**XYG**

* Data Scientist recognized for developing analytics-driven business methods/strategies that produce real value and minimize risks. Known as a self-starter who manages multiple project priorities such as mining data, leveraging machine learning algorithms, in fast-paced Big Data, and Cloud environments.

**Professional Summary:**

* IT professional with around 11 years of experience as a Data Scientist
* Extensive experience in Machine Learning, Statistics, Data Wrangling, Data Mining, Data Analysis and Data Visualization, Python Programming in Credit Risk management, strategic targeting, marketing campaign development, Tableau reporting experience.
* Experience in Generative AI models, including BERT and GPT, applied to Natural Language Processing tasks such as text summarization and sentiment analysis.
* Profound knowledge in Machine Learning Algorithms like Linear, Non-linear, Logistic Regression, Reinforcement Learning, Natural Language Processing, Fuzzy Logic, Random forests, Ensemble Methods, Decision tree, Gradient-Boosting, K-NN, SVM, Naïve Bayes, Clustering (K-means, DB Scan), Deep Learning.
* Good understanding of Agile and Test-driven development (TDD) methodologies.
* Well versed in various Natural Language Processing (NLP) methodologies including Word Embedding, Text Clustering and Classification, using Deep Learning and other conventional machine learning concepts.
* Experience in Acquisition of correct Datasets, Data Scrubbing to mine the target data, Data Engineering to extract features utilizing Statistical Techniques, Exploratory Data Analysis with an inquisitive mind, build diverse Machine Learning Algorithms for developing Predictive Models and design Stunning Visualizations to help the growth of Business Profitability.
* Outstanding pre-eminence in Data extraction, Data cleaning, Data Loading, Statistical Data Analysis, Exploratory Data Analysis, Data Wrangling, Predictive Modeling using, Python and Data visualization using Tableau.
* Proficient in design and development of various dashboards, reports utilizing Tableau Visualizations like Bar graphs, scatter plots, pie-charts, geographic visualization and other making use of actions, other local and global filters according to the end user requirement.
* Experience with common Machine Learning Algorithms like Regression, Classification and Ensemble models.
* Developed simple and elegant data visualizations designed to communicate predictions clearly.
* Develop and deliver key initiatives in the areas of data management, and analytics in line with respect to the timelines. Involved in the entire data science project life cycle and actively involved in all the phases including data cleaning, data extraction and data visualization with large data sets of structured and unstructured data, created ER diagrams and schema.
* Utilized both relational databases (MySQL, SQL Server) and NoSQL databases (e.g., MongoDB) for efficient data storage and retrieval, enhancing query performance and enabling flexible data handling for machine learning applications.
* Integrated ML models into large-scale production systems, implementing CI/CD pipelines to ensure continuous delivery. Applied ML Ops practices for model retraining, monitoring, and performance tracking in cloud environments.
* Designed, coded, and debugged scalable ML models, ETL processes, and SQL queries in production environments. Focused on maintainable and well-documented code for ease of future improvements and hand-offs.
* Executed rigorous data validation and cleansing processes as part of ML workflows, using Python and SQL to ensure data quality and consistency across structured and unstructured sources.
* Led multiple ML initiatives from strategic development through implementation, ensuring model accuracy, reliability, and alignment with business objectives. Managed iterative improvements based on performance metrics and stakeholder feedback.

