**Prayash T**

**Sr. Data Scientist**

****

Cell no: **(617)-982-2841**

Email: **tprayash91@gmail.com**

**PROFESSIONAL SUMMARY:**

**5+ years** of experience in IT and comprehensive industry knowledge on **Machine Learning**, **Artificial** **Intelligence**, **Statistical** **Modeling**, **Data** **Analysis**, **Predictive** **Analysis**, **Data** **Manipulation**, **Data** **Mining**, **Data** **Visualization** and **Business** **Intelligence** in various domains such as **Health Care**, **Banking**, **Manufacturing**, **Finance** and **Software Development** **Industry**

* Proficient at building robust **Machine** **Learning**, **Deep** **Learning** **models**, **Convolution** **Neural** **Networks** (**CNN**), **Recurrent** **Neural** **Networks** (**RNN**), **LSTM** using **Tensor** **Flow** and **Keras**. Adept in analyzing large datasets using Apache Spark, PySpark, Spark ML and Amazon Web Services (AWS).
* Adept in statistical programming languages like **Python**, **R** and **SAS** including Big Data technologies like **Hadoop**, **MapReduce** **jobs**, **HDFS**, **Apache** **Spark**, **Hive**, **Pig**, **Sqoop**, **Flume**, **Kafka** and **NoSQL** Based Databases. In depth knowledge in using of **spark** machine learning library **MLlib**.
* Knowledge and experience working in **Waterfall, SDLC** as well as **Agile** environments including the **Scrum** process and using Project Management tools like **ProjectLibre**, **Jira**/**Confluence** and version control tools such as **GitHub**/**Git**.
* Experience in providing the consistent computing environment/ CICD pipeline by installing and upgrading the required plugins.
* Hands on experience and in provisioning virtual clusters under **Amazon Web Service (AWS)** cloud which includes services like **Elastic compute cloud (EC2), S3, CFT and EMR.**
* Strong administrative and development knowledge of Cognos environment and tools IBM Cognos 8BI (Framework Manager, Report Studio, Query Studio, Analysis Studio, Event Studio), Report Net, Impromptu, Power Play, Impromptu Web Reports (IWR), Access Manager, Upfront and IWR servers.
* Experience in performing **Feature** **Selection**, **Linear** **Regression**, **Logistic** **Regression**, **Principle** **Component** **Analysis, LDA,** **k** - **Means** **Clustering**, **Classification**, **Decision** **Tree**, **Supporting** **Vector** **Machines** (**SVM**), **Naive** **Bayes**, **K**-**Nearest** **Neighbors** (**KNN**), **Random** **Forest**, and **Gradient** **Descent**, **Neural** **Network** algorithms to train and test the huge data sets.
* Experienced in **Data Integration Validation** and **Data Quality controls** for **ETL** process and Data Warehousing using **MS Visual Studio, SSAS, SSIS and SSRS.**
* Expert in creating **PL/SQL** Schema objects like Packages, Procedures, Functions, Subprograms, Triggers, Views etc. using **SQL Developer**.
* Performed dimensionality reduction using **principal component analysis**, **Multi-dimensional Scaling** and other dimensionality reduction and feature selection techniques to derive new insights from the data.
* Experienced with Recommendation Systems such as **Collaborative filtering** and **content-based filtering.**
* Hand on experience with **H2O (R packages) and** used **data cleansing, data manipulation, exploratory data analysis, feature engineering techniques** with **H2O.**
* Built variety of Executive Dashboards using Cognos Workspace (Business Insight) and Cognos Workspace Advanced (Business Insight Advanced) for a quick glance of various attributes/KPIs for Executive management.
* Good understanding of **Teradata SQL Assistant, Teradata Administrator** and data load/ export utilities like **BTEQ, Fast Load, Multi Load, and Fast Export**.
* Experience in **Text Analytics**, **Predictive Analytics**, **Optimization methods** **and Natural Language Processing(NLP), Time Series Analysis**, developing different **Statistical Machine Learning**, **Data Mining solutions** to various business problems and generating **data visualizations using Rstudio, Rstudio connect and Python.**
* Demonstrated analytical thinking by extensively leveraging statistical techniques, such as **T-test**, **P-value analysis**, **z-score analysis**, **ANOVA**, **Confidence Interval**, **Confusion Matrix**, **Precision**, **Recall**, **ROC / AUC** curve analysis, etc.
* Good at using Google Cloud Platform (GCP) including Docker Containers, Container Builder, Container Engine (Kubernetes), Network Load Balancer (Ingress & Node Port) with Certificates (HTTPS), Cloud Storage, Data Store (NoSQL), Data Proc (managed Hadoop), and Cloud Pub/Sub (managed Queues), Big Table (managed HBase) etc.
* Strong expertise in **ETL**, **Data warehousing**, Operational Data Store (ODS), **Data Marts**, **OLAP** and **OLTP** technologies.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Analytical Tool | Tableau, SSRS, SSAS, Power BI, Alteryx, Qlikview, Cognos |
| Data Integration (ETL) Tools | SSIS, Informatica |
| SQL Tools | Teradata SQL Assistant, PL/SQL Developer, Toad, SQL Server Management Studio, PostgreSQL, Oracle |
| Statistical Software | R, Python, SAS |
| Statistical Methods | sampling, under sampling, oversampling, SMOTE resampling methods, Hypothesis Testing, Confidence Interval, P-value, T-Test, ANOVA, Correlation, Feature Engineering / Feature Selection techniques, Time Series Analysis, Factor analysis etc. |
| Machine Learning | Linear Regression, Logistic Regression, Decision trees, Random forest, Association Rule Mining (Market Basket Analysis), Clustering (K-Means, Hierarchal), Gradient decent, SVM (Support Vector Machines), Deep Learning (CNN, RNN, ANN) using TensorFlow (Keras). |
| Ensemble Methods | Base Learning, Meta Learning, Majority Voting, Averaged Ensembles, Weighted Averaging, Stacked Ensemble, AutoML – Scikit-Learn, MLjar, etc.  |
| Programming  | R (libraries dplyr, dbplyr, tidyverse, stringR, data table caret, glmnet, xgboost, rpart, ggplot2), Python (libraries numpy, pandas, scikit-learn, seaborn, matplotlib, NLTK), SAS, NoSql, PySpark. PySpark sql, Sql, etc. |
| Big Data tools  | Hadoop Distributed File System, PIG, Hive, Map Reduce, Cloudera, MSBI, MLlib, Microsoft Azure Machine Learning Studio.  |
| Others | GIT, Windows, Linux, JIRA, HTML, Radar, Microsoft Azure, Spark, Google Cloud, Amazon AWS, Confluence, Git, Airflow |

