**ShaileshKhanal**

**santosh@metahorizon.com**

**Data Science**

**Professional Summary:**

* 7 years of working experience in designing, building and implementing analytical and enterprise application using machine learning, Python, R and Scala.
* Experience in Machine Learning algorithms such as Linear Regression, Logistic Regression, Naive Bayes, Decision Trees, K-Means Clustering and Association Rules.
* Experience in applying predictive modeling and machine learning algorithms for analytical reports.
* Experience using technology to work efficiently with datasets such as scripting, data cleaning tools, statistical software packages.
* Collaborated with engineers to deploy successful models and algorithms into production environments.
* Created various ETL applications using SSIS and Lavastorm to load the data from different sources, transform the data according to the business logic and load into different destinations.
* Developed predictive models using Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Cluster Analysis, and Neural Networks.
* Very Strong in Python, statistical analysis, tools, and modeling.
* Experienced in Machine Learning and Statistical Analysis with Python Scikit-Learn.
* Experienced in Big Data with Hadoop, HDFS, Map Reduce, and Spark.
* Strong programming skills in a variety of languages such as Python and SQL.
* Experience in processing real time data and building ML pipelines end to end.
* Very good experience working with large datasets and Deep Learning algorithms using apache spark and Tensor Flow.
* Experienced in developing Web Services with Python programming language.
* Worked on various applications using Python integrated IDEs Eclipse, PyCharm.
* Used Several Python modules and controls to rapidly build the application.
* Designed and developed the UI of the website using HTML, AJAX, CSS, jQuery and JavaScript.
* Worked on developing screens for delete, edit, enable and disable users in web application.
* Experience in Data Migrations, Data Cleaning, Transformation, Integration, Data Imports and Data Exports.
* Proficient code writing capability in a major programming language such as Python, R, and Scala.
* Experienced with machine learning algorithm such as logistic regression, random forest, Xgboost, KNN, SVM, neural network, linear regression, and k-means.
* Good Knowledge in Data Validation, Data Cleaning, Data Verification and Identifying data mismatch.
* Leveraging Big Data (Hadoop framework) to ingest, store, analyze and process big data.
* Good understanding of Apache Spark features & advantages over map reduce or traditional systems.
* Good knowledge in using all complex data types in Pig and MapReduce for handling the data and formatting it as required.
* Experience working on Integrated Development Environments (IDE) like PyCharm, Sublime Text and Eclipse.
* Experienced with Machine Learning, Regression Analysis, Clustering, Boosting, Classification, Principal Component Analysis and Data Visualization Tools.
* Developed highly scalable classifiers and tools by leveraging machine learning, Apache spark & deep learning.
* Developed MapReduce programs to perform Data Transformation and analysis.
* Delivered at multiple end-to-end Big data analytical based solutions and distributed systems like Apache Spark.
* Have hands on experience in applying SVM, Random Forest, K means clustering.
* Experienced in writing complex SQL Quires like Stored Procedures, triggers, joints, and Sub quires.
* Experience in analyzing the log files for Hadoop and eco system services and finding out root cause.

**Technical Skills:**

|  |  |
| --- | --- |
| **Programming Languages:** | Python, SQL, C, Java |
| **Scripting Languages:** | Python |
| **Markup Languages:** | HTML, XHTML |
| **Data Sources:** | HDFS, SQL Server, Excel |
| **Data Visualization:** | Tableau |
| **Predictive and Machine Learning:** | Linear Regression, Logistic regression, Principal Component Analysis (PCA), K-means, Random Forest, Decision Trees, SVM, K-NN, Deep learning, Time Series Analysis and Ensemble methods |
| **Big Data Ecosystems:** | Hadoop, MapReduce, HDFS, HBase, Hive, Spark |
| **Database Systems:** | Oracle, MySQL, DB2, Teradata |
| **Operating System:** | Linux, Windows, Unix. |