**Technical Skills:**

* Programming Language: Python, SQL, R, C, MATLAB
* Python Libraries: Pandas, Numpy, Matplotlib, Scikit-learn
* Cloud Services: AWS, GCP, Azure
* Operating System: Windows.
* Software Methodologies: SDLC - Waterfall, Agile, SCRUM
* Machine Learning and Statistics: Regression, Random Forest, Clustering, Time-Series Forecasting, Hypothesis testing (A/B test), Explanatory Data Analysis, ANOVA, t- test, chi- square.
* Machine Learning Algorithms: Linear Regression, Classification, Logistic Regression, Linear Discriminant Analysis (LDA), Decision Trees, Random Forests with Adaboost and Gradient Descent Boosting, Support Vector Machines (SVM), KNN, Decision Tree, Naïve Bayes, K - Nearest Neighbor, Hierarchical clustering, K-means clustering, Density based clustering (DBSCAN).
* Machine Learning Techniques: Principal Component Analysis, Single Value Decomposition, Data Standardization Techniques, L1 and L2 regularization, RMS prop, Hyper parameter tuning, Resampling Techniques like SMOTE, Cluster Centroid Methods, Ensemble Methods, Feature selection and Feature Engineering, Cross Validation Methods (K-fold).
* Databases/Web Technology: MS SQL Server, MS-Access , MySQL, GitHub 2, SAS, NoSQL, PostgreSQL, Hadoop, Spark, Hive, Pig.
* Documentation Tools: Office365 – Word, Excel, PowerPoint, Outlook, Access, one note
* Data Visualization Tools: Tableau.
* Other software’s: PyCharm, Jupyter, Jira, Excel.
* Data/Business Analysis Techniques: Data Visualization, Data Wrangling, Regression Analysis, Hypothesis testing, Dashboard Reporting, Trend Analysis, Reporting & Analytic Development, Exploratory Data Analysis
* Data Science: Deep Learning (CNN, RNN, LSTM, TensorFlow, PyTorch), Natural Language Processing (BERT), LLM, NLTK,LLM, SpaCy, GPT-3, Hugging Face's Transformers

**Professional Experience:**

**Verizon | Irving, TX Feb 2024 to Till Date**

**Data Scientist**

Project Description: Real-time data obtained from multiple resources can be used to improve the products offered by Verizon. Customer usage can be analyzed, and this will help in coming up with new product bundles which help in saving money and identifying and serving customer needs. Perform time series analysis for trading team interested in options markets for regular and trading around earnings, volatility modelling.

**Key Contributions:**

* Involved in Data mapping specifications to create and execute detailed system test plans. The data mapping specifies what data will be extracted from an internal data warehouse, transformed and sent to an external entity.
* Utilized PyTorch's dynamic computation graphs to experiment with custom neural network architectures for NLP tasks.
* Developed deep learning models using TensorFlow, leveraging Keras for model training and fine-tuning.
* Superintended usage of Python NumPy, SciPy, Pandas, Mat plot, Stats packages to perform dataset manipulation, data mapping, data cleansing and feature engineering. Built datasets using R and Python.
* Used common NLP pre-processing techniques, such as (tokenization, lemmatization/stemming, POS tagging, parsing) on text data for analytics over products and user reviews which in turn were used in search criteria using NLTK and SpaCy libraries.

**Responsibilities:**

* Used Pandas, NumPy, seaborn, SciPy, matplotlib, scikit-learn, NLTK, and spaCy in Python for developing various machine learning algorithms.
* Enforced model Validation using test and Validation sets via K- fold cross validation.
* Calculated Hadoop Storage for the project and deployed cluster in development environment and Installation of various Hadoop ecosystem components.
* Designed and developed data ingestion, aggregation and advanced analytics in Hadoop environment from MySQL and Oracle databases.
* Broadly designed easy to follow visualizations using Tableau software and published dashboards on web and desktop platforms.
* Designed and generated insightful data visualizations and reports using tools such as Tableau, providing actionable insights to support decision-making and business strategies.
* Utilize advanced data pre-processing techniques like n-grams generation and stemming to clean and pre-process customer feedback data.
* Developed sentiment analysis models using ensemble methods and neural networks to understand customer sentiments.
* Implemented lemmatization and translation techniques to pre-process textual documents from multiple languages.
* Designed neural network-based models and utilized BERT, GPT models for text summarization, condensing lengthy documents into concise summaries and FAQ generations.
* Fine-tuned algorithms to improve the accuracy and coherence of generated summaries, aiding in quick document understanding for research or business purposes.
* Spearheaded the implementation of Named Entity Recognition (NER) models to extract vital information from unstructured customer data, enhancing customer insights and service strategies.
* Collaborated within cross-functional teams to integrate NER and BERT models into Verizon's data processing pipelines, optimizing performance and efficiency in handling large-scale textual data.
* Implemented advanced data pre-processing techniques like lemmatization, n-grams generation, stemming, and translation for NLP tasks.
* Utilized and implemented Neural networks - CNN and deep - ANN, XGboost, Random Forest models to extract feature importance and gain insights for a multi class classification model Regarding Port in granularity to find more underlying patterns.
* Played a key role in making informed decisions regarding ML infrastructure, ensuring optimal model, data, and feature selection. Managed model training, hyper parameter tuning, and addressed challenges related to dimensionality, bias/variance, and validation.
* Developed and optimized production-ready modeling code capable of scaling efficiently to handle the demands of millions of calls and users, ensuring high performance and reliability in real-world scenarios.
* Developed linear models, ensemble models, and neural networks tailored for diverse NLP applications.
* Deployed NLP models in production environments, ensuring scalability and efficiency while comparing model performances to determine optimal solutions. Utilized advanced vectorization techniques and prompt engineering to optimize model behavior in NLP applications.

