**Name: Srikanth Gopu**

**Email: srikanthsg75@gmail.com**

**Contact:** 848-278-3371

**LinkedIn:** [**http://linkedin.com/in/srikanth-7-b45617194**](http://linkedin.com/in/srikanth-7-b45617194) **Senior Data Scientist / analyst**

**PROFESSIONAL SUMMARY**

* Over 9 years of extensive experience in Data Science, specializing in Python, R, and SQL for data handling.
* Proven expertise in Data Mining, Data Cleaning, Exploratory Data Analysis (EDA), Statistical Modeling, Dashboard Preparation, and implementation of machine learning models.
* Experience in integrating clustering into broader machine learning workflows, using clustering results as input features or as a basis for subsequent predictive modeling tasks.
* Proficient in Python 3.x for data analytics and visualizations, utilizing core libraries like NumPy, SciPy, Pandas, and Scikit-learn.
* Strong hands-on experience in data-wrangling, SQL server, and proficient in writing complex queries.
* Solid knowledge of Statistical methodologies, including Hypothesis Testing, ANOVA, Principal Component Analysis (PCA), and Time Series Analysis.
* Expertise in optimization techniques such as Gradient Descent and Stochastic Gradient Descent for regression models.
* Skilled in integrating Databricks with other tools and platforms, enabling a seamless end-to-end data science pipeline and enhancing the overall ecosystem for analytics and insights.
* In-depth understanding and application of Supervised and Unsupervised machine learning algorithms, including Ensemble Methods, Clustering, Classification, and Time Series models.
* Exposure to text mining using NLP, utilizing Natural Language Toolkit (NLTK).
* Skilled in effectively using and integrating PySpark with other components of the Spark ecosystem, such as Spark SQL, MLlib, and Spark Streaming, for a comprehensive data science workflow.
* Proficient in data transformation techniques, including log transformation, root-mean-square, skewness, normalization, and data aggregation.
* Strong knowledge of Dimensionality Reduction, Hyper-parameter tuning, Model Regularization (Ridge, Lasso), and Grid Search techniques to optimize model performance.
* Skilled in creating reports and dashboards using Tableau, incorporating advanced functionalities such as animations and interactive charts.
* Results-driven professional with the ability to take initiatives, handle multiple tasks, adapt to shifting priorities, and meet deadlines.
* Expertise in developing analytical approaches to meet business requirements.

**TECHNICAL EXPERTISE**

| **Category** | **Skills** |
| --- | --- |
| **Machine Learning** | Classification, Regression, Feature Engineering, One hot coding, Clustering, Regression analysis, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, KNN, Ensemble Methods, K-Means Clustering, Time Series Analysis, Confidence Intervals, Principal Component Analysis, Dimensionality Reduction |
| **Statistical Analysis** | Hypothesis testing, A/B analysis, ANOVA, MANOVA, normal distribution, mean, median, mode, standard deviation, regressions, f-test |
| **Exploratory Data Analysis** | Univariate/Multivariate analysis, Outlier detection, data cleansing, Missing value imputation, normalization, linearity assumption, Histograms/Density estimation, data visualization using libraries like seaborn from Python |
| **Programming Languages** | Python (pandas, seaborn, NumPy, Scikit-learn), R (dplyr, ggplot2), SQL (complex queries using window functions and aggregation) |
| **Big Data** | Hadoop Ecosystem - Hive, Pig, MapReduce, PySpark |
| **Databases** | MS SQL Server, Oracle, PostgreSQL, HDFS, HBase |
| **Data Visualization Tools** | Tableau 9.x/10.x, Seaborn/Matplotlib, Plotly, SSRS, Shiny, ggplot2 |
| **Cloud Data Systems** | AWS (RedShift), AWS S3 |
| **IDE** | Jupyter-lab, R Studio, Eclipse, Spyder, Atom, Notepad++, PyCharm |
| **Version Control Tools** | Git, GitHub, Bitbucket |

