**Vivek**

**Data Scientist**

**Email**: **vivek.dani27@gmail.com**

**Phone no:** **(469) 453-6340**

**PROFESSIONAL SUMMARY:**

* Professional qualified Data Scientist/Data Analyst with over 8+ years of experience in Data Science and Analytics including Machine Learning, Data Mining and Statistical Analysis
* Involved in the entire data science project life cycle and actively involved in all the phases including data extraction, data cleaning, statistical modelling and data visualization with large data sets of structured and unstructured data
* Experienced with machine learning algorithm such as logistic regression, random forest, Xgboost, KNN, SVM, neural network, linear regression, lasso regression and k-means
* Strong skills in statistical methodologies such as A/B test, experiment design, hypothesis test, ANOVA
* Used the version control tools like Git 2.X
* Passionate about gleaning insightful information from massive data assets and developing a culture of sound, data-driven decision making.
* Ability to maintain a fun, casual, professional and productive team atmosphere
* Experienced the full software life cycle in SDLC, Agile and Scrum methodologies.
* Skilled in Advanced Regression Modelling, Correlation, Multivariate Analysis, Model Building, Business Intelligence tools and application of Statistical Concepts.
* Proficient in Predictive Modelling, Data Mining Methods, Factor Analysis, ANOVA, Hypothetical testing, normal distribution and other advanced statistical and econometric techniques.
* Developed predictive models using Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Cluster Analysis, and Neural Networks.
* Experienced in Machine Learning and Statistical Analysis with Python Scikit-Learn.
* Experienced in Python to manipulate data for data loading and extraction and worked with python libraries like Matplotlib, Numpy, Scipy and Pandas for data analysis.
* Worked with complex applications such as R, SAS, Matlab and SPSS to develop neural network, cluster analysis.
* Understanding on **Hadoop** **MapReduce**, or any big data frameworks is added advantage.   
  Statistical skills.
* Experience in implementing data analysis with various analytic tools, such as Anaconda 4.0JupiterNotebook 4.X, R 3.0 and Excel
* Solid ability to write and optimize diverse SQL queries, working knowledge of RDBMS like SQLServer2008, NoSQL databases like MongoDB3.2
* Strong experience in Big data technologies like Spark 1.6, Spark sql, Pyspark, Hadoop 2.X, HDFS, Hive 1.X
* Experience in visualization tools like, Tableau9.X, 10.X for creating dashboards
* Excellent understanding Agile and Scrum development methodology,
* Proficiency in application of statistical prediction modeling, machine learning classification techniques and / or econometric forecasting techniques
* Proficiency in various type of optimization, Market Mix modeling, Segmentation, Time Series, Price Promo models etc.
* Familiar with systems programming language, such as Java, Python and C++
* Experience in the application of Neural Network, **Support Vector Machines** (SVM), and Random Forest.
* Identifies/creates the appropriate algorithm to discover patterns, validate their findings using an experimental and iterative approach.
* Applies advanced statistical and predictive modeling techniques to build, maintain, and improve on multiple real-time decision systems. Closely works with product managers, Service development managers, and product development team in productizing the algorithms developed.
* Experience in designing **star schema**, **Snow flake schema** for **Data** **Warehouse**, **ODS architecture**.
* Experience in **designing** stunning visualizations using **Tableau software** and publishing and presenting **dashboards**, Storyline on **web** and **desktop** platforms.
* Experience in working with relational databases (**Teradata, Oracle**) with advanced SQL programming skills
* In-depth knowledge of statistical procedures that are applied in Supervised / Unsupervised problems
* Basic-Intermediate level proficiency in SAS (**Base SAS, Enterprise Guide, Enterprise Miner**) & in UNIX
* Track record of applying machine learning techniques to marketing and merchandizing ideas
* Experience in Big Data platforms like Hadoop platforms (**Map-R, Hortonworks** & others) , Aster and Graph Databases.
* Excellent communication skills (verbal and written) to communicate with clients and team, prepare + deliver effective presentations.
* Mapping and tracing data from system to system in order to establish data hierarchy and lineage.
* Using Data Lineage and reverse engineering as a way to track back errors in data till the data source.

