**Rashi Arora**

**Data Scientist**

**Email:** [rashi297376@gmail.com](mailto:rashi297376@gmail.com)

**Mobile : 323-546-2729**

**Visa: H4EAD**

**Lnkedin:** <https://www.linkedin.com/in/rashi-a-59355517a/>

**Professional Summary:**

* Over **6 years** of Data Science and analytics experience in Various domains.
* Extensive IT experience with in-depth understanding of Business Re-engineering, Business Continuity plan, System Architecture, Problem Solving, Strategic Partnering, Strategic Planning, Implementation, Application Design, Information Security, Data Warehousing, Business Intelligence and Project Management.
* Experienced Data Scientist with hands on experience in developing Machine Learning models.
* Extensive experience in Strategic development of a Data Warehouse and in Performing Data Analysis and Data Mappings for operational Data Stores, Enterprise Data Warehouse , Data Marts, other databases.
* Expertise in conceptual, logical and physical designing and development of Data Models, Data Normalization, OLTP and OLAP Database and Data Warehouse.
* Experience in diverse set of procedures but not limited to methods such as machine learning, deep learning, Bayesian algorithm, Regressions, cluster analysis, decision trees, time series, resampling and regularization, NLP and other techniques.
* Strong working experience with various python libraries such as NumPy, SciPy for mathematical calculations, Pandas for data preprocessing/wrangling, Matplotlib, Seaborn for data visualization, Sklearn for machine learning, Theano, TensorFlow, Keras for Deep leaning and NLTK for NLP.
* Experience with common Machine Learning Algorithms like Regression, Classification and Ensemble models.
* Experience using machine learning models such as random forest, KNN, SVM, logistic regressions and used packages such as ggplot, dpylr, lm, e1071, rpart, randomForest, nnet, tree, PROC-(pca, dtree, corr, princomp, gplot, logistic, cluster), numpy, sci-kit learn, pandas, etc., in Python.
* Expertise in applying data mining techniques and optimization techniques and proficient in Machine Learning, Data/Text Mining, Statistical Analysis and Predictive Modeling.
* Expertise in SQL in creating and using Views, User Defined Functions, Triggers, Indexes, Stored procedures involving joins, sub-queries from multiple tables and established relationships between the tables using primary and foreign key constraints.
* Expert in ETL tools like Import/Export wizard and SSIS for transferring Data from different sources to a centralized SQL server.
* Hands on experience in Implementation of Enterprise Business Intelligence solutions using BI tools such as Tableau, Business Objects and Crystal Reports.
* Excellent in High Level Designing of SSIS Packages to perform ETL from heterogeneous sources such as Excel, CSV, Text Format Data by using multiple transformations provided by SSIS to generate underlying data for the reports and to export cleansed data.
* Experience with Data analysis and SQL querying with database integrations and data warehousing for large financial organizations.
* Strong knowledge of SQL, stored procedures, data analysis and data validation using SQL.
* Strong background in  Data Warehousing systems proficient on designing and implementing SQL queries for reporting / data validation using Oracle, SQL Server and DB2.
* Experienced in designing reports using SQL Server Reporting Services(SSRS) and Excel Pivot table based on OLAP cubes.
* Extensive experience in creating and maintaining source to target data mapping documents.
* Documented Traceability Matrices based on defined business rules in Microsoft Excel.
* Strong documentation and analytical skills, strong problem-solving skills.
* Facilitated weekly and monthly business review meetings to keep SME’s, Stakeholders, product managers, executive staff and team members apprised of goals, project status, resolving issues and conflicts and preparing product demonstrations.
* Solid business communication, and presentation skills, written correspondence, web-based products, reports,(project status reports, PowerPoint presentations, e-mails ).
* Organized, goal-oriented, self-starter, and ability to master new technologies, manage multiple tasks while following through from start to completion with limited supervision.
* Competent, confident and strategic decision maker organized with effective Time Management and Prioritizing skills in conflicting demands to execute business objectives.

