**Sana Batul**

**Data Scientist/Machine Learning Engineer**

**Email:sanabatul2409@gmail.com Phone No:**  571-386-9748

**Professional Summary:**

* 10+ years of experience in DataScience and Analytics including **MachineLearning, DataMining,** DataBlending&StatisticalAnalysis.
* Experience in **AWS** (AmazonWebServices)**EC2, VPC, IAM,S3**,**AWS Glue, Amazon Sagemaker, Kinesis Video Streams, Elastic Container Services[ECS] , Elastic Kubernetes Services[EKS**], **AWS Fargate,CloudFront, CloudWatch**, **CloudFormation, Glacier, RDSConfig, Route53, SNS, SQS, ElasticCache.**
* Having experience working with Docker Containers and Kubernetes.
* Having good experience in AmazonWebServices (**AWS**).
* AzureCloud Extensive full cycle **CloudAzure** experience with full **BigData, Elasticsearch and SOLR,Machine Learning and DeepLearning** development and deployment.
* Used **Pandas, NumPy, Seaborn, SciPy, Matplotlib, Scikit-learn, NLTK** in **Python** for developing various machine learning algorithms and utilized machine learning algorithms such as **Linear Regression**, **Multivariate Regression, Naive Bayes, Random Forests, K-Means, & KNN for Data Analysis**.
* Hands on experience leading delivery through **Agile** methodologies
* Skilled in using dplyr and pandas in **R** and **Python** for performing Exploratory **data analysis**.
* Experience in designing star schema, Snow flake schema for **Data Warehouse**, **ODS architecture**.
* Experienced with **machine learning** algorithms such as **logisticregression, randomforest, XGboost, KNN, SVM, neuralnetworks, linear regression, lassoregression,and k-means.**
* Expertise in synthesizing **Machine learning**, Predictive Analytics and **BigData** technologies like **Hadoop**, **Hive, Pig.**
* Proficient in **Python, R, SQL, PL/SQL** and **Automation** with **Python/Shell Scripting**.
* Discovering and Visualizing the Data to gain Insights using **Matplotlib, Seaborn,Tableau, PowerBI** and Data-Driven Documents (**D3/DC**) techniques.
* Strong skills in statistical methodologies such as **A/Btest**, experimentdesign, **hypothesistest**, **ANOVA**.
* Extensively worked on **R platform** for the data analysis and predictive modeling.
* Solid ability to write and optimize diverse **SQLqueries**, working knowledge of **RDBMS** like **SQLServer2008.**
* Experience in **BigData** technologies like **Spark1.6, Sparksql, pySpark, Hadoop2.X, HDFS, Hive1.X.**
* Experience in visualization tools **likeTableau9.X, 10.X, DataBlendingfor** creating dashboards.
* Proficient in PredictiveModeling, **DataMiningMethods, FactorAnalysis,** normal distribution and other advanced statistical and econometric techniques.
* Developed predictive models using **DecisionTree**, **RandomForest**, **NaïveBayes**, **LogisticRegression**, **ClusterAnalysis**, and **NeuralNetworks**.
* Excellent knowledge of **MachineLearning**, Mathematical Modeling and Operations Research. Comfortable with **R**, **Python**, **SAS**, **MATLAB**, **Relationaldatabases**. Deep understanding & exposure of **BigDataEco**-**system**.
* Worked with complex applications such as **R, SAS, Matlab,andSPSS** to developa neuralnetwork, cluster analysis.
* Experienced in **BigData** with **Hadoop, HDFS, MapReduce, and Spark**.
* Experienced in **DataIntegrationValidation** and **DataQuality** controls for **ETL** process.
* Proficient in **Tableau** and **R-Shinydatavisualization** tools to analyze and obtain insights into large datasets create visually powerful and actionable interactive reports and dashboards.
* Automated recurring reports using **SQL** and **R** and visualized them on **BIplatform** like **Tableau**.
* Worked in a development environment like **Git** and **VM**.

## Education:

Bachelor of Computer Science.

JNTU, Hyderabad.

Master in Data Analytics.

George Mason University, USA.