## EDUCATION:

Bachelor of Computer Science.

**TOOLS AND TECHNOLOGIES:**

|  |  |
| --- | --- |
| **Programming & Scripting Languages** | Python, Scala, Java, PHP, R |
| **Libraries** | Big Data Ecosystem Apache Hadoop, MapReduce, Spark, HDFS, HBase,  Hive, Pig, ETL, Storm, Flume, Sqoop  Python (NumPy, Panda, Scikit-learn, SciPy), Spark ML, Spark MLlib |
| **Databases** | MySQL, PostgreSQL NoSQL Cassandra, MongoDB BI and Visualization  ZOHO, SAS, Tableau, IBM SPSS, Microsoft SSAS, Microsoft SSRS,  Power BI IDE Jupyter, PyCharm, IntelliJ, Eclipse, NetBeans, Atom. |
| **Cloud Based Tools** | Data Bricks, AWS (Amazon Web Services) |
| **Scheduling/Workflow**  **Distributed Tools** | Apache Oozie  Apache Hadoop, Apache Elasticsearch, Apache Storm, Apache Kafka |
| **Tools and Utilities** | ZOHO, GOVWIN, Salesforce, FileZilla, FTP Servers, WordPress, Joomla,  SquareSpace, WampServer, Prezi, Drupal, WIX, MS Office (Excel,  Access, Word, PowerPoint, Mail Project), Visio, BNB Matrix, JIRA, SQL  Server Management Studio, SQL Server Enterprise Manager, SQL Server  Profiler, Import & Export Wizard, Microsoft Management Console,  Crystal Reports, Power Pivot, ProClarity, Microsoft Office, Excel Data  Explorer, Anaconda |

**PROFESSIONAL EXPERIENCE:**

|  |  |  |
| --- | --- | --- |
| **Client: EDF trading, Houston,TX** | | **Jan 2017 – Till Date** |
| **Role: Data Scientist** |  | |

**Description:**

EDF Trading Limited, together with its subsidiaries, operates in the power, gas, freight, liquefied natural gas (LNG), freight, and carbon emission credits markets in Europe, Asia, and North America. It engages in generation, sale, and wholesale of electricity and gas, as well as in offshore production, transportation, storage, and wholesale trading activities

**Responsibilities:**

* Utilized Spark, Scala, **Hadoop**, **HBase**, **Cassandra**, **MongoDB**, **Kafka**, Spark **Streaming**, MLLib, Python, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc. and Utilized the engine to increase user lifetime by 45% and triple user conversations for target categories.
* Developed **Spark/Scala**, Python for regular expression (regex) project in the **Hadoop/Hive** environment with **Linux/Windows** for big data resources. Used clustering technique K-Means to identify outliers and to classify unlabeled data.
* Evaluated models using Cross Validation, Log loss function, **ROC** curves and used **AUC** for feature selection.
* Analyze traffic patterns by calculating autocorrelation with different time lags.
* Ensured that the model has low False **Positive Rate**.
* Developed entire frontend and backend modules using **Python** on **Django** Web Framework.
* Implemented the presentation layer with **HTML**, **CSS** and JavaScript.
* Involved in writing stored procedures using Oracle.
* Involved in analysis of Business requirement, Design and Development of High level and Low level designs, **Unit and Integration testing**.
* Designed & Implemented the next generation Lambda Architecture to manage Real Time/Micro Batch/Batch Analytics use cases as well as IOT use cases using **Spark Core /Spark Streaming/Kafka/NIFI-Hortonworks Data Flow/Hortonworks HDP**
* Designed and created highly scalable and sub second response time data access patterns like Pivotal **HAWQ/HDP, Spark SQL/HDFS/S3 & MPP like Redshift**.
* Created roadmap, strategies & implementations around data governance, data lineage, data tagging, metadata management, security cutting across data lake platform using **Apache Atlas/Waterline Data**.
* Lead the effort on managing/designing/architecting/implementing use cases for the Data Lake initiative and do presentations with **C-level** Executives to build confidence, capability and to drive data analytics vision for the future
* Lead & Implemented various predictive analytics **POC** using **Data Science/Machine Learning** algorithms(**R, Python, Scala, SparkR, Spark MLib** ) to demonstrate capabilities around use cases like Diet Coke Market Share reduction/Shoping Basket Analysis/Customer
* **Segmentation/Upselling/Product Mix/Trade Promotions/Channel** Optimizations in the US market and build different predictive models for future forecasting.
* Lead the efforts to introduce the concept of Data Lake Architecture so that data from various sources like **Teradata/Social Media/Sensor Data/Survey** Data and Unstructured Data can be stored and process in **Hadoop/Spark** and provide a road map to provide timelines and steps to replace Data warehouse/Data Marts.
* Performed Exploratory Data Analysis and **Data Visualizations using R, and Tableau**.
* Perform a proper **EDA**, Univariate and bi-variate analysis to understand the intrinsic effect/combined.
* Worked with several R packages including **knitr, dplyr, SparkR, CausalInfer, spacetime**.
* Interacted with the other departments to understand and identify dataneeds and requirements.

