Bhavya Raj B

Data Scientist

**SUMMARY**

* Data Scientist 8+ Years of experience executing data - driven solutions with adept knowledge on Data Analytics, Text Mining, Machine Learning (ML), Predictive Modelling, and Natural Language Processing (NLP).
* Experience in productionizing Machine Learning pipelines on Google cloud platform which performs data extraction, data cleaning, model training and updating the model on performance basis.
* Utilized GCP resources namely BigQuery, Cloud Composer, compute engine, Kubernetes cluster and GCP storage buckets for building the production ML pipeline.
* Expertise in building batch and streaming data pipelines on pulling data from multiple sources into Google’s BigQuery. These pipelines are built using python, Kafka, Dataflow and DataProc.
* Expertise in building ML models to predict failure events over store self-checkout machines and provides root cause for those failure events.
* 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.
* Hands on solving problems which brings significant business value by building predictive models utilizing structured & unstructured data.
* Built a Machine Learning model to predict hourly sales (Orders, Invoices and Shipments) for an ecommerce platform.
* Hands - on experience in Machine Learning algorithms such as Linear Regression, GLM, CART, SVM, KNN, LDA/QDA, Naive Bayes, Random Forest, SVM, Boosting.
* Hands on experience in creating data visualizations, dashboards in a Tableau desktop.
* Expertise in performing time series analysis and built forecasting models to predict the temperature and humidity spikes inside cold storage warehouses.
* Expertise in building monitoring dash boards that visualizes the present and predicted health of the cold storage warehouses.
* Experience in building data warehouses, data marts and data cubes for creating power BI reports to visualize various key performance indicators of business.
* Utilized python libraries namely Pandas, matplotlib and plotly for performing data analysis, data visualizations and predicted unexpected reboot events on store self-checkout machines (POS systems).
* Built a facial recognition model which is being used to perform user authentication for employee work hours tracking system.
* Utilized python’s flask framework for building REST APIs on top of Data Lake (BigQuery, Cloud SQL).
* Using Docker and ansible, containerized virtual infrastructure’s configuration management tasks which are used to detect config drifts and change back to original configurations.
* Expertise in containerizing applications using Docker composes.
* Achieved Continuous Integration &Continuous Deployment (CI/CD) for applications using Concourse.
* Experience with Test driven development (TDD), Agile methodologies and SCRUM processes.
* Experience in version control and collaboration tools like Git and source tree.

**TECHNICAL SKILLS**

**Languages:**Python, R, SQL, Java Script, Java, C, C++,

**ML/AI:**Tensorflow, Kera’s, Scikit-learn, Prophet, PySpark, NLTK, Airflow, Pandas, OpenCV

**Data Base:**MySQL, SQL server, PostgreSQL, MongoDB

**Reporting Tools: Tableau**, Power BI, Wavefront

**Predictive and Machine Learning:**Regression (Linear, Logistic, Bayesian, Polynomial, Ridge, Lasso), Classification (Logistic Reg., two/multiclass classification, Boosted Decision Tree, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machines, k-Nearest Neighbours, Neural Network, and various other models), Clustering (K-means, Hierarchical), Anomaly Detection, LSTM, RNN

