolio: https://kapilsahukp.github.io

Education

Masters, Data Science **Sep 2021 – Dec 2023**

Northeastern University, Khoury College, Boston, MA

Bachelor of Engineering, Computer Science **Aug 2012 – Jul 2016**

Indore Institute of Science and Technology, India

Professional Experience

Data Scientist **May 2022 – Present**

Charles River Data, Boston, MA

Insurance Claims Verification using LLMs:

- Integrated **ChatGPT LLM** to build custom API to **summarize** & verify 100K insurance docs, **reducing manual workload** by **80%**.
- Developed an **efficient AI system** to flag irregularities in property damage valuation, **improving accuracy** by **20%**.

Flood Insurance Premium Prediction, Anomaly Detection & Geocoding Error Detection:

- Achieved **$1M/year savings** by reducing feature space of National Flood Insurance Program parameters using **QGIS**.
- Improved **accuracy** by **30%** via **Decision Trees** implementation for feature selection and using **PyOD** for outlier detection.
- **Reduced payment risk** and boosted geocoding efficiency by **20-30%** through **ranking** and engineering geospatial **features**.
- Implemented **NER** to **optimize purchase cost** by **40%** and built custom word embeddings for semantic similarity.

Revenue and Finance Management:

- **Redesigned** payments processing pipeline to achieve **40% latency reduction** by integrating multiple API endpoints.
- **Optimized API** hit rate to less than **1K hits/day** from 5K hits/day **saving** the client **$6000** of premium subscription.
- Improved **efficiency** by **70%, saving 720 hrs**. by implementing & **presenting** insightful **Tableau** dashboards to clients.

Software Engineer, Data Analytics **Dec 2016 – Oct 2020**

Zensar Technologies

Commercial Aviation Crew Leave and Payroll Management System:

- **Reduced** manual **workload by 60%** after **analyzing financial data** & developing automated payroll **ETL pipeline** using **Airflow**.
- **Enhanced** user experience by **30%** on designing a **highly scalable** feedback mechanism in Flight Plan (iPad app).
- Orchestrated client meetings and fostered collaboration across **cross-functional teams** throughout product development.
- Ensured uninterrupted flight operations by efficiently managing regular **hot fixes** deployment, **minimizing downtime**.
- Empowered a **team of 6** associates with comprehensive **Aviation** and **Financial** domain training on custom **KPIs** & **ROIs**

Technical Skills

Programming Languages & Databases:	Python, R, Java, C++, C, HTML/CSS, JavaScript, SQL (PostgreSQL, MySQL, SQL Server), NoSQL (MongoDB), **BigQuery**
Frameworks and Libraries:	Tensorflow, Pytorch, Pandas, NumPy, Sk-learn, **Plotly**, Spacy, Databricks, **Spark, Airflow**
Proficient in Tools and Topics:	**Git**, Bash, Unix, **Docker**, REST API, AWS(EC2, S3), **GCP**, **Tableau,** OOP Design, **NLP**, Database Design, **Machine Learning**, Predictive Modeling, Optimization, **CI/CD**, Deep Learning**, A/B Testing**, Statistical Analysis, **ETL**, Splunk, **MLOps**, Image Processing.

Personal and Academic Projects

FakeCheck (Image Forgery Detection):

- Developed and deployed an end-to-end image classifier on GCP to detect forged images and FAKE faces using **Streamlit** for API design and **CNN**, **VGG**, **DenseNet** Deep Learning Models and **GANs** to classify them with an **accuracy of 89%.**

Sentiment Analysis (Sarcasm Detection):

- Implemented RNNs, including **LSTMs**, and utilized **encoding techniques** such as bag-of-words with TF-IDF, Word2Vec, and GloVe to detect sarcasm in comments with an **accuracy of 82%**.

Question Answer Model:

- Implemented **seq-to-seq**, **IR model** and transformers **(BERT, DistilBERT, ALBERT Ensemble)** to create a Question-Answering Model based on SQuAD1.1 dataset, predicting correct answers with **81% accuracy** and **86% of F1 Score**.

Salary Predictor:

- Created an end-to-end salary predictor by training ML models on salary data scraped from LinkedIn and Glassdoor.
- Implemented **Multiple Linear Regression, Random Forest, XGBoost** to predict salary based on demographic data.