Sri Harsha Bokka
sriharshabokka3@gmail.com
732-902-8339

PROFESSIONAL SUMMARY:

- A data scientist professional with **8+ years** of IT Experience which includes **6 Years** of progressive experience in **Data Analytics, Statistical Modeling, Visualization** and **Machine Learning**. Excellent capability in collaboration, quick learning and adaptation.
- Experience in **Data mining** with large datasets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive modeling, Data Visualization.
- Experience in integrating data, profiling, validating and data cleansing transformation and data visualization using **R and Python**.
- Theoretical foundations and practical hands-on projects related to **(i) supervised learning (linear and logistic regression, boosted decision trees, Support Vector Machines, neural networks, NLP), (ii) unsupervised learning (clustering, dimensionality reduction, recommender systems), (iii) probability & statistics, experiment analysis, confidence intervals, A/B testing, (iv) algorithms and data structures**.
- Extensive knowledge on **Azure Data Lake** and **Azure Storage**.
- Experience in migration from heterogeneous sources including **Oracle** to **MS SQL Server**.
- Hands on experience in design, management and visualization of databases using **Oracle, MySQL and SQL Server**.
- In depth knowledge and hands on experience of **Big Data / Hadoop ecosystem (MapReduce, HDFS, Hive, Pig and Sqoop)**.
- Experience in **Apache Spark, Kafka** for **Big Data Processing & Scala Functional programming**.
- 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**.
- Intensive hands-on Boot camp on **Data Analytics** course spanning from Statistics to Programming including data engineering, data visualization, machine learning and programming in **R, SQL**.
- Experience in data analytics, predictive analysis like Classification, Regression, Recommender Systems.
- Good Exposure with **Factor Analysis, Bagging and Boosting algorithms**.
- Experience in Descriptive Analysis Problems like Frequent **Pattern Mining, Clustering, Outlier Detection**.
- Worked on Machine Learning algorithms like Classification and Regression with **KNN Model, Decision Tree Model, Naïve Bayes Model, Logistic Regression, SVM Model and Latent Factor Model**.
- Hands-on experience on **Python** and libraries like **Numpy, Pandas, Matplotlib, Seaborn, NLTK, Sci-Kit learn, SciPy**.
- Expertise and knowledge in **TensorFlow** to do machine learning/deep learning package in **python**.
- Good knowledge on **Microsoft Azure SQL, Machine Learning** and **HDInsight**.
- Good Exposure on **SAS analytics**.
- Good Exposure in deep learning with Tensor flow in **python**.
- Good Knowledge on **Natural Language Processing (NLP)** and **Time Series Analysis** and Forecasting using **ARIMA model** in **Python and R**.
- Good knowledge in **Tableau, Power BI** for interactive data visualizations.
- In-depth Understanding in **NoSQL databases** like **MongoDB, HBase**.
- Very good experience and knowledge in provisioning virtual clusters under **AWS** cloud which includes services like **EC2, S3, and EMR**.
- Experience and Knowledge in developing software using **Java, C++ (Data Structures and Algorithms)** technologies.
- Good exposure in creating pivot tables and charts in **Excel**.
- 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)**.
- Excellent **Database administration (DBA)** skills including user authorizations, Database creation, Tables, indexes and backup creation.

EDUCATION

Data Science Tech Institute, Paris, France
MSc in Applied Data Sciences and Big Data, Data

Indian Institute of Technology Bombay, India
M. Tech (Technology and Development)

Vellore Institute of Technology Vellore, India
B. Tech (Mechanical Engineering)

