**SRILATHA V**

Cell: 972-503-9357

Email- [srilathav0805@gmail.com](mailto:srilathav0805@gmail.com)

**PROFESSIONAL SUMMARY:**

* Over 6 years of professional experience as a Data Scientist working across Healthcare, Supply Chain Management, Insurance with a master's degree in Business Analytics.
* Experience in **Data Acquisition, Data Validation**, **Predictive modeling**, **Data Visualization.**
* Adept in statistical programming languages like **R** and **Python** including Big Data technologies like **Hadoop, Hive**.
* Experienced in Data Cleaning, Model Training and Building, Model Testing and Model Deployment.
* Knowledge in **SDLC** using Development, Component Integration, Performance Testing, Deployment, Support Maintenance.
* Worked in Hadoop platform using **Hive** and **Pig**. Hive is used to convert the series into **MapReduce** using the **Hive** Framework.
* Experienced in **SQL programming** and creation of relational **database** models. Experienced in creating cutting-edge data processing algorithms to meet project demands.
* Involved in writing the complex structured queries using **views, triggers**, and **joins.** Worke­d with packages like **Matplotlib, Seaborn, and pandas** in **Python.**
* Connected **python** with **Hadoop Hive** and **Spark** and performed **data analytics.** Worked on Large datasets of **structured**, **unstructured** and **semi-structured data**.
* Experienced in **Linear Regression, Logistic Regression, Random Forest, Decision Trees, Naïve Bayes, K-Means.**
* Worked in Current Techniques and approaches **in Natural Language Processing.** Better Understanding of Statistical Analysis and **Modeling**, **Algorithms** and **Multivariate Analysis** and familiar with **model selection**, **testing, comparison,** and **validations**.
* Worked with the **Amazon Web Services** Environment for database storage. Identify problems and provide solutions to business problems using **data processing**, **data visualization.**
* Worked with python libraries like **matplotlib, numpy, scipy and pandas** for **data analysis.** Connected python with Hadoop to perform **Hive** and **Spark** to perform **data analysis**.
* Ability to work independently and problem-solving skills as the part of the team. Excellent skills in using **pandas** in **python** and **dplyr** in **R** for performing **exploratory analysis.**
* Good understanding of installations, configurations, supporting and managing Hadoop clusters using **Amazon Web Services**(AWS)
* Good understanding of **Natural Language Processing.** Good knowledge of **analytical** and **problem-solving skills** and able to work within the team and as an individual.
* Extensive experience in **Data Visualization** using **tables, lists,** and tools like **Tableau.** Experience in Business Intelligence tools like **SSIS, SSRS,** and **ETL.**
* Proficient in design and development of various **Dashboards**, Reports utilizing **Tableau Visualizations** like bar graphs, scatter plots, pie-charts, Geographic's and other making use of actions, local and global filters, cascading filters, context filters, Quick filters, parameters according to the end user requirements
* Experience in **Apache Spark**, **Kafka** for Big Data Processing & Scala Functional programming. In depth knowledge and hands on experience of **Big Data/ Hadoop** ecosystem (**MapReduce, HDFS, Hive, Pig** and **Sqoop**).
* Experience in manipulating the large data sets with R packages like **tidyr, tidyverse, dplyr reshape, lubridate**, Caret and visualizing the data using lattice and ggplot2 packages. Experience in dimensionality reduction using techniques like PCA and LDA.
* Very good experience and knowledge in provisioning virtual clusters under AWS cloud which includes services like **EC2, S3,** and **EMR.**
* Experience in migration from heterogeneous sources including Oracle to **MS SQL Server**. Experience in writing **SQL queries** and working with **various databases** (MS Access, MySQL, Oracle DB).
* Experience in data analytics, predictive analysis like **Classification, Regression, Recommender Systems**. Experience in developing Custom Report and different types of **Tabular Reports**, **Matrix Reports**, **Ad hoc reports** and distributed reports in multiple formats using **SQL Server Reporting Services** (SSRS).

