**Milan**

**Data Scientist**

**Phone: (213) 423-1050**

**Email: vishwa.n@digitechnol.com**

PROFESSIONAL SUMMARY

Process oriented around **8 years of professional experience in Data Scientist/Analytics**. Mainly, IT industry and Expertise in Object Oriented Analysis including development and implementation of various applications using **PYTHON**, **R, SAS, SPARK, Data analyzing, Big data Technologies, machine learning methodologie**s, **Scala**, **AWS, Azure** and other web technologies. Demonstrated strong communication skills in collaboration with business partners throughout the **software development lifecycle** and **data life cycle development**. Flexible thinker and team player that is dedicated to supporting quality improvement initiatives of the organization.

* Experience in transforming raw data into actionable **strategic knowledge** to gain insight into business processes, and thereby guide and help businesses in their decision-making and run efficiently.
* Wide array of experiences in interpreting and analyzing data using **statistical techniques** and systems.
* Excellent knowledge in creating **Databases**, **Tables**, **Stored Procedure**, **DDL/DML Triggers**, **Views**, User defined data types, effective functions, Cursors and Indexes.
* Proficient in Statistical Modeling & Machine Learning techniques in Forecasting/ Predictive Analytics, Segmentation methodologies, Regression based models, Hypothesis testing, PCA, Ensembles.
* Implement and practice Machine learning techniques on structured and unstructured data with equal proficiency.
* Experience with Data Analytics, Data Reporting, Ad-hoc Reporting, Scales, Pivot Tables and OLAP reporting. Expert in data flow between primary DB and various reporting tools, Expert in finding Trends and Patterns within Datasets and providing recommendations accordingly.
* Ability to use dimensionality reduction techniques and regularization techniques.
* Expertise in creating Dashboards, Scales, Pivot Tables and OLAP reporting.
* Hands-on experience in **data structure**, design and analysis using **Machine Learning Technics and modules in PYTHON, R.**
* Experience in **Analyzing with Big Data technology**, Design, Development, Testing, Maintenance and Documenting of various Web-Applications, Client/Server and Distributed Applications.
* Experienced informing in memory data processing for batch and advanced analytic using **Apache Spark**(SparkSQL &Spark-Shell).
* Good experience in system monitoring, development and support related activities for **Hadoop** and Hadoop Admin, Java/J2EE Technologies and **Spark Streaming, Spark SQL**.
* Experience in working on **Apache Hadoop** ecosystem components like Map Reduce, HDFS, Hive, Pig, HBase, Flume, Sqoop, Oozie.
* Solid understanding of AWS (Amazon Web Services) S3, EC2, RDS and IAM, Azure ML, Apache Spark, Scala process and concepts.
* Experienced in Machine Learning Classification Algorithms like Logistic Regression, K-NN, SVM, Kernel SVM, Naive Bayes, Decision Tree & Random Forest classification.
* Extensive Knowledge in implementation of **NLP, Text Mining, Recommendation System, Deep learning, Scala**.
* Knowledge and experience with **cloud services** a must (**AWS using S3, EC2, Azure**).
* Proficiency in spark using Scala for loading data from the local file systems like HDFS, Amazon S3, Relational and NoSQL databases using **Spark SQL** and Import data into RDD.
* Strong expertise in ETL, Data warehousing, Operational Data Store (ODS), Data Marts, OLAP and OLTP technologies.
* Extensive experience in in-depth **data analysis on different DB and Data Extraction**. Strong knowledge in writing **SQL Queries**, **sub-queries**, and **joins**.
* Skilled in providing analytic support including **data importing/Extraction**, **data wrangling** and **data visualization**.
* Expert in Applying **Advance MS excel** and Adept in MS Excel with proficiency in **VLOOKUP’s**, **Pivot Tables** and understanding of **VBA Macros**.
* Extensive knowledge of financial accounting, Investment banking, fixed income products and financial system.
* Familiarity and good knowledge with data manipulation software **Alteryx**, SAS.
* Strong understanding of **HRMS/CRM databases/systems**, to query and analyze customer data. Experience with **Google Analytics**, **Tableau**, **MS EXCEL** and **SQL Server Reporting Services** (SSRS).
* Experienceindesigningstunningvisualizationsusing**Tableausoftware**andpublishingandpresentingdashboardsonweband desktop platforms.
* Proficient in trouble shooting various applications specific like JAVA, C, C++, JavaScript, HTML, and CSS related issues.
* Strong experience in **Software Development Life Cycle (SDLC)** including Requirements, Specifications Analysis/Design and Testing as per the Software Development Life Cycle.
* Excellentcommunication,teamorientedandinterpersonalskills,QuickLearner,ExceptionalTeamPlayerand I can work independently as well.

