**Prathap K**

**Data Scientist**

**Email: prathapkc17@gmail.com**

**Phone No: (513) 666-1820**

**Professional Summary:**

* Above 8+ years of experience in Machine Learning, Data Mining with large datasets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive Modelling, Data Visualization.
* Designing of Physical Data Architecture of New system engines.
* 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**.
* Having good experience in **NLP** with **Apache**, **Hadoop** and **Python**.
* Hands on **SparkMlib** utilities such as **Including Classification**, **Regression**, **Clustering**, **Collaborative Filtering**, **Dimensionality Reduction**.
* Proficient in **Statistical Modelling** and **Machine Learning** techniques (**Linear**, **Logistics**, **Decision Trees**, **Random Forest**, **SVM**, **K** - **Nearest Neighbours**, **Bayesian**, **XG Boost**) in **Forecasting**/ **Predictive Analytics**, **Segmentation Methodologies**, **Regression based models**, **Hypothesis testing**, **Factor analysis**/ **PCA**, **Ensembles**.
* 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.
* Developing **Logical Data Architecture** with **adherence** to **Enterprise Architecture**.
* 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.
* Adept in statistical programming languages like **R** and also **Python** including **Big Data** technologies like **Hadoop**, **Hive**.
* Strong experience in **Software Development Life Cycle** (**SDLC**) including **Requirements Analysis**, **Design Specification** and **Testing** as per Cycle in both **Waterfall** and **Agile** methodologies.
* Experience working with data modelling tools like Erwin, **Power Designer** and **ER Studio**.
* Skilled in using dplyr and pandas in **R** and **python** for performing exploratory data analysis.
* Experience in designing stunning visualizations using **Tableau** software and publishing and presenting **dashboards**, **Storyline** on web and desktop platforms.
* Experience in designing **star schema**, **Snow flake schema** for **Data Warehouse**, **ODS architecture**.
* Experience in designing and developing the **Tableau** and updating the **existing desktop**, developing **ad**-**hoc reports**, **scheduling** the processes and **administering** the **tableau activities** using **tableau**.
* Experienced in designing customized interactive **dashboards** in **Tableau** using **Marks**, **Action**, **Filters**, **Parameter** and **Calculations**.
* Good understanding of **Teradata SQL Assistant**, **Teradata Administrator** and data load/ export utilities like **BTEQ**, **Fast Load**, **Multi Load**, **Fast Export**.
* Experience and Technical proficiency in Designing, **Data Modelling Online Applications**, **Solution Lead** for **Architecting Data Warehouse**/**Business Intelligence Applications**.
* Experience in maintaining database architecture and metadata that support the **Enterprise Data warehouse**.
* **Prediction** - **Prediction** of a numerical value using **Regression** or **CART**.
* Experience with **Data Analytics**, **Data Reporting**, **Ad**-**hoc Reporting**, **Graphs**, **Scales**, **Pivot Tables** and **OLAP** reporting.
* Highly skilled in using visualization tools like **Tableau**, **ggplot2** and **d3**.**js** for creating dashboards.
* Highly skilled in using **Hadoop** (**Pig** and **Hive**) for basic analysis and extraction of data in the infrastructure to provide data summarization.

**Education:**

Bachelor of Computer Science & Technology.