TECHNICAL SKILLS

Languages	Core, Python, R
Python and R	Numpy, SciPy, Pandas, Scikit-learn, Matplotlib, Seaborn, ggplot2, caret, dplyr, purrr, readxl, tidyr, Rweka, gmodels, RCurl, C50, twitter, NLP, Reshape2, rjson, plyr, Beautiful Soup, Rpy2
Algorithms	Kernel Density Estimation and Non-parametric Bayes Classifier, K-Means, Linear Regression, Neighbors (Nearest, Farthest, Range, k, Classification), Non-Negative Matrix Factorization, Dimensionality Reduction, Decision Tree, Gaussian Processes, Logistic Regression, Naïve Bayes, Random Forest, Ridge Regression, Matrix Factorization/SVD
NLP/Machine Learning/Deep Learning	LDA (Latent Dirichlet Allocation), NLTK, Apache OpenNLP, Stanford NLP, Sentiment Analysis, SVMs, ANN, RNN, CNN, TensorFlow, MXNet, Caffe, H2O, Keras, PyTorch, Theano, Azure ML
Cloud	Google Cloud Platform, AWS, Azure
Web Technologies	JDBC, HTML5, DHTML and XML, CSS3, Web Services, WSDL
Data Modelling Tools	Erwin r 9.6, 9.5, 9.1, 8.x, Rational Rose, ER/Studio, MS Visio, SAP Power designer
Big Data Technologies	Hadoop, Hive, HDFS, MapReduce, Pig, Kafka
Databases	SQL, Hive, Impala, Pig, Spark SQL, Databases SQL-Server, My SQL, MS Access, HDFS, HBase, Teradata, Netezza, MongoDB, Cassandra.
Reporting Tools	Tableau, PowerBI, Business Intelligence, SSRS, Business Objects 5.x/ 6.x, Cognos7.0/6.0., MS Office (Word/Excel/Power Point/ Visio)
ETL Tools	Informatica Power Centre, SSIS.
Version Control Tools	SVM, GitHub
BI Tools	Tableau, Tableau Server, Power BI, SAP Business Objects, OBIEE, QlikView, SAP Business Intelligence, Amazon Redshift, or Azure Data Warehouse
Operating System	Windows, Linux, Unix, Macintosh HD, Red Hat

PROFESSIONAL EXPERIENCE

Vertex Pharmaceuticals – Boston, MA
Sep 2016 – Current
Sr. Data Scientist

Vertex is a global biotechnology company that invests in scientific innovation to create transformative medicines for people with serious and life-threatening diseases. Vertex work with leading researchers, doctors, public health experts and other collaborators who share our vision for transforming the lives of people with serious diseases, their families and society.

Responsibilities

- Analyze Data and Performed Data Preparation by applying historical model on the data set in **AZURE ML**.
- Perform **Data cleaning** process applied Backward - Forward filling methods on dataset for handling missing value.
- Perform **Data Transformation** method for **Rescaling** and **Normalizing Variables**.
- Develop a predictive model and validate **KNN model** for predict the feature label.
- Plan, develop, and apply **leading-edge analytic** and **quantitative tools** and **modeling techniques** to help clients gain insights and improve decision-making.
- Utilize **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.
- Apply 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**.

