**SUMMARY**

Sr Data Scientist/ (ML) 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.
* Built and maintained scalable data pipelines using Apache Beam and GCP services, supporting various data-driven applications and analytics platforms.
* Utilized python libraries namely **Pandas**, **matplotlib** and **plotly** for performing data analysis, data visualizations and predictions.
* Mentored junior data scientists on causal inference techniques and best practices in statistical analysis.
* 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**.

**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 )

**Azure Services:** Azure Machine Learning, Azure Databricks,

**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), Decision Tree, Naïve Bayes, LLM

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

**PROFESSIONAL EXPERIENCE:**

**VISYS CLOUD TECHNOLOGY, California**

**Sr. Data Scientist / ML Engineer 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.
* Developed and implemented causal models to analyze the impact of policy changes on user engagement and retention.
* Utilized regression discontinuity designs and difference-in-differences approaches to estimate causal effects.
* 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
* Designed and implemented data pipelines using GCP services such as Cloud Dataflow, BigQuery, and Cloud Pub/Sub, processing terabytes of data daily.
* Utilized Azure Databricks for data processing and model training, handling large-scale datasets efficiently.
* Leveraged Azure Synapse Analytics for big data processing and real-time analytics, enabling faster decision-making.  
   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.
* Developed robust ETL pipelines to extract, transform, and load large volumes of data from various sources into data warehouses and data lakes, ensuring data quality and integrity.
* Developed Docker images for application components, optimized Dockerfiles for efficient builds, and pushed images to container registries for use in AKS deployments.
* Utilized Kubernetes (Kubectl) for orchestrating the deployment, scaling, and management of containerized applications within AKS, ensuring high availability and reliability.
* Configured GitLab CI/CD pipelines to automate the building, testing, and deployment of applications to AKS, improving efficiency and reducing manual intervention.
* Integrated model deployment pipelines with Databricks Jobs and CI/CD processes, ensuring automated and reliable model deployment and updates.
* Designed and implemented both batch and stream processing solutions for data ingress and egress, ensuring timely and reliable data flow.
* Developed and integrated RESTful APIs for data ingress and egress, facilitating seamless data exchange between systems.
* Led the development of computer vision algorithms to enable augmented reality (AR) functionalities on eBay's mobile app, allowing users to visualize products in their living spaces.
* Implemented object detection and tracking algorithms to accurately place virtual 3D models of products within real-world environments, enhancing the shopping experience on eBay.
* Developed Spark code using Scala and Spark-SQL for faster processing and testing,
* Designed and implemented robust machine learning pipelines using Kubernetes(K8)/AKS with Argo Workflow orchestration, ensuring scalable and efficient end-to-end ML processes.
* Managed data pipelines for ML workflows on Kubernetes(K8)/AKS using Argo Workflows,
* 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%.

**FACEBOOK, California**

**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.

**BRIGHT MIND ENRICHMENT, New York**

**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.

**GRAVITE EDUVENTURE, India**

**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%.

**HCL TECHNOLOGIES, India**

**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

**J K T CONSULTING Pvt Ltd, India**

**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.

**XOVIAN, India**

**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.