**Shikhar Pratap Singh**[shikar.pratap88@gmail.com](mailto:shikar.pratap88@gmail.com)

Ph Number: (614) 362-0354

**Summary**

* Highly analytical and process-oriented data analyst with 8+ years of experience in data analysis and data management having proven ability to work efficiently in both independent and team work environments. Excellent team player, good communication and interpersonal skills with solid team leading capabilities.
* Expertise and experience in domains like **Healthcare, Banking, Insurance and e-commerce.**
* Proficient knowledge of the **SDLC** and extensive experience in **Agile (Scrum and XP**) and **Waterfall models.**
* Expertise in **Cost Benefit Analysis**, **Feasibility Analysis**, **Impact Analysis**, **Gap Analysis**, **SWOT** analysis and **ROI** analysis, **SCRUM**, leading **JAD** Sessions and Dashboard Reporting.
* Experience in data modeling, data analysis and working with **OLTP** and **OLAP** systems
* Worked with various **RDBMS** like **Oracle, MYSQL, SQL Server, DB2, Teradata** and expertise in creating tables, data population and data extraction from these databases.
* Expertise in **SQL Queries** to extract data from data tables along with creation of **tables, Sub queries, Joins, Views, Indexes, SQL Functions, Set Operators** and other Functionalities.
* Worked with **NoSQL** databases like **Apache Cassandra** to deal with stream processing/real time analysis regarding unstructured data and doing analytics on its data using **Pentaho ID.**
* Strong Experience in implementing Data warehouse solutions in **Amazon Redshift**, **Oracle** and **SQL Server**.
* Experience in extracting, transforming and loading (**ETL**) data from spreadsheets, database tables, flat files

and other sources using **Talend Open Studio** and **Informatica**.

* Having good knowledge in **Normalization** and **De-Normalization** techniques for optimum schema designing.
* Skilled in**Data profiling, Data Cleansing,Data mapping, creating workflows** and**Data Validation**using data integration tools like **Informatica** and **Talend Open Studio** during the **ETL** and **ELT** processes**.**
* Experience in Data warehousing concepts like **Star Schema** and **Snowflake Schema**, **DataMarts**,Kimball Methodology used in Relational and Multidimensional data modelling.
* Experience on **Apache Hadoop** Ecosystem with good knowledge of Apache Hadoop Distributed file system (HDFS), **Map Reduce, Hive, Pig, Python, HBase, Sqoop, Kafka, Flume, Cassandra, Oozie, Impala, Spark**.
* Experience with conceptual, logical and physical data modeling considering**Meta data** standards.
* Experience with DBA tasks involving **database creation**, **performance tuning**, creation of indexes,creating and modifying table spaces for optimization purposes.
* Knowledge of Machine Learning techniques like **Regression Models, Artificial Neural Networks, Clustering Analysis, Decision Tree, ANOVA, t-tests, Neural networks** and **SVM**.
* experience in Base **SAS/STAT, STATA, R, SQL, Tableau, Python, MS EXCEL (VLOOKUP, Pivot charts, Macros**).
* Expertise in Data Manipulations using **SAS/STATA** by using procedures like**SAS PROC SQL, PROC MEANS, PROC FREQ and PROC REPORTS, TABULATE, UNIVARIATE, Append, Array, DO loops, GPLOT** and **GCHART, Macros** and Merge procedures like **PROC APPEND, PROC DATASETS, PROC SORT**, and **PROC TRANSPOSE**.
* Expertise in creating **Tableau Dashboards** for data visualization and deploying it to the servers.

**Skills**

|  |  |
| --- | --- |
| **Analytical Techniques** | Predictive analysis, Machine Learning, Regression Modelling, Logistic Modelling, Time Series Analysis, Decision Tree, Neural Networks, Support Vector Machines(SVM). |
| **Analytical tool** | STATA, MEGASTAT, Rapid Data miner, Google analytics and IBM Watson, R &SAS |
| **DataVisualization Tool** | Tableau, Qlikview, Datawrapper, Microsoft PowerBI, Excel |
| **Data modeling** | Entity relationship Diagrams(ERD), Snowflake schema, Star schema |
| **Languages** | SQL, T-SQL, PL/SQL, HIVEQL, C, R, Python |
| **Database Systems** | SQL Server10.0/11.0/13.0, Oracle, MYSQL 5.1/5.6/5.7, Teradata, DB2, Amazon Redshift, |
| NOSQL Databases | HBASE, Apache Cassandra |
| **ETL Tools** | Microsoft SSIS, Pentaho ID, IBM Cognos, Talend Open Studio, Informatica Power House 9.0 |
| **Testing tool** | JIRA, HP Quality Check / HP ALM |
| **Big Data** | Apache Hadoop, HDFS, Sqoop, Flume, Kafka, Hive, Impala, MapReduce, Splunk ML-SPL, Splunk Hadoop Connect |
| **SDLC Methodology and Tools** | Waterfall, Agile / Scrum Methodology / XP, BIZAGI BPMN MODELER, SeeNowDoScrum. |