EDUCATION

Bachelor in Electronics & Communications Engineering.

TECHNICAL SKILLS

|  |  |
| --- | --- |
| **Databases** | Relational Database- Oracle, SQL Server, MySQL, HSQLDB, AWS RedShift, Cassandra, MongoDB, HBase. |
| **Programming** | Python, R, JAVA, SQL, HTML5, CSS3, C, C++, JavaScript, Bootstrap, PHP, jQuery, Angular.js, Hive, Scala. |
| **Operating System** | Linux, Windows XP/7/8/10, MAC OS. |
| **Business Intelligence Tools** | Tableau 9.1, MS Power BI, Google Refine, MS Excel - Analytical Solver, SSRS. |
| **Statistics** | Normal Distribution Plots, Inferential Statistics, Charts, Descriptive Statistics, Inferential Statistics, Histograms, ANOVA, Data Distributions, Standard Deviation and Variance, Hypothesis Testing (p-values) and Test for significance- t-test, Process Control Capability Analysis Hypothesis, Testing. |
| **Frameworks** | Apache Hadoop, Hive, Apache Spark. |
| **Machine Learning** | Feature Selection Supervised and Unsupervised Learning, Correlation Analysis, Regression, Logistics Regression, Clustering, Decision Tree, Support Vector Machines (SVM), Naive Bayes, K- Nearest Neighbors (KNN), Clustering, K-Means Clustering, Random Forest, Principal Component Analysis, Factor Analysis, Correlation Matrix. |
| **Development tool** | Eclipse, NetBeans, Visual Studio, People tools, Jupyter notebook, Spyder, Anaconda, AWS, AZURE |

**PROFESSIONAL EXPERIENCE:**

|  |  |
| --- | --- |
| **Client: SAP Americas, Newton Square, PA** | **Mar 2017-Till Date** |
| **Role: Data Scientist** |  |

**Description:**

The Americas' SAP Users' Group (ASUG) is the company's largest user group, with 100,000 individuals at 3,800 companies. ASUG members are professionals who work in more than 17 industries. SAP America, Inc. develops business software solutions for various industries.

**Responsibilities:**

* Analyze and Prepare data, identify the patterns on the dataset by applying historical models. Collaborating with Senior Data Scientists for understanding of data
* Perform data manipulation, data preparation, normalization, and predictive modeling. Improve efficiency and accuracy by evaluating model in R
* This project was focused on customer segmentation based on machine learning and statistical modeling effort including building predictive models and generate data products to support customer segmentation
* Used R and Python for programming for improvement of the model. Upgrade the entire models for improvement of the product
* Develop a pricing model for various product and services bundled offering to optimize and predict the gross margin
* Built price elasticity model for various product and services bundled offering
* Developed predictive causal model using annual failure rate and standard cost basis for the new bundled service offering
* Design and develop analytics, machine learning models, and visualizations that drive performance and provide insights, from prototyping to production deployment and product recommendation and allocation planning
* Utilized Spark, Scala, Hadoop, HBase, Kafka, Spark Streaming, MLLib, R, a broad variety of machine learning methods including classifications, regressions, dimensionality reduction etc.
* Worked with sales and Marketing team for Partner and collaborate with a cross-functional team to frame and answer important data questions
* prototyping and experimentation ML/DL algorithms and integrating into production system for different business needs
* Worked on Multiple datasets containing two billion values which are structured and unstructured data about web applications usage and online customer surveys
* Good hands on experience on Amazon Redshift platform
* Performed Data cleaning process applied Backward – Forward filling methods on dataset for handling missing values
* Design built and deployed a set of python modeling APIs for customer analytics, which integrates multiple machine learning techniques for various user behavior prediction and supports multiple marketing segmentation programs
* Segmented the customers based on demographics using K-means Clustering.
* Explored different regression and ensemble models in machine learning to perform forecasting
* Presented Dashboards to Higher Management for more Insights using Power BI
* Used classification techniques including Random Forest and Logistic Regression to quantify the likelihood of each user referring.
* Worked with using a different kind of compression techniques to save data and optimize data transfer over the network using LZO, Snappy, and GZip etc.
* Analyze large and critical datasets using Cloudera, HDFS, HBase, Map Reduce, Hive, HiveUDF, Pig, Sqoop, Zookeeper, &Spark.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Connected Tableau from client end with AWS IP addresses and view the end results
* Performed Boosting method on predicted model for the improve efficiency of the model
* Designed and implemented end-to-end systems for Data Analytics and Automation, integrating custom, visualization tools using R, Tableau, and Power BI.
* Collaborating with the project managers and business owners to understand their organizational processes and help design the necessary reports.

