# Data Scientist

# NAME: Saarib Sheikh PHONE: +1-330-481-7369

**EMAIL:** [**sheikhsaarib1@gmail.com**](mailto:sheikhsaarib1@gmail.com)

**LinkedIn**[**:**](https://www.linkedin.com/in/srujan-chilukuri-90341a25a) [www.linkedin.com/in/saaribsheikh1](http://www.linkedin.com/in/saaribsheikh1)

**CAREER HIGHLIGHTS**

* 8+ Years of Professional **Qualified Data Scientist** including **Machine Learning, Data Mining, and Statistical Analysis.**
* Expertise in transforming business requirements into analytical models, designing **algorithms**, building models, developing **data mining** and reporting solutions that scale across a **massive volume** of structured and unstructured data.
* Experienced with Machine learning algorithms. **Python (NumPy, Pandas, matplotlib and Sci-kit-learn), Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Linear Regression, Multiple regression, Cluster Analysis, Neural Networks, KNN, SVM, k-means** from Mathematical perspective.
* Experience in implementing data analysis with various analytic tools, such as **Anaconda 4.0, Jupiter Notebook 4.X, R (ggplot2, Caret, dplyr) and Excel.**
* Involved in the entire **data science project life cycle** and actively involved in all the phases including data extraction, data cleaning, statistical modelling, and data visualization with large data sets of structured and unstructured data.
* Skilled in **Advanced Regression Modelling, Correlation, Multivariate Analysis, Model Building, Business Intelligence tools** and application of Statistical Concepts.
* Ability to write and optimize diverse SQL queries, working knowledge of **RDBMS** like SQL server 2008, NoSQL databases like **MongoDB 3.2**.
* Experienced in Big Data with **Hadoop, Spark 1.6, PySpark, SparkSQL, HDFS, and Hive 1.X**.
* Extracted data from **HDFS** and prepared data for exploratory analysis using **data munging**.
* Extensive experience in **Text Analytics**, developing different Statistical Machine learning, Data Mining solutions to various business problems and generating data visualizations using R, Python.
* Performed Information Extraction using NLP algorithms coupled with **Deep learning (ANN and CNN, RNN, LSTM, Encoders, Embedders), Keras and Tensor Flow.**
* Trained convolutional neural network using Tensor flow to collect image features from images of items provided by each seller.
* Worked with complex applications such as, **MATLAB, and SPSS** to develop a neural network, cluster analysis.
* Performed Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python.
* Experienced in working with **ARIMA** parametric time series models.
* Knowledge on time series analysis using **AR, MA, GARCH and ARCH** models.
* Strong experience and knowledge in data visualization with **Tableau** creating line and scatterplots, Bar-charts, Histograms, Pie-chart, Dot-charts, Boxplots, Timeseries, Error Bars, Multiple Charts types, Multiple Axes, subplots etc.
* Assist in the creation of forecasting tools and models using mathematical and statistical techniques to determine continuous improvement opportunities.
* Automated recurring reports using **SQL and Python** and visualized them on BI platform like Tableau.
* Optimize code for performance and footprint, considering advantages of h/w platform and AI platform like o9.
* Excellent understanding of **SDLC, Agile and Scrum** development methodology.
* Worked as a member of the experience optimization team to design creative variants to be A/B tested Focused on developing, improving, and testing the subscriptions landing pages as well as the crosswords page.
* Updated python scripts to match training data with our database stored in **AWS Cloud Search**. So that we would be able to assign each documents a response label for further classification.
* Hands on experience and in provisioning virtual clusters under **Amazon Web Service (AWS) cloud which includes services like Elastic compute cloud (EC2), S3, and EMR, Sagemaker**.
* Performing Map Reduce jobs in Hadoop and implemented Spark analysis using Python for performing machine learning & predictive analytics on AWS platform.
* Experienced working with data modelling tools like Erwin, Power Designer and E-R studio.
* Efficient knowledge on microservices are a software development technique a variant of the service-oriented architecture.
* Used the version control tools like **Git 2.X and VM**.
* Passionate about gleaning insightful information from massive data assets and developing a culture of sound, data-driven decision making.
* Taking responsibility for technical problem solving, creatively meeting product objectives and developing best practices
* Excellent communication skills (verbal and written) to communicate with clients and team prepare + deliver effective presentations.
* Ability to maintain a fun, casual, professional, and productive team atmosphere.

