

**Sai Krishna**

**SUMMARY:**

**Above 6+ years** of experience as **Senior Data Scientist/Data Analyst** in Architecture, Identity Innovation Opportunities, Interpreting & Analyzing in a Fast-Paced Environment.

* Extensive experience in **Machine Learning**, **Data mining** with **large datasets** of **Structured** and **Unstructured data**, **Data Acquisition**, **Data Validation**, **Predictive modeling**, **Data Visualization**.
* Extensive experience in Text **Analytics**, developing different **Statistical Machine Learning**, **Data Mining** solutions to various business problems and generating **data visualizations** using **R**, **Python** and **Tableau**.
* Expertise in transforming business requirements into **analytical models**, **designing algorithms**, **building models**, developing **data mining** and **reporting solutions** that scales across massive volume of structured and unstructured data.
* Experience in **designing** stunning visualizations using **Tableau software** and publishing and presenting **dashboards**, Storyline on **web** and **desktop** platforms.
* **Designing** of **Physical Data Architecture** of New system engines.
* Experience with **NoSQL** databases, such as **MongoDB, Cassandra, HBase**.
* Hands on experience in implementing **LDA**, **Naïve Bayes** and skilled in **Random Forests**, Decision Trees, Linear and Logistic Regression, **SVM**, **Clustering**, **neural networks**, Principle Component **Analysis** and **good** knowledge on Recommender Systems.
* Proficient in **Statistical Modeling** and **Machine Learning** techniques **(Linear, Logistics, Decision Trees, Random Forest, SVM, K-Nearest Neighbors, MLbase, Bayesian, XG Boost)** in **Forecasting/ Predictive Analytics, Segmentation methodologies, Regression based models, Hypothesis testing, Factor analysis/ PCA, Ensembles**.
* Strong experience in Software Development Life Cycle (**SDLC**) including **Requirements Analysis**, **Design** Specification and **Testing** as per Cycle in both **Waterfall** and **Agile** methodologies.
* Adept in **statistical programming languages** like **R** and also **Python** including **Big Data** technologies like **Hadoop**, **Hive**.
* Skilled in using **dplyr** and **pandas** in **R** and **python** for performing exploratory **data analysis**.
* Experience with structured or un-structured **data analysis** and tools (**SQL, Hadoop**, **Spark**, **NoSQL**, **MYSQL**, **Hive**, **Pig**, **etc**.
* Adept in using libraries such as **pandas**, **numpy**, **seaborn**, **scipy**, **matplotlib**, **scikit-learn**, **Beautiful Soup**, **MDP**, **Orange**, **Rpy2**, **Lib SVM**, **neuro lab**, **NLTK**.
* Proficient in **managing** entire **data science** project life cycle and actively involved in all the phases of project.
* Skilled in performing **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.
* Good knowledge in **Data Analysis, Data Validation**, **Data Cleansing**, **Data Verification** and identifying **data mismatch.**
* Extensive experience in **data acquiring**, **merging**, **cleaning**, analyzing and **mining structured**, **semi**-**structured** and **unstructured datasets**.
* Good understanding of **Teradata SQL Assistant**, **Teradata Administrator** and **data load** / **export** utilities like **BTEQ, Fast Load**, **Multi Load**, **Fast Export**.
* Experienced in **SQL queries** and optimizing the queries **in Oracle, SQL Server**, **DB2**, and **Netezza** & **Teradata**.
* Experience in modeling with both **OLTP**/**OLAP** systems and **Kimball** and **Inmon Data Warehousing** environments.
* Good industry knowledge, **analytical** and problem solving skills and ability to work well within a team as well as an individual.

**Technical Skills:**

|  |  |
| --- | --- |
| **Data Science tool** | R3.4.2/3.x, Python3.0, and MATLAB. |
| **Database** | Oracle 11g/12c, MS Access, SQL Server 2012/2014, Sybase and DB2, Teradata14/15, Hive. |
| **Machine Learning** | Linear regression, Logistic regression, Decision tree, Random  Forest, K nearest neighbor, K mean, Avro, MLbase. |
| **Data Modeling Tools** | Erwin r9.6/9.5, ER/Studio 9.7, Star-Schema Modeling, Snowflake- Schema Modeling, FACT and dimension tables, Pivot Tables. |
| **BI Tools** | Tableau 7.0/8.2, Tableau server 8.2, Tableau Reader 8.1, SAP Business, Objects, Crystal Reports. |
| **Languages** | SQL, PL/SQL, ASP, Visual Basic, XML, Python, SQL, T-SQL, SQL Server, C, C++, JAVA, HTML, UNIX shell scripting, PERL. |
| **Applications** | Toad for Oracle, Oracle SQL Developer, MS Word, MS Excel, MS Power Point, Teradata. |
| **Big Data** | Hadoop, Spark, Hive, Cassandra, MongoDB, MapReduce, Sqoop, Hive. |
| **Operating Systems** | Microsoft Windows 9x / NT / 2000/XP / Vista/7 and UNIX. |
| **Methodologies** | Agile, RAD, JAD, RUP, UML, System Development Life Cycle (SDLC), Ralph Kimball and Bill Inmon's, Waterfall Model. |
| **Operating System** | Windows, UNIX, Sun Solaris. |