**WORK EXPERIENCE:**

 **United Health Care, MA Apr 2019- Present**

 **Sr. Data Scientist**

**Sr. Cloud/DevOps Engineer**

* Responsible in developing system **models, prediction algorithms, solutions to prescriptive analytics problems, data mining techniques, and/ or econometric model**.
* Built **Artificial Neural Network using Tensor Flow in Python** to identify the customer's probability of canceling the connections. (Churn rate prediction).
* Worked on Talend ETL to load data from various sources to Oracle DB. Used tmap, treplicate, tfilterrow, tsort and various other features in Talend.
* Developed custom airflow operators using Python to generate and load CSV files into GS from SQL Server and Oracle databases.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics . Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Developed Cognos Framework Manager Models, published packages on content stores, elevating reports from/to Dev/Test/Prod regions.
* Conducted a hybrid **of Hierarchical and K-means Cluster Analysis using IBM SPSS** and identified meaningful segments through a discovery approach.
* Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on GCP.
* Setting up and configuring Kafka Environment in Windows from the scratch and monitoring it.
* Created a data pipeline through Kafka Connecting two different clients Applications namely SEQUENTRA and LEASE ACCELERATOR
* Automation and orchestration of deployment to various environments using Jenkins, Airflow and Genie.
* Ensuring that the model has low False Positive Rate and Text classification and sentiment analysis for unstructured and semi-structured data.
* Built a data Warehouse system to improve overall quality and productivity of existing processes and deploying reliable and accurate information.
* Create various types of data visualizations using **Python**, **Tableau and Power BI**.
* Designed and implemented end-to-end systems for **Data Analytics and Automation**, integrating custom visualization tools using **Rstudio, Rstudio connect and Tableau**.
* Developed reports based on Cognos Framework Manager metadata model’s queries/stored procedures and by writing direct manual sql queries in reports.
* Built UNIX Linux shell scripts for running Informatica workflows, data cleansing, purge, delete, and data loading and for ELT process.
* Leveraged disparate data sources that provide deep customer insight including online transactional data, web data, payment and orders history and marketing campaigns exposure data using **Python**.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the AWS EMR Spark cluster and Cloud Dataflow on GCP.
* Built **data pipelines**, implemented code **modularization** involving package creation and developed **REST API’s** using **Flask** for production deployment.
* Created Talend Development Standards. This document describes the general guidelines for Talend developers, the naming conventions to be used in the Transformations and also development and production environment structures.
* Developed python code for different tasks, dependencies, SLA watcher and time sensor for each job for workflow management and automation using Airflow tool.
* Worked on google cloud platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.
* Handled importing data from various data sources, performed transformations using **Hive, Map Reduce, and loaded data into HDFS.**
* **Created Active Reports, List, Crosstab, Dashboard Reports and Drill-Through Reports using advanced reporting techniques in Cognos**
* Worked on **outlier detection** with data visualizations using boxplots, feature engineering using **Gaussian Mixture Models and KNN distances built using Pandas, NumPy.**
* Wrote, tested and **implemented Teradata Fast load, Multiload** and **BTEQ** scripts, DML and DDL.
* Adopted feature engineering techniques with 200+ predictors in order to find the most important instances for the models such as sequential feature selection.
* Extensively worked on **Data Cleansing**, **Manipulation**, **feature engineering**, **machine learning model**