**Education**

* **Masters in Computer Science**, Youngstown State University, OH
* **Bachelors in Computer Science**, Acropolis Institute of Technology and Research Indore, India **(December 2014)**

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Languages | Python, R, C, C++, Java, SQL, PL/SQL |
| Databases | Microsoft SQL Server MySQL, Oracle 12c, DB2, Teradata |
| Bigdata Ecosystems | Hadoop, HDFS, HBase, Hive, Spark, Kafka |
| Data Visualizations | Tableau, Power BI, Azure |
| Application Servers | Web Logic, Web Sphere |
| Cloud technologies | AWS, GCP |
| NO SQL Databases | Cassandra, MongoDB |
| Development Methodologies | Agile/Scrum, Waterfall |
| Version Controls Tools and Test | Git, SVM, GitHub |
| ETL Tools | Informatica Power Centre, SSIS |
| Reporting Tools | MS Office (Word/Excel/PowerPoint/ Visio/Outlook), Crystal Reports XI, SSRS |
| Operating Systems | All versions of UNIX, Windows, LINUX, Sun Solaris |

**WORK EXPERIENCE**

|  |
| --- |
| **Chatter communications, Dallas, Texas Jan 2022 – Present**  **Sr. Data Scientist** |

**Responsibilities:**

* Enhanced Data collection procedures to include information that is relevant for building analytic systems and created a value from data by performing advanced analytics and statistical techniques to determine to deepen insights, optimal solution architecture, efficiency, maintainability, and scalability which make predictions and generate recommendations.
* Support Sales and Engagement's management planning and decision making on sales incentives and production by, developing and maintaining financial models, reporting and sensitivity analysis by customer segment.
* Used Pandas, NumPy, Seaborn, SciPy, Matplotlib, Sci-kit-learn.
* Skilled in Advanced Regression Modelling, Correlation, Multivariate Analysis, Model Building, Business Intelligence tools and application of Statistical Concepts.
* Maintained and developed complex SQL queries, stored procedures, views, functions, and reports that qualify customer requirements using Microsoft SQL Server 2008 R2.
* Generated comprehensive analytical reports by running SQL queries against current databases to conduct data analysis.
* Resolved the data related issues such as: Assessing data quality, data consolidation, evaluating existing data sources.
* Proficient in Predictive Modelling, Data Mining Methods, Factor Analysis, ANOVA, Hypothetical testing, normal distribution and other advanced statistical and econometric techniques.
* Performed data manipulation, Data preparation, Normalization and Predictive modelling. Improved efficiency and accuracy by evaluating model in Python.
* Generated the reports and visualizations based on the insights mainly using Tableau and developed dashboards for the company insight teams.

**Environment:**Python, R, Machine learning, deep learning, NLP, Tableau, MapReduce, Tensor flow, Oracle.

|  |
| --- |
| **EBay, Salt Lake City, UT Jul 2019 – Dec 2021**  **Sr. Data Scientist(Machine Learning)** |

**Responsibilities:**

* Lead the full machine learning system implementation process: collecting data, model design, feature selection, system implementation, and evaluation.
* Data wrangling to clean, transform and reshape the data utilizing Python library Numpy, Pandas, Seaborn, SciPy, matplotlib, sci-kit-learn)
* Used text mining libraries (NLTK, Spacy, Gensim) and NLP techniques to rank AT&T complaints on Twitter for ‘authenticity’ to improve customer-care efficiency
* Implemented ARIMA and LSTM Models for Time Series Forecasting in Python to predict the volume of customer-care call
* Built, trained and deployed various machine learning and deep learning algorithm using Keras, TensorFlow and PyTorch in AWS SageMaker
* Utilized Spark, Scala, Hadoop, HBase, Kafka, Spark Streaming, MLLib and R a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
* Designed and maintained Tableau dashboards to present our findings
* Working on data extraction and creating features using SQL, Hive, Teradata, Oracle, Teradata, SQLServer and Excel, Flat, Complex Flat File, Cassandra, MongoDB and HBase files
* Developed unsupervised machine learning models in the Hadoop/Hive environment on AWS EC2 instance.
* Worked with datasets of varying degrees of size and complexity including both structured and unstructured data.
* Assisted in building the ETL Working with different sources such as Oracle, Teradata, SQLServer and Excel, Flat, Complex Flat File, Cassandra, MongoDB and HBase files.

