**Sucharitha Vallabhaneni**

**Phone: 314-528-1746**

**E-mail:** [**sucharithavallabhaneni300@gmail.com**](mailto:sucharithavallabhaneni300@gmail.com)

**SUMMARY:**

* Skilled in Python with almost 9 + years of IT experience and highly efficient Data Scientist with 6+ years of experience in Data Analysis, Machine Learning, Data mining with large data sets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive Modeling, Data Visualization.
* Proficient in managing data science project life cycle and actively involved in all the phases of the project life cycle including data acquisition, data cleaning, data engineering, features scaling, features engineering, and statistical modeling (decision trees, regression models, clustering).
* Expertise in transforming business requirements into analytical models, designing algorithms, building models, and developing data reporting solutions that scale across a massive volume of structured and unstructured data.
* Experienced in building ML models with Logistic regression, linear regression, Neural networks, Random Forest, Decision trees, Reinforcement learning, and other ML algorithms.
* Proficient in developing optimization models utilizing advanced techniques such as linear and integer programming.
* Experienced in using various packages such as pandas, NumPy, Seaborn, SciPy, Matplotlib, and sci-kit-learn.
* Extensive experience in Text Analytics, generating data visualizations using Python, and creating dashboards using tools like Tableau.
* Hands-on experience with big data tools like Spark, Pyspark, and Spark SQL.
* Good Knowledge in Proof of Concepts (PoCs), gap analysis, and gathering necessary data for analysis from different sources, prepared data for data exploration using data munging.
* Skilled in data engineering tasks using AWS services like S3, Glue, RedShift
* Responsible for data engineering functions including, but not limited to data extraction, transformation, loading, and integration in support of enterprise data infrastructures – data warehouse, operational data stores, and master data management.
* Experience in designing visualizations using Tableau software and publishing and presenting dashboards, and Storyline on web and desktop platforms.
* Experienced in using Amazon Sage Maker, S3, Glue, RedShift, EMR, EC2, Athena, Elastic Beanstalk, Cloud Formation, and version control like Git.
* Worked and extracted data from various database sources like Oracle, SQL Server, and DB2, regularly accessing the JIRA tool and other internal issue trackers for the Project development.

**Skills:**

|  |  |
| --- | --- |
| **Programming Languages** | Python, R(intermediate) |
| **Libraries & Frameworks** | Pandas, NumPy, Matplotlib, Keras, NLTK, PyTorch, Scikit-Learn, TensorFlow, Apache Spark |
| **Cloud architecture & frameworks** | AWS, S3, EC2, RedShift, Athena,  DynamoDB, Glue, SageMaker, Azure Data Factory, Synapse Analytics, Databricks |
| **Machine Learning Algorithms** | Linear Regression, Logistic Regression, Naïve Bayes, Random Forest, Decision trees, XG-Boost, K-Means, Neural  Networks |
| **Databases** | MySQL, NoSQL databases like DynamoDB, MongoDB |
| **Visualization tools** | Tableau |
| **Optimization software** | Gurobi |
| **Operating systems** | Windows, Linux, MacOS |

**WORK EXPERIENCE:**

**Halliburton, Houston, TX April 2023 - Present**

**Sr. Data Scientist**

Halliburton is a prominent multinational energy services company with a comprehensive range of services including drilling, completion, production, and digital solutions. It builds innovative technologies to support the exploration and production of oil and gas resources worldwide.

**Responsibilities:**

* Built an internal visualization platform at Halliburton, merging daily and historical data, and applying quantitative algorithms for product analysis and comparisons, using Python, Spark, and Tableau on AWS.
* Developed optimization models using linear programming and integer programming techniques, resulting in significant cost savings and streamlined logistics and supply chain operations.
* Applied various ML algorithms such as Decision Trees, Random Forest, and logistic regression to predict product outcomes, providing valuable insights for product improvement strategies.
* Utilized Amazon SageMaker for ML model development, AWS Glue for data preparation, and Amazon Redshift Spectrum for data analysis across AWS data warehouses and S3 resources.
* Managed S3 buckets, developed policies, and leveraged S3 and Glacier for efficient storage and backup solutions on the AWS platform.
* Processed third-party spending data with Excel macros and Python libraries and leveraged Jupyter Notebook and the scientific computing stack (NumPy, SciPy, pandas, and matplotlib) for data analysis and visualization.
* Wrote Python scripts for data extraction, parsing JSON documents, and loading data into databases, demonstrating proficiency in data handling and management.
* Leveraged Spark and SparkSQL for data integration and manipulation, including a Proof of Concept (POC) for creating a Docker image on Azure to optimize model deployment.Proficient in Continuous Integration (CI/CD) tools, build configuration, change management, and maintenance of source control management systems, showcasing strong skills in software development lifecycle management.