**Cloud:**Google Cloud Platform, Pivotal Cloud Foundry, Azure, AWS

**Cloud Resources**: Azure Data bricks, AWS Glue, GCP BigQuery, Cloud Composer, Dataflow, Data Proc

**Frame works:**Flask, Django, Falcon, Bottle

**Tools:** Apache Spark, Kafka, Docker, Git

**Operating System:**Linux, Windows, Unix, MacOS

**PROFESSIONAL EXPERIENCE**

**Ebay | San Jose, CA Oct 2022 – Present**

**Sr. Data Scientist / ML Engineer**

**Responsibilities:**

* Constructively been part of a talented research team of data scientists in the field of **Computer Vision** to innovate, analyse application requirements into data models, to support standardization & effective adoption of bleeding-edge scientific norms and practices with a vision to enable integration and collaboration of **AI/ML** into everyday workflow.
* Managed data pipeline using **RDF Graphs**, primitives from **Apache** **Beam** & **Apache** **Nifi** to build transparent and manageable data flow on **GCP** **Dataflow**, Google **BigQuery** platform for a practically fully automated solution alleviating daily routine.
* Applied **Probabilistic Graphical Methods** (Bayesian and Gaussian network) to create machine learning models.
* Implemented **classification models** various including Random Forest and Logistic Regression to quantify the likelihood of each member enrolment for the upcoming enrolment period.
* Used LinearRegression for predicting the member cost for the upcoming enrolment period.
* Facilitated and built **Deep Learning architectures** to achieve higher performance on classification tasks, object detection & localization tasks, image segmentation on variety of image data through rapid experimenting & customizing models from latest research studies via Transfer Learning and observed model’s performances over new image data to quantify data ambiguity.
* Worked on data cleaning and ensured data quality, consistency, integrity using **Pandas**, **NumPy**.
* Developed Spark code using **Scala** and **Spark-SQL** for faster processing and testing.
* Hands on expertise in working with different data formats such as JSON, XML and utilized machine learning algorithms using **Python**.
* Explored and analysed the specific features by using Matplotlib and ggplot2. Extracted structured data from **MySQL** databases and CRM systems like **SalesForce**, developing basic visualizations like analysing **A/B test** results.
* Performed data cleaning and feature selection using **MLlib** package in **PySpark** and worked with deep learning frameworks such as **Caffe**.
* Implement **Azure Cog Search** to enhance the search experience on online platforms.
* Developed Spark/Scala, R, and Python for regular expression (regex) project for text validation in the **Linux** environment with **Hadoop**/Hive for big data resources.
* Updated Python scripts to match training data with database stored in **AWS** **Cloud Search** to assign each document a response label for further classification.
* Responsible for reporting of findings that used gathered metrics to infer and draw logical conclusions from past and future behaviour.
* 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.
* Created an algorithm that can predict the type of the object in a typical house using Deep Learning. Used OpenCV for the image analysis and **keras** and **Tensorflow** for implementing artificial neural networks (ANN).
* Designed and Developed **NLP** models for Smart Inventory Management at eBay. The model established correlation between search keywords and checkouts. It helped forecast demand in various product categories using the search data. The model helped reduce 4% of inventory across 20% of highest selling Product categories maintaining demand fulfillment.
* Natural Language Processing (**NLP**) such as sentiment analysis, entity recognition, Topic Modeling and Text summarization was done using advanced **python library** such as NLTK, TextBlob, Spacy and Gensim.
* Applied Natural Language Processing (**NLP**) and ML to analyze and collect data, perform analyses, create hypotheses, design, and implement solutions, as well as conduct experiments to evaluate different data science algorithms to predict sentimental analysis, Topic Modeling, and Entity reorganization.
* Developed **NLP** **models** (Stemming, Lemmatization, Levenshtein distance) on search queries and established correlation between search keywords and checkouts.
* The model is deployed to forecast the demand in various product categories using the search data.
* Worked on multiple personalization initiatives such as generating recommendations using Collaborative **Filtering Algorithm**, Search Intent, Users interests.
* Led a project to increase the Clickthrough rate (CTR) of display ads on **eBay** using Logistic regression. Wrote complex **SQL** queries to get ML features for the project.
* Involved in story-driven agile development methodology and actively participated in daily scrum meetings.

**Environment:** Data Pipeline Management, AWS, Data Governance, Hybrid Cloud, GCP Dataflow, Google BigQuery, Apache Nifi, Scala, Spark, Apache Beam, Hadoop, ETL, OpenCV, Spark-SQL, MySQL, Azure Cloud, Azure ML Studio, Computer Vision (CV), Unix/Linux.