**Technical Skills:**

|  |  |
| --- | --- |
| **SDLC** | Agile (Scrum) |
| **BI Tools** | Tableau, MicroStrategy |
| **Data science** | Experiement design, A/B testing, Hypothesis testing, Supervised & Unsupervised Learning, Statistical Inference |
| **Programming languages** | Python, R |
| **MS Office** | Excel (pivot tables, v-lookups), Access, PowerPoint, Visio |
| **Database** | SQL Queries (SQL Server, MySQL, Db2) |
| **ETL Tool** | Informatica |
| **Task Management Tools** | Jira, Rally |
| **Others** | User Stories, Business Requirements, Functional Requirements, Use Cases, Activity Diagrams, Sequence Diagrams, Data Flow Diagrams |

**Professional Summary:**

**Client: Bayer Crop Science, Chesterfield, MO Apr 2019 – Till Date**

**Role: Data Scientist**

**Responsibilities:**

* Designed, prototyped and implemented “Incomplete environment testing design strategies” to lower cost by improving prediction error for yield estimates compared to randomized testing.
* Assisted in researching experimental design ideas using location and genetic clusterings driven by genotype andenvironmental interactions (GxE).
* Implemented data pipelines to gather data from various sources, clean, validated and prepared final dataset.
* Gathered millions of observations for daily weather data from aWHERE API using requests and pandas python packages.
* Performed timeseries forecasting (​**ARIMA**) ​to predict 2020 weather data while paying attention to domain requirements such as crop growth stage using ​**xts** in **R.**
* Performed EDA on multiple large datasets including genetic data, soil data, weather data and other environmental data using ​**dplyr​, ​sqldf**​ and other packages in **R.**
* Incorporated data management techniques to organize these datasets with millions of observations, hundreds of features that spanned tens of GBs using **doParallel**, **datatables** in **R**.
* Implemented machine learning algorithms like ​**random forest​ regressors, ​GBDTs**​ to compute feature importance, predict stability of GxE and perform feature selection. Used tools such as **DataRobot**​ (Auto ML framework)​ and ​**XGBoost**​ in python.
* Implemented and compared various weighted and unweighted clustering algorithms to compute environmental and genetic clusters.
* Applied generalized linear mixed model to produce incomplete environmental design for optimal testing and interpolated same information for missing blocks using recommender system such as​ item based collaborative filtering ​using ​**scikit-learn**​ in python and ​**lmer**​ in R.
* Researched and Implemented several validation and quality metrics for ​regression models​ and cluster computations. Used **pandas, numpy, seaborn, scipy, matplotlib, scikit-learn, NLTK** in Python for developing various machine learning algorithms.
* Coordinated with several teams and provided feedback on data quality and steps for remediation.
* Presented my work in several forums and to upper management.
* Improved development environment (dominos) by exploring new features.
* Data management on ​**AWS S3**​ and analysis using ​**spark-sql** on **AWS-EMR.**
* Built a mixed model and estimated the percentage of data points to produce a minimum number of locations to harvest for making an advancement using ​**lmer** in R​.
* Improved scalability by introducing parallel execution, cut down result generation time by **50x** using **doParallel** and multiple compute nodes in **AWS** **dominos.**
* Improved stability of the codebase and fixed several bugs.
* Generated reports and incorporated feedback to improve report generation mechanisms.

