☎ +513-548-8687 ✉ samskruthimusku97@gmail.com

SAMSKRUTHI MUSKU

Data Analyst

ABOUT ME

Results-driven Data Analyst with **3+ years** of experience in leveraging data insights to drive business decisions. Proficient in SQL, Python, Excel, and data visualization tools like Tableau and Power BI.

PROFILE SUMMARY

- Proficient in data analysis with **3+ years** of experience in collecting, cleaning, and interpreting datasets to drive actionable business decisions and Skilled in **Python** programming for data manipulation using libraries such as **Pandas**, **NumPy**, **Matplotlib**, and **Seaborn**.
- Expertise in writing complex **SQL queries** for data extraction, transformation, and reporting from relational databases such as **MySQL**, **PostgreSQL**, and **Oracle** and Experienced in developing interactive dashboards using **Tableau** and **Power BI** to visualize key performance indicators (KPIs).
- Strong knowledge of **Excel**, including advanced features like **VLOOKUP**, **Pivot Tables**, **Power Query**, and **Macros**.
- Hands-on experience with **Google Analytics** and **Adobe Analytics** for tracking website and application performance metrics and Adept in **data cleaning** and transformation processes using tools like **Alteryx** and **Knime**.
- Strong understanding of **machine learning concepts**, with practical application in predictive analytics using **scikit-learn**.
- Knowledge of cloud-based platforms such as **AWS Redshift**, **Google Big Query**, and **Azure Synapse Analytics** for data storage and analysis and Hands-on experience with **SAS** for statistical analysis in a corporate environment and Proficient in managing **data pipelines** using tools like **Apache Airflow** and **Talend**.
- Experienced in **real-time data analysis** with platforms like **Kafka** and **Spark Streaming** and Skilled in **geospatial analysis** with tools like **QGIS** and Python libraries such as **GeoPandas**.
- Familiar with **NoSQL databases**, including **MongoDB** and **Cassandra**, for handling unstructured data and understanding of **data encryption** and security measures for sensitive information and Adept in using **Jupyter Notebooks** for exploratory data analysis and presentations and Experience with **API integration** for data retrieval and automation tasks.
- Proficient in **ETL technologies** such as **Informatica**, **Talend**, **Apache Nifi**, and **SSIS** for designing, building, and maintaining efficient data pipelines to extract, transform, and load data from diverse sources and experienced in working with **Hadoop technologies**, including **HDFS**, **Hive**, **Pig**, and **MapReduce**, for distributed storage, processing, and analysis of large-scale datasets and Proficient in using **Looker** for creating dynamic dashboards, building custom **LookML** models, and delivering actionable insights through data visualization and exploration.
- Skilled in using **Snowflake** for cloud-based data warehousing, including writing optimized **SQL queries**, managing **virtual warehouses**, and performing advanced analytics on large datasets.
- Proficient in leveraging **AWS services** such as **S3**, **Redshift**, **Glue**, and **Athena** for efficient data storage, ETL workflows, and scalable analytics solutions in a cloud environment.
- Experienced in utilizing **Azure services** such as **Azure Data Factory**, **Azure Synapse Analytics**, **Azure Databricks**, and **Blob Storage** for building scalable data pipelines and performing advanced analytics.
- Proficient in leveraging **Google Cloud Platform (GCP)** services such as **big Query**, **Cloud Dataflow**, **Cloud Storage**, and **Looker Studio** for data processing, storage, and visualization in cloud-based analytics workflows.
- Expertise in writing complex **SQL queries** for data extraction, transformation, and analysis, including proficiency in stored procedures, joins, CTEs, window functions, and performance optimization.
- Skilled in developing and deploying **machine learning models** using **scikit-learn, TensorFlow, PyTorch, and XGBoost** for classification, regression, and clustering tasks.
- Experience in **CI/CD for ML models**, leveraging tools like **MLflow, Airflow, and Docker** for seamless model deployment and monitoring.

