ANVAY AWALGAONKAR

SUMMARY:

- Championed advanced statistical analysis, machine learning, and data modeling, showcasing expertise in predictive modeling for impactful insights, employing tools such as **TensorFlow 2.x** and **PyTorch**.
- Applied state-of-the-art machine learning frameworks and orchestration tools, including **Scikit-learn 0.24**, **Keras 2.x**, and **Apache Spark 3.x**, to seamlessly navigate end-to-end model development and deployment processes.
- Executed intricate feature engineering strategies, skillfully handling diverse structured and unstructured datasets using tools like **NumPy 1.x**, **pandas 1.x**, and **Scipy 1.x** to derive meaningful insights for informed decision-making.
- Demonstrated mastery in a multitude of programming languages, particularly **Python 3.x**, **R 4.x**, and **Scala 2.x**, wielding proficiency in key data science libraries like **NLTK 3.x**, **Beautiful Soup 4.x**, and **Gensim 4.x**.
- Validated success through a proven track record, applying advanced analytics to solve complex real-world problems, steering data-driven decision-making processes, and deploying models with **Docker 20.x** for efficient and reproducible workflows.
- Navigated the entire data science workflow, meticulously handling data exploration, cleansing, model training, evaluation, and deployment, utilizing tools like **Jupyter notebooks 6.x** and **RStudio 1.x**.
- Championed the utilization of diverse datasets, extracting actionable insights to directly influence business outcomes, utilizing technologies such as **SQL**, **HiveQL**, and **Pig 0.17** for advanced querying.
- Illustrated adeptness in data visualization techniques, effectively conveying findings to both technical and non-technical stakeholders, utilizing tools like **Matplotlib 3.x**, **Seaborn 0.11**, and **Tableau 2021.x**.
- Proficient in agile methodologies, actively contributing to collaborative work environments demanding dynamic problem-solving, using project management tools like **Jira 8.x** and **Trello**.
- Cultivated strong interpersonal and communication skills, fostering seamless collaboration within cross-functional teams, utilizing collaborative platforms like **Slack 4.x** and **Microsoft Teams**.
- Showcased rapid learning ability, adapting seamlessly to emerging technologies, and consistently delivering high-quality results within specified timelines, keeping abreast of technologies like **Quantum Computing** and **Explainable AI**.
- Maintained currency with the latest technology trends, ensuring continuous improvement in methodologies and toolsets, adopting cloud technologies such as **AWS (Amazon Web Services)**, **Azure 2021**, and **Google Cloud Platform 2021**.

Technical Skills

Data Analysis/Visualization: | Databases:

Tableau 2022.1 | - SQL Server 2022/2019/2017/2016
Matplotlib 3.4 | - MySQL 8.0 | PLSQL
Seaborn 0.11 | - MongoDB 5.x
Power BI 4.0 , Advanced Statistical Methods

Machine LearningAnalysis/Documentation:

Numpy 1.21- Requirements Engineering
TensorFlow | Business Process Modeling
Sci-Kit Learn 0.24| Gap Analysis
SPSS 28

OS/Scripting Language: | R Packages/Languages:

Linux (Kernel 5.x) | - Ggplot2 3.3
Excel 2021 (Pivot, Tables, Lookups) | -caret 6.0
MacOS 12| -wordcloud 2.8
Windows 10,11 | - quantmod 0.4 | UNIX
SharePoint|2019, BigQuery 3.x, Redshift 3.x, S3
Azure Cosmos DB, Apache Airflow, Teradata,
HBase, Hive, Apache Spark, Hadoop, Drill,

Advanced Analytics/ Devops/CI/CD

Time-Series, NLP, Neural Netwroks| Git 2.36
Apache Kafka 2.8, ELK Stack 7.17,Jupyter NB
Azure, AWS, IBM Cloud, GCP| Docker20.10
Kubernetes 1.22, Jenkins 2.319, Travis CI
Recommendation System(1.6.9), MLOps(1.2.5)
GCP Composer, Big Querry, Image Processing, Impala
@Scale, Impala

Collaboration/Communication/Other Tools
Restful APIs, GraphQL, Data Encryption, Network Security, OAuth2.0, Security Audits, Penetration Testing

Python Packages
Plotly(5.0), spaCy(3.2), Bokeh(2.3), XGBoost(1.5) Pandas(1.3), NLTK(3.6)