**PROFESSIONAL EXPERIENCE**

**Client: Fifth Third Bank, Evansville, IN Jun 2021 to Till date**

**Role: Sr. data scientist / analyst**

**Responsibilities:**

* Analyzing issues related to defective and non-defective medical devices in the context of banking technology.
* Utilizing Spark with Python, DataFrames, and Spark SQL API for accelerated data processing.
* Extensive hands-on experience with the Databricks platform, showcasing proficiency in navigating the interface, managing clusters, and leveraging collaborative features for data science projects.
* Ingesting, transforming, manipulating, and computing data using Stream Sets, Kafka, and Spark.
* Employing Spark Streaming to receive real-time data from Kafka.
* Proficient in developing and executing data science workflows using Databricks notebooks, allowing for interactive and collaborative coding environments.
* Storing streaming data in HDFS, leveraging Python and NoSQL databases like HBase and Cassandra.
* Working with Azure BLOB and Data Lake Storage.
* Loading data into Azure SQL Synapse Analytics (DW).
* Advanced proficiency in PySpark, demonstrating the ability to write efficient and scalable Python code for distributed data processing and analytics.
* Creating structured streaming notebooks for data ingestion using even hubs to delta tables.
* Utilizing Azure Synapse with JDBC push for real-time processing.
* Extensive experience in leveraging PySpark for distributed data processing tasks, allowing for the handling of large datasets and parallelized computation.
* Authoring PySpark scripts for custom UDFs, including row/column manipulations, merges, aggregations, stacking, data labeling, and cleaning tasks.
* Implementing Workload Management in Redshift to prioritize basic dashboard queries.
* Proficiency in selecting appropriate clustering algorithms based on the nature of the data and the objectives of the analysis, including K-means, hierarchical clustering, DBSCAN, and others.
* Enhancing business understanding and support by using data to model investment and business decision outcomes.
* Validating findings through an experimental and iterative approach.
* Integrating and preparing large, varied datasets for specialized databases and computing environments.
* Presenting findings to the business team in an understandable manner.
* Expertise in identifying and selecting relevant features for clustering tasks, ensuring that the chosen features contribute to meaningful patterns and insights.
* Exposing assumptions and validation work for easy comprehension by business counterparts.
* Working on identifying Critical Data Elements (CDE’s) relevant to the banking domain.

**Client: wellcare, Tampa, FL Feb 2019 to May 2021**

**Role: Sr.Data Scientist / analyst**

**Responsibilities:**

* Conducting statistical analyses such as ANOVA, linear regression, and logistic regression using Python and Advanced MS Excel on healthcare data.
* Collecting data from end clients in the healthcare sector.
* Demonstrated ability to optimize Databricks clusters for efficient data processing, ensuring optimal resource utilization and performance for data science tasks.
* Performing ETL using various platforms to define a uniform standard format.
* Implementing queries to retrieve healthcare data from SQL Server databases.
* Experience in collaborating with cross-functional teams within the Databricks environment, fostering efficient teamwork, version control, and knowledge sharing.
* Extracting sample datasets containing essential fields for analysis.
* Utilizing ensemble learning methods (Random Forest, Bagging, Gradient Boosting) for predicting the probability of customer enrollment in healthcare programs.
* Proven track record in using PySpark for feature engineering, enhancing the quality and relevance of input features for machine learning models.
* Selecting the final model based on metrics like confusion matrix, ROC, and AUC.
* Tuning hyperparameters of predictive models using Grid Search.
* Successful development and deployment of machine learning models using PySpark, encompassing tasks such as regression, classification, and clustering.
* Designing and implementing K-Fold Cross-validation to test and verify model significance in a healthcare context.
* Creating a Tableau dashboard and story displaying benchmarks and a summary of model measures in the healthcare domain.
* Skill in preprocessing data for clustering, including handling missing values, scaling features, and transforming data to create a suitable input for clustering algorithms.
* Creating database tables, constraints, views, synonyms, sequences, indexes, and object types relevant to healthcare data.
* Developing Entity Relation Diagrams for coding integrations using Microsoft Visio routines.
* Exploring sales of different healthcare categories using graphs in Matplotlib and Seaborn.
* Analyzing whether healthcare sales exhibit trending or seasonality patterns.
* Familiarity with and application of clustering evaluation metrics such as silhouette score, Davies-Bouldin index, or within-cluster sum of squares, to assess the quality of clustering results.
* Experimenting with predictive models such as Linear Regression, ARMA, and ARIMA to forecast healthcare sales amounts.
* Utilizing Scikit-learn and Stats-model library for model development.
* Developing and modifying database procedures and triggers to enhance and improve functionality in a healthcare context using T-SQL..