## Technical Skills:

|  |  |
| --- | --- |
| **Data Analytics Tools** | Python (numpy, scipy, pandas, Gensim, Keras), R (Caret, Weka, ggplot), MATLAB**.** |
| **Analysis &Modelling Tools** | Erwin, Sybase Power Designer, Oracle Designer, TOAD, Erwin, Rational Rose, ER/Studio, MS Visio, SAS, Django, Flask, pip, NPM, Node JS, Spring MVC. |
| **Data Visualization** | Tableau, Visualization packages, Microsoft Office**.** |
| **Machine Learning** | K-NN, K-Means, KernelSVM,Clustering,Softmax,GradientDescent,Decision Trees, Random Forest,Simple Linear Regression, Multi-Variate Linear Regression,Logistic Regression,Polynomial Regression, Backprop, Feed Forward ANN, CNN, RNN,and Word2Vec. |
| **Machine Learning Frameworks** | Spark ML, Kafka, Spark MiLB, Scikit-Learn & NLTK. |
| **Big Data Tools** | Hadoop, Map Reduce, SQOOP, Pig, Hive, NOSQL, Spark, Apache Kafka, Shiny, Yarn, Data Frames, pandas, ggplot2, Sklearn, Theano, Cuda, Azure, HD Insight, etc. |
| **OLAP Tools** | MS SQL Analysis Manager, DB2 OLAP, Cognos Power-play. |
| **Programming Languages** | C, C++, SQL, PL/SQL, T-SQL, XML, HTML, UNIX Shell Scripting, Microsoft SQL Server, Oracle PLSQL**,** Python, Scala, AWK, JavaScript. |
| **R Package** | dplyr, sqldf, data table, Random Forest, gbm, caret, elastic net and all sort of Machine Learning Packages. |
| **Databases** | SQL Server, Linked Servers, DTS Packages, SSIS, PL/SQL, MS SQL Server,DB2UDB,Teradata, Netezaa,HyperionEssbase, Sybase ASE, Informix, AWS RDS, Cassandra, and MongoDB, PostgreSQL. |
| **Operating Systems** | Windows 2007/8, UNIX (Sun-Solaris, HP-UX), Windows NT/XP/Vista, MSDOS. |

**Professional Experience-**

**Client:Charter Communications– Greenwood Village, CO**

**Role:Data Scientist/Machine Learning EngineerMay2019–Till date**

**Description:** Charter Communications, Inc., is an American telecommunications and mass media company that offers its services to consumers and businesses under the branding of Spectrum.Working as a part of the Charter Emerging Tech R&D team.

