SUMMARY

Data Scientist with 10 +Years of experience executing data - driven solutions with knowledge on Data Analytics, Text Mining, Machine Learning (ML), Predictive Modelling, forecasting and Natural Language Processing (NLP) in ecommerce, healthcare, finance, supply chain and social networking.

Proficiency in Python, R, SQL, and experience with cloud platforms such as AWS and Google Cloud Computing

* Experienced in **NLP** model development for Smart Inventory Management and sentiment analysis.
* Collaborative work with cross-functional teams and utilization of diverse technologies (**Python**, **Scala**, **TensorFlow**, **PyTorch**).
* Hands on solving problems which brings significant business value by building predictive & forecasting models utilizing **structured & unstructured data**.
* Hands - on experience in Machine Learning algorithms such as Linear Regression, GLM, CART, SVM, KNN, LDA/QDA, Naive Bayes, Random Forest, SVM, Boosting.
* Hands on experience in creating data visualizations, dashboards in a **Tableau** desktop.
* Experience in building data warehouses, data marts and data cubes for creating **PowerBI** reports to visualize various key performance indicators of business.
* Utilized python libraries namely **Pandas**, **matplotlib** and **plotly** for performing data analysis, data visualizations and predictions.
* Using **Docker** and **ansible**, containerized virtual infrastructure’s configuration management tasks which are used to detect config drifts and change back to original configurations.
* Expertise in containerizing applications using **Docker** composes.
* Utilized python’s flask framework for building REST APIs on top of Data Lake (**BigQuery**, **Cloud** **SQL**).
* Achieved Continuous Integration &Continuous Deployment (**CI/CD**) for applications using Git, Azure Devops.
* Experience with **Test driven development** (TDD), **Agile** methodologies and **SCRUM** processes.
* Experience in version control and collaboration tools like **Git** and **source** **tree**.
* Experience building **chatbot** using GENAI LLM models.

**EDUCATION:**

**Stevens Institute of Technology, Hoboken, NJ |** Master of Science in Information Systems (data science) | **Aug 2019 – May 2021**

**Uttar Pradesh Technical University, India |** Bachelor of Technology in Electronics and computer Engineering| **Sep 2009 – May 2013**

**SOFTWARE SKILLS**

**Languages**: Python, R, SQL, Java, JavaScript, Scala, C, C++, XML, Html5.

**Python Libraries**: Keras, TensorFlow, Numpy, Pandas, NTLK, SciPy, PyTorch, Pyro, OpenCV, Matplotlib, Geopandas, libpysal

**Frameworks, Libs, Version Control:** NodeJS, GraphQL, Presto Query Engine, Git, github, bitbucket

**Databases & Tools:** MongoDB, PostgreSQL, Hadoop, Airflow, Spark, Hive, Databricks, Snowflake, Salesforce, Teradata

**Cloud Technologies:** GCP (Pubsub,, Compute Engine, Spanner, BigQuery, Composer), AWS (Athena, Redshift, S3, Lambda )

**Data Visualization Tools:** Tableau, Looker, Qliksense , Quicksight

**Statistics: Inferential statistics,** Experimental Design, Hypothesis Testing (A/B Testing), Regression Analysis, Probability

**Machine Learning:** Regression, Classification, Clustering, Dimensionality Reduction, Ensemble Methods (Random Forest), Neural Nets, Deep Learning (CNN, RNN, LSTM), Natural Language Processing(BERT, NER), Decision Tree, Naïve Bayes, LLM, NER, Model design, Fine-tuning, hyperparameter tuning, ML flow, Data Augmentatio, Transfer Learnign. RAG(Retrieval-Augmented Generation)