**Environment:** Python Scikit-learn, NumPy, SciPy, Pandas, Matplot, Stats, Hadoop, SQL, SDLC-Agile/Scrum.

**Aetna | Dallas, TX May 2023 to Dec 2023**

**Data Science /NLP Engineer**

Project Description: The project involved the development of a classification model to predict whether clinics would engage in single or multiple deals with the company. The project aimed to optimize sales strategies by analyzing various datasets, including company deals, tickets, and sales activities, to forecast sales growth and improve business outcomes.

**Key Contributions:**

* Developed a Random Forest classification model with 79% accuracy to predict clinic engagement in multiple deals.
* Built and trained NLP models for text classification and sentiment analysis using PyTorch’s Autograd for automatic differentiation.
* Leveraged TensorFlow’s distributed computing capabilities to train models across multiple GPUs, improving training efficiency by 20%.
* Applied TensorFlow to design a neural network for classifying customer feedback, enhancing business insights.
* Cleaned and consolidated data from multiple sources, ensuring accuracy and consistency across datasets.
* Analyzed state-wise sales data to identify key regions for business expansion.
* Applied variable reduction techniques and removed high-correlation and high-null value variables to enhance model performance.
* Implemented feature importance analysis to identify significant factors influencing sales outcomes.

**Responsibilities:**

* Led data preparation efforts, including data cleansing, variable reduction, and encoding of categorical variables.
* Conducted exploratory data analysis to uncover trends and insights that informed sales strategies.
* Coordinated with team members to integrate additional datasets, such as Lunch-and-Learn metrics, into the analysis.
* Prepared visualizations and reports to communicate findings and recommendations to stakeholders.
* Assisted in forecasting sales growth and recommending strategies for market expansion and customer retention.

**Environment:**

Tools: Python, Pandas, Scikit-learn, Excel

Techniques: Data Cleaning, Data Consolidation, Variable Reduction, Label Encoding, Random Forest Modelling

Platform: Jupyter Notebook, Excel

**Value Labs | India Jan 2019 to Jul 2022**

**Data Scientist/NLP Engineer**

Project Description: Project was focused on customer clustering based on ML and statistical modeling effort including building predictive models and generate data products to support customer classification and segmentation. Also to Develop Estimation model for various product & services bundled offering to optimize and predict the gross margin, built sales model for various product and services bundled offering

**Key Contributions:**

* Conducted data analysis when necessary to determine root cause of data inconsistencies.
* Acquire and automate additional data sources (internal and external) as needed to improve operating efficiency or support business/product needs.
* Applied Dimensionality Reduction techniques like PCA, t-SNE to reduce the correlation between features and high dimensionality of the data to better use time and storage.
* Analyzed claims using NLP and worked on NLTK package for application development.
* Utilized Stemming, Lemmatization, Tokenization, TF-IDF, and Bag of Words Word2Vec to get most relevant and frequent words in the claimed documents.
* Developed and implemented various Machine Learning Algorithms and Statistical Modeling like Text Analytics, Sentiment Analysis, Decision Tree, Naive Bayes, and Logistic Regression for predicting the binary risk levels of applicants.
* Finalized the optimal model based on ROC & AUC and fine-tuned the hyper parameters of the above models using Grid Search.
* Employed Ensemble Learning techniques such as Random Forests and Ada Gradient Boosting to improve the model performance by 15%.