**Environment:** Python 2.x, CDH5, HDFS, Hadoop 2.3, Hive, Impala, Linux, Spark, Tableau Desktop, SQL Server 2012, Microsoft Excel, Matlab, Spark SQL, Pyspark.

|  |  |
| --- | --- |
| **Client: Amneal Pharmaceuticals, Brookhaven, NY** | **Nov 2015-Dec 2016** |
| **Role: Data Scientist** |  |

**Description:**

Amneal Pharmaceuticals LLC develops and produces generic drugs. It offers buprenorphine and naloxone tablets, ear drops, scalp and body oils, liquid syrups, folic acid tablets, oral solids, nasal sprays, and more. The company offers its products to customers and patients worldwide.

**Responsibilities:**

* Setup storage and data analysis tools in Amazon Web Services cloud computing infrastructure.
* Used pandas, numpy, seaborn, scipy, matplotlib, scikit-learn, NLTK in Python for developing various machine learning algorithms.
* Installed and used CaffeDeepLearning Framework
* Worked on different data formats such as JSON, XML and performed machine learning algorithms in Python.
* Worked as Data Architects and IT Architects to understand the movement of data and its storage and ER Studio 9.7
* Participated in all phases of data mining; data collection, data cleaning, developing models, validation, visualization and performed Gap analysis.
* Data Manipulation and Aggregation from different source using Nexus, Toad, Business Objects, Power BI and Smart View.
* Implemented Agile Methodology for building an internal application.
* Focus on integration overlap and Informatica newer commitment to MDM with the acquisition of Identity Systems.
* Good knowledge of Hadoop Architecture and various components such as HDFS, JobTracker, Task Tracker, NameNode, DataNode, Secondary NameNode, and MapReduce concepts.
* As Architect delivered various complex OLAP databases/cubes, scorecards, dashboards and reports.
* Programmed a utility in Python that used multiple packages (scipy, numpy, pandas)
* Implemented Classification using supervised algorithms like Logistic Regression, Decision trees, KNN, Naive Bayes.
* Utilized SQL, NoSQL databases, Python programing and API interaction
* Worked on DTS Packages, DTS Import/Export for transferring data between SQL Server 2000 to 2005
* Predictive modeling using state-of-the-art methods
* Build and maintain dashboard and reporting based on the statistical models to identify and track key metrics and risk indicators.
* Parse and manipulate raw, complex data streams to prepare for loading into an analytical tool.
* Migrating Informatica mappings from SQL Server to Netezza Foster culture of continuous engineering improvement through mentoring, feedback, and metrics
* Proven experience building sustainable and trustful relationships with senior leaders.

**Environment:** Regression, Logistic regression, Hadoop, Teradata, OLTP, Unix, Python, MLLib, SAS, random forest, OLAP, HDFS, NLTK, SVM, JSON and XML

|  |  |
| --- | --- |
| **Client: Acuity Insurance, Wisconsin, U.S** | **Jan 2014-Oct 2015** |
| **Role: Data Scientist** |  |

**Description:** Acuity Insurance is an insurance company with headquarters in Sheboygan, Wisconsin. The company is the 57th largest insurer in the United States.

