**PROFESSIONAL SUMMARY:**

* Around **10+** years of **IT experience** as a **Data Scientist**, including profound expertise and experience on statistical data analysis such as transforming business requirements into analytical models, designing algorithms, and strategic solutions that scales across massive volumes of data.
* Proficient in **Statistical Methods** like Regr**ession models, hypothesis testing, confidence intervals, principal component analysis** and **dimensionality** reduction.
* Proficient in building models using **R** and **Python**, with a strong understanding of **Machine Learning algorithms.**
* Hands on experience with big data tools like **Hadoop, Spark, Hive, Pig, Impala, PySpark** and **Spark SQL.**
* Professional working experience in **Machine Learning algorithms** such as, **linear regression, logistic regression, Naive Bayes, Random Forests, Decision Trees, Clustering** and **Principle Component Analysis.**
* Strong background in evaluating **machine learning** models using various metrics and methodologies to ensure robust and reliable predictive capabilities.
* Experienced with advanced **NLP** **methods**, including sentiment analysis and text classification, to gain deeper insights from text data.
* Adept in using **Hadoop Distributed File System (HDFS)** for scalable and fault-tolerant data storage, and leveraging **Map** **Reduce** for distributed data processing across large datasets.
* Demonstrated ability to translate complex data findings into actionable insights for non-technical stakeholders.
* Advanced skills in R for in-depth statistical analysis and visualization tasks.
* Excellent communication skills, facilitating collaboration with cross-functional teams to solve complex problems.
* Detail-oriented with a strong focus on data accuracy and integrity in all analytical processes.
* Hands on experience **AWS Services** such as **AWS S3, EC2, EMR, Step Functions, Lambda, Glue, AWS Dynamo DB** and **AWS Redshift.**
* Proficient in data visualization with **Python** libraries like **Matplotlib, Seaborne** and **Plotly**, enabling clear and insightful data representation.
* Proficient in creating interactive dashboards and visualizations using Tableau to drive business insights.
* Strong background in probabilistic and statistical modeling, enhancing decision-making processes.
* Experienced in experimental design and causal inference techniques to evaluate and optimize outcomes.
* Expertise in **Microsoft Azure Services** such as **Azure Data Factory Azure Data Lake Azure SQL Databases Azure** **Synapse Analytics** and **Azure Databricks.**
* Skilled in connecting to various data sources, including **SQL** databases, **Excel**, and cloud-based data warehouses, for seamless data integration.
* Skilled in creating complex data models and performing data blending to consolidate data from disparate sources, ensuring comprehensive and accurate reporting in **Tableau.**
* Expert in creating custom visuals and leveraging **Power BI's** built-in visualizations to present data in a meaningful and easily interpretable format for stakeholders.
* Knowledgeable in using **Dash** for creating data applications with real-time updates and user interactions, enhancing data accessibility and usability.
* Proficient in using **SpaCy** for advanced natural language processing tasks, including named entity recognition, part-of-speech tagging, and dependency parsing.
* Strong written and oral communication skills for giving presentations to non-technical stakeholders.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Programming Languages** | Python, R, , SQL, Scala |
| **Statistical Methods** | Regression models, Hypothesis testing, Confidence intervals, Principal component analysis (PCA), Dimensionality reduction |
| **Machine Learning** | Linear regression, Logistic regression, Naive Bayes, Random Forests, Decision Trees, Clustering, PCA, Supervised learning, Unsupervised learning |
| **Libraries & Frameworks** | Tensor Flow, Keras, PyTorch, Scikit-learn, SpaCy, NLTK, Matplotlib, Seaborne, Plotly, ggplot2, Shiny, Dash |
| **Big Data Tools** | Hadoop, Spark, Hive, Pig, Impala, HDFS, Spark SQL, PySpark, Map Reduce |
| **Natural Language Processing** | Sentiment analysis, Text classification, Named entity recognition, Part-of-speech tagging, Dependency parsing, Tokenization, Stemming, Lemmatization |
| **Clouds** | AWS, Microsoft Azure |
| **Data Visualization** | Tableau, Power BI,  |
| **Advanced Analytics** | Time series forecasting, RNNs, SVM, Clustering, Boosting, Bagging |
| **Version Control** | Git, GitHub, GitHub Actions, Jenkins |