**Responsibilities:**

* Visualized the categorical and continuous features via mean response, histogram and bar plot using Matplotlib and Seaborn packages.
* Monitored incoming data to ensure Russell Investments receives timely delivery and accurate data.
* Employed K-Fold cross-validation to test and verify the model accuracy.
* Fake claims, feedback was flagged using Recurrent Neural Networks (RNN) and Long Shot Term Memory models using Keras.

**Environment:** Python, NumPy, Pandas, Seaborn, Matplotlib, Scikit-Learn, Tableau, PostgreSQL, PySpark, Google collab.

**Gleam Technologies | India Nov 2016 to Dec 2018**

**Machine Learning Engineer**

**Responsibilities**:

* Used both supervised and unsupervised anomaly detection techniques such as DBSCAN, Isolation Forests, Local Outlier Factor, One-Class Support Vector Machines, and deep learning (Auto encoder) etc.
* Worked with time series analysis to forecast the sale of product using moving average, stationarity, autocorrelation, SARIMA, VAR, LSTM, and SARIMAX. Used LSTM to predict probability of failure at different time intervals compensating for independent variables reflecting states of wear.
* Optimized model performance, enhancing prediction probabilities by 20%, and balancing recall-precision to improve F1 score by 15%. Leveraged AWS services for data processing, storage, and machine learning model deployment.
* Applied various machine learning algorithms and statistical modelling like decision trees, text analytics, natural language processing (NLP), supervised and unsupervised, regression models, neural networks, deep learning, SVM, clustering to identify Volume of product using scikit-learn package in python and PySpark.
* Worked on Text Mining and Text Processing for prediction of the sentimental analysis of the customer data using BERT and the Universal sentence encoder.
* Developed and deployed Python modelling APIs, incorporating diverse ML techniques to predict user behaviour. This facilitated numerous marketing segmentation programs, boosting their effectiveness by 25%. Segmented the customers based on demographics, geographic, behavioural, and psychographic data using K-means Clustering. Designed and implemented end-to-end systems for Data Analytics and Automation, integrating custom visualization tools using Python and Tableau.
* Used AWS Machine Learning for building, training, and deploying machine learning models faster using drag-and-drop designer and automated machine learning.

**Rotech Info Systems | India Feb 2013 to Oct 2016**

**Data Analyst**

Project Description: Project was on developing, executing, tracking and analyzing targeted marketing campaigns. Utilized the social media campaign management application to develop and report on complex, multi-step campaigns. Analyze campaign performance, report on key business metrics and develop insights through customer analysis.

**Key Contributions:**

* Developed a statistical arbitrage strategy by applying machine learning algorithms including generalized linear models, boosted regression tress and support vector machines, tuned the hyper parameters and back-tested models using validation and out-of-sample data to build prediction models.
* Implemented and maintained scalable python code for daily automated data update, technical indicators generations which are used to build Ensemble models to predict the expected revenue of targeted companies
* Created Machine Learning tools that computes adjusted P/E values and few other custom visualizations to internally used application required for teams based on tkinter module in python

**Responsibilities:**

* Provided statistical insights into semi-deviation & skewness-to-kurtosis ratio to guide vendor decisions and inferences into optimum pricing for raw material order quantities.
* Conducted data preparation and outlier detection using Python.
* Used NumPy, SciPy, Matplotlib libraries for n-dimensional representation of data and plotting graphs.
* Performed Clustering with historical, demographic and behavioral data as features to implement the personalized marketing that offers right product to right person at the right time on the right device.
* Implemented Principal Component Analysis (PCA) in feature engineering to analyze high dimensional data. Perform data manipulation, data preparation, normalization, and predictive modelling. Improve efficiency and accuracy by evaluating models in Python and R.

**Environmen**t:

Tools: Python, Pandas, Scikit-learn, Excel

Techniques: Data Cleaning, Data Consolidation, Variable Reduction, Label Encoding, Random Forest Modelling

Platform: Jupyter Notebook, Excel