**development** and **model deployment** with Microsoft **Azure Machine Learning Studio.**

Environment: Python 2.x, R, HDFS, Hive, Linux, Spark, IBM SPSS, Tableau Desktop, SQL Server 2012, Microsoft Excel, MATLAB, Spark SQL, PySpark, Hadoop, Teradata 14, Tableau, GitLab, Azure Machine Learning Studio, Power BI, Rstudio, Rstudio Connect, Airflow

 **Cotiviti, GA Dec 2017- Nov 2018**

 **Data Scientist/ Tableau Developer**

**Sr. Cloud/DevOps Engineer**

* Responsible for developing and deploying risk-based decision tools and building knowledge-based systems to solve large scale computational problem.
* Involved in end-to-end implementation of Data warehousing projects, which include Business Requirements gathering, Analysis, System study, Prepare Functional & Technical specifications, Design (Logical and Physical model), Coding, Testing, Code migration, Implementation, System maintenance, Support, and Documentation.
* Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow and Oozie.
* Responsible for estimating the cluster size, monitoring and troubleshooting of the Spark databricks cluster.
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Involved in full life cycle of Cognos BI including installation, configuration of Cognos BI Products, development and maintenance of Cognos Reports, fixing issues, testing and migrated to production.
* Collaborated with data engineers and operation team to implement the ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
* Developed complex mappings using Informatica Power Center Designer to transform and load the data from various source systems from DB2, DB2/400, Oracle, SQL Server, Flat files to load into Staging, ODS, Enterprise Data warehouse, and subject oriented Data marts for Sales, Marketing and Finance departments.
* Automated by building workflows to extract data from various REST APIs and databases, processing responses, data transformations in python and R.
* Used Jenkins as a packaging and deployment tool for migrating ETL & Unix components.
* Created Procedure Document for Implementation procedures for every release for all the UNIX, Informatica Objects and if any catch-up process needed to be done.
* Generated percentile scores for capturing distance between recurring incidents and integrated with complaints effectiveness metrics dashboard in **Tableau and Power BI** to provide visual insights to business users.
* Partnered with infrastructure and platform teams to configure, tune tools, automate tasks and guide the evolution of internal big data ecosystem; serve as a bridge between data scientists and infrastructure/platform teams.
* Tested classification methods such as **Random Forest, Logistic Regression and Gradient Boosting; performed Cross Validation** for hyper-parameter tuning to optimize the models for unseen data.
* Participated in product redesigns and enhancements to know how the changes will be tracked and suggest product direction based on data patterns.
* Responsible for **data aggregation**, **data pre-processing**, **missing value imputation** and **descriptive** and **inferential** analysis with **R**, **Python**, **Tableau prep**, **Tableau desktop and Power BI**.
* Developed Advanced Dashboards using advanced features available in Cognos Workspace 10.2.1 like tabbed-dashboards, Action Button Widget, Global Area, Freeze Column, Row Headings, Expand/Collapse Column/Row headings and Cognos Workspace Advanced for Accessing advanced functions.
* Performed **hypothesis testing**, collected evidences to prove **null hypothesis** wrong and came up with most optimal solutions based on statistical test results conducted using **Rstudio** and **R** Programming Language.
* Used Talend Admin Console Job conductor to schedule ETL Jobs on daily, weekly, monthly and yearly basis (Cron Trigger)
* Implemented Natural language processing techniques with R libraries - NLP, Tidyverse, stringR, Magrittr, RWeka and Wordcount.
* Used Kafka Streams to Configure Spark streaming to get information and then store it in HDFS.
* Designed Data Marts and extensively worked on performance tuning which includes Partitioning and other techniques to improve the query performance of reports
* Efficiently Worked with all of the following tools/Instances but not limited to including: Kafka, Zookeeper, Console Producer, Console Consumer, Kafka Tool, File Beat, Metric Beat, Elastic Search, Logstash, Kibana, Spring Tool Suite, Apache Tomcat Server etc.
* Extensively used different data sources for pulling data, cleansing, manipulation, feature engineering and machine learning model with **H2O AI driverless** tools.
* Assisting in building Docker container clusters managed by Kubernetes Linux, Bash, GIT, Docker, on GCP (Google Cloud Platform) . Utilized Kubernetes and Docker for the runtime environment of the CI / CD system to build, test deploy.