**PROFESSIONAL EXPERIENCE:**

**Client: Kansas State Department of Education, Topeka, KS Oct 2022 – Present**

**Role: Sr. Data Scientist**

**Responsibilities:**

* Utilized popular **NLP libraries** such as **NLTK, SpaCy**, and **Genism** for various tasks such as text preprocessing, feature extraction, and topic modeling.
* Developed and implemented a **churn prediction** **model** using supervised learning algorithms such as **logistic regression, decision trees, and random forests** to identify customers at risk of churn.
* Designed and developed analytics, **machine learning** models, and visualizations that drive performance and provide insights, from prototyping to production deployment.
* Created interactive dashboards in **Tableau** to visualize key performance metrics and track business performance.
* Applied probabilistic and **statistical modeling techniques** to derive insights and support data-driven decision-making processes.
* Utilized **R Markdown** to generate dynamic reports that combine code, analysis, and visualizations.
* Developed continuous integration and delivery pipelines with **Jenkins** to automate testing and deployment processes for machine learning models and data pipelines.
* Integrated **Dask** with **machine learning** libraries to scale out computations and optimize performance for large datasets.
* Implemented **Spark MLlib’s** advanced **machine learning** algorithms for large-scale data analysis and modeling.
* Created reusable **Python** modules and libraries for common data science tasks, improving development efficiency and code maintainability.
* Developed and maintained complex SQL queries to extract, manipulate, and analyze large datasets for actionable insights.
* Utilized Python for statistical analysis, building predictive models and performing data cleaning and preprocessing tasks.
* Designed and executed experiments to test hypotheses and assess causal relationships in various business scenarios.
* Collaborated with cross-functional teams to translate business requirements into data analysis projects and actionable recommendations.
* Communicated complex data findings effectively to non-technical stakeholders, ensuring clarity and actionable outcomes.
* Conducted thorough data exploration and analysis to identify trends, anomalies, and opportunities for optimization.
* Utilized **Flask** for deploying **machine learning** models as web services, facilitating easy integration with web applications and **APIs.**
* Performed data analysis by using **Hive** to retrieve the data from **Hadoop cluster, SQL** to retrieve data from **Redshift.** Optimized **SQL queries** to perform data extraction to fit the analytical requirements.
* Deployed and managed containerized applications using **Kubernetes**, ensuring efficient scaling, load balancing, and high availability.
* Leveraged **AWS services** for data science projects, including **AWS S3** for data storage, **AWS EC2** for scalable computing resources and **Amazon Redshift** for data warehousing.
* Created multiple custom **SQL queries** in **Teradata SQL Workbench** to prepare the right data sets for **Tableau** dashboards.
* Configured **Docker** networking to enable communication between containers and external systems while ensuring security and performance.
* Configured and managed **Kubernetes** secrets and **Config Maps** to securely store and manage sensitive information and configuration data.
* Developed and deployed **RESTful APIs** using **AWS API Gateway** and **AWS Lambda** for real-time data access and integration with machine learning models.
* Implemented **NLTK**-based text preprocessing techniques, including tokenization, stemming, and lemmatization, for **NLP tasks.**
* Configured **Docker containers** to provide isolated environments, enhancing scalability and flexibility for application deployment.
* Developed **Jenkins** pipelines using pipeline as code principles, including **Jenkins** file scripting, to manage complex build and deployment workflows.
* Utilized **GitHub** pull requests to review code, facilitate discussions, and merge changes into main branches.

**Environment:** NLTK, SpaCy, Genism, logistic regression, decision trees, random forests, R Markdown, Jenkins, Dask, Machine Learning Spark MLlib, Python, Flask, APIs, Hive, SQL, Redshift, AWS S3, AWS EC2, Amazon Redshift, Teradata SQL Workbench, Docker, Kubernetes secrets, Config Maps, Restful API, AWS API Gateway, AWS Lambda, NLTK, Jupyter notebooks, AWS Glue Crawlers, Scala, Spark SQL, Hadoop HDFS, GitHub.

