**Swetha Varadarajan**

**Data Scientist**

**Phone:+1(925)5921860**

[**swethavrajan2411@gmail.com**](mailto:swethavrajan2411@gmail.com)

Professional SUMMARY

* Professional qualified **Data Scientist/Data Analyst** with **around 12+ years** of experience in **Data Science and Analytics including Data Mining** , **Deep Learning/Machine Learning** and **Statistical Analysis**
* Involved in the entire data science project life cycle and actively involved in all the phases including **data cleaning, data extraction** and **data** **visualization** with large data sets of structured and unstructured data, created **ER diagrams** and schema.
* Experienced with machine learning algorithm such as **logistic regression, KNN, SVM, random forest, neural network, linear regression**, **lasso regression** and **k-means**
* Implemented Bagging and Boosting to enhance the model performance.
* Experience in implementing data analysis with various analytic tools, such as **Anaconda Jupiter Notebook 4.X, R 3.0 (ggplot2, , dplyr, Caret)** and **Excel**
* Solid ability to write and optimize diverse SQL queries, working knowledge of RDBMS like **SQL Server, NoSql databases** like **MongoDB.**
* Excellent understanding **Agile** and **Scrum** development methodology
* Used the version control tools like Git 2.X and build tools like Apache Maven/Ant
* Passionate about gleaning insightful information from massive data assets and developing a culture of sound, data-driven decision making
* Ability to maintain a fun, casual, professional and productive team atmosphere
* Experienced the full software life cycle in **SDLC**, **Agile, DevOps** and **Scrum methodologies** including **creating requirements, test plans**.
* Skilled in Advanced **Regression Modeling**, **Correlation**, **Multivariate Analysis**, **Model Building**, **Business Intelligence tools** and **application** of Statistical Concepts.
* Developed predictive models using **Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Social Network Analysis, Cluster Analysis,** and **Neural Networks.**
* Experienced in Machine Learning and Statistical Analysis with **Python Scikit-Learn**.
* Experienced in Python to manipulate data for data loading and extraction and worked with **python libraries like Matplotlib, Scipy, Numpy** and **Pandas** for data analysis.
* Worked with complex applications such as **R, R Shiny, SAS, Plotly, ArcGIS, Matlab** and **SPSS** to develop neural network, cluster analysis.
* Strong SQL programming skills, with experience in working with **functions**, **packages** and **triggers**.
* Expertise in transforming business requirements into **designing algorithms, analytical models, building models, developing data mining** and **reporting solutions** that scales across massive volume of structured and unstructured data.
* Skilled in performing **data parsing, data manipulation, data architecture, data ingestion** and **data preparation** with methods including describe data contents, compute descriptive statistics of **data, regex, split and combine, merge, Remap, subset, reindex, melt** and **reshape**.
* Worked with **NoSQL** **Database** including **Hbase**, **Cassandra** and **MongoDB**.
* Experienced in **Big Data** with **Hadoop, MapReduce, HDFS** and **Spark.**
* Experienced in Data **Integration Validation** and **Data Quality controls** for **ETL process** and **Data Warehousing** using **MS Visual Studio, SSAS**,  **SSIS** and **SSRS**.
* Proficient in **Tableau** and **R-Shiny** data visualization tools to analyze and obtain insights into large datasets, create visually powerful and actionable interactive reports and dashboards.
* Automated recurring reports using **SQL and** **Python** and visualized them on **BI platform** like **Tableau**.
* Worked in development environment like **Git** and **VM**.
* Excellent communication skills. Successfully working in fast-paced multitasking environment both independently and in collaborative team, a self-motivated enthusiastic learner.