**Responsibilities:**

* Evaluated business requirements and prepared detailed specifications that follow project guidelines required to develop written programs.
* Configured Hadoop cluster with Name node and slaves and formatted HDFS.
* Used Oozie workflow engine to run multiple Hive and Pig jobs.
* Launching AmazonEC2 Cloud Instances using Amazon Images (Linux) and Configuring launched instances with respect to specific applications.
* 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.
* Used Hive to partition and bucket data.
* Performed MapReduce Programs those are running on the cluster.
* Developed multiple MapReduce jobs for data cleaning and preprocessing.
* 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.
* Sequenced, assembled and annotated the transcriptome (**51 base paired reads, millions of data points**).
* Created Perl programs for combining and analyzing transcriptome and small **RNA** **datasets** to identify novel and conserved **micro** **RNAs**
* Established the core bioinformatics infrastructure for the university department.
* 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.
* Wrote Pig Scripts to perform ETL procedures on the data in HDFS.
* Created HBase tables to store various data formats of data coming from different portfolios.
* Worked on improving performance of existing Hive Queries.
* Leveraged working with cloud computation (**AWS, UC Farm cluster**), thus reducing the establishment costs
* Provided leadership and support to colleagues and research teams in data analysis and techniques at UC Davis and with collaborative teams

**Environment:**Linux, Hadoop, Python, MapReduce, HDFS, Python, GIT, Hive, SQL, Pig, Scoop, Flume, AWS, EC2, Twitter API, Oozie

**Client: Conair Corporation, Connecticut, U.S Oct 2012- Dec 2013**

**Role: Data Scientist**

**Description:** Conair Corporation is an American company which sells small appliances, personal care products, and health and beauty products for both professionals and consumers. It was founded in 1959 and has since expanded to include ten product divisions.

**Responsibilities:**

* Involved with Data Analysis primarily Identifying Data Sets, Source Data, Source Meta Data, Data Definitions and Data Formats
* Performance tuning of the database, which includes indexes, and optimizing SQL statements, monitoring the server.
* Wrote simple and advanced SQL queries and scripts to create standard and ad hoc reports for senior managers.
* Collaborated the data mapping document from source to target and the data quality assessments for the source data.
* Used Expert level understanding of different databases in combinations for Data extraction and loading, joining data extracted from different databases and loading to a specific database.
* Co-ordinate with various business users, stakeholders and SME to get Functional expertise, design and business test scenarios review, UAT participation and validation of financial data.
* Manipulating/mining data from database tables (**Redshift, Oracle, Data Warehouse**)
* Create automated metrics using **complex databases.**
* Providing analytical network support to improve quality and standard work result.
* Root cause research to identify process breakdowns within departments and providing data through use of various skill sets to find solutions to breakdown
* Foster culture of continuous engineering improvement through mentoring, feedback, and metrics
* Broad knowledge of programming, and scripting (**especially in R / Java / Python**)
* **Implemented Event Task for execute Application Automatically.**
* Troubleshot and resolved bugs in **.NET** **applications** to ensure optimal development environment.
* Created views, queries, and data warehouse reports using SSRS providing management with financial information from **SQL** Server production databases.
* Involved in developing Patches & Updates Module.

**Environment:** Machine learning, AWS, MS Azure, Cassandra, Spark, HDFS, Hive, Pig, Linux, Python (Scikit-Learn/Scipy/Numpy/Pandas), R, SAS, SPSS, Mysql, Eclipse, PL/SQL, SQL connector, Tableau.

**Client: Pidilite Industries, India Jan 2011-Sep 2012**

**Role: Python/ AWS Developer**

**Description:** Pidilite Industries Limited is an Indian-based adhesives manufacturing company. It also sells art material, construction chemicals and other industrial chemicals. Pidilite markets the Fevicol range of adhesives.