**Technical Skills:**

|  |  |
| --- | --- |
| **Machine Learning** | Regression, Polynomial Regression, Random Forest, Logistic Regression, Decision Trees, Classification, Clustering, Association, Simple/Multiple linear, Kernel SVM, K-Nearest Neighbours (K-NN). |
| **Data Modeling Tools** | Erwin r, ER/Studio, Star-Schema Modeling, Snowflake-Schema Modeling, FACT and dimension tables, Pivot Tables. |
| **Big Data Technologies** | Hadoop, Hive, HDFS, Map Reduce, Pig, Kafka. |
| **OLAP/ BI / ETL Tool** | Business Objects 6.1/XI, MS SQL Server 2008/2005 Analysis Services (MS OLAP, SSAS), Integration Services (SSIS), Reporting Services (SSRS), Performance Point Server (PPS), Oracle 9i OLAP, MS Office Web Components (OWC11), DTS, MDX, Crystal Reports 10, Crystal Enterprise 10(CMC) |
| **BI Tools** | Tableau, Tableau server, Tableau Reader, SAP Business Objects, OBIEE, Qlik View, SAP Business Intelligence, Amazon Redshift, or Azure Data Warehouse. |
| **Packages** | ggplot2, caret, dplyr, Rweka, gmodels, RCurl, tm, C50, twitter, NLP, Reshape2, rjson, plyr, pandas, numPy, seaborn, sciPy, matplot lib, scikit-learn, Beautiful Soup, Rpy2, sqlalchemy. |
| **Web Technologies** | JDBC, HTML5, DHTML and XML, CSS3, Web Services, WSDL, |
| **Tools** | Erwin r 9.6, 9.5, 9.1, 8.x, Rational Rose, ER/Studio, MS Visio, SAP Power designer. |
| **Languages** | SQL, PL/SQL, ASP, Visual Basic, XML, SAS, Python, SQL, T-SQL, SQL Server, C, C++, JAVA, HTML, Shell Scripting, PERL, R, Matlab, Scala. |
| **Databases** | SQL, Hive, Impala, Pig, Spark SQL, Databases SQL-Server, MySQL, MS Access, HDFS, HBase, Teradata, Netezza, Mongo DB, Cassandra, SAP HANA. |
| **Reporting Tools** | MS Office (Word/Excel/Power Point/ Visio), Tableau, Crystal reports XI, Business Intelligence, SSRS, Business Objects 5.x/ 6.x, Cognos7.0/6.0. |
| **ETL Tools** | Informatica Power Centre, SSIS. |
| **VersionControlTools** | SVM, GitHub. |
| **Project Execution**  **Methodologies** | Ralph Kimball and Bill Inmon data warehousing methodology, Rational Unified Process (RUP), Rapid Application Development (RAD), Joint Application Development (JAD). |
| **Operating System** | Windows, Linux, Unix, Macintosh HD, Red Hat. |

**Professional Experience:**

|  |  |
| --- | --- |
| **Client: Schlumberger Limited, Houston, Texas.** | **S Sep 2017 - Till Date** |
| **Role: Data Scientist/ Machine Learning Engineer** |  |

**Description:** Schlumberger Limited is the world's largest oilfield services company. Schlumberger employs approximately 100,000 people representing more than 140 nationalities working in more than 85 countries.

**Responsibilities:**

* Built models using **Statistical techniques** like Bayesian **HMM** and **Machine Learning** classification models like **XG Boost**, **SVM**, and **Random Forest**.
* A highly immersive **Data Science** program involving **Data Manipulation**&**Visualization**, **Web** Scraping, **Machine Learning**, **Python** programming, **SQL**, **GIT**, **Unix** Commands, **MySQL**, **Mongo DB**, **Hadoop**.
* Designing and develop **Tableau Reports**, **Documents**, **Dashboards** for **specified requirements** and **timelines**.
* Setup storage and **data analysis** tools in **Amazon Web Services** cloud computing infrastructure.
* Used **pandas**, **numpy**, **sea born**, **scipy**, **matplotlib**, **scikit**-**learn**, **NLTK** in **Python** for developing various **machine learning algorithms**.
* Installed and used **Caffe Deep Learning 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**.
* Purchasing, Setting up and configuring a **Tableau Server** and **MS**-**SQL 2008 R2 server** for **Data warehouse purpose**.
* Preparing **Dashboards** using **calculations**, **parameters** in **Tableau**.
* Designed, developed and implemented **Tableau Business Intelligence reports**.
* 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.
* Good knowledge of **Hadoop Architecture** and various components such as **HDFS**, **Job Tracker**, **Task Tracker**, **Name Node**, **Data Node**, **Secondary Name Node**, and **Map Reduce** 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**.
* Used **Teradata15** utilities such as **Fast Export**, **MLOAD** for handling various tasks data migration/ETL from **OLTP** Source Systems to **OLAP** Target Systems
* Experience in Hadoop ecosystem components like **Hadoop Map Reduce**, **HDFS**, **HBase**, **Oozie**, **Hive**, **Sqoop**, **Pig**, Flume including their installation and configuration.
* Updated **Python scripts** to match training data **with** our database stored in **AWS Cloud** Search, so that we would be able to **assign** each document a response label for further classification.
* **Data transformation** from various resources, **data organization**, features extraction from raw and stored.
* Validated the machine learning classifiers using ROC Curves and Lift Charts.
* Extracted data from **HDFS** and prepared data for exploratory analysis using **data mugging**.

**Environment:**ER Studio 9.7, Tableau 9.03, AWS, Teradata 15, MDM, GIT, Unix, Python 3.5.2,, Machine learning, MLLib, SAS, regression, logistic regression, Hadoop, NoSQL, Teradata, OLTP, random forest, OLAP, HDFS, ODS, NLTK, SVM, JSON, XML, Map Reduce.

|  |  |
| --- | --- |
| **Client: BBVA Compass, Birmingham, Alabama.** | **Jun 2016 - Aug 2017** |
| **Role: Data Scientist/Machine Learning Engineer** |  |

**Description:** BBVA Compass Bancshares, Inc. (formerly Compass Bancshares) is a bank holding company headquartered in Birmingham, Alabama. It has been a subsidiary of the Spanish multinational Banco Bilbao VizcayaArgentaria since 2007 and operates chiefly in Alabama, Arizona, California, Colorado, Florida, New Mexico, and Texas.