**Responsibilities:**

* Involved in extracting value from IoT data and build models based on collected data or data available in the public domain.
* Utilized knowledge of custom developed prototype to develop an automated pipeline for data preprocessing.
* Managed a team of two to develop a prediction model for cargo trains with help of custom developed prototype.
* Experience with working in **Agile/SCRUM** software environments
* Proficient in conducting comprehensive research, developing datasets, and crafting insightful visualizations to effectively communicate findings.
* Leading the testing team for conducting any test all new features and any new releases by the DevOps team for launching of Saint Petersburg client POC for the integration of illuminating concept and Vaisala with the gingersnap platform
* Worked on Data Analytics processes from data collection, cleaning, transformation, and enrichment of IoT data before storing it in a data store for querying and analysis.
* Working on Launchingdifferent phases of Hitachi client POC for the integration of Video analytics use‐cases for security and tracking of various unidentified object or persons within the multiple camera frames.
* Used Agile methodology and Scrum process for project developing.
* Conducted several demos for both Internal and external teams with several use‐cases using the combination MVS﴾Multi‐video search﴿ and HVA ﴾Hitachi video analytics﴿ software.
* Providing requirements to the Hitachi team for Phase2 development and enhancements.
* Successfully finished Building the hardware and network fragment for integration of Hitachi software on 2 con‐current system using 4 camera each.
* Leading and Maintaining the resources deployed for ML stack in AWS and performing software updates and cleaning up resources whenever necessary.
* Working with the Development team to help them release the requirements for St.Pete smart city POC project Worked with the team to release St.Pete smart city POC platfrom 1.0.1 and 1.1.0
* Interacted with other team members and helped on evaluating the **Request for Information [RFI]** responses received from various companies and selected the Top Vendors supporting the IoT Smart City Platforms
* Worked with rest of the team members in developing the **Request for Proposal [RFP]**for Spectrum Smart Cities Questionnaire for **Data Management Layer, Business Logic/Analytics Layer.**
* Developed a Metrics Evaluation tool for the RFP.
* Developed algorithms to analyze data, gather insights and analytics for Object Detection using **TensorFLow, OpenCVand detection models like SSD, R-CNN, Faster R-CNN and YOLOv3**and Python to train, test, evaluate and finally integrate and deploy the algorithms.
* Involved in Developing PoC in our labs then do a technology transfer to the production environment
* Worked on **AWS Glue**toautomate much of the effort in building, maintaining, and running ETL jobs and **AWS Glue to** crawl data sources, identifies data formats, and suggests schemas and transformations**.**
* Developed algorithms, train, tune and validate these based on different datasets (internal/ external data sources) using the **AWS Sagemaker.**
* Worked on PoC for Computer Vision Algorithm & Video Analytics configuration using **Kinesis Video Stream, EC2** instances in AWS
* Developed to provide the capability of capturing, encoding, and streaming video to AWS Kinesis and subsequently subscribing and processing video for vehicle tracking metrics which are streamed to Core IoT.
* Worked on a PoC for Data Management Layer using different Services **DynamoDB,S3**, **AWS Glue ETL Jobs, SageMaker, Jupyter Notebook, Amazon Forecast**and Performed data visualization and Designed dashboards using **AWS Quicksight**
* Currently Working on developingPoC with **Elastic Container Services[ECS] , Elastic Kubernetes Services[EKS**], **AWS Fargate**to seamlessly deploy Docker Containers and Applications into the AWS environment.
* Working on Orchestrating the Docker files and Containers into Kontena Lens IDE used to for Kubernetes.
* Demonstrated ability to write clean, concise, and functional Scala code, leveraging features such as higher-order functions, pattern matching, and immutability to create robust and maintainable solutions.
* Havinghands-on experience with popular Scala libraries and frameworks such as Akka, Play Framework, and Cats, showcasing the ability to develop scalable, reactive, and type-safe applications for various domains.
* Demonstratedin-depth knowledge of Scala's concurrency model, including concurrent data structures, actors, and parallel collections, enabling the design and implementation of efficient and concurrent systems capable of utilizing multicore architectures effectively.

**Client:Principal Global Investors – Des Moines,IA.Aug 2018 – April 2019**

**Role:Data Scientist/Machine Learning Engineer**

**Description:**Principal Global Investors is a limited liability company and a wholly owned, indirect subsidiary of the Principal Financial Group® (“The Principal®”). Our investment expertise spans multiple asset classes and global markets, to help you achieve your investment objectives.