- Perform data cleaning and feature selection using **MLlib package** in **PySpark** and working with deep learning frameworks such as **Caffe, Neon**.
- Leverage the most appropriate algorithms and be prepared to justify your decisions.
- Work closely with key stakeholders in product, finance and operations to form deep understanding of growth and marketplace dynamics, including product and pricing patterns, outlier detection, forecasting, and imputation.
- Collaborate with product and engineering to integrate various sources of data.
- Apply **strict sampling, statistical inference, and survey techniques** to derive insights from small samples of data.
- Utilize **Sqoop** to ingest real-time data. Used analytics libraries **Sci-Kit Learn, MLLIB and MLxtend**.
- Extensively use **Python's** multiple data science packages like **Pandas, NumPy, matplotlib, Seaborn, SciPy, Scikit-learn and NLTK**.
- Work on data pre-processing and cleaning the data to perform feature engineering and performed data imputation techniques for the missing values in the dataset using **Python**.
- Implement **machine learning model (logistic regression, XGBoost, SVM)** with **Python Scikit- learn**.
- Work on different data formats such as **JSON, XML** and applied **machine learning algorithms** in **Python**.
- Performed Exploratory **Data Analysis**, trying to find trends and clusters.
- Develop rigorous **data science models** to aggregate inconsistent real-time signals into strong predictors of market trends.
- Automate and own the end-to-end process of modeling and **data visualization**.
- Collaborate with **Data Engineers and Software Developers** to develop experiments and deploy solutions to production.
- Create and publish multiple dashboards and reports using **Tableau server**.
- Work on **Text Analytics, Naive Bayes, Sentiment analysis, creating word clouds** and retrieving data from **Twitter** and other **social networking platforms**.
- Work on data that was a combination of unstructured and structured data from multiple sources and automate the cleaning using **Python scripts**.
- Perform data analysis by using **Hive** to retrieve the data from **Hadoop cluster, SQL** to retrieve data from **Oracle database**.
- Create **Data Quality Scripts** using **SQL** to validate successful data load and quality of the data.
- Create data visualizations using **Python and Tableau**.
- Extract data from **HDFS** and prepared data for exploratory analysis using **data munging**.
- Interface with supervisors, artists, systems administrators, and production to ensure production deadlines are met.
- Extensively perform large data read/writes to and from **csv and excel** files using **pandas**.
- Tasked with maintaining **RDD's** using **SparkSQL**.
- Communicate and coordinate with other departments to collection business requirement.
- Tackle highly imbalanced Fraud dataset using under sampling with ensemble methods, oversampling and cost sensitive **algorithms**.
- Improve fraud prediction performance by using random forest and gradient boosting for feature selection with **Python Scikit-learn**.
- Optimize **algorithm** with **stochastic gradient descent algorithm** Fine-tuned the **algorithm parameter** with manual tuning and automated tuning such as **Bayesian Optimization**.
  Write research reports describing the experiment conducted, results, and findings and also make strategic recommendations to technology, product, and senior management.

Environment: Python 2.x, R, HDFS, Hadoop 2.3, Hive, Linux, Spark, IBM SPSS, Tableau Desktop, SQL Server 2012, Microsoft Excel, Matlab, Spark SQL, Pyspark.

Capital One– McLean, VA **Jan 2015 – Aug 2016**

Data Scientist

Capital One is diversified bank that offers a broad array of financial products and services to customers, small businesses and commercial clients. A fortune 500 company. Capital one has one of the most widely used recognized brands in America. As one of the nation's top 10 largest banks based on deposits, Capital One serves banking customers through branches.

Responsibilities:

- Implemented **Data Exploration** to analyze patterns and to select features using **Python SciPy**.