Environment: Python, Tableau, Machine Learning algorithms, Power BI, GitLab, Linux, Rstudio Connect, Rstudio, Airflow

**Deerwalk Inc, MA Sep 2016- Nov 2017**

**Machine Learning Engineer**

**Sr. Cloud/DevOps Engineer**

* Design and develop state-of-the-art deep-learning / machine-learning algorithms for analyzing the image and video data among others. Used **R Studio(R) and Spyder(Python)** for statistical analysis and building machine learning algorithms.
* Utilized **Spark, Scala, Hadoop, HBase, Kafka, Spark Streaming, Caffe, Tensor Flow, MLLib, Python**, a broad variety of **machine learning** methods including classifications, regressions, dimensionally reduction etc.
* Develop necessary connectors to plug **ML** software into wider data pipeline architectures.
* Worked on POC to check various cloud offerings including Google Cloud Platform (GCP).
* Compared Self hosted Hadoop with respect to GCPs Data Proc, and explored Big Table (managed HBase) use cases, performance evolution.
* Load and transform data into HDFS from large set of structured data /Oracle/Sql server using Talend Big data studio.
* Built predictive models including support **Vector Machine, Decision tree, Naive Bayes Classifier, Neural Network** plus ensemble methods of the models to evaluate the likelihood of Customer segments being open to additional purchase opportunities.
* Supported in full-scale system break-fix maintenance, troubleshooting and problem solving in UNIX (Lucent QIP, DNS, SSH, PHP, .NET, VMWare, Sun VirtualBox and DHCP design and implementation), Sun Solaris 10 and Red-Hat Linux Enterprise environments.
* Implemented statistical **modeling** with **XGBoost** machine learning software package using **Python** to determine the predicted probabilities of each **model**.
* Extracted data from **HDFS** using **Hive**, **Presto** and performed data analysis using **Spark** with **Scala**, **pySpark**, **Redshift,** and feature selection and created nonparametric models in **Spark**.
* Created and designed reports that will use gathered metrics to infer and draw logical conclusions of past and future behavior with cloud-based products like **Azure** ML Studio.
* Developed predictive models by using **Naïve Bayes**, **Random Forest**, **Support Vector Machine, Decision Tree** to evaluate how the likelihood to recommend of customer groups would change in different set of service by using **Python**, **Scikit-Learn, AWS SageMaker, and AWS Machine Learning**.
* Extracted data from **Oracle**, **SQL server**, **Hadoop** file system and created workflow for data science project which helped team to work faster and increased efficiency of overall work.
* Use SAS Functions to **manipulate character, Reshape** and numeric data values
* Performance analysis of Spark streaming and batch jobs by using Spark tuning parameters.
* Developed Active Reports for providing an interactive analytics experience in a self-contained Cognos Business Intelligence application for browsing and exploring data offline.
* Managed 4 Junior data scientist & analyst intern and evaluated their weekly based report and guided them how to solve complex data science problem.
* Developed Reports with Multiple Cognos Viewers and with multiple Cognos Reports Integrated onto a Single Page Report
* Wrote production level Machine Learning classification models and ensemble classification models from scratch using **Python** and **PySpark** to predict binary values for certain attributes in certain time frame.
* Migrated **SSRS**, **Google Analytics**, **Power BI Dashboards** and **Rmarkdown**, **Jupyter report**s into **Tableau** **Interactive Dashboard**.
* Successfully secured the Kafka cluster with Kerberos Implemented Kafka Security Features using SSL and without Kerberos. Further with more grain-fines Security I set up Kerberos to have users and groups this will enable more advanced security features.