Debugging Tools
Python Debugger(pdb), Python 3.9
Jupyter Notebook Debugger
Logging Libraries

Methodologies
CRISP-DM, AgileDataScience, Scrum, KDD
Design Thinking

PROFESSIONAL EXPERIENCE:

Client: ADP.Inc, NewYork **Sept 2023 - Current**
Job Title: Data Engineer
Project: Modernizing Data Infrastructure for Enhanced Analytics

Description:
This project involves leveraging advanced technologies, methodologies, and best practices to enhance the organization's ability to collect, store, process, analyze, and derive insights from data effectively. As an experienced data scientist with a focus on upgrading data architecture and expanding analytics capabilities, have significantly advanced our organization's data strategy and analytics capabilities. Leading initiatives to provide real-time analytics, optimize data infrastructure, and provide stakeholders throughout the business with relevant insights has been aided by my job.

Responsibilities:

- **Led the migration of our data infrastructure to AWS, Azure, and Google Cloud Platform (GCP) environments, leveraging services like Amazon S3, Microsoft Azure Blob Storage, and GCP Cloud Storage for scalable storage solutions.**
- **Designed and implemented a scalable data lake architecture using Hadoop ecosystem components such as HDFS (Hadoop Distributed File System version: 3.x)**, **Apache Hive**, **Apache Drill**, **Apache Impala**, and **HBase**, while integrating GCP tools like **Cloud Composer**, **BigQuery**, and **Dataproc 5.9.3**, resulting in a 30% reduction in data storage costs and improved data accessibility.
- **Implemented real-time data streaming solutions with Apache Kafka 2.8**, enabling seamless capture and processing of streaming data from IoT devices and transactional systems.
- **Developed and maintained streaming analytics pipelines with technologies like Apache Flink 1.13**, allowing for real-time monitoring of key performance indicators and proactive decision-making.
- **Designed and automated end-to-end data pipelines using tools like Apache Airflow 2.1**, while programming in **Java** and **Python 3.8**, and leveraging **Apache Beam 2.5.6.0**, reducing manual effort and enhancing efficiency in data ingestion, transformation, and loading processes.
- **Optimized data workflows and processing pipelines with Apache Spark 3.1**, resulting in a 25% reduction in data processing time and improved data quality and consistency.
- **Developed predictive models using Python 3.9** libraries such as **scikit-learn 0.24** and **TensorFlow 2.6** to forecast sales, customer churn, and product demand, leading to a 15% increase in revenue and improved customer retention.
- **Implemented and optimized SQL 8.0.33** and **PLSQL 12.2** queries for data extraction, transformation, and loading processes, ensuring high performance and accuracy in data handling.
- **Utilized Teradata** for advanced data warehousing solutions, enabling efficient large-scale data management and analytics.
- **Developed and maintained UNIX** shell scripts for automating data processing tasks and system maintenance.
- **Leveraged BigQuery** for large-scale data analysis and storage, enhancing the ability to derive actionable insights from data.

- **Utilized strong Python** programming skills to develop, optimize, and maintain data processing and machine learning pipelines.
- **Worked with relational databases** and designed scalable data architectures ensuring data accessibility, storage, and scalability.
- **Implemented data management**, **data quality standards**, and **data governance** policies.
- **Applied statistical techniques** and developed machine learning models to derive business insights.
- **Leveraged Knowledge Discovery in Data (KDD) tools and applications like SQL**, **Oracle**, and **Apache Mahout** for advanced analytics.
- **Used MS Excel** for data analysis and reporting.
- **Engaged in data accessibility** and **scalability** practices to enhance overall data infrastructure.
- **Worked with @Scale** for handling large-scale data operations and ensuring efficient processing.

Environment:

Linux (Ubuntu 20.04) and **Windows Server OS**, **AWS**, and **Azure** cloud platforms. Utilizing tools like **Apache Hadoop**, **AWS S3**, **Amazon Redshift**, **Apache Spark (version: 3.1)**, **Python (NumPy 1.21, Pandas, scikit-learn (version: 0.24), TensorFlow (version: 2.6))**, **SQL 8.0.33**, **PLSQL 12.2**, **Jira**, **Slack**, **Jenkins**, **Apache Airflow 2.1**, **Apache Kafka 2.8**, **Cloud Composer**, **Unix Scripting**, **GCP**, **Dataproc 5.9.3**, **Apache Beam 2.5.6.0**, **Teradata**, **Oracle**, **Apache Mahout**, and **MS Excel**.