**Environment:** Python, ML Algorithms, Gurobi, Spark SQL, Plotly, AWS services like Sage Maker, EC2, S3, EMR, Redshift, Glue

**American Airlines, Dallas, TX Jan 2022 - Mar 2023**

**Data Science Engineer**

American Airlines, a major legacy carrier, boasts a rich history as one of the oldest operating airlines globally, with its corporate headquarters located in Fort Worth, Texas. Taking on the responsibility of contributing to the Crew Modernization team, one is involved in the entire product lifecycle, from software design to development, aiming to create solutions that captivate our users. This involves the transformation of data from various systems to offer business stakeholders valuable insights and analytics.

**Responsibilities:**

* Worked with Machine learning algorithms like Regressions (linear, logistic, etc.) SVMs, Decision trees for classification of groups and analyzing most significant variables such as FTE, waiting times of purchase orders, and capacities available and applied process improvement techniques.
* Participated in features engineering such as feature creating, feature scaling, and One- Hot encoding with Scikit-learn.
* Converted raw data to processed data by merging, and finding outliers, errors, trends, missing values, and distributions in the data.
* Created SQL tables and developed advanced queries using stored procedures and functions using SQL server management studio.
* Used Pandas, NumPy, seaborn, SciPy, Matplotlib, Scikit-learn, NLTK in Python for developing various machine learning algorithms and utilized machine learning algorithms such as linear regression, multivariate regression, naive Bayes, Random Forests, K-means, & KNN for data analysis.
* Used Amazon Sage Maker to build, train, and tune ML models, Amazon Athena to query S3 data with SQL expressions, and Amazon Redshift Spectrum to analyze data that is stored across AWS data warehouses and S3 resources.
* Supported AWS Cloud instances and used AWS command-line Interface to manage and configure various AWS products.
* Written spark scripts for data transformation.
* Worked on ETL process using AWS glue.
* Designed and built data pipelines using AWS services like EC2, Redshift, and S3
* Used packages like ggplot2 in R Studio for data visualization and generated scatter plots and low graphs to identify relations between different variables.
* Worked on Business forecasting, segmentation analysis, and Data mining and prepared management reports defining the problem; documenting the analysis and recommending courses of action to determine the best outcomes.
* Worked on various Statistical models like DOE, hypothesis testing, Survey testing, and queuing theory.
* Experience with risk analysis, root cause analysis, cluster analysis, correlation and optimization, and K-means algorithm for clustering data into groups.
* Coordinate with data scientists and senior technical staff to identify client's needs and document assumptions.

**Environment:** SQL, Jupyter Notebook, Python, MongoDB, HBase, Microsoft Office, AWS services like Sage Maker, S3, Glue, EMR, Redshift, EC2.

**Stanley Black & Decker Inc., New Britain, CT Jan 2020 - Dec 2021**

**Data Scientist**

Stanley Black & Decker Inc. is a multi-national company which focusses on manufacturing industrial tools, hardware, and security products like monitoring systems.

**Responsibilities:**

* Developed System design documents, Data Model, System build guides, Support guides, User support documents, infrastructure documentation and security processes, integration documentation.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, SQL, and U-SQL Azure Data Lake Analytics.
* Created Pipelines in AD using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool, and backward.
* Involved in developing data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL. I also worked with Cosmos DB (SQL API and Mongo API).
* Migrated complex Map Reduce programs into Spark RDD transformations using PySpark.
* Develop dashboards and visualizations to help business users analyze data as well as provide data insight to upper management with a focus on Microsoft products like SQL Server Reporting Services (SSRS)
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool, and backward.
* Hands-on experience in developing SQL Scripts for automation purposes.
* Created Linked Services for multiple source systems (i.e.: Azure SQL Server, ADLS, BLOB, Rest API).
* Worked with the SCRUM team in delivering agreed user stories on time for every Sprint.
* Worked on analyzing and resolving production job failures in several scenarios.
* Responsible for maintaining quality reference data in source by performing operations such as cleaning, transformation, and ensuring Integrity in a relational environment by working closely with the stakeholders & solution architect.
* Ingested data in mini-batches and performed RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Data bricks.
* Worked on complex U-SQL for the data transformation and loading table and report generation.
* Working with complex SQL, Stored Procedures, Triggers, and packages in large databases from various servers.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that processes the data using the SQL Activity.