**Responsibilities:**

* Extracted data from **HDFS** and prepared data for exploratory analysis using **data munging**
* Built models using Statistical techniques like Bayesian HMM and Machine Learning classification models like **XG Boost**, **SVM**, and **Random Forest.**
* Participated in all phases of **data mining, data cleaning, data collection, developing models, validation, visualization,** and **performed Gap analysis.**
* A highly immersive Data Science program involving **Data Manipulation** &**Visualization**, **Web Scraping**, **Machine Learning, Python programming, SQL**, **GIT**, **Mongo DB, Hadoop**.
* Setup storage and data analysis tools in **AWS cloud computing infrastructure**.
* Installed and used Caffe **Deep Learning 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
* Used **pandas, numpy, sea born, matplotlib, scikit-learn, scipy, NLTK** in **Python** for developing various machine learning algorithms.
* Data Manipulation and Aggregation from different source using **Nexus, Business Objects, Toad, 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.
* Coded proprietary packages to analyse and visualize **SPC** file data to identify bad spectra and samples to reduce unnecessary procedures and costs.
* Programmed a utility in Python that used multiple packages (**numpy, scipy, pandas**)
* Implemented Classification using supervised algorithms like **Logistic Regression, Decision trees, Naive Bayes, KNN.**
* As Architect delivered various complex **OLAP databases/cubes, scorecards, dashboards** and **reports**.
* Updated **Python** scripts to match training data with our database stored in **AWS Cloud Search**, so that we would be able to assign each document a response label for further classification.
* Used Teradata utilities such as **Fast Export, MLOAD** for handling various tasks data **migration/ETL** from **OLTP** Source Systems to **OLAP Target Systems**
* Data transformation from various resources, data organization, features extraction from raw and stored.
* Validated the machine learning classifiers using **ROC Curves** and **Lift Charts.**

**Environment:** SQL/Server, Oracle 10g/11g, MS-Office, Teradata, Informatica, ER Studio, XML, R connector, Python, R, Tableau 9.2.

**Client: Signet Accel, Columbus, Ohio. Feb 2015 - May 2016  
 Role: Data Scientist**

**Description:** Signet Accel, LLC operates AVEC, a commercial federated data integration platform for the healthcare industry worldwide. Its platform enables query and analysis of distributed healthcare data, and protects the ownership and control of data at each site.

**Responsibilities:**

* Utilized Spark**, Scala, Hadoop, HQL, VQL, oozie, pySpark, Data Lake, Tensor Flow, HBase, Cassandra**, **Redshift, Mongo DB, Kafka, Kinesis, Spark Streaming, Edward, CUDA, MLLib, AWS, Python**, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
* Utilized the engine to increase user lifetime by 45% and triple user conversations for target categories.
* Application of various machine learning algorithms and statistical modelling like **decision trees, text analytics**, natural language processing (**NLP**), supervised and unsupervised, regression models, social network analysis, neural networks, deep learning, **SVM,** clustering to identify Volume using **scikit-learn** package in **python, Mat lab.**
* Worked onanalysing **data from Google Analytics, Ad Words, Facebook, etc.**
* Evaluated models using Cross Validation, Log loss function, **ROC** curves and used **AUC** for feature selection and elastic technologies like **Elastic Search, Kibana**.
* Performed Data Profiling to learn about behaviour with various features such as traffic pattern, location, Date and Time etc.
* Categorized comments into positive and negative clusters from different social networking sites using Sentiment Analysis and Text Analytics
* Performed **Multinomial Logistic Regression, Decision Tree**, **Random forest, SVM** to classify package is going to deliver on time for the new route.
* Performed data analysis by using Hive to retrieve the data from **Hadoop cluster**, **Sql** to retrieve datafrom Oracle database and used **ETL** for data transformation.
* Performed Data Cleaning, features scaling, features engineering using **pandas** and **numpy** packages in **python**.
* Exploring DAG's, their dependencies and logs using Air Flow pipelines for automation
* Performed data cleaning and feature selection using **MLlib** package in **PySpark** and working with deep learning frameworks such as **Caffe, Neon**.
* Developed **Spark/Scala, R, 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 un-labelled data.
* Tracking operations using sensors until certain criteria is met using **Air Flow technology**.
* Responsible for different Data mapping activities from Source systems to Teradata using utilities like **TPump, FEXP,BTEQ, MLOAD, FLOAD** etc.
* Analyse traffic patterns by calculating autocorrelation with different time lags.
* Ensured that the model has low False Positive Rate and Text classification and sentiment analysis for unstructured and semi-structured data.
* Addressed over fitting by implementing of the algorithm regularization methods like **L1 and L2**.
* Used Principal Component Analysis in feature engineering to analyse high dimensional data.
* Used **MLlib, Spark's 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.
* Created and designed reports that will use gathered metrics to infer and draw logical conclusions of past and future behaviour.
* Developed **Map Reduce** pipeline for feature extraction using **Hive** and **Pig**.
* Created Data Quality Scripts using **SQL** and **Hive** to validate successful data load and quality of the data. Created various types of data visualizations using **Python** and **Tableau**.
* Communicated the results with operations team for taking best decisions.
* Collected data needs and requirements by Interacting with the other departments.