**Professional Experience**

**J.P Morgan Chase Columbus, OH**

**Senior Data Analyst Oct 2017 – Till date**

The project was pertaining to identify patterns which leads to gaining operational intelligence on MiFID violations of J.P Morgan Chase that provides personal and institutional wealth management solutions. Leveraged the enterprise Splunk platform and search processing language to develop insightful reports and visualization for proactive monitoring of MiFID regulations.

* Collaborated with product owner, key stakeholders and subject matter experts in brainstorming sessions to identify state of reporting structure, identify gaps, and shape the need for event correlation over log data
* Participated in as-is to-be analysis between the available **system information and event monitoring(SIEM)** enterprise tool and **Splunk enterprise** platform to prepare the business case
* Assisted the Product Owner in creating a Proof of Concept by analyzing months of syslog data in the proposed Splunk platform in terms of reporting, monitoring and alerts.
* Built models in machine learning tool kit provided in the Splunk platform to identify those crucial patterns that lead to **MiFID** violations.
* Expertise knowledge on various components within **Splunk (indexer, forwarder, search head, deployment server**), Heavy and Universal forwarder, Parsing, Indexing, Searching concepts, Hot, Warm, Cold, Frozen bucketing, License model.
* Created data models and lookup knowledge objects for day-to-day operational analysis
* Coached a team of data analyst on how **Splunk** can be leveraged for machine data analysis at scale
* Shell scripting and extensively used **Regular expressions (Regex)** in search string and data anonymization.
* Utilized the in-built **Search Processing Language (SPL)** to analyze massive amount of time-series events and identify event correlations
* Used **DB connect** to create lookups into operational **MySQL** database for key management day to day decisions.
* Assisted the Information Security team to build visualizations into operational **syslog’s** to identify security anomaly and detect outliers
* Collaborated with subject matter experts and project manager identified data inputsources, did input data profiling, and documented analysis undertaken
* Authored technical case documentation for **Splunk Hadoop Connect** provided as an input to thearchitectural runway setup for the project

**Environment**: Apache Hadoop Distribution 2.7.X, HDFS, Splunk Hadoop Connect, JIRASuite, MS Office Suite, Splunk ML-SPL (Machine Learning toolkit), MySQL

**DaVita Nashville, TN**

**Data Analyst / Dec 2016 - Sep 2017**

The project was to collect the data of the IOT devices of the Hospital infrastructure on the real time basis (stream data) and to analyze it through the Hadoop framework to significantly reduce the billing process of hospice & ambulatory services, administrative costs, effective logistic services and coordinated healthcare services.