|  |
| --- |
| **Environment: MS SQL Server, R/R studio, SQL Enterprise Manager, Python, Redshift, MS Excel, Power BI, Tableau, T-SQL, ETL, MS Access, XML, MS Office, Outlook, AS E-Miner, Power BI, K-means.** |

|  |  |
| --- | --- |
| **Client: Humana ,Louisville, KY** | **Oct 2015 - Feb 2017** |
| **Role: Data Scientist** |  |

**Description:**

Humana Inc., headquartered in Louisville, Ky., is a leading health care company that offers a wide range of insurance products and health and wellness services that incorporate an integrated approach to lifelong well-being. By leveraging the strengths of its core businesses, Humana believes it can better explore opportunities for existing and emerging adjacencies in health care that can further enhance wellness opportunities for the millions of people across the nation with whom the company has relationships.

**Responsibilities:**

* Used various approaches to collect the business requirements and worked with the business users for ETL application enhancements by conducting various JRD sessions to meet the job requirements
* Performed exploratory data analysis like calculation of descriptive statistics, detection of outliers, assumptions testing, factor analysis, etc., in R
* Worked exclusively on making applications more scalable and highly available system in AWS (load balancing) with full automation.
* Extracted data from the database using SAS/Access, SAS/SQL procedures and created SAS datasets for statistical analysis, validation, and documentation.
* Extensively understanding BI, analytics focusing on consumer and customer space
* Innovate and leverage machine learning, data mining and statistical techniques to create new, scalable solutions for business problems.
* Extensive experience in working with Tableau Desktop, Tableau Server, and Tableau Reader in various versions of Tableau9.2 and 10 as a Developer and Analyst.
* Analyzed different types of data to derive insights about relationships between locations, statistical measurements and qualitatively assess the data using R/R Studio
* Performed Data Profiling to assess data quality using SQL through complex internal database
* Improved sales and logistic data quality by data cleaning using NumPy, SciPy, Pandas in Python
* Designed data profiles for processing, including running SQL, PL/SQL queries and using R for Data Acquisition and Data Integrity which consists of Datasets Comparing and Dataset schema checks
* Used R to generate regression models to provide statistical forecasting
* Used drill downs, filter actions and highlight actions in Tableau for developing dashboards in Tableau.
* Applied Clustering Algorithms such as K-Means to categorize customers into certain groups
* Performed data management, including creating SQL Server Report Services to develop reusable code and an automatic reporting system and designed user acceptance test to provide end with an opportunity to give constructive feedback
* Used Tableau and designed various charts and tables for data analysis and creating various analytical Dashboards to showcase the data to managers.
* Applied association rule mining & chain model to identify hidden patterns and rules in remedy ticket analysis which aid in decision making.
* Created AMI images of critical ec2 instances as backup using AWS CLI and GUI.
* Created AWS Cloud formation templates on creating IAM Roles & total architecture deployment end to end (Creation of EC2 instances & its infrastructure)
* Isolating customer behavioral patterns by analyzing millions of customer data records over a period and correlating multiple customers' attributes.
* Empowered decision makers with data analysis dashboards using Tableau and Power BI.

|  |
| --- |
| **Environment: R/R Studio, SAS, SSRS, SSIS, Oracle Database 11g, Oracle BI tools, Tableau, MS-Excel, Python, Naive Bayes, SVM, K- means, ANN, Regression, MS Access, SQL Server Management Studio, SAS E-Miner.** |

|  |  |
| --- | --- |
| **Client: TUFT’S MEDICAL CENTER Boston, MA**  **Role: Data Scientist** | **Dec 2014 - Sep 2015** |

**Description:** The project was to build a recommendation system that accurately suggest customer health symptoms and habit based on the historical data, data collected from four areas within healthcare, claims and cost data, pharmaceutical and research and development (R&D) data, clinical data (collected from electronic medical records (EHRs)), and patient behavior and sentiment data. The results were used by the higher authorities to create appropriate strategies. Further, developed BI reports that provided **predictive analytics**, future prediction of healthcare.