**Client: Centene Corporation, MO Dec 2017 to Jan 2019**

**Role: Sr .Data Scientist / analyst**

**Responsibilities:**

* Demonstrating expertise in extracting, transforming, and loading (ETL) operational data using SQL queries.
* Successfully shortening the average processing time by 10 minutes.
* Expertise in automating data science workflows on Databricks, utilizing features such as job scheduling, automated data ingestion, and pipeline orchestration for streamlined processes.
* Collaborating with cross-functional teams in the insurance domain to identify data requirements.

Expertise in optimizing PySpark code for performance, including techniques such as tuning Spark configurations, leveraging broadcast variables, and optimizing transformations.

* Delivering business insights to teams within specified timelines.
* Identifying, measuring, and recommending improvement strategies for Key Performance Indicators across all business areas.
* Achieving a 7% increase in revenue through effective data-driven strategies.
* Creating financial reports and dashboards using Tableau, incorporating advanced functionalities such as animations and interactive charts.
* Providing visual insights for better decision-making in the insurance sector.
* Ability to interpret and communicate the results of clustering analyses to non-technical stakeholders, providing actionable insights and understanding the implications of identified clusters.
* Accomplishing data collection, cleansing, modeling, profiling, and queries to analyze data from various insurance sources.
* Generating 50+ Business Intelligence dashboards and ad-hoc reports using data visualization tools to uncover hidden trends, seasonality, and patterns.
* Performing exploratory and descriptive data analysis using Excel, including pivot tables, formulas, and functions.
* Reporting findings to the senior management team in the insurance domain.
* Addressing a wide range of challenging problems in the insurance sector using techniques from applied statistics, machine learning, and data mining fields.
* Applying advanced analytical methods to derive actionable insights for insurance-related challenges.

**Client: Genpact , Hyd Dec 2015 to Sep 2017**

**Role: Data Scientist**

**Responsibilities:**

* Understanding the client business problems and analysing the data by using appropriate Statistical models to generate insights.
* Participated in all phases of data mining, data cleaning, data collection, developing models, data validation and visualization
* Worked on data that was a combination of unstructured and structured data from multiple sources and automated the cleaning using Python scripts.
* Developed and implemented predictive models using Machine learning algorithms such as linear regression, classification, multivariate regression, Naive Bayes, Random Forests, K-means clustering, KNN, PCA and regularization for data analysis.
* Implemented Classification using supervised algorithms like Logistic Regression, Decision Trees, KNN, Naïve Bayes.
* Participated in all phases of Data Mining, Data Collection, Data Cleaning, Developing Models, Validation, Visualization. Extensively used joins and sub queries for complex queries involving multiple tables from different databases
* Performance tuning of stored procedures and functions to optimize the query for better performance and successfully implemented indexes on tables for optimum performance.

**Client: HSBC, HYD Jun 2014 to Nov 2015**

**Role: Jr. Data Analyst/Scientist**

**Responsibilities:**

* Involved in reviewing and understanding the Business requirements. Involved in the entire lifecycle of the projects including Design, Development, and Deployment, Testing and Implementation and support.
* Developed web services that make database calls for inserts, updates, and select queries.
* Performed Data Analysis and Data Validation by writing SQL queries using SQL assistant.
* Implemented procedures for extracting Excel sheet data into the mainframe environment by connecting to the database using SQL.
* Worked PL/SQL in Oracle database for writing queries, functions, stored procedures, and triggers.
* Developed and executed various MYSQL database queries from python using Python-MySQL connector and MySQL dB package.
* Test-Driven Development approach was used and developed merge jobs in Python to extract and load data into MySQL database. Worked on SQL Server Integration Services (SSIS) and SQL Lite.
* Worked on specific features by using Matplotlib, Seaborn in Python and implemented dashboards in Tableau Partnered with modelling experts to develop data frame requirements for projects.

**EDUCATIONAL QUALIFICATION**

* Bachelors’ specialization in Computer Science from JNTUH University, India. Year of Passing May 2014