AirPrice **- Newark, NJ** **June 2022- -June 2023**

Job Title: Data Engineer
Project: Description:
This project involves utilizing historical flight data, current market trends, and various features such as departure time, destination, airline, and more to forecast future prices accurately. The development of a sophisticated flight fare prediction system, employing advanced machine learning algorithms and feature engineering techniques to accurately forecast ticket prices.

Responsibilities:

- Developed and deployed data processing pipelines using technologies like **Dataflow**, **Apache Beam 2.5.6.0**, and **Cloud Storage** to handle and analyze historical flight data efficiently.
- Leveraged **BigQuery** to perform complex queries and analyze flight data insights, aiding in accurate prediction of future ticket prices.
- Orchestrated workflow automation and task scheduling with **Cloud Composer 2.2.8**, ensuring timely execution of data processing tasks for flight fare prediction.
- Integrated **PubSub** and **GCP PubSub** for real-time data streaming and messaging, enabling capture and processing of streaming flight data for immediate fare updates.
- Utilized **Kafka 2.8**, **Pulsar**, and **RabbitMQ** as stream processing and messaging systems to handle real-time flight data and facilitate dynamic fare adjustments.
- Leveraged **Dataproc 5.9.3** for running **Apache Spark 3.1** tasks at scale, enabling efficient processing of large volumes of flight data to enhance prediction accuracy.
- Implemented flight fare prediction models using programming languages like **Java** and **Python 3.8**, incorporating machine learning algorithms and frameworks like **scikit-learn 0.24** and **TensorFlow 2.6** to forecast ticket prices accurately.
- Utilized **Python 3.9** for developing and optimizing data processing pipelines and machine learning models.
- Orchestrated data workflows with **GCP Composer 2.2.8**, enhancing automation and efficiency in data processing tasks.
- Performed complex data analysis and queries using **BigQuery** for deep insights into flight data.
- Automated data pipelines and workflow scheduling with **Apache Airflow 2.1**, ensuring efficient data processing.
- Implemented and optimized **SQL 8.0.32** queries for **data extraction, transformation, and loading processes**.

- Leveraged **Teradata 17.20** for advanced data warehousing solutions, enabling efficient large-scale data management and analytics.
- Developed and maintained **UNIX** shell scripts for automating data processing tasks and system maintenance.
- Architected and scaled relational database systems to provide high data accessibility, optimized storage solutions, and robust scalability.
- Established comprehensive **data management practices, ensuring data quality and governance** are upheld across all processes.
- Devised and implemented advanced statistical methods(**Linear Regression, Multivariate Analysis, Time Series, Bayesian Statistics, Bayesian Inference)** and machine learning models(**Random Forests, Support Vetor Machines)** to provide critical business insights.
- Utilized **Knowledge Discovery in Data** (KDD) tools and platforms such as **Oracle**, and **Apache Mahout** for performing complex data analytics.
- Executed detailed data analysis and reporting using **MS Excel** for clear and actionable insights.
- Enhanced data infrastructure through strategies focused on improving data accessibility and ensuring scalability.
- Managed and streamlined large-scale data operations using **@Scale** to maintain processing efficiency.

Environment:
The development environment for this project leveraged **Python 3.9** alongside a comprehensive suite of libraries including **Pandas (version: 1.3.3)**, **NumPy (version: 1.21.2)**, **scikit-learn (version: 0.24.2)**, **TensorFlow (version: 2.7.0)**, and **PyTorch (version: 1.9.1)** for robust data preprocessing, analysis, and model development, complemented by **Apache Spark** for distributed data processing, **Docker** for containerization, **Kubernetes** for orchestration, and **AWS** or **Azure** for scalable deployment. **Kafka**, **Pulsar**, **RabbitMQ**, **Apache Airflow**, **Teradata**, **SQL**, **GCP**, **BigQuery**, **Unix Scripting**, **Java**, **Hadoop**, **Hive**, **Spark**, **Drill**, **Impala**, **HBase**, **Data Management**, **Data Quality Standards**, **Data Governance**, **Data Accessibility**, **Data Storage**, **Data Scalability**, **Knowledge Discovery in Data (KDD) tools**, **Oracle**, **Apache Mahout**, **MS Excel**, and **statistical techniques**.