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Languages | R, SQL, Python,shell scripting,Java |
| IDE | R studio, Jupyter, Eclipse,NetBeans,Atom |
| Databases | Oracle 11g,SQL Server,MS Access,MySQL,MangoDB,PL/SQL, Cassandra,T-SQL,ETL. |
| Big data Ecosystems/Data science | Hadoop,MapReduce,HDFS,HBase,Hive,Pig,Spark MLLib,ETL,Statistical analysis, Machine Learning. |
| Operating systems | Windows XP 7/8/10, Linux, Unix |
| Packages | ggplot2, caret, dplyr, RWeka, gmodels, RCurl, tm, C50, Wordcloud, Kernlab, Neuralnet, twitter, NLP, Reshape2, rjson, plyr, pandas, numpy, seaborn, scipy, matplotlib, scikit-learn, Beautiful Soup, Rpy2, Tensorflow, pytorch, CNN, RNN |
| Web Technologies | HTML, CSS, PHP, JavaScript. |
| Data Analytics Tools | R console,Python(numpy,pandas,scikit-learn,scipy),SPSS |
| Version Controls | GIT, SVN |
| Machine Learning Algorithms | Linear,Logistic, K means clustering,Random forests,support vector machine,Decision Trees,Naïve Bayes. |
| BI and Visualization tools | Tableau,QuickView. |

**WORK EXPERIENCE:**

**Data Scientist January 2018 – present**

**Centene, St.Louis, MO**

**Responsibilities:**

* **Compiled**, **defined,** and **analyzed data** requirements for projects. Prepared and presented **data reports** and offered **analytical** and **statistical interpretations** using **Machine Learning Algorithms.**
* Performed data pre-processing and cleaning to prepare the data sets for further statistical analysis.
* Developed **Machine Learning** algorithms to find the number of health insurance claimed and provided insights for even smarter **healthcare.**
* Analyzed **supervised** and **unsupervised** data, developed **predictive models**, **statistical modeling** by applying **machine learning algorithms.**
* Performed Data Analysis using **Pandas, NumPy, Seaborn, SciPy, Matplotlib, Scikit-learn** and **NLTK** and developed various machine learning algorithms such as **linear regression**, **multivariate regression, naïve Bayes, K-means, KNN and random forest.**
* Experimented with predictive models including Logistic Regression, Support Vector Machine (SVC), Random Forest Provided by Scikit-Learn, XGBoost, LightGBM.
* Re-built the existing model and increased its accuracy from **68%** to **75%** using Advanced MI Algorithms.
* Managed entire data science project life cycle and actively involved in all the phases of project life cycle including **data acquisition**, **data cleaning**, **data engineering**, **features scaling**, **features engineering**, **statistical** **modeling**, **dimensionality reduction** using **Principal Component Analysis** and **Factor Analysis**, **testing** and **validation using ROC plot, K- fold cross validation** and **data visualization**.
* Performed data parsing, data manipulation and data preparation with methods including describe data contents, compute descriptive statistics of **data, regex, split and combine, Remap, merge, subset, re index, melt and reshape.**
* Designed and implemented **Cross-validation** and **statistical tests** including **k-fold**, **stratified k-fold**, hold-out scheme to test and verify the models' significance.
* Implemented **Natural Language Processing**(NLP) based classification to categorize various insurances.
* Worked on **large datasets**, **acquired data** and **cleaned the data**, analyzed trends by making visualizations using matplotlib. Created reports to show the insights, helped to make correct decisions that advance patient care, show the customers graphs and reduce the price and improve health.
* Implemented, tuned and tested the model on **AWS Lambda** with the best performing algorithm and parameters.
* Designed and Implemented cross-validation and statistical tests including k-fold, stratified k-fold, hold-out schema to test and verify the model’s significance.
* Imported data using Sqoop to load data from **MySQL to HDFS** on regular basis. Developed Scripts and Batch Job to schedule various **Hadoop** Program.
* Used Spark platform for analysis using **Pyspark** library and performed splitting of date into clusters on **AWS**.
* Developed the load script for **Extracting**, **Transformation** and **Loading** the data into **Tableau** applications.
* Designed and developed the new interface elements and objects as required and **SET ANALYSIS** to provide functionality using **Tableau**.
* Created analytical reports by running **SQL queries** against databases and presented to both BA teams.
* Designed, developed and maintained daily and monthly summary, trending and benchmark reports in **Tableau Desktop.**