**Responsibilities:**

* Involved in extensive adhoc reporting, routine operational reporting, and data manipulation to produce routine metrics and dashboards for management.
* Created parameters, action filters and calculated sets for preparing dashboards and worksheets in **Tableau**.
* Interacting with other data scientists and architects, custom solutions for data visualization using tools like a **tableau** and Packages in **Python**.
* Involved in running **MapReduce** jobs for processing millions of records.
* Written complex SQL queries using joins and **OLAP** functions like **Count, CSUM, and Rank etc.**
* The building, publishing customized interactive reports, report scheduling and dashboards using Tableau server.
* Worked on Design and implement Internal Intelligent Ticketing Tool by using Text Mining Technics Such as **NLTK, NLP**, and **Google Word2Vec model.D**
* eveloped and implemented machine learning Model for purchasing pattern recognition with Association **Rule/Market BasketAnalysis** technics (Apriori and CrossTab models).
* Used Python to train, test, evaluate and finally integrate and deploy the various **Machine Learning** models.
* Worked on Time Series Model building for Future Sales forecasting with **Arima and Kera's Deep Learning** techniques.
* Collaborating with marketing professionals for writing market insights reports; **Data QualityAssurance**, **Data Cleaning, Data visualization and Production**.
* Wrote a python-based **SQL** generator that helped speed up a weekly reporting from several days effort to just running the automation and getting the report done in a few hours.
* Optimize the **Data Capture** and Data Process Methods by **SQL** Tuning and Multiprocessing techniques.
* Developed in **R programs** for manipulating the data reading from various **SQL** and convert them as one **CSVFiles**.
* Performing statistical data analysis and data visualization using **R.**
* Worked on creating filters and calculated sets for preparing dashboards and worksheets in **Tableau**.
* Created data models in **Splunk** using pivot tables by analyzing the vast amount of data and extracting key information to suit various business requirements.
* Implemented data refreshes on **TableauServer** for biweekly and monthly increments based on a business change to ensure that the views and dashboards were displaying the changed data accurately.
* Developed normalized Logical and Physical database models for designing an OLTP application.
* Knowledgeable in **AWS** Environment for loading data files from on prim to Redshiftcluster.
* Performed **SQLTesting** on **AWS Redshift** databases.
* Developed **Teradata SQL scripts** using **OLAP** functions like rank and rank over to improve the query performance while pulling the data from large tables.
* Developed and implemented **SSIS, SSRS** and **SSAS** application solutions for various business units across the organization.
* Analyzed **Data Set** with **SAS programming**, **R and Excel**.
* Publish Interactive dashboards and schedule auto-data refreshes
* Maintenance of large data sets, combining data from various sources by **Excel**, **Enterprise**, and **SASGrid**, Access and **SQL** queries.
* Created Hive queries that helped market analysts spot emerging trends by comparing incremental data with SQL reference tables and historical metrics.
* Design and development of **ETL** processes using Informatica **ETL** tools for dimension and fact file creation.
* Develop and automate solutions for a new billing and membership Enterprise **data Warehouse** including **ETL routines**, tables, maps, materialized views, and stored procedures incorporating Informatica and **OraclePL/SQL toolsets**.
* Performed analysis of implementing **Spark uses Scala** and wrote spark sample programs using **PySpark**.

**Environment:** SQL/Server, Oracle 10g/11g, MS-Office, Spyder, Jupyter, Anaconda, Rstudio, Tableau, Excel,PowerBI, Flask, D3 and DC, Teradata, Informatica, ER Studio, XML, R connector, Python, R, Tableau 9.2.

**Client: Geneia- Philadelphia, PennsylvaniaMay 2017 - Jul 2018**

**Role: Data Scientist/Machine Learning Engineer**

**Description:**Geneia provides healthcare data, technology, business and education solutions to excel the business performance and care quality.We believe in the healing power of personalized, patient-centered care, where all patients have a meaningful and trusted relationship with a physician and care team; where healthcare systems are designed and function optimally in support of improved health; and where physicians and care teams are inspired and find joy in practice.