**Roles and Responsibility:**

* Identified the **Edge node**(s), an **IOT gateway** or a cloud aggregator and a back-end **data analytics engine** operating on the aggregated data for trend analysis, anomaly detection, etc.
* Explored and classified all the IOT devices of the healthcare infrastructure and performed the **data profiling** through **Informatica** to get the better understanding of it.
* Used the **Kafka** Interfacing software of **Apache Hadoop framework** to get the data from all the **IOT** devices of the healthcare network into the **Apache Hadoop Spark system**
* Processed the stream data in the **Apache Spark streaming** by breaking the stream data into the **micro batches** and later processed by the spark’s core, which results in lower latency.
* Stored the processed data in the HDFS and generated the reports through **Spark’s SQL** queries.
* Performed data analytics on the output data obtained from the spark’s core by using **spark’s MyLib,** an advanced machine Learning tool of **Apache Spark**
* Performed the descriptive analysis on the data like **correlations and Scatter plots** to understand the current performance of the IOT devices of the healthcare and to improve the efficiency and optimizing the usage of it.
* Partitioned the data set into training, testing and validation set to use it in the supervised learning processes.
* Performed the predictive analysis through popular machine learning algorithms like **Linear regression, logistic regression and Artificial Neural networks**.
* Visualized the model performance through the **ROC curve (Receiver operating characteristic Curve)** by plotting **sensitivity** against **specificity** at different thresholds.
* Measured the predictive ability of a classifier by the **Area under the curve (AUC).** Area under the curve greater than 0.75 was the model acceptance criteria.
* Performed the **data reporting** and created the **dashboard** and shared with all the major stake holders
* Performed the statistical analysis to improve the system processes to give better and quality healthcare. Performed the process modelling through **BIZAGI BPMN MODELER** business process modeler to remove the inefficient tasks and processes in the systems
* Experience in documenting all the analytical results and findings of the **Apache Sparks** by **Apache Zeppelin.**
* Ensured the proper process was followed to demonstrate to the monitoring government entity that the data provided to them had gone through a stringent **data governance** process.

**Environment:** Apache Hadoop, HBASE, Apache Spark, Kafka, Apache Zeppelin, Informatica, BIZAGI BPMN MODELER, Spark’s Mylib, Tableau, R, SAS/STAT, Spark’s SQL, Predictive analysis, Machine Learning, MS office suite

**Gilead Sciences Foster city, CA  
Senior Data Analyst / July 2015 - Dec 2016**

The project was about the sentiment analysis of the public opinions shared at twitter and Facebook by the company to synchronize their research, manufacturing and marketing of their products with the public opinionand sentiments of the people to increase their customer base (Sources: Twitter and Facebook).

**Roles and Responsibility:**

* Worked on the **Facebook developers** and **twitter developers** to extractthe user tweets and comments through their **API’s** and use it for the analytical purposes.
* By using the **R** platform, created R scripts and merging it with the Twitter’s and Facebook **API’s** to make the connections for the data extraction part.
* Made the **Dataframes** of the extracted data through call function in **R**, to get the data in the tabular form.
* Performed the **data cleansing** through the **R** functions **sapply** and **gsub** by removing the **emoticons** and **URL’s**. By doing this, processes were streamlined, and data quality has been improved steadily over time.
* Performed the **lexical Analysis** on the cleaned Twitter tweets and FB comments to analyze the sentiment of the tweets and comments and to convert it into the numerical values.
* Performed the **scan analysis** on each tweet and comments to find the number of positive and negative words through a **Scan** function through its negative and positive words repository.
* Quantification of the sentiments in terms of **Positive Score, Negative score** and **overall score**.
* Performed the descriptive analysis on the data **correlations, Scatter plots** and **measures of central tendency.**
* Visualized the results of the Lexical analysis over the **Tableau by making the Histograms, Bubble charts, Pie charts** and creating the **Dashboard for the reporting purposes.**
* Assisted the departments by giving data driven solutions, based on lexical analysis findings in taking key decisions regarding manufacturing and marketing of key products.
* Optimized data collection procedures and proposed solutions to improve system efficiencies and reduce total expenses through **BIZAGI BPMN MODELER** tool.
* Ensured the proper process was followed to demonstrate to the monitoring government entity that the data provided to them had gone through a stringent data governance process.

**Environment:** R, Lexical Analysis, Tableau, MS office suite, Microsoft PowerBI, R, SAS/STAT, SQL, MYSQL.

**West Hills Hospital Reno, NV  
Data Analyst/ Nov 2014- June 2015**

As a part of increasing the quality of care and efficiency of preventive care on the guidelines of Center for Medicare and Medicaid services (CMS), hospital management launched a project to address the issue of Readmission rates.