**Kaiser Permanente | Oakland, CA Jan 2020 – Sep 2022**

**Sr. Data Scientist**

**Responsibilities:**

* Applied Supervised Machine Learning Algorithms for the **predictive modelling** to tackle various types of problems: Successful Transition from Skilled nursing facility, identify predictors for medicare advantage members, lower the cost for mitigating homelessness, issues management.
* Explored and analysed the customer specific features by using **Matplotlib, Seaborn** in **Python** and dashboards in **Tableau**.
* Built a datawarehouse by utilizing **ETL** processes with tools such as **Apache Nifi** and **Apache Hadoop** to gather all the business data related to doctors, patients, prescriptions, orders and calls from different sources.
* Developed doctor report cards for real-time insights into their performance over the years. Using **Apache** **Kafka** for data ingestion and **Tableau** integrated with **Hadoop**/**Spark** for creating the reports.
* Developed predictive models like disease risk, readmission risk using **advanced machine learning** algorithms, ensemble models, and deep learning architectures.
* Used **Pandas, NumPy, Scikit-learn** in **Python** for developing various like emergency department wait time, chronic disease progression machine learning models and utilized algorithms such as Linear regression, Logistic regression, Gradient Boosting, SVM and KNN.
* Developed pipeline using **Hive** (**HQL**) to retrieve the data from **Hadoop** cluster, **SQL** queries to retrieve data from **MySQL** database and used **ETL** for data transformation.
* Derived data from relational databases to perform complex data manipulations and conducted extensive data checks to ensure data quality. Performed Data wrangling to clean, transform and reshape the data utilizing **NumPy** and **Pandas** library.
* Implemented installation and configuration of **multi-node cluster** on Cloud using **AWS**.
* Utilized IOT sensors for collecting health information of cold storage and build streaming data pipeline into **GCP**’s **BigQuery** with the help of **Apache** **Airflow**.
* Productionized machine learning pipelines to gather data from **BigQuery** using **Apache** **Airflow** and build forecasting models to predict temperature and humidity spikes inside the cold storage.
* Built monitoring dash boards by employing visualization tools such as **Tableau** or **Power** **BI**, visualizing both the current state and predictive health of cold storage warehouses.
* Applied **NLP** techniques for sentiment analysis on customer feedback and reviews.
* Implemented **NLP** algorithms to analyse customer interactions and provide personalized health and wellness recommendations.
* Utilized **NLP** for understanding and extracting information from clinical documents and medical records.
* Led the development and deployment of machine learning models on **GCP** **Vertex** **AI** tailored for healthcare applications, including predictive analytics for patient outcomes and disease progression.
* Utilized **Vertex AI**'s **AutoML** capabilities to fine-tune models for medical image analysis, ensuring high accuracy in tasks such as radiology image interpretation.
* Designed end-to-end machine learning pipelines on **GCP** **Vertex** **AI** with a focus on security and compliance, ensuring that healthcare data handling adheres to regulatory standards like HIPAA.
* Implement encryption and access controls within **Vertex** **AI** to safeguard patient data during model training, evaluation, and inference stages.
* Led the development of Deep Learning models, utilizing **PyTorch** and **Tensorflow**, to address intricate challenges and enhance predictive capabilities.
* Leveraged **Python**, **PyTorch**, and **Tensorflow** to design and implement cutting-edge models, enhancing the organization's capabilities in applied research and data-driven decision-making.
* Utilized **Azure tech stack** for secure storage and management of Electronic Health Records (EHRs).
* Applied **OpenAI**'s natural language processing capabilities to analysed and understand unstructured clinical notes.
* Implemented **Hugging** **Face**'s Transformers library to enhance natural language understanding in customer interactions.
* Utilized pre-trained language models to **improve chatbot** responses and customer support services.
* Implemented the application of **retrieval augmented generation** techniques to optimize customer support chatbots. Implemented advanced methods to retrieve historical customer interactions and generate context-aware responses, significantly improving the accuracy and responsiveness of the chatbot in addressing user queries.