**Professional Experience:**

**CVS Health, Florham Park, NJ Sep 2016-Till date**

**Data Scientist**

**Responsibilities:**

* Responsible for data identification, collection, exploration, and cleaning for modeling, participate in model development.
* Visualize, interpret, report findings, and develop strategic uses of data by python Libraries like Numpy, Scikit-learn, MatPlotLib.
* Performed Data Cleaning, features scaling, features engineering.
* Responsible for loading, extracting and validation of client data.
* Creating statistical models using distributed and standalone models to build various diagnostics, predictive and prescriptive solution.
* Missing value treatment, outlier capping and anomalies treatment using statistical methods, deriving customized key metrics.
* Performed analysis using industry leading text mining, data mining, and analytical tools and open source software.
* Understanding and implementation of text mining concepts, graph processing and semi structured and unstructured data processing.
* Evaluated models using Cross Validation, Log loss function, ROC curves and used AUC for feature selection.
* Setup storage and data analysis tools in Amazon Web Services cloud computing infrastructure.
* Dummy variables where created for certain datasets to into the regression.
* Developed triggers, stored procedures, functions and packages using cursors and ref cursor concepts associated with the project using Pl/SQL.
* Developed logging for ETL load at the package level and task level to log number of records processed by each package and each task in a package using SSIS.
* Used Meta data tool for importing metadata from repository, new job categories and creating new data elements.
* Used MLlib, Spark's Machine learning library to build and evaluate different models.
* Worked with several R packages including knitr, dplyr, SparkR, Causal Infer, space-time.
* Build multiple features of machine learning using python, Scala and java based on need.
* Developed Map Reduce pipeline for feature extraction using Hive.
* Expertise in Big Data architecture like Hadoop (Azure, Hortonworks, Cloudera) distributed system, MongoDB, NoSQL
* Creating data pipelines using big data technologies like Hadoop, spark etc.
* Installed and configured Nifi on Hortonworks, CentOS, Ubuntu and Windows 10 operating systems.
* Strong skills in data visualization like matPlotLib and seaborn library.
* Create different charts such as Heat maps, Bar charts, Line charts etc.,
* Developed, Implemented & Maintained the Conceptual, Logical & Physical Data Models using Erwin for forwarding/Reverse Engineered Databases.
* Collaborating with team to enhance modeling process and develop project.
* Developed Hive queries for ad hoc analysis.
* Analyzed the partitioned and bucketed data and compute various metrics for reporting.
* Exported the result set from Hive to MySQL using Sqoop after processing the data.

**Nepal Investment Bank, Kathmandu, Nepal Sep 2014-July 2016**

**Data Scientist**

**Responsibilities:**

* Data mining using state-of-the-art methods
* Extending company's data with third party sources of information when needed
* Enhancing data collection procedures to include information that is relevant for building analytic systems
* Processing, cleansing, and verifying the integrity of data used for analysis
* Doing ad-hoc analysis and presenting results in a clear manner
* Creating automated anomaly detection systems and constant tracking of its performance
* Strong command of data architecture and data modelling techniques.
* Hands on experience with commercial data mining tools such as Splunk, R, Map reduced, Yarn, Pig,Hive, Floop, Oozie, Scala, HBase, Master HDFS, Sqoop, Spark, Scala (Machine learning tool) or similar software required depending on seniority level in job field.
* Developed scalable machine learning solutions within a distributed computation framework (e.g. Hadoop, Spark, Storm etc.).
* Develop, Unit test and implement Informatica ETL mappings and workflows.
* Design and Develop ETL processes to populate Enterprise Data Warehouse and data marts using Informatica, Pentaho, Oracle PL/SQL, & Unix shell scripts, Python.
* Utilizing NLP applications such as topic models and sentiment analysis to identify trends and patterns within massive data sets.
* Knowledge in ML& Statistical libraries (e.g. Scikit-learn, Pandas).
* Having knowledge to build predict models to forecast risks for product launches and operations and help predict workflow and capacity requirements for TRMS operations
* Having experience with visualization technologies such as Tableau
* Draw inferences and conclusions, and create dashboards and visualizations of processed data, identify trends, anomalies
* Generation of TLFs and summary reports, etc. ensuring on-time quality delivery.
* Participated in client meetings, teleconferences and video conferences to keep track of project requirements, commitments made and the delivery thereof.
* Solved analytical problems, and effectively communicate methodologies and results
* Worked closely with internal stakeholders such as business teams, product managers, engineering teams, and partner teams.
* Created automated metrics using complex databases.
* Foster culture of continuous engineering improvement through mentoring, feedback, and metrics.
* Environment: Erwin r9.0, Informatica 9.0, ODS, OLTP, Oracle 10g, Hive, OLAP, DB2, Metadata, MS Excel, Mainframes MS Visio, Rational Rose, Requisite Pro. Hadoop, PL/SQL, etc.

**Nepa Auto Insurance, Kathmandu, Nepal June 2011-August 2014**

**Data Scientist**

**Responsibilities:**

* Responsible for reporting of findings that will use gathered metrics to infer and draw logical conclusions from past and future behavior.
* Performed Multinomial Logistic Regression, Random Forest, Decision Tree, SVM to classify package is going to deliver on time for the new route.
* Used Principal Component Analysis & Factor Analysis in feature engineering to analyze high dimensional data in MatLab.
* Performed data analysis by using SQL to retrieve the data from Hadoop cluster.
* Used R 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.
* Collected data needs and requirements by Interacting with the other departments.
* Created various types of data visualizations using Python and Tableau schemas.
* Communicated the results with operations team for taking best decisions.
* PL/SQL, Python 2.x, HDFS, Hadoop, Impala, Linux, Tableau Desktop, MatLab, SQL Server 2012, Microsoft Excel, MS Visio, Rational Rose.
* Utilized Python to develop and execute Big Data Analytics and Machine learning applications, prototyped machine Learning use cases under SciKit-Learn.
* Led technical implementation of advanced analytics projects and defined the time-series approaches.
* Develop new and effective analytics algorithms and wrote the key pieces of mission-critical source code.
* Implemented advanced machine learning algorithms including regression trees, kernel PCA, among other methods in Scala, Hadoop 1.x, R and in other tools and languages as needed.