**Environment:**Python, R, Machine learning, deep learning, NLP, Tableau, MapReduce, Tensor flow, Oracle.

|  |
| --- |
| **Elevence Health, OH Aug 2016 – June2019**  **Data scientist** |

**Responsibilities:**

* Used pandas, NumPy, Seaborn, SciPy, matplotlib, sci-kit-learn in python for developing various machine learning algorithms.
* Participated in all phases of datamining; data collection, data cleaning, developing models, validation, visualization and performed Gap analysis.
* Extracted company relations from unstructured news to form a relation map surrounding a company.
* Managed with portfolio risk modeling and diversified system of mutual funds that are grouped together to provide an expected return with a corresponding amount of risk.
* Performed the risk at funds which are invested in the model portfolio of payroll and taken the future contributions which are invested in the model portfolio from what we have chosen.
* Maintained the model portfolio’s balance and rebalanced regularly to return it to the target allocation.
* Created Revenue optimization algorithm to divert click traffic to different advertiser throughout the day to maximize Revenue.
* Monitored the portfolio as often or as infrequently as we can wish via on-line access to the accounts to avoid risk by applying ML algorithms.
* Applied distant supervision, CNN model and attention mechanism to obtain relation label for each company and achieved 91% accuracy in relation classification.
* Performed Information Extraction using NLP algorithms coupled with Deep learning (ANN and CNN), Keras and TensorFlow.
* Implemented Classification using supervised algorithms like Logistic Regression, Decision trees, KNN, Naive Bayes, Bayesian learning.
* Implemented Agile Methodology for building an internal application.
* Worked on predictive and what-if analysis using R from HDFS and successfully loaded files to HDFS from Teradata and loaded from HDFS to HIVE.
* Prepared data visualization reports for the management using R.
* Experience working in the Python 2.7 and 3 environments using the following libraries Anaconda, NumPy, Pandas, Scipy, iPython etc. Excellence in at least one of these is highly desirable.
* Designed dashboards with Tableau provided complex reports including summaries, charts, and graphs to interpret findings to team and stakeholders.
* Evaluated models using Cross Validation, Log loss function, ROC curves and used AUC for feature selection
* Loaded and saved the models by using the Pytorch and Reinforcement learning (RL)
* Developed MapReduce/Spark, R modules for machine learning & predictive analytics in Hadoop on AWS. Implemented an R-based distributed random forest.
* Knowledge on time series analysis using AR, MA, ARIMA, GARCH and ARCH model.
* Work with the capacity planning director in building and maintaining highly advanced capacity plans for the Individual Client Solutions team.
* Handled importing data from various data sources, performed transformations using Hive, MapReduce.
* Project experience in data mining, Segmentation analysis, business forecasting and association rule mining using large data sets with Machine learning.
* Worked as a member of the experience optimization team to design creative variants to be A/B tested Focused on developing, improving, and testing the subscriptions landing pages as well as the crosswords page.
* Created financial package that supports 3-Year financial plan for all AWS cloud services infrastructure expenses.
* Researched, evaluated, architected, and deployed new tools, frameworks, and patterns to build sustainable Big Data platforms for the clients.
* Identifying and executing process improvements, hands-on in various technologies such as Oracle, Informatica, and Business Objects.
* Updated python scripts to match training data with our database stored in AWS Cloud Search. So that we would be able to assign each documents a response label for further classification.
* Used Data quality Validation techniques to validate Critical Data Elements (CDE) and identified various anomalies.
* Analyzed & processed complex data sets using the advanced query, visualization, and analytics tools.
* Identified process improvements that significantly reduce workloads or improve quality.
* Analyzed the email user click history and third-party data for pattern recognition and to support and change targeting algorithms.
* Supported client by developing Machine Learning Algorithms on Big Data using PySpark to analyze transaction fraud, Cluster Analysis etc.

**Environment:**Python, R, TensorFlow, Machine learning Algorithms, Deep learning, Oracle 12c, Hive, Hadoop, Spark, Tableau, MapReduce, AWS, E-R studio, MS Excel.

|  |
| --- |
| **HSBC, India Jul 2014 – Dec 2015**  **Data Analyst** |

**Responsibilities:**

* Utilized Spark, HBase, Mongo DB, MLLib, Python, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc. and Utilized the engine to increase user lifetime by 45% and triple user conversations for target categories.
* Performed Data Profiling to learn about behavior with various features such as traffic pattern, location, and time, Date and Time etc.
* Worked with sales and marketing team for Partner and collaborate with a cross-functional team to frame and answer important data questions.
* Application of various machine learning algorithms and statistical Modelling like decision trees, regression models, neural networks, SVM, clustering to identify Volume using Sci-kit-learn package in python.
* Performed Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python.
* Performed data visualization with Tableau and generated dashboards to present the findings.
* Recommended and evaluated marketing approaches based on quality analytics of customer consuming behavior
* Evaluated models using Cross Validation, Log loss function, ROC curves and used AUC for feature selection.
* Analyse traffic patterns by calculating autocorrelation with different time lags.
* Ensured that the model has low False Positive Rate.
* Addressed overfitting by implementing the algorithm regularization methods like L2 and L1.
* Used Principal Component Analysis in feature engineering to analyze high dimensional data.
* Created and designed reports that will use gathered metrics to infer and draw logical conclusions from past and future behavior.
* Performed Logistic Regression, Random forest, Decision Tree, SVM to classify package is going to deliver on time for the new route.
* Used MLLib, Spark's Machine learning library to build and evaluate different models.
* Implemented rule-based expertise system from the results of exploratory analysis and information gathered from the people from different departments.
* Communicated the results with operations team for taking best decisions.
* Developed MapReduce pipeline for feature extraction using Hive.
* Interaction with Business Analyst, SMEs and other Data Architects to understand business needs and functionality for various project solutions.

**Environment:**Python, R, Machine learning, CDH5, HDFS, Hive, AWS, Impala, Linux, Spark, Tableau Desktop, SQL Server 2012, Microsoft Excel, MATLAB, Spark SQL, PySpark.