**Operating System:** Linux, Windows, MacOS

**PROFESSIONAL EXPERIENCE:**

**VISYS CLOUD TECHNOLOGY, California | Role: Data Scientist May 2023 - CURRENT**

* Topic Modeling: Develop topic modeling framework to identify issues faced by our merchants. Help product teams strategize to identify crucial areas for product development.
* Propensity Modeling: Build ML models to identify indicators of attrition. Develop strategies for retention and better merchant experience.
* Extracted indicators of service variability that caused bad customer experience by clustering similar tickets using NLP techniques such as document embeddings and topic modeling.
* Designed a predictive module to forecast the Time to Resolve (TTR) for tickets using an ensemble model built on multiple Gradient Boosting trees and Logistic Regression models trained for high precision.
* Analyzed customer survey data to compare the effect of TTR for **chat** support vs self-service and benchmark it’s effect on customer experience and presented it to the CIO and his staff  
   Discovered negative effect of improper staffing of service groups on TTR which was a lagging indicator of customer experien.
* Built and deployed data pipeline on a serverless architecture on AWS using S3, Lambda and Sagemaker
* **Chatbot** Development: Develop algorithms and deploy solutions capable of catering to the needs of millions of merchants using LLM technologies. Fine-tune, train and deploy transformer architectures to serve multiple use-cases via a **Chatbot** interface.
* Apply natural language processing on conversations and classify them to identify topics.Use various model interpretation techniques to understand the influence of words on topics.
* Performed regression (linear regression, Logistic regression ) , time series analysis for prediction modelling
* Built models with high dimensional microbiome data to predict diseased state of fields with an accuracy of 79%
* Provided predictive analytics to customers in p»n regime of ML models to determine efficacy of biological products and crop treatments, predict disease risk, & assess nutrient availability (Carbon, Organic Matter, etc.) in specialty and row crops
* Clustered soil samples into 4 Management Zones based upon important soil health indicators with an accuracy of 78%
* Created mail-classifier for TallyCare incoming service mailers using word-to-vector training model with TF-IDF tweaking to improve accuracy to 86% and thus decreased significant workload on TallyCare representatives
* Developed Python scripts for ETL processes, using Pandas, Pyspark to automate data ingestion, cleansing, transform, loading tasks.
* Implemented, maintain Pyspark applications to extract raw payload from external systems through REST and SOAP APIs, applying transformations to ensure high-quality input for dimensional, fact models.
* Developed ETL, ingesting ~5Tb/day from iceberg, hive using PySpark, performed Spark optimizations to reduce runtime by 66%.
* **Utilized google vertex AI (GEN AI ) studio tune models with own data and deploy to applications.**
* **Creating a chatbot using a large language model (LLM) like GPT-4**
* **Utilized RAG(Retrieval-Augmented Generation) tools like Langchain , Liamaindex ,verba, Haystack to  enhance LLMs with the ability to query and retrieve information from various data sources.**
* Constructively been part of a talented research team of data scientists in the field of Computer Vision to innovate, analyse application requirements into data models, to support standardization & effective adoption of bleeding-edge scientific norms and practices with a vision to enable integration and collaboration of AI/ML into everyday workflow.
* Implemented Fully Convolutional Networks (FCNs), Convolutional Neural Networks (CNNs), and Deep Neural Networks (DNNs) for various computer vision tasks such as image classification, object detection, and semantic segmentation.
* Performed data cleaning and feature selection using MLlib package in PySpark, working with deep learning frameworks such as Caffe with considerations for MLOps.utilized CircleCI for model deployment.
* Deployed, tested and Monitored ML model in production
* Integrated CI/CD pipelines with Argo Workflows and AKS to automate the deployment of updated machine learning models, ensuring continuous delivery and integration.
* Created an algorithm that can predict the type of the object in a typical house using Deep Learning. Used OpenCV for the image analysis and keras and Tensorflow for implementing artificial neural networks (ANN).
* Achieved Continuous Integration &Continuous Deployment (CI/CD) for applications using Git, Azure Devops,Azure Synapase analytics
* Developed robust ETL pipelines to extract, transform, and load large volumes of data from various sources into data warehouses and AZURE data lakes, ensuring data quality and integrity.
* Designed and implemented both batch and stream processing solutions for data ingress and egress, ensuring timely and reliable data flow.
* Led customer data migration project of ETL codes from SAS to Python using Azure Databricks. Utilized Kafka for data ingestion
* Successfully transitioned legacy SAS scripts to Python, enhancing scalability, flexibility, and maintainability of data processing workflows .
* Optimized model serving infrastructure on Databricks for low-latency inference, utilizing features such as model caching, distributed serving, and parallel processing.
* Performed feature engineering (PCA,  Feature **Tools:** Using libraries like Featuretools for automated feature generation. AutoML **Tools), performed. Dimensionality Reduction** **Principal Component Analysis (PCA): Reducing dimensionality by transforming the original features into a smaller set of uncorrelated components**
* **Performed Model design and deployment (MLflow cycle), Developed text analysis and Predictive, sentiment analysis deep learning model utilizing Transformers like BERT (Bidirectional Encoder Representations from Transformers). Fine-tuned BERT for specific tasks by adding a simple output layer. So, it can be adapted to various NLP tasks such as text classification, named entity recognition (NER), question answering, and more with minimal task-specific architecture changes.** **NLP with BERT (using Hugging Face’s Transformers library)**
* **Performed image classification using transfer learning by Selecting a model that has been pre-trained on a large dataset, such as ImageNet. Performed Hyperparameter tuning using learning rate.**
* **Performed Data augmentation for computer vision, natural language processing (NLP) models. using libraries like  TensorFlow/Keras: ImageDataGenerator and tf.image modules.PyTorch: torchvision.transforms module. Albumentations: A fast and flexible image augmentation library.**

