**Kusum Singh** [**www.linkedin.com/in/kusum-singh-8997b350/**](http://www.linkedin.com/in/kusum-singh-8997b350/)

**Data Scientist/Data Analyst** <https://learn.microsoft.com/en-us/users/kusumsingh-4060/>

**Email:** kusumsingh209@gmail.com<https://public.tableau.com/app/profile/kusum.singh8561> **Phone:** (339) 205-4770 <https://trailblazer.me/id/kusumsingh209>

[**https://github.com/kusum1984**](https://github.com/kusum1984)

**PROFESSIONAL SUMMARY**

* Highly experienced Data Scientist with over 12 years of experience in **Data Extraction, Data Modelling, Data Wrangling, Statistical Modeling**, **Data Mining, Machine Learning** and **Data Visualization**.
* Domain knowledgeand experience in**Retail, Banking** and **manufacturing**industries.
* Expertise in transforming business resources and requirements into **manageable data formats** and **analytical models**, **designing algorithms**, **building models**, and **developing data mining**and **reporting solutions** that scale across a massive volume of structured and unstructured data.
* Proficient in managing the entire data science project life cycle and actively involved in all the phases of the project life cycle including **data acquisition, data cleaning, data engineering, features scaling, features engineering, statistical modeling**, **testing and validation** and **data visualization.**
* Proficient in **Machine Learning algorithms** and **Predictive Modeling** including **Regression Models**, **Decision Trees**, **Random Forests**, **Sentiment Analysis, Naïve Bayes Classifier**, **SVM**, and **Ensemble Models**.
* Proficient in **Statistical Methodologies** including **Hypothetical Testing**, **ANOVA**, **Time Series**, **Principal Component Analysis**, **Factor Analysis**, **Cluster Analysis**, and **Discriminant Analysis**.
* Knowledge of **time series analysis**using Univariate (ARIMA/SARIMA)/Multivariate (VARMA/SVARMA)
* Knowledge of **Natural Language Processing (NLP)** algorithms and **Text Mining**.
* Worked in **large-scale database** environments like **Hadoop** and **MapReduce**, with the working **mechanism** of Hadoop clusters, nodes, and **Hadoop Distributed File System (HDFS).**
* Strong experience with **Python** to develop **analytic models** and **solutions**.
* Proficient in **Python** with **SciPy Stack** package including **NumPy**, **Pandas**, **SciPy**, **Matplotlib and IPython**.
* Very good experience and knowledge in provisioning virtual clusters under the **AWS cloud** which includes services like **EC2, S3, Sagemaker** and **EMR.**
* Proficient in data visualization tools such as **Tableau**, **Python Matplotlib**, and **R Shiny**to create visually powerful and actionable interactive reports and dashboards.
* Excellent **Tableau Developer**, with expertise in building, and publishing customized interactive **reports** and **dashboards** with customized parameters and user filters using **Tableau**.
* Experienced in **Master Data Management (MDM)**
* Experienced in **pyspark.**
* Experience in working with **Azure** cloud platforms **(HDInsight, DataLake, Databricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH and Data Storage Explorer)**.
* Experienced with direct marketing such as email or direct mail.
* Strong business sense and ability to communicate data insights to both technical and non-technical clients.

**EDUCATION**

**Saint Peters University, New Jersey, U.S.A 2020-2022**

Master of Science, Data Science GPA 3.95/4.00

**Sikkim Manipal University, Gangtok, India 2010-2012**

Master of Business Administration (MBA) Grade A

**TECHNICAL SKILLS**

**Skills:**  Natural Language Processing, Classification, Regression, Clustering, Hypothesis testing, Time-series modeling, Deep learning, forecasting, Data modeling.

**Programming Languages:**  Oracle PL/SQL, My-SQL

**Scripting Languages:**  Python (NumPy, SciPy, Pandas, Keras), R (ggplot, Caret, Weka)

**Machine Learning:**  Logistics Regression, Naive Bayes, Decision Tree, Random Forest, KNN,

Linear Regression, Lasso, Ridge, SVM, Regression Tree, K-means, Apriori

**Libraries:**  Scikit-learns, Keras, TensorFlow, NumPy, Pandas, seaborn,

Matplotlib, ggplot2, pyspark.