**Client:** **Focus1 Insurance, Portland, OR Mar 2018 – Apr 2019**

**Role: Data Scientist**

**Responsibilities:**

* Managed multiple projects of small to large scale through sprints.
* Worked closely with stakeholders and subject matter experts to elicit and gather business requirements.
* Documented business and functional requirements, user stories, acceptance criteria, test cases, and wireframes/mockups.
* Worked directly with the Dev team to ensure product backlog was understood to the level needed and to keep sprint on track.
* Tracked Velocity, Capacity, Burn Down Charts, and other metrics during iterations.
* Created Data flow diagrams.
* Prioritized the backlog to ensure on time delivery.
* Facilitated Agile team ceremonies including Daily Standup, Backlog Grooming, Sprint Review, Sprint Planning etc...
* Developed a machine learning system that predicted purchase probability at a particular offer based on customer’s real time location data and past purchase behavior.
* Used Pandas, NumPy, seaborn, SciPy, Matplotlib, Scikit-learn, NLTK in Python for developing various machine learning algorithms and utilized machine learning algorithms such as linear regression, multivariate regression, naive Bayes, Random Forests, K-means, & KNN for data analysis.
* Assisted Data Scientist team to Gathered the Requirements, Develop Process Model and detailed Business Policies and modified the business requirement document.
* Analyzing Data Processing implementing Data Mart applications mainly transformation processes using ETL tool and understanding of ETL workflows and ETL mappings and monitoring the Production Environment
* Performed legacy application data cleansing, data anomaly resolution and developed cleansing rule sets for ongoing cleansing and data synchronization.
* Involved in project cycle plan for the data warehouse, source data analysis, data extraction process, transformation and loading strategy designing.
* Using Python, I have automated a process to extract data and various document types from a website, save the documents to specified file path, and upload documents into an excel template.
* Performed data analysis and data profiling using SQL on various sources systems including Oracle and Teradata.
* Extensively used Star Schema methodologies in building and designing the logical data model into Dimensional Models.
* Maintain and enhance data model with changes and furnish with definitions, notes, reference values and check lists.
* Utilized MicroStrategy and Tableau (BI) reporting to create, test and validate various visualization, reports (ad-hoc), dashboards, and KPI’s.
* Operated the MicroStrategy’s command manager essential in providing descriptions for key performance indicators (KPIs).
* Ran SQL queries to test and validate data in database (SQL Server, DB2).
* Developed SQL queries /scripts to validate the data such as checking duplicates, null values, truncated values and ensuring correct data aggregations within an ETL process testing cycle.
* Involved with data model and reporting requirements for Tableau Reports with the Data warehouse / ETL and Reporting teams.
* Worked with Users to develop Test cases for user acceptance testing.
* Drove User Acceptance Testing.
* Designed and published visually rich and intuitively interactive Tableau/Excel workbooks and dashboards for executive decision making.

**Client:** **Thrive Market, Los Angeles, CA Jan 2017 – Feb 2018**

**Role: Data Scientist**

**Responsibilities:**

* Serves as the primary liaison between client and technical teams to ensure successful, on-time project delivery using the appropriate methodologies.
* Analyzed and compared materials and their documentation to ensure they are a match and will work for a specific product.
* Used Pandas, NumPy, seaborn, SciPy, Matplotlib, Scikit-learn, NLTK in Python for developing various machine learning algorithms and utilized machine learning algorithms such as linear regression, multivariate regression, naive Bayes, Random Forests, K-means, & KNN for data analysis.
* Analyzed large data sets apply machine learning techniques and develop predictive models, statistical models and developing and enhancing statistical models by leveraging best-in-class modeling techniques.
* Utilized Scala, Hadoop, HBase, Kafka, Spark Streaming, MLlib, Python, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
* Hands on expereince with Hadoop, Hbase, Spark deployed on AWS EMR.
* Prepared, gathered and submitted regulatory documents to various agencies to achieve certifications for specific projects.
* Evaluated data received from contract research organizations.
* Used Excel spreadsheets to manipulate numerical data and formulas.
* Organized and filed received documentation physically and electronically.
* Resolved the data related issues such as assessing data quality, data consolidation, evaluated existing data sources.
* Worked with renowned statistician using Python and Excel to analysis marketing and consumption data.
* Created and reviewed mapping documents based on data requirements.
* Assisted in creating fact and dimension table implementation in Star Schema model based on requirements.
* Created Analysis Model and Data Model and used these artifacts to develop the Design Model: Class Diagrams, Collaboration, and Sequence Diagrams.
* Identify source systems, their connectivity, related tables and fields and ensure data suitably for mapping.
* Developed normalized Logical and Physical database models to design OLTP system for finance applications. Created dimensional model for reporting system by identifying required dimensions and facts using Erwin Data Modeler
* Performed Data Analysis and extensive Data validation by writing several complex SQL queries.
* The reporting was created with SSRS and utilized the cube and MDX to create datasets to perform reporting and included interactive drill down reports, report models and dashboard reports.
* Analyzed and created business intelligence reports using SSRS based on the business requirements.