**TECHNOLOGIES:**

|  |  |
| --- | --- |
| **BigData/Hadoop Technologies** | Hadoop, HDFS, YARN, MapReduce, Hive, Pig, Impala, Sqoop, Flume, Spark, Kafka, Storm, Drill, Zookeeper and Oozie |
| **Languages** | HTML5,DHTML, WSDL, CSS3 ,C, C++, XML,R/R Studio, SAS Enterprise Guide, SAS ,R (Caret, Weka, ggplot) , Perl, MATLAB, Mathematica, FORTRAN, DTD, Schemas, Json, Ajax, Java, Scala, Python (NumPy, SciPy, Pandas, Gensim, Keras), Java Script, Shell Scripting |
| **NO SQL Databases** | Cassandra, HBase, MongoDB, MariaDB |
| **Business Intelligence Tools** | Tableau server, Tableau Reader, Tableau, Splunk, SAP Business Objects, OBIEE, SAP Business Intelligence, QlikView, Amazon Redshift, or Azure Data Warehouse |
| **Development Tools** | Microsoft SQL Studio, IntelliJ, Eclipse, NetBeans. |
| **Development Methodologies** | Agile/Scrum, UML, Design Patterns, Waterfall |
| **Build Tools** | Jenkins, Toad, SQL Loader, Maven, ANT, RTC, RSA, Control-M, Oziee, Hue, SOAP UI |
| **Reporting Tools** | MS Office (Word/Excel/Power Point/ Visio/Outlook), Crystal reports XI, SSRS, cognos 7.0/6.0. |
| **Databases** | Microsoft SQL Server 2008,2010/2012, MySQL 4.x/5.x, Oracle 11g, 12c, DB2, Teradata, Netezza |
| **Operating Systems** | All versions of Windows, UNIX, LINUX, Macintosh HD, Sun Solaris |

Professional Experience

**Client: Intel Corporation, Santa Clara, CA**  **Mar 2022 - Till Date**

**Data Scientist**

**Description**: We create world-changing technology that improves the life of every person on the planet. Intel put the silicon in Silicon Valley. For more than 50 years, Intel and our people have had a profound influence on the world, driving business and society forward by creating radical innovation that revolutionizes the way we live.

**Responsibilities**:

* Extracted data from **HDFS** and prepared data for exploratory analysis using **data munging**
* Built models using Statistical techniques like Bayesian HMM and Machine Learning classification models like **XG** **Boost**, **SVM**, and **Random Forest.**
* Participated in all phases of **data mining, data cleaning, data collection, developing models, validation, visualization** and **performed Gap analysis.**
* A highly immersive Data Science program involving **Data Manipulation** & **Visualization**, **Web Scraping**, **Machine Learning, Python programming, SQL**, **GIT**, **MongoDB, Hadoop**.
* Setup storage and data analysis tools in AWS **cloud computing infrastructure**.
* Installed and used Caffe **Deep Learning Framework**
* Worked on different data formats such as **JSON, XML** and performed machine learning algorithms in **Python**.
* Worked as Data Architects and IT Architects to understand the movement of data and its storage and ER Studio 9.7
* Used **pandas, numpy, seaborn, matplotlib, scikit-learn, scipy, NLTK** in **Python** for developing various machine learning algorithms.
* Data Manipulation and Aggregation from different source using **Nexus, Business Objects, Toad, Power BI** and **Smart View.**
* Implemented **Agile Methodology** for building an internal application.
* Focus on integration overlap and Informatica newer commitment to **MDM** with the acquisition of Identity Systems.
* Coded proprietary packages to analyze and visualize **SPC** file data to identify bad spectra and samples to reduce unnecessary procedures and costs.
* Programmed a utility in Python that used multiple packages (**numpy, scipy, pandas**)
* Implemented Classification using supervised algorithms like **Logistic Regression, Decision trees, Naive Bayes, KNN.**
* As Architect delivered various complex **OLAP** **databases/cubes, scorecards, dashboards** and **reports**.
* Updated **Python** scripts to match training data with our database stored in **AWS Cloud Search**, so that we would be able to assign each document a response label for further classification.
* Used Teradata utilities such as **Fast Export, MLOAD** for handling various tasks data **migration/ETL** from **OLTP** Source Systems to **OLAP Target Systems**
* Data transformation from various resources, data organization, features extraction from raw and stored.
* Validated the machine learning classifiers using **ROC Curves** and **Lift Charts.**

**Environment:** Unix, Python, MLLib, SAS, regression, logistic regression, Hadoop, NoSQL, Teradata, OLTP, random forest, OLAP, HDFS, ODS, NLTK, SVM, JSON, XML and MapReduce.