**Environment**: AWS (S3, EC2), Python, Tableau, scikit-learn, Linear Regression, KMeans, Random forest, KNN, Spark.

**Data Scientist**  **January 2017- January 2018**

**Insync, San Francisco, CA**

**Responsibilities:**

* Worked as **data identification**, **collection,** and **exploration**, **cleaning the model**, **participate in model identification.**
* Worked with both **supervised** and **unsupervised** data algorithms and evaluated the models, **tested** and **validated** before selecting the best fit model for predictions.
* Working with large sets of **complex** datasets that include **structured, semi-structured** and **unstructured** data and discover meaningful business insights.
* Utilized **Machine** **Learning** **Algorithms** such as **Linear** **Regression**, **Logistic Regression**, **Naive Bayes**, **Random** **Forests,** **K-means**, & **KNN** for **data** **analysis**.
* Skilled in **Advanced Regression Modelling**, **Correlation**, **Multivariate Analysis**, **Model Building**, **Business Intelligence tools** and application of **Statistical Concepts**.
* **Extracted, Transformed** and **loaded** the data into a database using python scripts. Involving in gathering information while uncovering and defining multiple dimensions.
* Enforced **F-Score**, **AUC/ROC**, **Confusion Matrix**, **Precision**, and **Recall** evaluating different model's performance.
* Participated in features engineering such as feature generating, **PCA**, **Feature normalization** and label encoding with **Scikit-learn preprocessing**. Data Imputation using variant methods in Scikit-learn package in Python. Worked with **Tableau** Reports to test and validate data integrity to the reports. Evaluated the performance of various **models** based on real datasets.
* Completed many tasks from collecting the data and exploring the data and interpreting the **statistical** information.
* Identifying the **data needs** and **requirements** and work with other members of the IT- organization to deliver proper **Data** **Visualization** and reporting solutions to those needs.
* Created **classification models** to recognize web request with product association to classify the orders and scoring the products for analytics which improved the online sales percentage by 13%.
* Used **pruning algorithms** to cut away the connections and perceptron’s to significantly improve the performance of back-propagation algorithm.
* Customer segmentation based on their behavior or specific characteristics like age, region, income, geographical location and applying **Clustering algorithms** to group the customers based on their similar behavior patterns.
* Conducting studies, rapid plots and using **advanced data mining** and **statistical modeling techniques** to build a solution that optimizes the quality and performance of data.
* Worked with several outlier algorithms like **Z-score, PCA, LMS**, and **DBSCAN** to better process the data for higher accuracy.
* Developed the model with ~1.4million data points and used the elbow method to find the optimal value of K using the **Sum of Squared error** as the **error measure**.
* Implemented **Pearson's Correlation** and **Maximum Variance techniques** to find the key predictors for the Regression models.
* Used **Tableau** to convey the results by using dashboards to communicate with team members and with other data science teams, marketing, and engineering teams. Enforced model **Validation using test** and **Validation sets** via **K- fold cross validation**, **statistical significance testing**.
* Designed predictive models using the **Machine** **Learning** platform. Used query languages such as **SQL, HiveQL, Piglatin.** Worked with a large dataset and deep learning class using **TensorFlow**.
* Worked with numerous **data visualization** tools in python like **matplotlib, seaborn, ggplot, pygal.**

**Environment:** Apache Spark 2.0.2, Tableau9.3, logistic regression, random forest, neural networks, SVM, JSON, XML, MLLib, Tensorflow.