**Work Experience:**

**Care ATC, Tulsa, OK Apr 17 – Till Date**

**Sr. Data Scientist**

**Responsibilities:**

* Worked as **Data Scientist** and used predictive modeling, statistics, **Machine Learning, Data Mining**, and other aspects of **data analytics** techniques to collect, explore, and extract insights from structured and unstructured data.
* Responsible for data identification, collection, exploration & cleaning for modeling, participate in model development.
* Researched, evaluated, architect, and deployed new tools, frameworks and patterns to build sustainable **Big Data platforms** for our clients.
* Utilized **machine learning algorithms** such as **linear regression, multivariate regression**, **Naive Bayes, Random Forests, K-means, & KNN** for data analysis.
* Advanced Text analytics using **Deep learning** techniques such as **Convolution neuralnetworks** to determine the sentiment of texts.
* **Natural Language Processing** to understand text content on including reviews, descriptions and interactions between users on our marketplace.
* Developed **POC** in **Spark** to **Query** the dataset.
* Understand transaction data and develop Analytics insights using Statistical models using **Machine learning.**
* Involved in gathering requirements while uncovering and defining multiple dimensions. Extracted data from one or more source files and Databases.
* Used **Spark** for test data analytics using **MLLib** and Analyzed the performance to identify bottlenecks.
* **Designed** data processing pipelines with a combination of the following technologies: **Hadoop**, **MapReduce**, **Spark**, **Hive**, **Kafka**, **Avro**, **SQL** and **NoSQLdatawarehouses**.
* Converted the **unstructured** data into **structured** data using **ApacheAvro**.
* Designed predictive models using the machine learning platform -**H2O, Flow UI.**
* Used the **Agile Scrum methodology** to build the different phases of Software development life cycle.
* Developed **MapReduce/Spark** modules for machine learning & predictive analytics in **Hadoop** on **AWS**
* Architected overall **master data hub** for data elements that are used by multiple IT systems.
* Defined the architecture and various phases for the implementation of the transactional and data warehouse systems.
* Utilized various new supervised and unsupervised **machine learning** algorithms/software to perform **NLP** tasks and compare performances.
* Worked with **(Tableau) Report Writers** to Test, Validate Data Integrity of Reports.
* Captured Modeling requirements from **Senior Stakeholders** to Bench functional requirements for **SAS**.
* Created reports with **Crystal Reports** and scheduled to run on a daily basis.
* Accomplished multiple tasks from collecting data to organizing data and interpreting statistical information.
* Evaluate the performance of various **algorithms/models/strategies** based on the real world data sets.
* Interacted with the other departments to understand and identify data needs and requirements and work with other members of the IT organization to deliver data visualization and reporting solutions to address those needs.

**Environment**: Erwin 9.6, Apache Spark 2.0.2, Tableau9.3,MLbase, Oracle 12c, Teradata 15, SAS, ODS, Agile, MapReduce, regression, logistic regression, random forest, neural networks, Avro, Topic Modeling, NLTK, XML, MLLib& Json.

**Coca Cola - Atlanta, GA Apr 15 – Feb 17**

**Data Scientist**

**Responsibilities:**

* Participated in all phases of **datamining**; **data collection**, **data cleaning**, **developing models**, **validation**, **visualization** and performed **Gap analysis**.
* Performed **K**-**means clustering**, Multivariate analysis and **Support Vector Machines** in **Python**.
* Used **Pandas**, **Numpy**, **Seaborn**, **Scipy**, **Matplotlib**, **Scikit**-**Learn**, **and NLTK** in **Python** for developing various **machine learning algorithms (MLbase).**
* Developed **Clustering** algorithms and **Support Vector Machines** that improved Customer segmentation and Market Expansion.
* Data Story teller, **Mining** Data from different Data Source such as **SQL**
* **Server**, **Oracle**, **Cube Database**, **Web Analytics**, **Business Object** and **Hadoop.**
* Data Manipulation and Aggregation from different source using **Nexus**, Toad**, Business Objects**, **Power BI** and Smart **View.**
* Developed **Clustering** algorithms and **Support Vector Machines** that improved Customer segmentation and Market Expansion.
* Developed analytics and strategy to integrate **B2B** analytics in outbound calling operations.
* Programmed a utility in Python that used multiple packages **(scipy, Numpy, pandas).**
* Implemented **analytics delivery** on **cloud**-based **visualization** such as Business Object and Google analytics platform.
* **SPOC Data Scientist** and **predictive analyst** to create annual and quarterly Business forecast reports**.**
* Responsible for defining the testing procedures, test plans, error handling strategy and performance **tuning** for **mappings**, Jobs and interfaces.
* Created entity relationship diagrams and multidimensional **data models**, reports and diagrams based on the requirements.
* Gathering all the data that is required from multiple data sources and creating datasets that will be used in **analysis**.
* Perform **data cleaning** and transformations that is suitable for applying models.
* Perform a proper **EDA**, Univariate and bi-variate analysis to understand the intrinsic effect/combined effects.
* Worked on **multiple predictive models** to predict future electricity and gas bills based on the current usage patterns.
* Automate the process of performing stepwise regression and model training based on **AIC** values and **P-**values.
* Perform cross-validation to check MAPE values to self-train the model.
* Performed **statistical tests** and applied regression techniques to predict the power.
* Created dash boards and visualization on regular basis using **ggplot2** and **Tableau.**
* Creating customized business reports and sharing insights to the management.
* Responsible for planning & scheduling new product releases and promotional offers.
* Parsing **data**, producing concise conclusions from raw data in a clean, well-structured and easily maintainable format.
* Worked on **Text Analytics**, **NaiveBayes**, Sentiment analysis, creating word clouds and retrieving data from Twitter and other social networking platforms.