**Client: Amazon Seattle, WA Mar 2021 - Feb 2022**

**Data Scientist**

**Description**: Amazon is guided by four principles: customer obsession rather than competitor focus, passion for invention, commitment to operational excellence, and long-term thinking. We are driven by the excitement of building technologies, inventing products, and providing services that change lives. We embrace new ways of doing things, make decisions quickly, and are not afraid to fail. We have the scope and capabilities of a large company, and the spirit and heart of a small one.

**Responsibilities**:

* Utilized Spark**, Scala, Hadoop, HQL, VQL, oozie, pySpark, Data Lake, TensorFlow, HBase, Cassandra**, **Redshift, MongoDB, Kafka, Kinesis, Spark Streaming, Edward, CUDA, MLLib, AWS, Python**, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
* Utilized the engine to increase user lifetime by 45% and triple user conversations for target categories.
* Application of various machine learning algorithms and statistical modeling like **decision trees, text analytics**, natural language processing (**NLP**), supervised and unsupervised, regression models, social network analysis, neural networks, deep learning, **SVM,** clustering to identify Volume using **scikit-learn** package in **python, Matlab.**
* Worked on analyzing **data from Google Analytics, AdWords, Facebook etc.**
* Evaluated models using Cross Validation, Log loss function, **ROC** curves and used **AUC** for feature selection and elastic technologies like **ElasticSearch, Kibana**.
* Performed Data Profiling to learn about behavior with various features such as traffic pattern, location, Date and Time etc.
* Categorized comments into positive and negative clusters from different social networking sites using Sentiment Analysis and Text Analytics
* Performed **Multinomial Logistic Regression, Decision Tree**, **Random forest, SVM** to classify package is going to deliver on time for the new route.
* Performed data analysis by using Hive to retrieve the data from **Hadoop cluster**, **Sql** to retrieve datafrom Oracle database and used **ETL** for data transformation.
* Performed Data Cleaning, features scaling, features engineering using **pandas** and **numpy** packages in **python**.
* Exploring DAG's, their dependencies and logs using AirFlow pipelines for automation
* Performed data cleaning and feature selection using **MLlib** package in **PySpark** and working with deep learning frameworks such as **Caffe, Neon**.
* Developed **Spark/Scala, R Python** for regular expression (regex) project in the **Hadoop/Hive** environment with **Linux/Windows** for big data resources.
* Used clustering technique **K-Means** to identify outliers and to classify unlabeled data.
* Tracking operations using sensors until certain criteria is met using **AirFlow** **technology**.
* Responsible for different Data mapping activities from Source systems to Teradata using utilities like **TPump, FEXP, BTEQ, MLOAD, FLOAD** etc
* Analyze traffic patterns by calculating autocorrelation with different time lags.
* Ensured that the model has low False Positive Rate and Text classification and sentiment analysis for unstructured and semi-structured data.
* Addressed over fitting by implementing of the algorithm regularization methods like **L1 and L2**.
* Used Principal Component Analysis in feature engineering to analyze high dimensional data.
* 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.
* Created and designed reports that will use gathered metrics to infer and draw logical conclusions of past and future behavior.
* Developed **MapReduce** pipeline for feature extraction using **Hive** and **Pig**.
* Created Data Quality Scripts using **SQL** and **Hive** to validate successful data load and quality of the data. Created various types of data visualizations using **Python** and **Tableau**.
* Communicated the results with operations team for taking best decisions.
* Collected data needs and requirements by Interacting with the other departments.

**Environment:** Python 2.x, CDH5, HDFS, Hadoop, Hive, Impala, AWS, Linux, Spark, Tableau Desktop, SQL Server, Microsoft Excel, Matlab, Spark SQL, Pyspark.

**Client: Transamerica - Plano, TX Jan 2019 - Feb 2021**

**Data Scientist**

**Description:** Transamerica is the US-based brand of Aegon, a Dutch financial services firm. Aegon is one of the world's leading providers of life insurance, pensions and asset management and is helping approximately 30 million customers globally to achieve a lifetime of financial security.