**Data Analyst November 2014 – December 2016**

**GEICO,India**

**Responsibilities:**

* Collaborated with my manager to gather all the information needed for **data analysis** and **databases** and analyzed the raw data.
* Implemented and designed predictive models using Natural Processing Language Techniques and machine learning algorithms such as **linear**, **logistic** and **multivariate** **regression**, **random** **forests**, **k** **means** **clustering**, **KNN**, **PCA** for **data** **analysis**.
* Involved in all aspects like **data** **collection**, **data** **cleaning**, **developing models**, **visualization.**
* Maintenance of **large data sets**, combining datasets from various sources like **Excel** and **SQL Queries**. Writing SQL Scripts to select the data from the serves and modify the data as the need of **python pandas** and stored back to the different **database servers**.
* Created **action filters**, **parameters** and calculated sets for dashboards and worksheets in the tableau. Published **customized reports** and **dashboards**, **report scheduling** using tableau server and used **Teradata** **SQL Queries** using **Teradata SQL Assistant.**
* Performed **data cleaning**, **exploring analysis** and **feature engineer** using R. Performed **data visualization** with **tableau** and generated the findings and enchanted **customer** **satisfactions**.
* Programmed in python that used in packages like **numpy, pandas** and **scipy**. Developed the content involving in data manipulation, visualization, **Machine Learning,** and **SQL.**
* Implemented classification using supervised algorithms like Decision trees, KNN, Logistic Regression, Naive Bayes.
* Good knowledge of **Hadoop Architecture** and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Secondary Name Node, and **MapReduce** concepts.
* Understanding and analyzing the data using appropriate **statistical models** to generate insights.

**Environment**: R, Python, Tableau 6.0, ML Lib, PL/SQL, HDFS, Teradata 12, HADOOP (HDFS), HIVE, AWS.

**SQL Developer October 2013 – November 2014**

**Info Edge Solutions,India**

**Responsibilities:**

* Experience in Transforming logic into database design and maintaining into **SQL tools** like **Stored Procedures**, **Views** and **User Defined Functions**.
* Worked with **Database Analysts** and Architect for the design of **Summary Tables** required.
* Used **T-SQL** for Stored Procedures to transfer the data from the databases to initial stages and transfer into **data warehouse**.
* Formated **report** and **sub-reports** layout using **Expressions**, **Global Variables** and **Functions**.
* Worked in **SQL** and **T-SQL Views**, **Joins**, **Stored Procedures** and **Indexes**. Tuned and Optimized **SQL Queries** using **Execution plans**.
* Used **Joins** and **Sub-Queries** to simplify the **complex queries** involving **multiple tables**. Worked with **Stored Procedures**, **Triggers,** **Functions**, **User-Defined Functions**, **Indexes**.
* Analyzed User Requirements and Specifications for various database applications and Identified Issues and resolved them with the database.
* Developed **charts** and **graphs** like **line charts**, **pie charts**, **bar charts** by using chart expert.
* Enhanced the database by creating the **clustering** and **non-clustering** indexes and indexes views.
* Extract, Transform and Load(**ETL**) development using SQL server and SQL integrated services 2008.
* Analyzed data using complex **SQL queries** across various databases.
* Used **SSIS** and **T-SQL** stored procedures to transfer data from **OLTP** databases to staging area and finally transfer into data marts.
* Performed maintenance of stored procedures to improve performance of different front-end applications.
* Scheduled the reports to run on the daily basis and weekly basis in Report Manager and analyst to review the excel sheets.
* Actively involved in **Software Development Life Cycle** (SDLC) involving Systems Initiation, Analysis, Plan, Design and Development, Implementation.

**Environment**: SQL SERVER 2008, VISUAL STUDIO 2008, Microsoft office 2010 PowerPoint office.