**Roles and Responsibility:**

* Explored, identified, and aggregated all the meaningful data sources of the healthcare group and performed the data profiling through **Informatica**.
* Performed the extraction from **OLAP Cube** into the **Python** environment.
* By using the **Pycharm** platform, and using the **ODBC** drivers, created the connections between **OLAP** and **PYcharm.**
* Making the **Data frames** of the extracted data through the **Pandas** library.
* Performed exploratory data analysis (EDA) to analyze the data sets to summarize their main characteristics in terms of Box plot, Histogram, Multi-vari chart, Run chart, Pareto chart, Scatter plot, and Odds ratio.
* Analyze, understand, interpret and explain complex medical and pharmacy trends through **Python functions.**
* Performed the **logistic** and **multiple linear regression algorithms** to find the dependency rate of readmissions in the hospitals over the other dependent variables.
* Performed the **Anova test**, individual **t-tests** and **F-test** to explore the significance of variables which are affecting the readmissions in the Hospitals and not useful for the model.
* Provided recommendations for cost and performance improvement to internal management and clients through the documentation and data reporting by **Tableau.**
* Visualizing the results of the analysis over the **Tableau** by making the **Histograms, pie charts, box plots and bubble charts.**
* And other chartsand creating the **Dashboard** and deployed it on the servers.
* Trusted with management of confidential professional and personal information / **HIPAA** Compliance.

**Environment:** OLAP, Pycharm, Tableau, Excel, Microsoft PowerBI, Python, R, SAS/STAT, SQL, MYSQL, Predictive analysis, Neural Networks.

**Bajaj Allianz Bangalore, India**

**Data Analyst / Dec 2013- Oct 2014**

The project was to avoid the churns of the customers of the organization by doing the predictive analysis with the help of machine learning algorithms, so that it can providing insight in churn behaviorand help the company in optimizing their services for customer attraction, customer development and customer retention.

**Roles and Responsibility:**

* Collaborated with the different stakeholders and identified the potential variables from the broad categories like Demographic data, Policy-related data, Claims and Complaints related variables.
* Explored customer behavior through the data available on the social media & networking sites, website logs, blog posts, surveys etc., to get the 360o degree view.
* Identified and addressed the key issues that organization was facing from the customer side in terms of lack of feedback, sudden inactivity and friction while accessing the services.
* Performed the data profiling using **k-means clustering** to make the clusters of the population and checked the anomalies in it and did the cleansing of the data.
* Balanced the dataset through under and over sampling techniques, by using the **R** platform.
* Performed the Comparative analysis of churningand non-churning profiles to generalize the model by using the **Hypothesis testing.**
* Performed the **predictive analysis** of the extracted data by using **machine learning algorithms** like **Regression analysis, Support vector machines, decision tree** and **neural networks**to predict the churn**.**
* Partitioned the data set into training and testing sets and executed on each model.
* Measured the performance of the models through a **confusion matrix, classification accuracy, sensitivity, specificity, precision, AUC, ROC, AUK** and used it for the best model selection.
* Performed the **cost benefit analysis (CBA)** to investigate the models and to identify the minimum percentage of churners to be contacted for the profitability.
* Visualized the predictive analysis results over the **tableau** and created the **Dashboard** of key indicators in the customer retention analysis and deployed it on the servers.
* Documented the findings and provided data driven recommendations to the decision makers.

**Environment:** R platform**,** Tableau, Excel, SAS/STAT, SQL, MYSQL, Machine Learning, Time Series Analysis, Neural Networks.

**Kotak Mahindra bank Pune, India**

**Data Analyst / July 2012- Nov 2013**

The project was to design and develop an accurate, useful and stable credit risk model to assess the chances of a customer’s default on payment based on rating an applicant’s characteristics and their attributes accessed from bank records and social media information.

**Roles and Responsibility:**

* Defined the data sources and the methodology for the project and collaborated with key stakeholders.
* Performed the data integration through **Talend Open StudioETL** tool on data coming from both internal and external sources through data merging, concatenation and manipulation of relational tables.
* **Data segmentation** through **Decision tree**on **SAS/STAT** to better understand the riskiness of borrowers.
* Developed a **Probability of Default (PD)** Model through **Logistic regression** algorithm to estimate the probability of a loan whether it will get repaid or it fall into default.
* Developed a **Loss given Default(LGD)** Model through **Artificial Neural Networks (ANN)** algorithm to estimate the economic loss incurred if an obligor goes into default, expressed as a percentage of exposure.
* Created **behavioral scorecards** and **application scorecards** to know the applicant in a better way.
* **Data cleaning** through **KNN imputation and Multiple imputation** to remove outliers and missing values.
* Performed the variable selection through **stepwise selection** method to remove irrelevant and redundant variables from the data set, with the aim of improving the performance of regression techniques.
* Developing the credit scorecard model by using **logistic regression, model validation,** and **scaling processes.**
* Assessed the **model accuracy** by complexity, Mean Absolute Error, RMS Error, Confusion Matrix, etc.
* Visualized the model performance through the **ROC curve** by plotting **sensitivity** against **specificity** at different thresholds and Measured the predictive ability of a classifier by the **Area under the curve (AUC).**
* Performed **Segmentation** and **Reject Interference** to eliminate the selection bias.
* Addressing the problem of overfitting through **Bootstrapping** and cross-validation frameworks.
* Visualizing the results of the credit score modelling over the **Tableau** and creating the **Dashboard.**