EDUCATION

UNIVERSITY OF CINCINNATI, Masters in Information Technology, USA from Aug 2023 – Dec 2024

TECHNICAL SKILLS

- **Programming Languages:** SQL, Python, R, JavaScript
- **Data Analysis & Statistical Tools:** Excel (Advanced), SAS, SPSS
- **Data Visualization:** Tableau, Power BI, Looker, QlikView
- **ETL Tools:** Informatica, Alteryx, Talend, Apache Nifi
- **Big Data Technologies:** Hadoop, Spark, Snowflake, Apache Hive
- **Machine Learning:** Scikit-learn, TensorFlow, PyTorch, Keras
- **Databases & Data Warehousing:** MySQL, PostgreSQL, Oracle, Teradata, Amazon Redshift
- **Cloud Platforms:** Microsoft Azure, Snowflake.
- **Statistical Analysis:** SciPy, Stats models.
- **Data Wrangling:** Pandas, NumPy, Dplyr
- **Web Scraping & APIs:** Beautiful Soup, Selenium, Postman
- **Automation & Scripting:** VBA (Excel), Power Automate, Bash
- **Version Control:** Git, GitHub
- **Collaboration & Project Management:** JIRA, Confluence, Microsoft Teams
- **Dashboarding & Reporting:** Google Data Studio, Zoho Analytics
- **Natural Language Processing (NLP):** NLTK, spaCy
- **Data Encryption & Security:** SSL/TLS, AWS KMS, GDPR
- **Data Automation & Workflow Orchestration:** Apache Airflow, Luigi
- **Data Cleaning & Preprocessing:** Open Refine, Trifacta

WORK EXPERIENCE

Huntington Bancshares| Data Analyst II| Columbus, Ohio Feb 2024 – Present

Description: Huntington Bancshares Incorporated is a regional bank holding company offering a comprehensive suite of financial services to consumers, small and middle-market businesses, corporations, municipalities, and other organizations. Analysing data, identifying trends, and providing actionable insights to improve business operations and designing and developing dashboards, ensuring data accuracy.

Responsibilities:

- Analyse and interpret large financial datasets related to loans, transactions, customer accounts, and risk assessment, utilizing **SQL** for complex data querying and reporting from databases like **Oracle** and **SQL Server**.
- Develop **data models and dashboards** using **Tableau** or **Power BI** to visualize financial trends, fraud detection insights, and operational performance for business users.
- Conduct in-depth analysis on **loan default trends, customer spending behaviour, and credit risk assessment**, performing statistical analysis using **R** or **Python** to derive actionable insights from financial data.
- Perform **predictive analytics and machine learning modelling** using tools like **scikit-learn** or **TensorFlow** to identify fraud patterns, customer segmentation, and credit risk scoring, enabling proactive decision-making.
- Apply **machine learning algorithms** to analyse banking data for **fraud detection, churn prediction, and credit scoring**, improving risk management and customer retention strategies.
- Leverage **SAS** or **SPSS** for advanced statistical modelling and support integration of third-party financial data sources such as credit bureaus, transaction history, and external risk assessments.
- Utilize **Jenkins** and **CI/CD pipelines** to automate **data transformation and reporting workflows**, ensuring seamless updates and continuous monitoring of financial data.
- Create **automated alerts** for fraud detection, suspicious transactions, and credit risk changes using **SQL triggers** and scheduled jobs.
- Utilize **Hadoop** and **Spark** for processing and analysing large volumes of **banking transaction data**, enabling scalable and efficient insights from big data sources.
- Leverage **Azure Data Factory** for building and orchestrating **ETL pipelines**, integrating financial data from multiple sources into **Azure Data Lake** for advanced analytics and compliance reporting.
- Utilize **Apache Spark** for distributed **financial data processing** and **Apache Kafka** for real-time transaction monitoring and fraud detection, ensuring efficient handling of high-velocity banking data.