**Responsibilities:**

* Implemented end-to-end systems for **Data Analytics, Data Automation** and integrated with custom visualization tools using **R, Hadoop and MongoDB, Cassandra**.
* Building predictive models using tools such as **SAS, R** with very granular data stored in **big data** platform.
* Worked on different data formats such as **JSON, XML** and performed **machine learning** algorithms in **R** and used **Spark** for test data analytics using **MLLib** and Analyzed the performance to identify bottlenecks.
* Involved working with Machine Learning Algorithms such as Decision Trees, Random Forest, Gradient Boosting, Support Vector **Machines, K Mean Clustering**, Naïve Bayes, Bayesian Belief Networks and **Artificial Neural Networks.**
* Developed Predictive models, Machine learning (Supervised and non-Supervised) using **R for Machine Motor.**
* Creating various **B2B** Predictive and descriptive analytics using **R and Tableau** and performed data cleaning and data preparation tasks to convert data into a meaningful data set using **R.**
* Used **R** to verify the results of **Mahout** on small data sets. Developed missing but important features of **ML** algorithms to the **Mahout**.
* Utilized **Spark, Scala, Hadoop, HBase, Kafka, Spark Streaming, MLLib, Python**, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc
* Created partitioned and bucketed tables in **Hive**. Involved in creating **Hive** internal and external tables, loading with data and writing hive queries which involves multiple join scenarios.
* Performed **K-means** clustering, Multivariate analysis and Support **Vector Machines in** R.
* Create analytical models using analytics algorithms like regression, decision trees, clustering, text mining etc. and leveraging tools like **R, Tableau** etc. to deliver actionable insights and recommendations.
* Developed multiple **Spark** jobs using **Scala** for data cleaning and preprocessing
* Designed the **schema**, configured and deployed **AWS Redshift** for optimal storage and fast retrieval of data.
* Used External Loaders like **Multi Load, T Pump and Fast Load** to load data into **Teradata14.1Database**.
* Involved in Troubleshooting and quality control of data transformations and loading during migration from **Oracle** systems into **Netezza EDW.**
* Used **S3 Bucket** to store the jar's, input datasets and used **Dynamo DB** to store the processed output from the input data set.
* Worked on classification/**scripting** of multiple attribute models by applying text-mining, **NLP, SVM** and Regular Expressions given product features like title, description etc. & predicting product attribute values using **Python/R**
* Worked on different data formats such as **JSON, XML** and performed machine learning algorithms in R.
* Used **Spark** for test data analytics using **MLLib** and Analyzed the performance to identify bottlenecks and used Supervised learning techniques such as classifiers and neural networks to identify patters in these data sets
* Developed **Tableau** visualizations and dashboards using **Tableau Desktop. Tableau workbooks** from multiple data sources using **Data Blending.**
* Developing new data warehousing system based on **spark 2.x** and **spark** streaming, utilizing Scala and Java 8
* Strong Knowledge on concepts of **DataModelingStar Schema/Snowflake** modeling, **FACT**& Dimensions tables and Logical&Physical data modeling.

**Environment:**R3.x, Erwin, MDM, QlikView, MLLib, PL/SQL, Tableau, Teradata, JSON, HADOOP (HDFS), MapReduce, SQL Server, MLLib, Scala NLP, SSMS, ERP, CRM, Netezza, Pandas, SAS, SPSS, Java, IDE, Cassandra, SQL, PL/SQL, AWS, SSRS, Informatica, PIG, Spark, Azure, R Studio, MongoDB, MAHOUT, JAVA, HIVE, AWS Redshift.

**Client: Cummins Inc., Columbus, IN.May 2016 - Apr 2017**

**Role: Data Scientist/R Developer**

**Description:** Cummins Inc. is an American Fortune 500 corporation that designs, manufactures, and distributes engines, filtration, and power generation products. Cummins also services engines and related equipment, including fuel systems, controls, air handling, filtration, emission control and electrical power generation systems.

**Responsibilities:**

* The conducted analysis in assessing customer consuming behaviors and discover the value of customers with **RMFanalysis**, applied customer segmentation with clustering algorithms such as **K**-**MeansClustering** and **HierarchicalClustering**.
* Collaborated with data engineers to implement the **ETLprocess**, wrote and optimized **SQLqueries** to perform **dataextraction** and merging from **Oracle**.
* Involved in managing backup and restoring data in the live **CassandraCluster**.
* Used **R**, **Python**, and **Spark** to develop a variety of models and algorithms for analytic purposes.
* Performed **dataintegritychecks**, **datacleaning**, exploratory analysis and feature engineer using **R** and **Python**.
* Developed personalized product recommendation with **Machinelearningalgorithms**, including **GradientBoostingTree** and **Collaborative** filtering to better meet the needs of existing customers and acquire new customers.
* Used **SpringFramework** at **BusinessTier** and uses **Spring** container for initializing services and used Spring Core Annotations for Dependency Injection such as **Bean**, **Component**, **Repository**, **Autowired** etc.
* Used **Python** and **Spark** to implement different **machinelearningalgorithms**, including GeneralizedLinearModel, RandomForest, SVM, Boosting and NeuralNetwork.
* Evaluated parameters with **K-Fold**CrossValidation and optimized performance of models.
* Worked on benchmarking **CassandraCluster** using the **Cassandra**stresstool.
* A highly immersive Data Science program involving **DataManipulation** and **Visualization**, **WebScraping**, **MachineLearning**, **GIT**, **SQL**, **UNIX**Commands, **Python**programming, **NoSQL**.
* Worked on datacleaning, datapreparation,and feature engineering with**Python**, including**Numpy**, **Scipy**, **Matplotlib**, **Seaborn**, **Pandas**, and **Scikit**-**learn**.
* Identified risk level and eligibility of new insurance applicants with **MachineLearning** algorithms.
* Determined customer satisfaction and helped enhance customer using **NLP**.
* Utilized **SQL** and **HiveQL** to query, manipulate data from variety data sources including **Oracle** and **HDFS**, while maintaining data integrity.
* Performed datavisualization and Designeddashboards with **Tableau** and **D3.js** and provided complexreports, includingcharts, summaries, and graphs to interpret the findings to the team and stakeholders.