**Responsibilities**:

* Responsible for analyzing large **data sets** to develop multiple custom models and algorithms to drive innovative business solutions.
* Perform Data profiling, preliminary data analysis and handle anomalies such as missing, duplicates, outliers, and imputed irrelevant data.
* Remove outliers using Proximity Distance and Density based techniques.
* Involved in Analysis, Design and Implementation/translation of Business User requirements.
* Experienced in using supervised, unsupervised and regression techniques in building models.
* Performed Market Basket Analysis to identify the groups of assets moving together and recommended the client their risks
* Experience in determine trends and significant data relationships using advanced Statistical Methods.
* Implemented techniques like forward selection, backward elimination and step wise approach for selection of most significant independent variables.
* Performed Feature selection and Feature extraction dimensionality reduction methods to figure out significant variables.
* Used RMSE score, Confusion matrix, **ROC**, Cross validation and **A/B testing** to evaluate model performance in both simulated environment and real world.
* Performed Exploratory **Data Analysis** using **R**. Also involved in generating various graphs and charts for analyzing the data using **Python Libraries**.
* Involved in the execution of multiple business plans and projects Ensures business needs are being met Interpret data to identify trends to go across future data sets.
* Developed interactive dashboards, created various **Ad** **Hoc** reports for users in **Tableau** by connecting various data sources.

**Environment:** Python, SQL server, Hadoop, HDFS, HBase, MapReduce, Hive, Impala, Pig, Sqoop, Mahout, LSTM, RNN, Spark MLLib, MongoDB, Tableau, Unix/Linux.

**Client: Crane Payment New York, NY Dec 2017 - Dec 2018**

**Data Scientist/ Data Engineer**

**Description:** Crane Payment Innovations is the leader in world-class, automated payment solutions. We partner with businesses to provide payment experiences that set new and innovative standards for consumer convenience, security, and reliability. We are excited to explore new ways every day to improve how millions of people across the globe can make their payments in a way that fits with their lives.

**Responsibilities**:

* Created new features based on information from million transaction records and training models using **Machine-Learning** techniques such as Gradient **Boosting Tree** and **Deep Learning.**
* Analyzed and determined a cutoff point for accepting/ declining transactions to minimize fraud losses and increase customer experience by using various machine learning algorithms such as Logistic Regression, Classification, Random Forests and Clustering in **SAS, R and Python**.
* Used **Pandas, Numpy, Seaborn, Scipy, Matplotlib, Scikit-learn, NLTK** in **Python** for implementing various machine learning algorithms.
* Used **SAS, SQL, Oracle, Teradata** and **MS Office** analysis tools to complete analysis requirements. Created **SAS** data sets by extracting data from Oracle database and flat files
* Used **Proc SQL, Proc Import, SAS Data Step** to clean, validate and manipulate data.
* Performed updating data by weekly and monthly; maintained, manipulated the data for database management. Used the **SAS** **Macro** and **Excel Macro** for the monthly production.
* Experienced in **SQL** queries to retrieve and validate data, prepared for data mapping document.
* Experienced in creating dashboards using **SSAS** and building matrix and tabular reports using reporting services
* Worked on **RDBMS** like **MySQL** and **NoSQL** databases like **MongoDB.**
* Used the **Agile Scrum** methodology to build the different phases of software development life cycle.

**Environment:** SAS9.4, Base SAS, SAS Macros SAS Graph, SAS Access, SAS STAT, SAS ODS, SAS SQL, SAS/ETL, SAS/Stat, SAS ENTERPRISE Miner, Python, PL/SQL, Oracle 9i, Hadoop, MongoDB

**Client: CONCHO - Midland, TX Feb 2014 - Nov 2017**

**Data analyst**

**Description:** Concho pursues oil and gas exploration and development exclusively in this part of Texas and New Mexico, one of the most prolific producing oil and gas regions in the U.S. Our deep knowledge of the area enables us to seek high-quality assets while capitalizing on the strategic benefits of being one of the largest operators in the region