**Environment:** - Hadoop3.0, Map Reduce, Azure Data bricks, HBase 1.2, Hive2.3, Pig0.17, Sqoop1.4, Spark SQL, Python, Talend, Azure BLOB Storage, Azure Data Lake storage, Pyspark, Kafka1.1, GIT, SQL, JIRA, NoSQL, Azure SQL, Oozie4.3, Scala

**Zeta Global, NY Oct 2017 - Dec 2019**

**Data Scientist**

Zeta Global is a data-powered marketing technology company specializing in personalized customer experiences and targeted campaigns with advanced data management and analytics solutions.

**Responsibilities:**

* Spearheaded Data Exploration and Factor Analysis at Zeta Global, utilizing Python SciPy to uncover patterns and select features for optimized customer segmentation models, leading to more targeted marketing strategies.
* Designed and executed an A/B experiment to evaluate the performance of a new recommendation system, enabling data-driven decision-making and refinement of customer targeting strategies.
* Leveraged Amazon Sage Maker and AWS Glue for ML model development and data preparation, facilitating the seamless integration of advanced analytics solutions and enhancing customer engagement strategies.
* Orchestrated the development of streaming pipelines and employed Azure analytics for anomaly detection, ensuring robust data processing and quality assurance measures within the marketing analytics framework.
* Demonstrated expertise in Agile methodology and SCRUM processes, fostering iterative development and efficient deployment of data-driven marketing solutions, resulting in improved campaign performance and customer engagement strategies.
* Conducted knowledge transfer (KT) sessions with clients to understand data management systems, ensuring the alignment of data analytics solutions with client objectives and business goals for enhanced customer insights and segmentation.
* Applied reinforcement learning models to optimize automation in manufacturing processes, contributing to streamlined operations and improved product quality through data-driven insights and process enhancements.

**Environment:** Unix, Python (SciPy, NumPy, Pandas, Plotly), MySQL, Excel, AWS services, Tableau, SVM, Random Forests, Naïve Bayes Classifier, A/B testing, Git, Agile/SCRUM., MLLib, SAS, regression, logistic regression, Hadoop, NoSQL, random forest, HDFS, NLTK, SVM, MapReduce

**Data Solutions, Hyderabad, India June 2014 - Sept 2017**

**Data Analyst**

Data Solutions is a consulting firm that provides different clients across various domains like health care, finance, and networking with technical personnel to build innovative software.

**Responsibilities:**

* Extensively used Python's multiple data science packages like Pandas, NumPy, matplotlib, Seaborn, SciPy, Scikit-learn, and NLTK.
* Performed Exploratory Data Analysis, trying to find trends and clusters.
* Built models using techniques like Regression, Tree-based ensemble methods, Time Series forecasting, KNN, and Clustering.
* Worked on data that was a combination of unstructured and structured data from multiple sources and automated the cleaning using Python scripts.
* Extensively performed large data read/writes to and from CSV and Excel files using pandas.
* Improved fraud prediction performance by using random forest and gradient boosting for feature selection with Python Scikit-learn.
* Implemented machine learning models (logistic regression, XG-Boost) with Python Scikit-learn.
* Summarizing the data at the customer level by joining the datasets of customer transactions, dimensions, and from 3rd party sources.
* Used Python to preprocess data and attempt to find insights.
* Iteratively rebuild models dealing with changes in data and refining them over time.
* Created and published multiple dashboards and reports using Tableau server.
* Extensively used SQL queries for legacy data retrieval jobs.
* Gained expertise in Data Visualization using Matplotlib.
* Responsible for maintaining and analyzing large datasets used to analyze risk by domain experts.
* Developed Hive queries that compared new incoming data against historic data. Built tables in Hive to store large volumes of data.
* Used big data tools Spark (SparkSQL, MLLib) to conduct real-time analysis of credit card fraud cases.

**Environment:** Spark, Hadoop, AWS, SQL, Python (sci-kit-learn, pandas, NumPy), Machine Learning (logistic regression, XG Boost), Gradient Descent algorithm, Bayesian optimization, Tableau.