**Client: Wintrust Financial Corporation, Rosemont, IL Aug 2019 – Sep 2022**

**Role: Sr. Data Scientist**

**Responsibilities:**

* Performed model enhancement using **Random forest, Bagging** and **Boosting** implementations.
* Designed and deployed a set of **Python** modelling **APIs** for customer analytics, which integrate multiple machine learning techniques for various user behavior prediction.
* Incorporated various machine learning algorithms and advanced statistical analysis like **decision trees, regression** **models, SVM, clustering** using **Scikit-learn** package in **Python.**
* Performed technical build out, created **Tableau dashboards** to enable reports on data analysis and modeling. Implemented data acquisition and manipulation based on **SQL.**
* Implemented **RNNs** for sequential data analysis, including applications in time series forecasting and natural language processing.
* Utilized **R Shiny** to build interactive web applications for data visualization and user engagement. Involved in database design, optimization on **SQL server.**
* Developed and implemented probabilistic and statistical models to analyze complex datasets and derive actionable insights.
* Collaborated with cross-functional teams to communicate complex analyses and insights to non-technical audiences effectively.
* Leveraged R for statistical modeling and hypothesis testing, contributing to high-quality research outcomes.
* Maintained a strong attention to detail in data analysis, ensuring the accuracy and reliability of results.
* Created dashboards with **Matplotlib** and **Seaborne**, to communicate insights and trends to stakeholders effectively.
* Collaborated with data engineering teams to integrate **Databricks** workflows with **Azure Data Factory** and **Azure SQL** **Database** for seamless data movement and processing.
* Utilized SQL for efficient data extraction, manipulation, and modeling to support data-driven decision-making.
* Conducted experimental design and causal inference studies to evaluate the impact of various interventions and strategies.
* Utilized **Docker Compose** and **Kubernetes** for orchestrating multi-container applications.
* Developed and fine-tuned neural network architectures using **Tensor Flow** and **PyTorch** for complex data patterns. Managed **Dask** **clusters** for optimal resource utilization and performance in distributed computing environments.
* Leveraged **Python’s Flask framework** to develop **RESTful APIs** for deploying **machine learning** models and integrating them into applications.
* Created interactive dashboards in Tableau to visualize data trends and present findings to stakeholders.
* Employed Python for advanced statistical analysis, data cleaning, and visualization, enhancing the overall data processing workflow.
* Utilized **Tableau Prep** to clean, shape, and merge data sources for more effective analysis and visualization, ensuring high data quality and consistency across all reporting platforms.
* Created custom **Scala**-based **Spark** transformations and actions to meet specific analytical requirements and optimize data workflows.
* Implemented automation scripts and tools to streamline **Kubernetes** deployments and rollbacks, improving deployment efficiency and reliability for machine learning models and data pipelines.
* Configured **GitHub** Actions workflows for continuous integration and continuous deployment, automating build, test, and deployment processes.
* Managed and processed large datasets using **Hadoop** and **Spark**, leveraging distributed computing frameworks to handle and analyze big data efficiently.
* Implemented **natural language processing (NLP)** solutions using **NLTK** and **SpaCy** for text analysis, sentiment analysis, and language modeling.

**Environment:** Random Forest, Python, Machine Learning, Decision Trees, Regression Models, SVM, Scikit-learn, Tableau, SQL, RNN, Time Series Forecasting, R Shiny, SQL Server, Matplotlib, Seaborne, Azure Data Factory, Azure SQL Database, Databricks, Scala, Jenkins, Docker, Synapse Analytics, Hadoop Map Reduce, Azure Databricks, Tensor Flow, PyTorch, Dask, Flask, RESTful APIs, Tableau Prep, Kubernetes, GitHub, Spark, NLP, NLTK, SpaCy.