**Responsibilities**:

* Involved in moving legacy data from Sybase data warehouse to **Hadoop Data Lake** and migrating the data processing to lake.
* Responsible for creating Data store, Datasets and Virtual Warehouse in the lake and then creating Spark and Hive refiners to implement the existing **SQL** Stored Procedures.
* Created Java based Spark refiners to replace existing **SQL** Stored Procedures.
* Created Hive refiners for simple **UNIONS and JOINS.**
* Have experience in executing Hive Queries using **Spark SQL** that integrates Spark environment.
* Implemented near real time data pipeline using framework based on **Kafka, Spark** and **MemSQL**.
* Used **REST** services in **Java** and **Spring** to expose data in the lake.
* Automated the triggering of **Data Lake** **REST API** calls using Unix Shell Scripting and PERL.
* Created reconciliation jobs for validating data between source and lake.
* Used Scala to test Dataframe transformations and debugging issues with data.
* Redesigned and implemented Scala **REPL** (read-evaluate-print-loop) to tightly integrate with other **IDE features in Eclipse.**
* Developed multiple **POCs** using Scala and deployed on the **Yarn cluster**, compared the performance of **Spark**, with Hive and **SQL**/**Teradata**.
* Used Avro format for staging data and **ORC** for final repository.
* Worked on the data modeling service which is our own tool (i.e. **PURE MODEL**). I have used the data from data lake virtual warehouse and I have exposed the output of data model to java web services and which has been accessed by the end users.
* Implemented Daily **Cron jobs** that automate parallel tasks of loading the data into **HDFS** and pre-processing with **Pig** using **Oozie** co-coordinator jobs.
* Used **Sqoop** import and export functionalities to handle large **data set** transfer between **Sybase** database and **HDFS**.
* Experience in tuning **Hive** Queries and **Pig scripts** to improve performance.
* Involved in creating **Oozie** workflow and Coordinator jobs to kick off the jobs on time and data availability.
* Knowledge on handling **Hive** queries using **Spark SQL** that integrate **Spark** environment.
* Used **Eclipse** and Ant to build the application.
* Performed unit testing and integration testing using **Junit framework.**
* Configured build scripts for multi module projects with **Maven** and **Jenkins.**
* Responsible to manage data coming from different sources and involved in HDFS maintenance and loading of structured and unstructured data
* Designing technical architecture and developed various **Big Data** workflows using custom **MapReduce**, Pig, Hive and **SQOOP.**
* Responsible for creating Data store, Datasets and **Virtual Warehouse** in the lake and then creating **Spark** and Hive refiners to implement the existing **SQL Stored Procedures**.
* Involved in moving legacy data from **Sybase** ASE data warehouse to **Hadoop Data Lake** and migrating the data processing to lake.
* Built re-usable **Hive UDF** libraries for business requirements which enabled various business analysts to use these **UDF's** in **Hive** querying.
* Used **Maven** extensively for building jar files of **MapReduce** programs and deployed to Cluster.
* Assigned the tasks of resolving defects found in testing the new application and existing applications.

**Environment:** Hadoop, HDFS, Pig, Hive, Spark, Scala, Oozie, Sqoop, HBase, Sybase, Java, Kafka, UNIX, Maven, Junit, SVN, MapR.

**Client: Echidna Software pvt ltd - Bengaluru, IN Oct 2012 – Jul 2013**

**Java Developer**

**Description:** Echidna began when a small group of eCommerce leaders knew there had to be a better way to do eCommerce. So they branched off and created a new kind of agency one that combines amazing UX, enterprise-level technology implementation, and value-added marketing and analytics services.

**Responsibilities**:

* Designed the application by implementing Struts based on **MVC** Architecture, used simple **Java Beans** as a **Model, JSP-UI** Components as **View** and **Action** **Servlet as a Controller**.
* Implemented **EJB's** Container Managed Persistent strategy.
* Requirement **gathering, Design Analysis** and **Code development**.
* Implemented Struts framework based on the **Model View Controller** design paradigm.
* Implemented the **MVC** architecture using **Struts MVC**.
* Used **JDBC** for data access from Oracle tables.
* Worked on triggers and stored procedures on Oracle database.
* Worked on Eclipse IDE to write the code and integrate the application.
* Application was deployed on WebSphere Application Server.
* Coordinated with testing team for timely release of product.
* Apache **ANT** was used for the entire build process.
* **JUnit** was used to implement test cases for beans.

**Environment**: Java, JSP, Servlets, JMS, JavaScript, Eclipse, WebSphere, PL/SQL, Oracle, Log4j, JUnit, ANT, Clear-case, Windows.

**EDUCATIONAL DETAILS**

1.Colorado State University, PhD, May 2022

2.Masters - Colorado State University, 2017

3.Bachelors - SASTRA University, 2013