**Environment:** Python, R, HDFS, ODS, OLTP, Power BI, Tableau, Hive, AWS, BigQuery, Apache Airflow, Hadoop, MySQL, OLAP, DB2, Metadata, Tera Data, MS Excel, Mainframes MS Vision, Spark, Map-Reduce, SQL and MongoDB

**Capital One | McLean, VA Sep 2018 – Dec 2019**

**Machine Learning Consultant**

**Responsibilities:**

* Built an ML model to automate the process of finding the root cause over failed events on store self-checkout machines (POS systems). Integrated the ML model by utilizing **Flask API**.
* Decreased the enterprise service now tickets by 15% in building a service named Back up as a service by utilizing **AWS Backup** which gives the ability for a customer to initiate backups, restores on servers.
* Constructed a machine learning model for Capacity Planning by collecting historical CPU and Disk usage data from on-premises infrastructure, preprocessing the data, engineering features, and selecting suitable algorithms, such as **LSTM** networks, to predict resource utilization.
* Using **Apache** **Airflow**, built data pipelines to gather data from store-checkout devices into **BigQuery**.
* Used **classification techniques** including Random Forest and Logistic Regression to quantify the likelihood of each user referring.
* Utilizing **Cloud Composer**, **BigQuery** and **GCP** **storage buckets** on Google cloud platform productionized machine learning pipelines to performs data extraction, data cleaning, model training and updating the model on performance basis.
* Utilized python’s **Django** framework and **Cloud Bolt**, to build a web user interface where users can perform backups and restores on servers.
* Designed and developed an automation process, that helps enterprise to maintain common configurations and detect configuration drifts across the enterprise virtual infrastructure using **Docker**.
* Applied 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.
* Visualized data using MS PowerBI, ggplot, Seaborn and matplotlib.
* Utilized python and **Kafka** to build data pipelines for pulling data from multiple sources (vCenters, data bases, store devices) into Google’s **BigQuery**.
* Utilized machine learning algorithms such as logistic regression, multivariate regression, K-means, & Recommendation algorithms to extract the hidden information from the data.
* Used **Pandas**, **NumPy**, **Scikit-learn** in Python for developing various machine learning models and utilized algorithms such as Linear regression, Logistic regression, Gradient Boosting, SVM and KNN.
* For serving data, built **REST APIs** on the data lake (**BigQuery**, cloud **SQL**).
* Wrote **Dataflow** jobs for moving data across the google cloud platform.
* Using python, built an alerting application that sends alerts (email, slack messages) and create tickets for critical **Rubrik** (Backup & recovery management system) failure events.
* Using python to build a data pipeline that pulls the performance metrics of all the virtual infrastructure into **Wavefront** to create dashboards.
* Deployed all the applications, **REST APIs** to google app engine, pivotal cloud foundry and on-premises Linux servers.

**Environment:** BigQuery, Google Cloud Platform, Apache Airflow, Django, Docker, PowerBI, ML algoritms, Pyhton Libraries, REST API, MySQL, Hadoop Ecosystem, Wavefront, Rubrik , HDFS, Hive, Hive QL, Pig, Map Reduce, Regression, Time-Series Forecasting, Predictive Analytics, Clustering, Text Mining, NLP, Unix/Linux.