**DW-BI tools:**  Tableau, ggplot2, RShiny, Microsoft Power BI, QlikView, Dash Plotly

**Databases/Tools:**  MySQL, Oracle, PostgreSQL, MS SQL Server, Alteryx, Google Big Query, Excel

**Clouds:**  Google Cloud Platform, AWS, Azure(databrick)

**Professional Experience**

## **Data Scientist and Analytics**

## **Saint Peter’s University, New Jersey**

## **(Qtech US inc) Dec 2022-Present**

**Responsibilities**

* Leverage Apache Spark (Pyspark) for large-scale batch processing.
* Experienced in building models on generative AI for text auto-generation.
* Work closely with technology stakeholders to understand their requirements.
* Collaborate with architects, engineers, information analysts, and business stakeholders.
* Conduct regular reviews of existing architectures for **optimization** opportunities.
* Collaborate with the platform, APIM, Azure DevOps, and support teams.
* Experienced in LLM for enhancing customer experiences in Teradata.
* Spearheaded the development and implementation of predictive models to forecast.
* Conducted exploratory data analysis (EDA) to identify key variables and patterns influencing predictive outcomes.
* Experienced in machine learning algorithms for predictive analytics.

**DiD agency group LLC (Part of Lucid Group)**

**Pennsylvania, USA** **March 2022 –Dec 2022**

**Data Scientist**

**Responsibilities:**

* Experienced with impala and hive.
* Performed **data exploration** and analysis of Sales and marketing data with various BI tools.
* Experience working in **AWS** environment using **S3, Lambda, IAM, Glacier, EC2 and Experienced in Google Cloud Platform (GCP) such as cloud storage and database, cloud AI, cloud IAM, compute engine, billing, SDK, cloud shell, Google big query** etc.
* Revitalized the use and promoted the importance of data-derived insights to support strategic planning and tactical decisions. Created and socialized the use of key metrics/**KPI**s.
* **Data modeling, data wrangling and forecasting with Python and R Markdown**
* Application of various **Artificial Intelligence (AI)/machine learning** algorithms and statistical modeling 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 sci-kit-learn package in python.
* Experience with time series forecasting – Univariate (ARIMA/SARIMA)/Multivariate (VARMA/SVARMA)
* Developed complex **SQL** queries as to retrieve dimension counts from Olap query log to retrieve dimension usage and queries to retrieve measure count usage from **Google Big query** and **Snowflake**.
* Perform validation on machine learning output from **R Markdown**
* Build reports and dashboards with **R-shiny**.
* **Sagemaker**
* **Experience on Alteryx for data cleaning, data modeling and workflow.**
* **Used Alteryx/python for Predictive modeling such as Logistic Regression and time series analysis.**

**Projects worked on**

1) Xolair (Peanut Allergy)

2) CeraVe (Quarterly sales analysis and comparison with other competitors, Walgreens, CVS etc)

3) Symtuza (HIV-1)

**Digitas Health,**

**New Jersey, Sep 2021 - Feb2022**

**Data Scientist (G.R.A)**

**Responsibilities:**

* Cultural Health Moments: A Search Analysis During Times of Heightened Awareness to Identify Potential Interception Points with Digital Health Consumers
* Analysis of Digital Search Behaviors of Digital Health Consumers to Identify Trends in Highly Personalized Health Search.
* Used **Python**, **tableau**, **Power BI**
* **Used R for designing statistical modeling.**
* Created **Pipelines** in **Azure Data Factory (ADF)** using **Linked Services**, **Datasets**, and **Pipeline** to extract, transform and load data from different sources like **Azure SQL**, **Blob storage**, and **Azure SQL** **DW**.
* Created Automated ETL jobs in Talend and pushed the data to Azure SQL data warehouse.
* Used Azure Synapse to manage processing workloads and served data for BI and prediction needs.
* Experience in developing MLOPS processes/pipelines and in ML Model deployment.