**Responsibilities:**

* Design, develop, test, deploy and maintain the website.
* Created a custom CMS built using Django, **AWS(Amazon web service**), **REST**, **Json** with JavaScript, JQuery, AJAX components, worked closely with designer. Converted Photoshop images into XHTML/CSS.
* Used **AW** stat and Google Analytics to analyze web traffic.
* Designed and developed the UI of the website using **HTML, AJAX, CSS** and JavaScript.
* Created scripts for system administration and AWS using languages such as BASH and Python.
* Designed and developed data management system using **MySQL**.
* Rewrite existing **Python/Django** modules to deliver certain format of data.
* Used **Django** Database API's to access database objects.
* Wrote python scripts to parse XML documents and load the data in database.
* Handled all the client side validation using JavaScript.
* Worked extensively with **Bootstrap, Java script, and jQuery** to optimize the user experience.
* Used Selenium Library to write fully functioning test automation process that allowed the simulation of submitting different we requests from multiple browser to web application.
* Used Python and Django to interface with the jQuery UI and manage the storage and deletion of content.
* Involved in development of Web Services using SOAP for sending and getting data from the external interface in the **XML format**.
* Responsible for debugging the project monitored on JIRA (Agile).
* Used **JQuery for all client side** Java script manipulation.
* Scalable, database-driven web application development using a variety of frameworks: ASP.NET on C#, Flask on **Python, and PHP**.
* Created unit test/regression test framework for working/new code.
* Using Subversion control tool to coordinate team-development.
* Built development environment with **JIRA.**
* Implemented multiple high performance **MongoDB** replica sets on EC2 with robust reliability.
* Developed entire frontend and backend modules using Python on **Django Web Framework**.
* Responsible for debugging and troubleshooting the web application. .

**Environment:** R 3.0, Erwin 9.5, Tableau 8.0, MDM, QlikView, MLLib, PL/SQL, HDFS, Teradata 14.1, JSON, HADOOP (HDFS), MapReduce, PIG, Spark, R Studio, MAHOUT, JAVA, HIVE, AWS.

**Client: Sun Pharmaceutical, Mumbai, India Jun 2009 -Dec 2010**

**Role: Data Analyst**

**Description:** Sun Pharmaceutical Industries Limited is an Indian multinational pharmaceutical company headquartered in Mumbai, Maharashtra that manufactures and sells pharmaceutical formulations and active pharmaceutical.

**Responsibilities:**

* Involved in analysis of Business requirement, Design and Development of High level and Low level designs, Unit and Integration testing
* Performed data analysis and data profiling using complex **SQL** on various sources systems including Teradata, **SQL Server.**
* Developed the logical data models and physical data models that confine existing condition/potential status data fundamentals and data flows using **ER Studio**
* Created the conceptual model for the data warehouse using Erwin data modeling tool.
* Reviewed and implemented the naming standards for the entities, attributes, alternate keys, and primary keys for the logical model.
* Performed second and third normalizations for **ER** **data** **model** of **OLTP system**
* Worked with data compliance teams, Data governance team to maintain data models, Metadata, Data Dictionaries; define source fields and its definitions.
* Involved in the daily maintenance of the database that involved monitoring the daily run of the scripts as well as troubleshooting in the event of any errors in the entire process.
* Investigate data from multiple systems, at different timescales and in complex formats to discover hidden relationships and useful information.
* Data analysis may encompass a range of analytical capabilities including aggregations, statistical analysis, regression, time-series analysis, clustering, network analysis and other statistical modeling techniques.
* Document implementation and design details for model choices, rationale and reasoning.
* Demonstrate and deliver rapid prototypes for the models and work with QA to ensure quality.
* Primary focus will be in applying data mining techniques, doing statistical analysis, and building high quality prediction model to support various **Business**/service functions.
* 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
* Building forecast model and build predictive models with the relevant information
* Support different functions **like Sales, Finance, Supply chain, competitive intelligence** etc. on their analytics
* Implementing and developing data visualization models for the business presentations
* Research, design, implement and validate cutting-edge algorithms to analyze diverse sources of data to achieve targeted outcomes.

**Environment:** Python 2.7 and 3.4, Django, HTML5/CSS, MS SQL Server 2013, MySQL, Java script, Eclipse, Linux, Shell Scripting, JQuery, GitHub, Angular.JS, Jira.