**Environment:**R, MATLAB, MongoDB, exploratory analysis, feature engineering, K-Means Clustering, Hierarchical Clustering, Machine Learning), Python, Spark (MLlib, PySpark), Tableau, Micro Strategy, SAS, Tensor Flow, regression, logistic regression, Hadoop 2.7, OLTP, random forest, OLAP, HDFS, ODS, NLTK, SVM, JSON, XML and MapReduce.

**Client: Cenvien Technologies, Hyderabad, India Mar 2014 - Dec 2015**

**Role: Data Scientist**

**Description:** Fidelity Investments Inc., commonly referred to as Fidelity, earlier as Fidelity Management & Research or FMR, is a multinational financial services corporation. At Fidelity, we are constantly innovating and challenging the status quo in order to deliver more value to our customer. It's core to who we are. Our founder said: "Take intelligent risks, rather than follow the crowd."

**Responsibilities:**

* Provided the architectural leadership in shaping strategic, business technology projects, with an emphasis on application architecture.
* Utilized domain knowledge and application portfolio knowledge to play a key role in defining the future state of large, business technology programs.
* Participated in all phases of datamining, datacollection, datacleaning, developingmodels, validation, and visualization and performed Gapanalysis.
* Developed **MapReduce/Spark, R** modules for machine learning & predictive analytics in Hadoop on AWS. Implemented a R-based distributed randomforest.
* Utilized **Spark, Scala, Hadoop, SparkStreaming**, **MLLib**, Rabroadvariety of machinelearning methods includingclassifications, regressions, dimensionallyreduction etc. and utilized the engine to increase user lifetime by 45% and triple user conversations for target categories.
* Well versed with: **CloudIaaS** and **PaaS** implementations in both private and publicclouds like **VMware**, Openstack, **AmazonAWS** and Cloudfoundry (**Pivotal and HPStackato**).
* Used **R** and **h2o.ai** libraries for developing various machinelearningalgorithms and utilizedmachinelearningalgorithms such as linearregression, multivariateregression, **naiveBayes, RandomForests, K-means**, &**KNN** for **dataanalysis**.
* Worked on databasedesign, relationalintegrityconstraints, **OLAP, OLTP**, Cubes, andNormalization (**3N0F**) and De-Normalization of the database.
* Worked on customersegmentation using an unsupervised learning techniqueclustering.
* Utilized **Spark, Scala, Hadoop, HBase, Kafka, SparkStreaming, MLlib, Python**, a broad variety of machinelearningmethods including classifications, regressions, dimensionallyreduction etc.
* Data analysis, reporting using TableauPerform numerous data pulling requests using **SQLserver2012.**
* Designed and implemented system architecture for **AmazonEC2** based cloud-hosted solution for the client.
* Tested Complex **ETLMappings** and Sessions based on business user requirements and business rules to load data from source flat files and RDBMS tables to target tables.
* Hands on experience in **Hadoop**ecosystem with components**HadoopMapReduce, HDFS, Oozie, HiveQL, Sqoop, HBase, MongoDB, Zookeeper, Pig, andFlume.**

**Environment:**R, SQL, Oracle 12c, Netezza, SQL Server, Informatica, Java, SSRS, PL/SQL, T-SQL, Tableau, MLlib, regression, Cluster analysis, Scala NLP, Spark, Kafka, MongoDB, logistic regression, Hadoop, Hive, Teradata, random forest, OLAP, Azure, MariaDB, SAP CRM, HDFS, ODS, NLTK, SVM, JSON, Tableau, XML, Cassandra, MapReduce, AWS.