**Responsibilities:**

* Extract health care claims, provider, and enrollment data from the Colorado All Payer Claims Database (CO APCD) in support of the triple aim for Better health, better quality, and lower costs.
* Performing data analysis and build machine learning models like [Customer Lifetime Value (LTV) Model](https://wiki.corp.adobe.com/display/CustomerAnalytics/Customer+Lifetime+Value+%28LTV%29+Model), Maximum Likelihood estimation.
* Analyze customer, **behavior data, symptoms data**, **transaction data** and **campaign data** to identify trends and patterns of data in different visualization techniques like **Seaborn library** in PYTHON.
* Designed and developed some of the complex modules of the system using **AWS, S3**.
* Extraction of large amounts of data for analysis and reporting. Responsible for documentation of all analysis as well as data discrepancies in both **Spark** and **Python**.
* Working closes with other analysts to reconcile all issues related to data production, data extraction and delivery to ensure the integrity of the data and the reporting that it is used for.
* Determined the missing data, outlier and invalid data and applied appropriate data management techniques.
* Worked on data manipulation and raw marketing data of different formats from multiple sources and prepared the data for **Sentiment analysis** of all the customer medical issue data using packages like **NLTK (Natural Language Processing with Python / Analyzing Text with the Natural Language Toolkit).**
* Wrote **script in python** to predict number of people getting effect of some diseases, by collecting set of predicted (symptoms) data from all medical sectors and evaluated with outcome data and Make the aware of people using **Machine Learning** Module like **logistic regression**.
* Wrote simple and advanced **SQL queries** and **scripts** to create standard reports for senior managers.
* Designed custom reports, charts, tables and dashboards using **Power BI** and for marketing and operations teams for their decision-making process.
* Create focused **reports** and **dashboards** on **content/channel performance**, **lead generation** and **conversion rates**.
* Assisted in writing a wide range of documents including work plans, monthly, quarterly and annual progress reports, and provider/grantee guidance materials and manual.

|  |
| --- |
| **Environment: Jupyter Notebook, PYTHON, AWS, Sentiment analysis, My SQL, CRM, Tableau, MS Access, Power BI.** |

|  |  |
| --- | --- |
| **Client: UNIVERSITY OF MASSACHUSETTS, Boston, MA**  **Role: Data Analytics** | **Nov 2013 -Nov 2014** |

**Description:**

Worded on projects were to **detect the future (new) Ham or scam emails and identify and Search famous Restaurants with set of reviews by given address and radius using both spark and python in AWS**. During this project, we have encountered whether the new email is ham or span with historical data by trained model. We analyzed and developed a text mining mechanism to identify suspects with high percentage of word repeated in **Wikipedia**. My responsibilities were to perform text mining in python and parse text documents to compute the TF\*ID by using HashingTF and IDF libraries in Spark.

**Responsibilities:**

* Gathering **business requirements** and collected all set of required data from different sectors and we clean the data.
* Developed two generic categories like Spam and ham classifier using Python.
* 17,000 emails were processed and categorized into ham and spam classifier in **AWS** Cloud Service.
* Loaded our training data into a pandas Data Frame of ham and spam emails, implemented **sklearn with family naïve Bayes** to train a module with ham and spam classifier using **Multinomial NB.**
* Naive Bayes Methods in Machine Learning techniques to identify and to predict the New email is whether it is ham or scam.
* Identified entities and emotions in a sentence and use these to determine if the entity is being viewed positively or negatively using NLTK Package.
* Identified set of restaurants in given radius with longitude and latitude and Extract review snippets.
* Gathered set of Wikipedia documents and Worked on **analysis** in python to analyze the TF-IDF rate of each words in Rating in each restaurant.
* Compiled set of 30 reviews (yelp data) each of ten neighborhood restaurants and Achieved to Identify best restaurants around any place. Implemented in both **Python** and **Spark**.