**Environment:**Python 2.x, CDH5, HDFS, Hadoop 2.3, Hive, Impala, AWS, Linux, Spark, Tableau Desktop, SQL Server 2014, Microsoft Excel, Matlab, Spark SQL, Pyspark.

**Client: Grand Canyon Education Inc, Phoenix, Arizona. Jun 2013 - Jan 2015**

**Role: Data Analyst/Data Architecture**

**Description:** Grand Canyon Education, Inc. provides postsecondary education services in the United States and Canada. It focuses on offering graduate and undergraduate degree programs in education, healthcare, business, and liberal arts disciplines.

**Responsibilities:**

* Worked with **BI team** in gathering the report requirements and also **Sqoop** to export data into **HDFS** and **Hive**
* Involved in the below phases of Analytics using **R, Python** and **Jupyter** notebook. a. Data collection and treatment: **Analysed** existing internal data and external data, worked on entry errors, classification errors and defined criteria for missing values b. Data Mining: Used cluster analysis for identifying customer segments, Decision trees used for profitable and non-profitable customers, Market Basket Analysis used for customer purchasing behaviour and part/product association.
* Developed multiple **Map Reduce** jobs in **Java** for data cleaning and pre-processing.
* Assisted with data capacity planning and node forecasting.
* **Installed, Configured** and **managed** Flume Infrastructure.
* Administrator for **Pig**, **Hive** and **HBase** installing updates patches and upgrades.
* Worked closely with the claims processing team to obtain patterns in filing of fraudulent claims.
* Worked on performing major upgrade of cluster from **CDH3u6 to CDH4.4.0**
* Developed Map Reduce programs to extract and transform the data sets and results were exported back to **RDBMS** using **Sqoop**.
* Patterns were observed in fraudulent claims using text mining in **R** and **Hive**.
* Exported the data required information to **RDBMS** using **Sqoop** to make the data available for the claims processing team to assist in processing a claim based on the data.
* Developed Map Reduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the **EDW**.
* Created tables in Hive and loaded the structured (resulted from **Map Reduce** jobs) data
* Using **Hive QL** developed many queries and extracted the required information.
* Created Hive queries that helped market analysts spot emerging trends by comparing fresh data with EDW reference tables and historical metrics.
* Was responsible for importing the data (mostly log files) from various sources into **HDFS** using **Flume**
* Enabled speedy reviews and first mover advantages by using Oozie to automate data loading into the Hadoop Distributed File System and **PIG** to pre-process the data.
* Provided design recommendations and thought leadership to sponsors/stakeholders that improved review processes and resolved technical problems.
* Managed and reviewed **Hadoop** log files.
* Tested raw data and executed performance scripts.

**Environment:** HDFS, PIG, HIVE, Map Reduce, Linux, HBase, Flume, Sqoop, R, VMware, Eclipse, Cloud era, and Python.

**Client: Trade India, New Delhi, India Jan 2012 - May 2013**

**Role: Data Analyst/ Data Modeler**

**Description:** Trade India is an online [Business to Business](https://en.wikipedia.org/wiki/Business-to-business) (B2B) [portal](https://en.wikipedia.org/wiki/Web_portal) for small businesses based in India and around the globe. The portal was started in 1996 by [Bikky Khosla](https://en.wikipedia.org/wiki/Bikky_Khosla) and is maintained and promoted under the flagship company, Infocom Network Ltd.