- Built **Factor Analysis** and **Cluster Analysis models** using **Python SciPy** to classify customers into different target groups.
- Built **predictive models** including **Support Vector Machine, Random Forests** and **Naïve Bayes Classifier** using **Python Scikit-Learn** to predict the personalized product choice for each client.
- Using **R's dplyr** and **ggplot2 packages**, performed an extensive graphical visualization of overall data, including customized graphical representation of revenue reports, specific item sales statistics and visualization.
- Designed and implemented cross-validation and statistical tests including **Hypothetical Testing, ANOVA, Auto-correlation** to verify the models' significance.
- Designed an **A/B experiment** for testing the business performance of the new recommendation system.
- Supported **MapReduce Programs** running on the **cluster**.
- Evaluated business requirements and prepared detailed specifications that follow project guidelines required to develop written programs.
- Configured **Hadoop cluster** with **Namenode** and **slaves** and formatted **HDFS**.
- Used **Oozie workflow** engine to run multiple **Hive and Pig jobs**.
- Participated in **Data Acquisition** with **Data Engineer** team to extract historical and real-time data by using **Hadoop MapReduce** and **HDFS**.
- Performed **Data Enrichment** jobs to deal missing value, to normalize data, and to select features by using **HiveQL**.
- Developed multiple **MapReduce** jobs in **java** for data cleaning and pre-processing.
- Analyzed the partitioned and bucketed data and compute various metrics for reporting.
- Involved in loading data from **RDBMS** and **web logs** into **HDFS** using **Sqoop and Flume**.
- Worked on loading the data from **MySQL** to **HBase** where necessary using **Sqoop**.
- Developed **Hive queries** for Analysis across different banners.
- Extracted data from **Twitter** using **Java** and **Twitter API. Parsed JSON** formatted twitter data and uploaded to database.
- Launching **Amazon EC2 Cloud** Instances using **Amazon Images (Linux/ Ubuntu)** and Configuring launched instances with respect to specific applications.
- Developed **Hive queries** for analysis, and exported the result set from **Hive** to **MySQL** using **Sqoop** after processing the data.
- Analyzed the data by performing **Hive queries** and running **Pig scripts** to study customer behavior.
- Created **HBase** tables to store various data formats of data coming from different portfolios.
- Worked on improving performance of existing **Pig** and **Hive Queries**.
- Created reports and dashboards, by using **D3.js** and **Tableau 9.x,** to explain and communicate data insights, significant features, models scores and performance of new recommendation system to both technical and business teams.
- Utilize **SQL, Excel** and several **Marketing/Web Analytics tools** (**Google Analytics, AdWords**) in order to complete business & marketing analysis and assessment.
- Used **Git 2.x** for version control with **Data Engineer team** and **Data Scientists colleagues**.
- Used **Agile methodology** and **SCRUM** process for project developing.

Environment: HDFS, Hive, Scoop, Pig, Oozie, Amazon Web Services (AWS), Python 3.x (SciPy, Scikit-Learn), Tableau 9.x, D3.js, SVM, Random Forests, Naïve Bayes Classifier, A/B experiment, Git 2.x, Agile/SCRUM.

Capital Fortunes Pvt. Ltd, - **Hyderabad, India** **June 2012 – July 2013**
Data Analyst/Data Scientist

Responsibilities:
- Developing **Data Mapping, Data Governance, Transformation** and Cleansing rules for the **Master Data Management (MDM) Architecture** involving **OLTP, ODS** and **OLAP**.
- Providing source to target mappings to the **ETL** team to perform initial, full, and incremental loads into the target data mart.
- Conducting **JAD sessions**, writing meeting minutes, collecting requirements from business users and analyze based on the requirements.
- Involved in defining the source to target **data mappings**, business rules, and data definitions.
- Transformation on the files received from clients and consumed by **Sql Server**.

- Working closely with the **ETL, SSIS, SSRS** Developers to explain the complex Data Transformation using Logic.
- Worked on **DTS Packages, DTS Import/Export** for transferring data between **SQL Server 2000 to 2005**.
- Performing **Data Profiling, Cleansing, Integration** and extraction tools
- Defining the list codes and code conversions between the source systems and the data mart using **Reference Data Management (RDM)**.
- Applying data cleansing/data scrubbing techniques to ensure consistency amongst data sets.
- Extensively using **ETL methodology** for supporting data extraction, transformations and loading processing, in a complex **EDW**.

Environment: MS Excel, Agile, Oracle 11g, Sql Server, SOA, SSIS, SSRS, ETL, UNIX, T-SQL, HP Quality Center 11, RDM (Reference Data Management).