- Use **Alteryx** for data blending, financial analytics, and workflow automation, streamlining **ETL processes** to enhance regulatory compliance and risk management.
- Leverage **Snowflake** as a cloud-based data warehouse to store, manage, and analyse large-scale banking and financial datasets with enhanced security and scalability.
- Write and optimize **complex SQL queries** to extract and manipulate **banking datasets**, ensuring data accuracy and efficiency for regulatory reporting, customer insights, and operational analysis.
- Implement **machine learning models** using **scikit-learn, TensorFlow, or PyTorch** to enhance fraud detection, credit risk prediction, and customer behaviour analysis, enabling data-driven decision-making in financial operations.
- **Developed operational indicators** by identifying process improvement opportunities and integrating disparate data sources to enhance decision-making.

Environment: Oracle, SQL Server, Tableau, Power BI, R, Python, Tableau, Power BI, machine learning, scikit-learn, TensorFlow, SAS, Jenkins, CI/CD pipelines, Hadoop, Spark, big data, Azure Data Factory, ETL pipelines, Azure Data Lake, Informatica, Talend Apache NiFi, pandas, NumPy, Matplotlib, Apache Kafka, Alteryx, Snowflake.

Exxon Mobil's | Data Analyst |Mumbai, India| Sep 2022 – Jul 2023

Description: Exxon Mobil Corporation, commonly known as multinational oil and gas corporation. Utilized the software and data management tools to collect, track, analyse, and visualize energy data for a variety of internal and external audiences and analysing complex datasets from various sources like drilling operations, production wells, pipelines, and equipment sensors to identify trends, patterns, and potential issues, providing actionable insights to optimize production, improve efficiency, predict maintenance needs.

Responsibilities:

- Analyse large volumes of oil exploration, drilling, and production data using **SQL, Hive, or Snowflake** to identify patterns and optimize resource allocation and build and optimize predictive models with **Python** or **R** to forecast equipment failures, production rates, and oil price fluctuations, enhancing operational efficiency.
- Design and maintain **data pipelines** using tools like **Apache Airflow, Informatica, or Talend** to streamline geospatial, seismic, and drilling data integration and Leverage **Tableau, Power BI, or QlikView** to create interactive dashboards for production monitoring, supply chain optimization, and equipment maintenance tracking.
- Perform **ETL operations** with **Azure Data Factory** or **Google Dataflow** for efficient data transformation and loading of exploration and refinery data and conduct in-depth analysis of **unstructured data** such as sensor logs, geological reports, and drilling logs using **NLP tools like SpaCy, NLTK, or Hugging Face Transformers**.
- Implement **machine learning algorithms** using **TensorFlow or scikit-learn** to predict equipment failures, optimize drilling strategies, and enhance reservoir management and Perform **data enrichment** by integrating external datasets such as weather patterns, commodity price indices, and geopolitical events using **APIs or tools like Alteryx**.
- Utilize **Hadoop, Spark, or Databricks** for processing massive geophysical and production datasets in distributed environments and monitor real-time sensor data from drilling rigs, pipelines, and refineries using **Kafka, Kinesis, or Azure Event Hubs** to enable proactive maintenance.
- Conduct statistical analysis and hypothesis testing with **SAS or Excel** for optimizing refining processes and supply chain logistics and Work with **NoSQL databases** like **MongoDB or Cassandra** to manage and analyses unstructured refinery and geological data.
- Develop scenario-based forecasting models for **oil spill impacts,** equipment failures, and energy market trends using **time-series analysis libraries** such as **Prophet or Stats models** and Maintain **version control** for data processing scripts and predictive models using **Git or Bitbucket** to ensure collaboration and reproducibility.
- Provide actionable insights to stakeholders by presenting findings in visually compelling formats using **Matplotlib or Plotly** and Develop and manage data workflows using **Apache Airflow, Luigi, or Prefect** to schedule and monitor complex oilfield and refinery data processes.
- Use **Domo and Sisense** for embedding analytics into business processes and creating executive-level performance dashboards for asset management.
- Employ **Pyspark or Spark SQL** for distributed processing of **seismic, drilling, and production datasets**, enabling large-scale data analytics.