**Client:** **Medikabazaar, Mumbai, India Apr 2016 – Feb 2018**

**Role: Data Scientist**

**Responsibilities:**

* Designed and implemented end-to-end systems for Data Analytics and Automation, integrating custom visualization tools using **R** and **Power BI.**
* Optimized machine **learning** models through **hyper parameter tuning** and **grid search techniques**. Applied **regression** **models** to predict continuous outcomes and evaluate the relationship between variables.
* Integrated **Scala** with **Hadoop ecosystem** tools to manage and process data across distributed computing clusters.
* Applied classification algorithms to categorize data into **predefined classes, including decision trees, random forests,** **support vector machines** and **k-nearest neighbors.**
* Developed data processing workflows using **Apache Spark** for large-scale data transformation and analysis tasks.
* Developed and implemented **machine learning** models using **Scikit-learn, Tensor Flow, Keras, PyTorch** and to drive actionable insights and optimize business processes.
* Managed **SQL** databases with **MySQL** and **PostgreSQL**, optimizing query performance and ensuring data integrity.
* Used **Power BI’s** custom visuals marketplace to create unique and tailored visual representations of data.
* Developed interactive **storytelling techniques** to present complex data insights in a clear and engaging manner, aiding decision-making processes.
* Utilized **AWS Lambda** for server less data processing tasks, integrating with **AWS** services such as **AWS** **S3** and **AWS** **Dynamo DB** for real-time analytics.
* Integrated **NLTK** with other **Python libraries** for advanced **NLP** tasks such as named entity recognition.
* Integrated **Matplotlib** with **Python** data analysis tools to produce high-quality visual reports and dashboards. Implemented data validation checks using **Pandas** to ensure the reliability of data inputs for analysis.
* Created and managed **Kubernetes** resources such as pods, services, deployments, and namespaces to ensure effective resource utilization.
* Created **Docker Compose** files to define multi-container **Docker** applications with complex service dependencies.
* Employed **AWS Step Functions** to design and monitor complex data workflows, coordinating tasks across multiple **AWS services** to streamline data processing pipelines.
* Conducted data manipulation and visualization with **R** packages like **ggplot2** and **dplyr** to analyze and present data findings effectively.
* Implemented **Spark SQL** and **Data Frames** for efficient querying of structured data in big data environments.
* Utilized Hadoop tools like **Hive** and **Pig** for data querying and scripting, enhancing data analysis capabilities. Developed and modified database procedures, triggers to enhance and improved functionality using **T-SQL.**
* Configured **Jenkins** to work with **Git** and other version control systems for automatic triggering of build jobs based on code changes.
* Applied advanced data manipulation and analysis techniques using **Pandas** and **Dask**, enhancing the efficiency of data processing and analysis pipelines.

**Environment:** R, Power BI, Machine Learning, Hyper parameter Tuning, Regression Models, Decision Trees, Random Forests, Vector Machines, K-Nearest Neighbors, Spark, Scikit-learn, Tensor Flow, Keras, PyTorch, MySQL, PostgreSQL, Power BI, AWS Lambda, AWS S3, Dynamo DB, NLTK, Python, Matplotlib, Pandas, Kubernetes, Docker, Scala, Hadoop, Step Functions, Spark SQL, Data Frames, Hive, Pig, T-SQL, Jenkins, Git, Pandas, Dask.

**Client: Reliance Retail, Mumbai, India Jul 2014 – Mar 2016**

**Role: Data Analyst**

**Responsibilities:**

* Developed and designed interactive and visually compelling dashboards and reports using **Power BI.**
* Designed and developed new reports and maintained existing reports using **Microsoft SQL Reporting Services (SSRS)** and **Microsoft Excel** to support the firm's strategy and management.
* Developed automated reporting systems to generate periodic performance reports and alerts using **Python.**
* Used **SAS SQL** to pull data out from databases and aggregate to provide detailed reporting based on the user requirements.
* Implemented parameterized pipelines in **Azure Data Factory** to enhance the flexibility and reusability of **ETL** processes across different environments and datasets.
* Performed data analysis and data profiling using complex **SQL** queries on various sources systems including **Oracle** and **SQL Server.**
* Automated data extraction and transformation processes through **Python** scripts to streamline data workflows.
* Designed and implemented **ETL** pipelines using **Azure Data Factory** to automate data integration and transformation from various sources into **Azure Data Lake** and **Azure SQL Databases.**
* Integrated **Excel** with external data sources using **Power Query** and other tools to consolidate and analyze data from multiple systems.
* Developed and maintained **ETL** processes using **Pig** and **Sqoop** to move data between **Hadoop** and relational databases.
* Utilized **HDFS** for scalable data storage and **Hive** for querying and managing data within the **Hadoop ecosystem.**

**Environment:** Power BI, Microsoft SQL Reporting Services (SSRS), Excel, Python, SAS SQL, Azure Data Factory, Oracle, SQL Server, Excel Power Query, Pig, Sqoop, Hadoop, HDFS, Hive, Azure Data Lake, Azure SQL Databases.

 **EDUCATION:** Jawaharlal Nehru Technology University, Hyderabad, TS, India

 BTech in Computer Science and Engineering, June 2010 - May 2014