**FACEBOOK (contract via EPITEC), California | Role: Data Analyst 2 March 2022 - March 2023**

* Designed the bad actor entity recognition pipeline, built its data model from scratch communicating with cross-functional partner.
* Developed Performance matrix using SQL, Python, Tableau to monitor KPIs, maintained SLA, Performed UATs for recommendation.
* Thrashed the landing-time of several pipelines by 30% by tuning, optimizing SQL queries and building ad-hoc data pipelines.
* Designed Star, Snowflake Schema Data Models for Data Warehouse, created data dictionaries and maintained documents.
* Built Prediction model using python (regression, SVM) to forecast customer behavior, provided stakeholders insights using Tableau.
* Analyzed performance of new hardware using Python & Presto SQL to identify & bucket requests into categories, for feature launch.
* Utilized META GENAI studio for lab products testing
* Applied Supervised Machine Learning Algorithms for the **predictive modelling** to tackle various types of problems
* Deployed **LLMs** in customer interaction systems to enhance virtual assistants and chatbots,

**BRIGHT MIND ENRICHMENT, New York |** **Role: Data science Engineer** **September2021 - March 2022**

* Developed Python scripts to automate data validation, cleaning process using pandas, improved process by 70%.
* Designed, developed scalable data pipelines using Databricks, Airflow, AWS to integrate, process data from systems (Workday, Jira).
* Created and developed Spark Streaming applications to process real time data from Kafka with direct approach, and processed data with both stateless and stateful transformations, and then stored data in SQL Server.
* Developed a customer churn prediction model with 84% accuracy using a Random Forest classifier that helped the organization send timely emails to retain customer subscriptions and increased ROI by ₹ 200k every month
* Collaborated with the cross-functional (engineer, designers, Product managers) to define business requirements, documented SOPs.
* Deployed LLMs in patient interaction systems to enhance virtual assistants and chatbots, providing patients with accurate and natural language responses, improving communication, and offering personalized healthcare information.
* Developed doctor report cards for real-time insights into their performance over the years. Using Apache Kafka for data ingestion and Tableau integrated with Hadoop/Spark for creating the reports.
* Developed predictive models like disease risk, readmission risk using advanced machine learning algorithms, ensemble models, and deep learning architectures, integrating MLOps practices for model deployment and monitoring.
* Used Pandas, NumPy, Scikit-learn in Python for developing various like emergency department wait time, chronic disease progression machine learning models and utilized algorithms such as Linear regression, Logistic regression, Gradient Boosting, SVM and KNN, incorporating MLOps for efficient model development and deployment.
* Developed pipeline using Hive (HQL) to retrieve the data from Hadoop cluster, SQL queries to retrieve data from MySQL database and used ETL for data transformation.
* Derived data from relational databases to perform complex data manipulations and conducted extensive data checks to ensure data quality. Performed Data wrangling to clean, transform and reshape the data utilizing NumPy and Pandas library.
* Implemented model versioning and A/B testing strategies on Databricks for evaluating model performance and conducting experiments to improve model accuracy and effectiveness.
* Utilized A/B testing to refine the appointment scheduling interface within Kaiser's online platforms, improving usability and ensuring a seamless scheduling experience for patients.
* Utilized IOT sensors for collecting health information of cold storage and build streaming data pipeline into GCP’s BigQuery with the help of Apache Airflow.
* Employed Docker in the deployment of deep learning architectures, providing a consistent runtime environment for image analysis tasks using OpenCV, Keras, and TensorFlow.
* Productionized machine learning pipelines to gather data from BigQuery using Apache Airflow and build forecasting models to forecast and predict temperature and humidity spikes inside the cold storage.
* Built monitoring dash boards by employing visualization tools such as Tableau or PowerBI, visualizing both the current state and predictive health of cold storage warehouses.
* Leveraged Kubeflow pipelines to automate end-to-end machine learning workflows for Kaiser’s applications, enhancing reproducibility and scalability in analyzing medical data and deriving insights, incorporating MLOps for streamlined pipelines.
* Applied NLP techniques for sentiment analysis on customer feedback and reviews.