Enivironment: Python, Oracle, SQL server, Hadoop, Tableau, Apache, HIVE, NoSQL, AWS S3, NLP, EC2, Scala, Cassandra, Git, scikit-learn, Power BI, GitLab, Rstudio, Rstudio connect

**Verisk Analytics, NJ May 2015- Aug 2016**

 **Senior Data Analyst**

**Sr. Cloud/DevOps Engineer**

* Responsible for data aggregation, data cleansing, missing value imputation and descriptive and inferential analysis to calculate the relevance of existing search engine.
* Implemented big data processing applications to collect, clean and normalization large volumes of open data using Hadoop ecosystems such as **PIG, HIVE, and HBase**.
* Built model on **Azure platform** using **Python** and **Spark** for the model development and Dash by **plotly** for visualizations.
* Clean data and processed third party spending data into maneuverable deliverables within specific format with Excel macros and python libraries such **as NumPy, SQLAlchemy and matplotlib**.
* Used **Spark and SparkSQL** for data integrations, manipulations. Worked on a POC for creating a docker image on azure to run the model.
* Built Data marts for Bankruptcy, Pre-Sales, Foreclosure, Delinquency, Loss Mitigation, Document Management, Mortgage Resolution Team subject areas using Kimball Methodology .
* Managed **tableau dashboard** in Tableau server, create user permission, set permission, optimized dashboard performance.
* Performed data cleaning and imputation of missing values and data validation using data pre-processing techniques in **R**
* Leveraging the different (Git, Jenkins and Maven) tools and provide the CICD pipelines for the different teams based on their build and release structure.
* Launched multi-node kubernetes cluster in Google Kubernetes Engine (GKE) and migrated the dockerized application from AWS to GCP.
* Communicated and presented default customers profiles along with reports using Python, Tableau, Power BI, analytical results and strategic implications to senior management for strategic decision making.
* Implemented big data processing applications to collect, clean and normalization large volumes of open data using Hadoop ecosystems such as **PIG, HIVE, and HBase**.
* Exported the data required information to **RDBMS** using **Hadoop Sqoop** to make the data available for the claims processing team to assist in processing a claim based on the data.

Environment: Python, R, Hadoop, PIG, HIVE, HBase, Tableau, Spark, SparkSQL, Azure platform, Power BI, GitLab, Rstudio

**Leapfrog Technology Inc, Nepal Nov 2014- April 2015**

**Data Analyst**

* Collaborated with data engineers and operation team to implement the ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
* Worked on InformaticaPowerCenter tool - Source Analyzer, Data Warehousing designer, Mapping &Mapp let Designer and Transformation Designer and Developed Informatica mappings and tuning of mappings for better performance.
* Used the dynamic SQL to perform some pre-and post-session task required while performing Extraction, transformation and loading.
* Wrangled data, worked on large datasets (acquired data and cleaned the data), analyzed trends by making visualizations using **python**
* Conducted analysis and patterns on customers' shopping habits in a different location, different categories and different months by using time series Modeling techniques.
* Loading data from various sources like **OLEDB**, flat files to **SQL Server 2012** database Using SSIS Packages and created data mappings to load the data from source to destination.
* Built **SSIS** packages, to fetch file from remote location like FTP and SFTP, decrypt it, transform it, mart it to data warehouse and provide proper error handling and alerting
* Modified existing **SAS/SQL** programs and created new programs using **SAS** macro variables to improve ease and speed of modification as well as consistency of results.
* Created monthly and quarterly business monitoring reports by writing complex SQL queries to include System Calendars, Inner Joins and Outer Joins to retrieve data from multiple tables.
* Extensively used calculated field in **Tableau** and Power BI with window function, table calculation, forecasting with trend seasonality and without trend, seasonality.

Environment: Python, Tableau, SAS, Informatica, SQL Developer, SQL Server 2012, Power BI, GitLab

**EDUCATION**

Master of Science in Information Technology– Valparaiso University 2020