Environment: SQL, Hive, Snowflake, Python, Apache Airflow, Informatica, Talend, Tableau, Power BI, QlikView, Azure Data Factory, SpaCy, NLTK, TensorFlow, scikit-learn, Alteryx, Kafka, Kinesis, SAS, Excel, MongoDB, Cassandra, Prophet, Git, Bitbucket, Matplotlib, Apache Airflow, Luigi, Domo, Sisense, Pyspark.

Vertex Pharmaceuticals | Data Analyst |Mumbai, India| May 2021 – Aug 2022

Description: Vertex Pharmaceuticals is a biopharmaceutical company discovering, developing, and commercializing transformative medicines for serious diseases. Developed the predictive models to forecast sales trends, patient outcomes, and potential safety concerns and creating comprehensive data visualizations (dashboards, graphs, charts) to communicate insights effectively to stakeholders.

Responsibilities:

- Developed and optimized **ETL workflows** using **Azure Data Factory, SQL Server Integration Services (SSIS),** and **Databricks** to ensure seamless **data integration** and **transformation** for pharmaceutical research and development.
- Built and managed scalable **data lakes** using **Azure Data Lake Storage (ADLS)** and integrated with **Azure Synapse Analytics** for advanced analytics, enabling efficient storage and analysis of clinical trial and patient data.
- Automated data workflows using **Python, Pyspark**, and **SQL**, ensuring efficient handling of large datasets related to drug development, clinical trials, and patient outcomes.
- Deployed interactive **dashboards** using **Tableau** and **Power BI** to visualize key performance indicators for drug efficacy, sales, and clinical trial results.
- Established **data warehousing solutions** with **Snowflake** and **Redshift** for robust storage and querying of experimental results, clinical data, and regulatory reports.
- Monitored and optimized **Hadoop** and **Spark clusters** for big data processing of genomic data, patient information, and experimental results, ensuring high availability and performance.
- Managed **SQL** and **NoSQL databases** such as **MongoDB** and **Cassandra** to support diverse **data models** related to patient records, clinical trials, and research data.
- Developed and deployed **machine learning models** using **Azure Machine Learning Studio** for **predictive analytics** in drug discovery, patient risk modelling, and clinical trial outcome predictions.
- Streamlined **deployment processes** using **CI/CD pipelines** with **Jenkins, Azure DevOps**, and **Git** for version control, ensuring smooth delivery of data-driven insights and models.
- Orchestrated **containerized applications** using **Docker** and **Kubernetes** for scalable deployments of analytical tools and applications in drug development projects.
- Performed advanced data processing and analysis using **Hive, Impala**, and **Spark SQL** for high-performance insights into clinical trial data, drug efficacy, and patient outcomes.
- Developed and maintained **data workflows** using **Airflow** for scheduling and orchestrating complex data pipelines, ensuring timely and reliable delivery of pharmaceutical data.
- Conducted advanced data transformation using **Scala** and **Spark Structured Streaming** for **real-time analytics** of clinical trial data, allowing for quick decision-making.
- Managed **real-time data processing pipelines** with **Apache Flink** for high-throughput, low-latency use cases in the monitoring of clinical trials and laboratory experiments.
- Employed **Terraform** and **Ansible** for **infrastructure as code (IaC)** to automate provisioning and deployment of **data environments**, streamlining the setup of environments for drug research.
- Optimized batch and real-time data jobs using **Apache NiFi** for **data ingestion** and processing of clinical trial data, improving efficiency and data flow.
- Configured **Redis** and **Memcached** for **caching** and optimizing **data retrieval** in low-latency applications, enhancing real-time access to clinical data and research results.

Environment: Azure Data Factory, SQL Server, Python, Pyspark, Kafka, Tableau, Power BI, Snowflake, Hadoop, Spark, machine learning, CI/CD, Jenkins, Azure DevOps, Git, Docker, Kubernetes, Hive, Impala, Scala, Airflow, Apache Flink, Terraform, Ansible, Apache NiFi, Redis, Memcached.