* Led the development and deployment of machine learning models on GCP Vertex AI tailored for Kaiser’s applications, including predictive analytics for patient outcomes and disease progression.
* Designed end-to-end machine learning pipelines on GCP Vertex AI with a focus on security and compliance, ensuring that data handling adheres to regulatory standards like HIPAA.
* Utilized Vertex AI's AutoML capabilities to fine-tune models for medical image analysis, ensuring high accuracy in tasks such as radiology image interpretation.
* Integrated Docker and Kubeflow into Kaiser’s data science workflows, allowing for efficient collaboration and sharing of machine learning models and experiments within the research team, with a focus on improving analytics and MLOps practices.
* Led the development of Deep Learning models, utilizing PyTorch and Tensorflow
* GRAVITE EDUVENTURE, India | Role: Decision scientist February 2018 - June 2019
* Developed Marketing Mix Models for ecommerce clients using Bayesian algorithm, Analyzed the impact promotional tactics across 6 channels. Tailored program recommendations based on customer preference, helped boosting the ROI by ~9%
* Forecasted linear and digital traffic and ad sales for a media giant using an ensemble of xgboost and prophet in R. The models provided a guideline towards pricing the ad spots on the website bringing a +1.5M as savings to the org
* Implemented statistical modelling (SAS, R), A/B testing, built reports for stakeholders to identify pain points, product improvements.
* Leveraged Tableau for real time KPI reporting based on demand, substitutability, and loyalty to help the category managers with better negotiation strategies improving client revenue by 5% annually.
* Collaborated with the strategy managers to build and schedule the dataflow architecture for refreshing Tableau dashboards using Python, Shell scripts and Spark SQL to hold merchants accountable towards their core policy.
* Helped identify critical issues and collaborated with stakeholders to execute an improvement strategy which led to a 15% increase in customer satisfaction. Tracked down the performance of products, based on sales & spend during the 2018 holiday season leading to a quarterly growth rate of 50%
* Used Google Kubernetes Engine, Pubsub, BigQuery for data warehousing to deliver business insights, increasing efficiency by 15%
* Experienced in NLP model development for Smart Inventory Management and sentiment analysis.
* Collaborative work with cross-functional teams and utilization of diverse technologies (Python, Scala, TensorFlow, PyTorch).
* Hands on solving problems which brings significant business value by building predictive & forcasting models utilizing structured & unstructured data.
* Hands - on experience in Machine Learning algorithms such as Linear Regression, GLM, CART, SVM, KNN, LDA/QDA, Naive Bayes, Random Forest, SVM, Boosting.
* .

**HCL TECHNOLOGIES, India| Role: Consultant, Data Science July 2016 - January 2018**

* Developed a customer churn model (logistic regression, SVM, Decision tree, Naïve bayes) using python, PCA with .84 accuracy.
* Optimized Advanced SQL query by indexing and normalization to improve data retrieval, reduced processing time by 25%.
* Developed ETL (Extract, Transform, Load) pipeline to scrape data using Pyspark REST APIs in JSON format, loaded them in MongoDB.
* Classified sentiment for an e-commerce client using NLP on reviews using rating Models used: Logistic Regression, Naïve Bayes.
* Built Google Studio dashboards combining key performance indicators to better understand customer acquisition.
* Used python and SQL to create Customer Tags based on their buying behavior to personalize user experience.
* Created regression and classification models in python to predict customer conversion with an accuracy of 92%
* Performed Exploratory Data Analysis (EDA), Predictive analysis, fraud analysis using Excel, reported by building tableau dashboard
* Led the development of Deep Learning models, utilizing PyTorch and Tensorflow, to address intricate challenges and enhance predictive capabilities.
* Leveraged Python, PyTorch, and Tensorflow to design and implement cutting-edge models, enhancing the organization's capabilities in applied research and data-driven decision-making.
* Utilized Python programming language and cloud-native technologies to refactor SAS scripts, optimizing them for deployment on cloud platforms such as AWS and Azure.
* Constructed a machine learning model for Capacity Planning by collecting historical CPU and Disk usage data from on-premises infrastructure, preprocessing the data, engineering features, and selecting suitable algorithms, such as LSTM networks, to forecast resource utilization.
* Implemented automated data ingestion pipelines using cloud based ETL tools like AWS Glue or Azure Data Factory to streamline the process of extracting, transforming, and loading (ETL) data into the dashboard.