**Kmart Corporation**,

**New Jersey, USA** **Sep 2020-May 2021**

**Data Scientist (G.R.A)**

**Responsibilities:**

* To create a working **prototype** of a visual search tool that can eventually be integrated into the Kmart website to improve customers' experience.
* Developed a reverse image search engine of more than 200k images, applying NLP techniques for product recommendation. Work involves the application of Tools: **Python: TensorFlow, Keras, PyTorch, Tableau, R Platform: Google Cloud, AWS, Lambda, Sagemaker S3, Redshift**
* Collaborated with data engineers and operation team to implement ETL process, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
* User **R** for data wrangling and analysis.
* Developed RNN-based models for sequence-to-sequence tasks, demonstrating proficiency in capturing temporal dependencies for applications such as natural language processing.
* Developed CNN-based object detection models, enhancing the accuracy and efficiency of identifying and localizing objects within images.
* Applied expertise in time series forecasting, utilizing techniques such as ARIMA, SARIMA, Prophet, and other univariate and multivariate models to optimize predictions. Designed and deployed **time series forecasting model** Prophet on Azure cloud, a staff engagement system.
* 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**.
* Ingested data to one or more **Azure** Services - (**Azure Data Lake**, **Azure Storage**, **Azure SQL**, **Azure DW**) and processed the data in Azure **Databricks**.
* Engineered ETL pipeline for Single View of Customer
* Experience in building Pipelines in Azure Data Factory using Linked Services/Datasets/Pipelines to Extract, Transform and load data from different sources like SQL, Blob storage, SQL Data warehouse, and write back tool.
* Experienced in **Databricks** integration with popular machine learning libraries like scikit-learn, TensorFlow, **PyTorch** and MLflow. Can build, train, and deploy machine learning models at scale.
* Used Alteryx for **Data Integration,** **Data Cleansing,** **Data Transformation,** **Geospatial Analysis,** **Predictive Analytics,** **Data Profiling,** **Cloud Integration,** **Data Governance**, **Reporting and Visualization.**
* Explored and analyzed the customer-specific features by using Spark SQL.
* Performed univariate and multivariate analysis of the data to identify any underlying pattern in the data and associations between the variables.
* Performed data imputation using the Scikit-learn package in Python.
* Participated in features engineering such as feature intersection generating, feature normalization and label encoding with Scikit-learn preprocessing.
* Used Python 3.X (NumPy, SciPy, pandas, scikit-learn, seaborn) and Spark 2.0 (PySpark, MLlib) to develop a variety of models and algorithms for analytic purposes.
* Developed and implemented **predictive models** using machine learning algorithms such as linear regression, classification, multivariate regression, Naive Bayes, Random Forests, K-means clustering, KNN, PCA and regularization for data analysis.
* Conducted analysis on assessing customer consuming behaviors and discovering the value of customers with RMF analysis; applied customer segmentation with clustering algorithms such as K-Means Clustering and Hierarchical Clustering.
* Built regression models including Lasso, Ridge, SVR, and XGboost to predict Customer Lifetime Value.
* Built classification models including Logistic Regression, SVM, Decision Tree, and Random Forest to predict Customer Churn Rate.
* Used **Azure MLFlow** platform for running machine learning models and virtual delivering virtual apps.
* Used F-Score, AUC/ROC, Confusion Matrix, MAE, and RMSE to evaluate different Model performances.
* Designed and implemented recommender systems that utilized Collaborative filtering techniques to recommend courses for different customers and deployed to the AWS EMR cluster.