**Client: Big Savings Insurance Inc, San Francisco, CA Jul 2015 – Dec 2016**

**Role: Data Analyst**

**Responsibilities:**

* Work with users to identify the most appropriate source of record and profile the data required for sales and service.
* Document the complete process flow to describe program development, logic, testing, and implementation, application integration, coding.
* Involved in defining the business/transformation rules applied for sales and service data.
* Define the list codes and code conversions between the source systems and the data mart.
* Worked with internal architects and, assisting in the development of current and target state data architectures
* Worked with project team representatives to ensure that logical and physical ER/Studio data models were developed in line with corporate standards and guidelines
* Involved in defining the source to target data mappings, business rules, business and  data definitions
* Responsible for defining the key identifiers for each mapping/interface
* Responsible for defining the functional requirement documents for each source to target interface.
* Document, clarify, and communicate requests for change requests with the requestor and coordinate with the development and testing team.
* Involved in configuration management in the process of creating and maintaining an up-to-date record of all the components of the development efforts in coding and designing schemas
* Developed the financing reporting requirements by analyzing the existing business objects reports
* Interact with computer systems end-users and project business sponsors to determine, document, and obtain signoff on business requirements.
* Responsible in maintaining the 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.
* Evaluated data profiling, cleansing, integration and extraction tools(e.g. Informatica)
* Coordinate with the business users in providing appropriate, effective and efficient way to design the new reporting needs based on the user with the existing functionality
* Remain knowledgeable in all areas of business operations in order to identify systems needs and requirements.

**Company: SOPRA GROUP, Noida, India Jun 2012 – Sep 2014**

**Data Analyst**

**Responsibilities:**

* Involved in extensive DATA validation using SQL queries and back-end testing
* Used SQL for Querying the database in UNIX environment
* Involved with Design and Development team to implement the requirements.
* Defects tracking, review, analyze and compare results using Quality Center.
* Defined the Scope for System and Integration Testing
* Prepares and submit the summarized audit reports and taking corrective actions
* Involved in Uploading Master and Transactional data from flat files and preparation of Test cases, Sub System Testing.
* Document and publish test results, troubleshoot and escalate issues
* Preparation of various test documents for ETL process in Quality Center.
* Involved in Test Scheduling and milestones with the dependencies
* Functionality testing of email notification in ETL job failures, abort or data issue problems.
* Identify, assess and intimate potential risks associated to testing scope, quality of the product and schedule
* Created and executed test cases for ETL jobs to upload master data to repository.
* Responsible to understand and train others on the enhancements or new features developed

**EDUCATION SUMMARY:**

**University of Southern California**, M.S. in Biostatistics, GPA: 3.80

Relevant Courses: Machine Learning, Spatial Statistics, Statistical Computing, Data Analysis, Experimental Design

**RGPV**, Bhopal, M. Tech. (Hons.), Biotechnology, GPA: 3.60

Relevant Courses: Engineering Mathematics, Comprehensive Analysis

**Amity University**, Noida, B. Tech., Bioinformatics, GPA: 3.75

Relevant Courses: Data mining and warehousing, Statistics, Engineering Mathematics, Introduction to algorithms