**Responsibilities:**

* Worked as **DataExpert** on a data mining **ETL** development project using **SAS** Enterprise Guide.
* Created test plan documents for all back-end database modules.
* Worked with large amounts of structured and unstructured data.
* Responsible for data collection, cleansing, and**ANOVA**. Designed technical solution roadmap to deal with noise in sales data.
* Worked on loading the data from **MySQL** to **HBase** where necessary using **Sqoop**.
* Handled end-to-end project from data discovery to model deployment
* Knowledge in Business Intelligence tools and visualization tools such as **BusinessObjects**, **Tableau**, **ChartIO**, etc.
* Knowledge in Machine Learning concepts (**GeneralizedLinearmodels**, **Regularization**, **RandomForest**, **TimeSeriesmodels**, etc.).
* Deployed **GUI** pages by using **JSP, JSTL, HTML, DHTML, XHTML, CSS, JavaScript, and AJAX**.
* Configured the project on **Web Sphere 6.1** application servers.
* Implemented the online application by using **Core Java, JDBC, JSP, Servlets and EJB 1.1, Web Services, SOAP, WSDL.**
* Monitoring the automated loading processes.
* Communicated with other Health Care info by using Web Services with the help of **SOAP, WSDLJAX-RPC.**
* Used Singleton, factory design pattern, **DAO** Design Patterns based on the application requirements
* Used **SAX** and **DOM** parsers to parse the raw **XML** documents
* Used **RAD** as Development **IDE** for web applications.
* Used **Log4J** logging framework to write Log messages with various levels.
* Involved in fixing bugs and minor enhancements for the front-end modules.
* Implemented **Microsoft Visio** and Rational Rose for designing the Use Case Diagrams, Class model, Sequence diagrams, and Activity diagrams for **SDLC** process of the application.
* Maintenance in the testing team for **System testing/Integration/UAT**.
* Guaranteeing quality in the deliverables.
* Conducted Design reviews and Technical reviews with other project stakeholders.
* It was a part of the complete life cycle of the project from the requirements to the production support.
* Implemented the project in Linux environment.

**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: Tessolve, Bangalore, India. Mar 2010 - Dec 2011**

**Role: Data Architecture/Data Analyst**

**Description:** Tessolve is a multinational semiconductor engineering [solution](https://en.wikipedia.org/wiki/Solution) provider based in [Bangalore](https://en.wikipedia.org/wiki/Bangalore), [India](https://en.wikipedia.org/wiki/India). The company offers engineering expertise in the areas of [Semiconductor](https://en.wikipedia.org/wiki/Semiconductor) IC Design, Test & Product Engineering, [PCB design](https://en.wikipedia.org/wiki/Printed_circuit_board#Design), Failure Analysis and Systems design.

**Responsibilities:**

* Implementation of **Metadata Repository**, **Maintaining Data Quality**, **Data Cleanup procedures**, **Transformations, DataStandards**, DataGovernanceprogram, **Scripts**, **StoredProcedures**, **triggers** and execution of test plans.
* Developed Internet traffic scoring platform for **adnetworks**, **advertisers**, and **publishers** (rule engine, site scoring, keyword scoring, lift measurement, linkage analysis).
* Responsible for communication and negotiation with project related aspects of project loading, construction budget, design alterations, and unexpected events on the project.
* Responsible for defining the key identifiers for each mapping/interface.
* Clients include **eBay, Click Forensics, Cars.com, Turn.com, Microsoft, and Looksmart**.
* Designed the architecture for one of the first analytics 3.0. Online platforms: all-purpose scoring, with **on-demand, SaaS, API** services.
* Web crawling and text mining techniques to score referral domains, generate keyword taxonomies and assess commercial value of bid keywords.
* Used **RAD** as Development **IDE** for web applications.
* Developed new hybrid statistical and data mining technique known as hidden decision trees and hidden forests.
* Reverse engineering of keyword pricing algorithms in the context of pay-per-click arbitrage.
* Performed data quality in **Talend Open Studio**.
* Coordinated meetings with vendors to define requirements and system interaction agreement documentation between client and vendor system.
* Automated bidding for advertiser campaigns based either on keyword or category (run-of-site) bidding.
* Creation of multimillion bid keyword lists using extensive web crawling. Identification of metrics to measure the quality of each list (yield or coverage, volume, and keyword average financial value).
* Enterprise Metadata Library with any changes or updates.
* Document data quality and traceability documents for each source interface.
* Establish standards of procedures.
* Generate weekly and monthly asset inventory reports.

**Environment:** Erwin r7.0, SQL Server 2000/2005, Windows XP/NT/2000, Oracle 8i/9i, MS-DTS, UML, UAT, SQL Loader, OOD, OLTP, PL/SQL, MS Visio, Informatica.