**ICICI Bank Ltd**

**Mumbai, India**

**Data Scientist Nov 2014 - April 2019**

**RESPONSIBILITIES:**

* Experienced in collecting, summarizing, and predicting customer databases to detect fraudulent activities.
* Experienced in customer sentiment analysis NLP.
* Implemented a project using Python, scikit-learn, NumPy, Pandas, SciPy and UNIX for detecting handwritten digits using artificial neural networks (Natural Language Processing) and convolutional neural networks.
* Designed and developed an advanced recommendation system for real estate and finance customers utilizing R and Python.
* Experienced in **Time Series** forecasting/Prediction.
* Experienced with impala and hive.
* Used Python, R, and SQL to create Statistical algorithms involving Multivariate Regression, Linear Regression, Logistic Regression, PCA, Random Forest models, Decision trees, and Support Vector Machines for estimating the risks of welfare dependency.
* Identified and targeted welfare high-risk groups with Machine learning/deep learning algorithms.
* Performed data cleaning and feature selection using MLlib, and SQL Alchemy package in PySpark.
* Using Airflow to keep track of job statuses in repositories like MySQL and PostgreSQL databases.
* Prediction of Multiple Regression analysis.
* Performed data discovery and built a stream that automatically retrieves data from a multitude of sources (SQL databases, external data such as social network data, user reviews) to generate KPIs using Tableau.
* Conducted campaigns and ran real-time trials to determine what works fast and track the impact of different initiatives.
* Experienced with **recommendation systems such as content filtering, and collaborative filtering.**
* Used text-mining process of reviews to determine customer’s concentrations.
* Experienced in creating powerful recommendation engines that can create upsell/cross-sell opportunities for banks.

**Data Analyst (Data Science & Analytics Group) April 2009 – Oct 2014**

**Responsibilities:**

* Gathering data from a variety of data sources, performing meaningful analysis (**Predictive, prescriptive**, **descriptive**) and presenting the analysis to management.
* **Analyzing large data** sets to reveal underlying patterns and trends quantitatively.
* Create **reports and dashboards** to closely monitor performance metrics and provide insights.
* Develop **advanced analytics** solutions including customer segmentation, optimization, testing & and design of experiments and **prescriptive analytics** to solve business problems. Operate as a subject matter expert on statistical analysis, test and design of experiments, analysis methodology**, modeling** & and application and financial impact analysis.
* Experienced in developing business reports by writing complex **SQL queries** using views, and volatile tables.
* Extensively use SAS procedures like **means, frequency** and other statistical calculations for Data validation.
* Performed Data Analysis and extensive **Data validation** by writing several complex SQL queries.
* Involved in the **design and development** of standard and ad-hoc reporting using SQL.
* Responsible for data lineage, maintaining data dictionary, naming standards and data quality.
* Performed Data Manipulation using MS **Excel Pivot** Sheets, **and VLOOKUP** and produced various charts for creating the mock reports.
* Used **Alteryx** for accessing, cleaning, testing and combining data.
* Used SQL Server and **advanced MS Excel** daily to manipulate the data for business intelligence reporting needs.
* Extracted data from different sources like Oracle and text files using SAS/Access, and SAS SQL procedures and created SAS datasets.
* Conducted **data mining** and **data modeling** in coordination with the finance manager.
* Experienced with **CRM** (Customer Relationships Management) tools for managing bank clients.

#### IBM Dakash Business Process Services Pvt Ltd,

#### India Oct 2008 - May 2009

**Technical Support**

**Responsibilities:**

* Troubleshooting internet-related issues of Canadians over the phone or by remote control.
* Selling and setting up new internet connections.
* Handling modem and **antivirus**-related issues.
* Configuring email and **troubleshooting** email-related issues Troubleshooting and software repair of internet-related issues

**Publications**

**Cultural Health Moments: A Search Analysis During Times of Heightened Awareness to Identify Potential Interception Points with Digital Health Consumers**

[**https://j.ideasspread.org/index.php/jems/issue/view/97**](https://j.ideasspread.org/index.php/jems/issue/view/97)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**\_**

**CERTIFICATIONS**

Introduction to Supervised & Unsupervised Machine Learning

Google Analytics Individual Qualification

Introduction to Artificial Intelligence

Data Science with R Programming

Power BI for Beginners

Tableau