**Environment:** Erwin 9.5, Tableau 8.0, Json, MDM, MLLib, PL/SQL, Teradata 14.1, Spark, R Studio, MAHOUT, AWS, HDFS, NoSQL.

**ATOS - Tarrytown, NY Aug 12 – Mar 15**

**Sr. Data Analyst/Data Scientist**

**Responsibilities:**

* Worked with Business users for requirements gathering, business analysis and project coordination.
* Data blending implemented on databases and generated interactive **dashboards** in **Tableau** Portal
* Worked on data related to enterprise data quality in order to analyze data flow models in **JD Edwards’s environment.**
* Researched on Multi-layer classification algorithms as well as building **Natural Language Processing** model through ensemble.
* Executed and validated data transformations in target system with **SQL**.
* Published Workbooks by creating user filters so that only appropriate teams can view it.
* Work with customers to define reporting needs and implement **Tableau** reports to satisfy needs
* Worked on data cleaning and reshaping, generated segmented subsets using **Numpy** and **Pandas** in **Python**.
* Developed **Python scripts** to automate data sampling process.
* Ensured the data integrity by checking for completeness, duplication, accuracy, and consistency.
* Worked on model selection based on confusion matrices, minimized the Type II error.
* Generated cost-benefit analysis to quantify the model implementation comparing with the former situation.
* Continuously collected business requirements during the whole project life cycle.
* Identified the variables that significantly affect the target.
* Used cascaded parameters to generate a report from two different Data Sets.
* Involved with **Query Optimization** to increase the performance of the Report.
* Employed Filtered Indexes to improve query performance and reduce the storage and maintenance cost.
* Generated data analysis reports using **Matplotlib, Tableau**, successfully delivered and presented the results for **C-level** decision makers.
* Design and create daily, monthly and **ad-hoc** reports for **BI** and Analytics customers  
  Provide recommendations for data presentation within reports/dashboards
* Support for end users connected to servers in various departments regarding connection and data access issues.
* Involved in defining the business/transformation rules applied for sales and service data.
* Played a major role in production support of **SSAS** cube and **SSIS** jobs.
* Involved in writing **MDXqueries** and performance optimization of the **SSAS** cubes.
* Implemented Incremental load, used Event Handlers to clean the data from different data sources
* Wrote and optimized **complex SQL queries** involving multiple joins and advanced analytical functions to perform data extraction and merging from large volumes of historical data stored in **Oracle 11g,** validating the **ETL** processed data in target database.
* Conducted model optimization and comparison using stepwise function based on AIC value
* Data blending implemented on **databases** and generated interactive dashboards.
* Published Workbooks by creating user filters so that only appropriate teams can view it.
* Embedded **Tableau** views in to SharePoint
* Analyzed the source data and handled efficiently by modifying the data types.
* Generated tableau dashboards for sales with forecast and reference lines.
* Created **SSIS** packages to clean and load data to **data warehouse.**
* Created package to transfer data between **OLTP** and **OLAP databases.**
* Created **SSIS** Packages using **Pivot Transformation, Fuzzy Lookup, Derived Columns, Condition Split, Term extraction, Aggregate, Execute SQL Task, Data Flow Task, and Execute Package Task** etc. to generate underlying data for the reports and to export cleaned data from **Excel Spreadsheets, CSV files** to **data warehouse.**
* Applied various **machine learning algorithms** and statistical modeling like **decision tree, logistic regression, Gradient Boosting Machine** to build predictive model using **scikit-learn package** in **Python**.

**Environment:** Python, Oracle, MS Excel, SSIS, SSAS, Tableau, SQL, ETL, decision tree, logistic regression, Gradient Boosting Machine, scikit-learn, OLAP & OLTP, SQL\*Loader.