**J K T CONSULTING Pvt Ltd,India|** **Role: Analytics Engineer June 2015 - July 2016**

* Performed data wrangling & modelling on large data sets using python, SQL and monitored KPI using Power to provide business insights. Performed ANOVA using Excel (Pivot table, index, array), VBA, to generate data insights.
* Implemented analytics adoption dashboard compiling multiple reports utilizing DAX queries on Power BI for global audit team.
* Automated 40% manual steps using Advanced SQL queries, analyzed data using python(pandas), DA for 150 Product categories.
* Designed ETL workflows in SSIS and Talend to integrate data from different data sources, MySQL, SQL Server, PostgreSQL.
* Optimized model serving infrastructure on Databricks for low-latency inference, utilizing features such as model caching, distributed serving, and parallel processing.
* Utilized python and Kafka to build data pipelines for pulling data from multiple sources (vCenters, data bases, store devices) into Google’s BigQuery.
* Perform Quality Analysis testing and validation internally using Django and reformulate models to ensure accurate prediction of outcomes of interest and end to end API testing with dummy data and actual data before launch of the actual product with the Engineering team.
* Utilized machine learning algorithms such as logistic regression, multivariate regression, K-means, & Recommendation algorithms to extract the hidden information from the data.
* Used Pandas, NumPy, Scikit-learn in Python for developing various machine learning models and utilized algorithms such as Linear regression, Logistic regression, Gradient Boosting, SVM and KNN.
* For serving data, built REST APIs on the data lake (BigQuery, cloud SQL).
* Provided Agile coaching and training to teams, ensuring a common understanding of Agile principles, practices, and ceremonies for efficient project delivery.

**XOVIAN, India|** **Role: Software Engineer January 2014 - May 2015**

* Executed SQL stored procedure, designed data profiling packages in SSIS to uncover data quality issues, Developed ETL test scripts.
* Scheduled tests in Selenium using JUnit for regression testing of Actimize Watch List Filtering (WLF) in all environments.
* Apply feature selection algorithms to models such as ANOVA (analysis of variance), decision trees using PySpark’s Mllib package and hyper tune the parameters based on interest and to predict the outcomes.
* Developed and maintained Tableau dashboards used by Transaction Monitoring, and FIU departments for reporting essential Anti-Money laundering (AML) transactional metrics, improving workstream by 17%.
* Revamped PostgreSQL/SQL to Hive queries, increasing processing speed by 40% and reducing deployment time by three weeks.
* Examined the behaviour of millions of Asset Transfer Party, Journal, and Trade Execution transactions from the L1 layer to the L4 layer in the Enterprise Analytics Platform to understand the attributes that impact Feature Calculation.
* Reviewed 120 AML Data Quality rules consistent with the Business Requirement Documents (BRD) and Functional requirement documents (FRD) and made updates based on the upstream rule change requirements.
* Experienced in handling large datasets using Partitions, spark in-memory capabilities, Broadcasts in spark, effective and efficient Joins, Transformations, and others during the ingestion process.
* Automated data governance for asset management data using Python scripts, reducing manual efforts by 60%.
* Optimized SQL queries on AML transaction monitoring rules by automating data cleansing, extraction, and analysis processes, resulting in improved performance and data accuracy by 50%.
* Incorporated data assessment by reviewing and monitoring Anti-Money laundering (AML) transactions consistent with global standards and procedures.
* Engineered highly scalable backend REST APIs to efficiently collect and aggregate data from a Azure Data Lake, ensuring seamless data flow for analytics and reporting.