**Environment:** Talend Open Studio, Tableau, R, SAS/STAT, SQL, MYSQL, Machine Learning, Time Series Analysis, Neutral Networks, MS Office 2010

**R. K. Singhal & Company Private Limited New Delhi, India**

**Data Engineer / Dec 2010- Jun 2012**

The project was toimplementthe Oracle Datawarehouse in the organization to add the value to its business operations in the field of Taxation, Business consultancy and new ventures, Auditing, Company law matters and Accounting.

* Created connections to access data from relational databases and other data sources. Created the connection to create data objects, preview data, run profiles, and run mappings.
* Imported metadata to create data objects for sources and targets for the mapping. Used data objects to define the input and output of the mapping.
* Executed the profile to analyze the structure and content and quality of the data, and to determine the quality of your data. Informatica applies the profiling rules and runs the profile through **Informatica.**
* Developed mappings to implement data integration tasks through **Informatica.** Linked the sources and targets with transformation objects that define the rules for **data transformation**.
* Created a workflow to define a sequence of events, tasks, and decisions based on a business process architected through **BIZAGI BPMN MODELER** tool.
* Deployed the workflow to the **Informatica** and run the workflow.
* Experience in monitoring the workflow instance run on the Monitoring tab of the **Informatica**Administrator tool.
* Extensive working knowledge with different types of data load strategies and scenarios like **Historical Dimensions, Surrogate keys, Summary facts** etc.,
* Experience in visualization of data with the help of data visualization tool **Excel and Tableau**.
* Understand the business requirements and identified gaps in different processes and implemented process improvement initiatives across the business improvement model.
* Maintained and updated all data archives and Conducted periodic internal audit.

**Environment:** BIZAGI BPMN MODELER, Tableau, Informatica, Windows XP/NT/2000, SQL Server 2005/2008, SQL, MYSQL, Microsoft Visio 2009, MS Office 2010, MS Access 2010.

**Acme Telepower Limited Gurgaon, India**

**Data Support Engineer / June 2009-Dec 2010**

I worked on the ERP implementation project in the organization to integrate day to day activities of the organizational units like manufacturing unit, R&D and sales unit to reduce the redundancy, cost and integratingthe business operations.

* Provided Data analytics support for high-priority decisions, customer service improvement and organizational realignment through Data reporting and analysis on Excel.
* Worked on **RAMCO Systems**Enterprise Resource Planning Systems (**ERP**)to generate and manage the Management information systems reports (**MIS**).
* Developed database **objects**, including **tables** and **views** to normalize our data and to secure its integrity and **materialized views** using SQL queries on **MYSQL** database.
* Developed SQL Queries with multiple table joins, functions, subqueries, set operations and T-SQL stored procedures and user defined functions for data analysis.
* Coordinated statistical data analysis, design, and information flow.
* Analyzed and interpreted consumer behavior, market opportunities and conditions, marketing results, trends, and investment levels through **Content** and **Collaborative** analysis to do the **market segmentation**.
* Work directly with internal users on implementing **ERP** system to meet business needs.
* Designed data model based on the requirement, interacted with end users to understand the business logics.

**Environment:** Ramco Systems ERP, Windows XP/NT/2000, SQL Server 2005/2008, SQL, MYSQL, Microsoft Visio, MS Office 2010, MS Access 2010.

**Education**

**Master of Science, University of Nevada, Reno, US.**

* Major: Information Systems; GPA: 3.7/4
* Worked as Graduate Teaching Assistant in 'Mathematics & Statistics department' and 'MSIS’ departments'.

**Bachelor of Technology,BIET, Jhansi, India.**

* Major: Electronics and communication Engineering; GPA: 3/4
* Internshipinwirelesscommunication (GPRS)fromBharatSancharNigamlimitedin2005 and 2006.