|  |
| --- |
| **Environment: Python, Spark, Machine learning modules, Apache Spark,Jupyter notebook, AWS.** |

|  |  |
| --- | --- |
| **Client: BSNL Bangalore, India**  **Role: Data Analyst AS Intern** | **Feb 2011 - Oct 2013** |

**Description: B**SNL Mobile is an [Indian mobile network operator](https://en.wikipedia.org/wiki/Mobile_network_operators_of_India), operated by the public enterprise [BSNL](https://en.wikipedia.org/wiki/BSNL). BSNL Mobile has a pan-India presence with presence in all the 21 [telecom circles](https://en.wikipedia.org/wiki/Telecom_circle) in [India](https://en.wikipedia.org/wiki/India). I worked as Intern, analyzing and building variety of **business reports** to support critical business decisions. My role is to collect data that are crucial for business operations from appropriate sources, analyze, manipulate and present them to senior management in the form of Reports and Dashboards.

**Responsibilities:**

* Worked with different teams to gain insights about the **data concepts** behind their business.
* **Research**, **update**, and **validate data** underlying spreadsheet production; strategically fill gaps.
* Create **pivot tables** and modify spreadsheets to achieve **analytical goals**.
* Wrote simple and advanced **SQL queries** for extracting data and created **dashboard** and stories for senior managers.
* Delivered file in various **file formatting** system (ex. Excel file, Tab delimited text, Comma separated text, Pipe delimited text etc.)
* Analyzed data using complex SQL queries, across various databases.
* Verify entered account data by reviewing, correcting, deleting, or reentering data, combining data form both systems when account information is incomplete, purging files to eliminate duplication of data.
* Performed **data analysis** primarily identifying data sets, source data, source Meta data, data definitions and data formats.
* Used **R Programming language** to simple and complicate data analysis. Mainly, implement a wide variety of statistical and graphical techniques, including **linear and nonlinear modeling**, classical statistical tests, simple **time-series analysis using IDE- R studio.**
* Created required SQL queries to be integrated with database to meet the requirement for generating the user defined formulary reports.
* Assisted and **support to application development team** members as required for back end development while getting interaction between application and database (MYSQL).
* Prepares source data for computer entry by compiling and sorting information; establishing entry priorities to set up and complete build of materials.
* Ensured **best practices** are applied and integrity of data is maintained through security, documentation, and change management

|  |
| --- |
| **Environment: R, R-studio, SQL, Microsoft Office, Microsoft Excel, MS Access, Statistical techniques, MYSQL.** |

|  |  |
| --- | --- |
| **Client: Siddaganga Institute of Technology Tumkur, India**  **Role: Web Developer/ Academic Projects** | **July 2009 - Jan 2011** |

**Description:** Worked on Project to developed Event Organization website using PYTHON and front end development technologies using HTML CSS, Bootstrap and JavaScript. Used MYSQL for backend database. Meanwhile, help the team with technology issues including writing SQL statements for data interaction. Eventually, worked with different team for technology suggestion, manly with HTML/CSS templates and any Development issues.

**Responsibilities:**

* Build the web site mainly for research department of mammals and data management. (This website has been placed in my GitHub drive)
* Developed the UI Screens using **HTML5, Bootstrap, JavaScript and CSS3** for complex page layouts while adhering to code standards.
* Perform design, architect, implement, and maintain real estate investment and financial management web applications.
* Set up, configure, and maintain **DNS** for web application domain name.
* Provided **HTML** 5 validations, Java script validations and **Java** validations on the controller side.
* Created and deployed new features to sustain and amend existing applications.
* Developed dynamic **e-mails using JavaScript**, and hand coding of **HTML, XHTML, and CSS**.
* Implemented the all social sites and twitter connection with review of each mammals.
* Implemented the **MYSQL database** to interact with application and database through SQL queries.
* Played an active role in testing the application **by writing test cases** for different scenarios.
* Created use cases, class diagrams, activity diagrams and collaboration diagrams.
* Involved in System documentation, **object documentation and User documentation**.

|  |
| --- |
| **Environment: MYSQL, SQL, Microsoft Office, Microsoft Excel, MS Access, HTML5, Bootstrap, JavaScript and CSS3.** |