**Mastercard | Purchase, NY Jun 2017 – Aug 2018**

**Data Scientist**

**Responsibilities:**

* **Identify business problems** or management objectives that can be addressed through data analysis and propose creative solutions and strategies to existing business challenges.
* Analyse, manipulate, and process massive amounts of data using statistical software to discover trends, patterns, and insights via Jupyter, Sci-kit learn and **Tableau**.
* Automate the entire collection process pipeline by identifying valuable data sources by using **ETL** tools like **Apache** **Nifi**, **Apache** **Beam**.
* Apply feature selection algorithms to models such as ANOVA (analysis of variance), decision trees using **PySpark’s** **Mllib** package and hyper tune the parameters based on interest and to predict the outcomes.
* Design and implement predictive learning models and machine learning algorithms like K-NN, Logistic regression etc for financial applications like- Transaction Classification and Risk Modelling and combine them through **ensemble learning**.
* Present data using data visualization techniques by using tools like **Tableau** and libraries like matplotlib, ggplot2, seaborn by creating graphs, charts, or other visualizations to convey the Results of data analysis.
* Read research and scientific articles, conference papers, or other sources to **identify emerging analytic trends** and technologies to improve efficiency of existing models.
* Perform Quality Analysis testing and validation internally using **Django** and reformulate models to ensure accurate prediction of outcomes of interest and end to end **API** testing with dummy data and actual data before launch of the actual product with the Engineering team.
* Prepare documentations and deliver oral or written presentations of the results of mathematical modelling and data Analysis to stakeholders.
* Applied **mean-variance optimization algorithms**, such as **Markowitz portfolio theory**, using optimization libraries like **scipy.optimize** in Python, to construct efficient investment portfolios balancing risk and return.
* Implemented anomaly detection algorithms, such as isolation forests and autoencoders, with Python libraries like **scikit-learn** and **TensorFlow/Keras** to detect and prevent fraudulent activities in financial transactions.
* Utilized natural language processing (NLP) techniques, including sentiment analysis and named entity recognition, with NLP libraries like **NLTK** and **spaCy** in Python, to analyze regulatory texts and ensure compliance.
* Employed clustering algorithms, such as k-means and hierarchical clustering, using Python libraries like **scikit-learn**, to segment customers based on demographic and behavioural data for targeted marketing campaigns.
* Developed credit risk models using **gradient boosting machine algorithms**, such as XGBoost and LightGBM, in **Python**, to assess creditworthiness and predict default probabilities for loan applicants.
* Employed model evaluation techniques, including cross-validation and performance metrics like ROC-AUC and F1 score, using Python libraries like **scikit-learn**, to continuously improve predictive models based on feedback and data updates.
* Utilized version control systems like **Git** and collaboration platforms like Jira to facilitate seamless collaboration with cross-functional teams and track project progress efficiently.

**Environment:** Python, PySpark, Django, Tableau, PySpark, Apache Nifi, Apache Beam, ML algorithms, Jupyter.

**Brillio, HYD Sep 2014 – Oct 2016**

**Data Scientist**

**Responsibilities:**

* Conducted qualitative and quantitative research to gather data from **data mart**.
* Responsible for data identification, collection, exploration & cleaning for **modelling**, participated in model development.
* Visualized, interpreted, report findings, and developed strategic uses of data.
* Understood transactional data and developed analytics insights using **Statistical models** using Machine learning.
* Involved in **gathering requirements** while uncovering and defining multiple dimensions. Extracted data from one or more source files and Databases.
* Collected database of sales of items in all aspects. Cleaned, filtered, and transformed data to specified format.
* Designed various intelligent reports using various reporting tools.
* Cleaned data using **R**, then visualized the data, and derived **statistical modelling** plots.
* Performed data visualization via **ggplot2** in **R** and **matplotlib** in **Python**.
* Worked in **Amazon Web Services** cloud computing environment.
* Responsible for **providing reports**, analysis, and insightful recommendations to business leaders on key performance metrics pertaining to sales & marketing.
* Gathered all the data that is required from multiple data sources and creating datasets that will be used in analysis.
* Used R to identify product performance via Classification, tree map and regression models along with visualizing data for interactive understanding and decision making.
* Created Intelligent dashboards and visualization on regular basis using **ggplot2** and **Tableau** **Tabpy**. Reserve.
* Accomplished multiple tasks from collecting data to organizing data and interpreting statistical information.
* Created dynamic linear models to perform trend analysis on customer transactional data in R.
* Conducted exploratory and descriptive data analysis of large data sets.
* Expertise in **Business Intelligence** and data visualization using **R** and **Tableau**.
* Performed Exploratory Data Analysis and Data Visualizations using **R**, and **Tableau**.
* Applied concepts of probability, distribution, and statistical inference on given dataset to unearth interesting findings through use of comparison, T-test, F-test, R-squared, P-value.

**Environment**: Machine learning, AWS, Spark, HDFS, Hive, Pig, Linux, Python (Scikit-learn/SciPy/NumPy/Pandas/PyTorch), R, SAS, SPSS, MySQL, Eclipse, PL/SQL, SQL connector, Tableau.

Education

**California State University, East Bay**  2016 – 2018

**Master of Science, Business Analytics**

Relevant courses-Data warehousing & Business intelligence, Big Data Tech & Applications, Global Supply Chain, Data Mining.

**SRM University, India** 2010 – 2014

**Bachelor of Technology, Information Technology**

Relevant courses- IoT, Data Structures, Cloud Computing, Object-Oriented Analysis and Design.