Ford India Pvt Ltd - **Chennai, India** **June 2011 – May 2012** *Data Scientist*

Responsibilities:

- Involved in complete **Software Development Life Cycle (SDLC)** process by analyzing business requirements and understanding the functional work flow of information from source systems to destination systems.
- A highly immersive **Data Science program** involving **Data Manipulation & Visualization, Web Scraping, Machine Learning, Python programming, SQL,Unix Commands, NoSQL, Hadoop**.
- Used **pandas, numpy, seaborn, scipy, matplotlib, scikit-learn, NLTK** in **Python** for developing various machine learning algorithms.
- Worked on different data formats such as **JSON, XML** and performed machine learning algorithms in **Python**.
- Analyzed sentimental data and detecting trend in customer usage and other services.
- Analyzed and Prepared data, identify the patterns on dataset by applying historical models.
- Collaborated with **Senior Data Scientists** for understanding of data.
- Used **Python and R scripting** by implementing machine algorithms to predict the data and forecast the data for better results.
- Used **Python and R scripting** to visualize the data and implemented machine learning algorithms.
- Experience in developing packages in **R** with a **shiny interface**.
- Used predictive analysis to create models of customer behavior that are correlated positively with historical data and use these models to forecast future results.
- Predicted user preference based on segmentation using **General Additive Models**, combined with feature clustering, to understand non-linear patterns between user segmentation and related monthly platform usage features (time series data).
- Perform **data manipulation, data preparation, normalization,** and **predictive modeling**.
- Improve efficiency and accuracy by evaluating model in **Python and R**.
- Used **Python and R script** for improvement of model.
- Application of various machine learning algorithms and statistical modeling like **Decision Trees, Random Forest, Regression Models, neural networks, SVM, clustering** to identify Volume using **scikit-learn package** .
- Performed **Data cleaning process** applied Backward - Forward filling methods on dataset for handling missing values.
- Developed a predictive model and validate **Neural Network Classification model** for predict the feature label.
- Performed Boosting method on predicted model for the improve efficiency of the model.
- Presented Dashboards to Higher Management for more Insights using **Power BI and Tableau**.
- Hands on experience in using **HIVE, Hadoop, HDFS** and **Bigdata** related topics.

Environment: R/R studio, Python, Tableau, Hadoop, Hive, MS SQL Server, MS Access, MS Excel, Outlook, Power BI.

Cognizant Technology Solutions Ltd, Chennai, India **Sep 2006 - May 2009**
ETL Developer

Responsibilities:

- Involved in full **SDLC** of **BI Project** including **Data Analysis, Designing, Development of Data Warehouse environment**.
- Used **Oracle Data Integrator Designer** to develop processes for extracting, cleansing, transforming, integrating, and loading data into **data warehouse database**.
- Experience in Developing and customizing **PL/SQL packages, procedures, functions, triggers** and **reports** using **Oracle SQL Developer**.
- Responsible for designing, developing and testing of the **ETL strategy** to populate the data from various source systems **(Flat files, Oracle)**.
- Worked with the Business units to identify data quality rule requirements against identified anomalies.
- Develop **Data Mapping, Join and queries** – Validation, and addressing/fixing data queries raised by project team in a timely manner.
- Worked closely with **Business analyst** and interacted with the Business users to gather new business requirements and to understand the accurate business and current requirements.
- Created **Repositories, Agent, Contexts** and both of **Physical & Logical Schema** in **Topology Manager** for all the source and **target schemas**.
- **Data mapping, logical data modeling,** created **class diagrams** and **ER diagrams** and used **SQL queries** to filter data within the **Oracle database**.
- Installed and Setup **ODI Master Repository, Work Repository, Execution Repository**.
- Used **Topology Manager** to manage the data describing the information systems physical and **logical architecture**.
- Extensively worked and utilized **ODI Knowledge Modules (Reverse Engineering, Loading, Integration, Check, Journalizing and service).**
- Created various procedures and variables.
- Created **ODI Packages, Jobs of various complexities** and automated **process data flow**.
- Configured and setup **ODI, Master repository, Work repository, Project, Models, sources, targets, packages, Knowledge Modules, Interfaces, Scenarios, filters, condition, metadata**.

Environment: R/R studio, Python, Tableau, Hadoop, Hive, MS SQL Server, MS Access, MS Excel, Outlook, Power BI.