DataPulse: Newark,NJ **Sept 2021 – May 2022**
Job Title: Data Engineer- Newark, NJ
Description:
This project aimed to develop a tweet classification system to differentiate between expert and non-expert COVID-19-related content on Twitter, utilizing advanced machine learning techniques and addressing ethical considerations. By accurately categorizing tweets, the system aimed to enhance the reliability of information dissemination during the pandemic.

Responsibilities :

- **Data Collection and Streaming**: Orchestrated the development of robust data collection pipelines utilizing the **Twitter API** and **Tweepy (version: Latest)** library for real-time ingestion of COVID-19-related tweets, ensuring high throughput and low latency.
- **Text Preprocessing and Feature Extraction**: Implemented advanced text preprocessing techniques using **NLTK (version: Latest)** and **SpaCy (version: Latest)** for tokenization, stemming, and lemmatization, extracting relevant features such as sentiment scores and topic embeddings.
- **Machine Learning Model Development**: Collaborated closely with data scientists to build and optimize machine learning models using algorithms like **Naive Bayes**, **Support Vector Machines (SVM)**, and **Bidirectional Encoder Representations from Transformers (BERT)** for expert and non-expert tweet classification.
- **Model Training and Evaluation**: Designed and executed scalable model training pipelines using **Apache Spark (version: 3.1)** and **TensorFlow (version: 2.7.0)**, incorporating techniques like cross-validation and hyperparameter tuning to ensure optimal model performance.
- **Deployment and Integration**: Led the deployment of trained classification models into production environments using containerization tools like **Docker (version: Latest)** and **Kubernetes (version: Latest)**,

seamlessly integrating with real-time data streaming platforms such as **Apache Kafka (version: 2.8)** for efficient tweet classification.

Environment:

The environment for the Social Media (Twitter) COVID-19 Tweet Classification project utilized Python 3.9 along with key technologies including Twitter API, Tweepy, NLTK, SpaCy, TensorFlow (version: **2.7.0**), Scikit-learn (version: **0.24.2**), Docker, Kubernetes, Apache Kafka (version: **2.8**), Apache Spark (version: **3.1**), Prometheus, Grafana, and Jupyter Notebook for comprehensive data processing, machine learning model development, and monitoring.

Client: VitalHealthAnalytics - San Francisco, CA **Sept 2020 –June 2021**

Job Title: Data Engineer

Description:

The project aimed to develop a comprehensive healthcare data analytics platform to improve patient outcomes by leveraging advanced data engineering techniques and cloud-based solutions. The platform enabled real-time data integration and analysis, providing actionable insights for healthcare professionals.

Responsibilities:

- Developed robust ETL pipelines utilizing **Apache Airflow (version: 2.0)** and **GCP Composer** for the ingestion of healthcare data from various sources, ensuring seamless data flow and transformation using **Python (version: 3.9)**.
- Orchestrated the design and implementation of scalable data warehousing solutions using **Google BigQuery**, optimizing query performance and storage costs for large datasets.
- Created complex SQL queries and stored procedures for data extraction, transformation, and loading (ETL) processes, optimizing them for performance and accuracy on **Teradata** and **BigQuery** platforms.
- Employed **UNIX Shell Scripting** for automating data processing tasks, ensuring high data quality and consistency across the pipeline.
- Collaborated with data scientists to develop predictive models using **TensorFlow (version: 2.5)** and **Scikit-learn (version: 0.24.2)**, focusing on patient readmission risk and disease outbreak prediction.
- Implemented robust scheduling and monitoring solutions using **Prometheus** and **Grafana** for real-time insights into data pipeline performance and system health.
- Led the deployment of data processing and machine learning workflows in production environments, utilizing **Docker (version: 20.10)** for containerization and ensuring continuous integration and delivery (CI/CD) pipelines.

Environment:

The environment for the Healthcare Data Analytics project utilized **Python (version: 3.9)** along with key technologies including **Apache Airflow (version: 2.0)**, **GCP Composer**, **Google BigQuery**, **TensorFlow (version: 2.5)**, **Scikit-learn (version: 0.24.2)**, **Teradata**, **UNIX Shell Scripting**, **Docker (version: 20.10)**, **Prometheus**, and **Grafana**.

Client: Tech-Mahindra Business Services **September 2019 – July 2020**

Job Title: Data Engineer

Project: Implementation of Apriori and Brute Force Algorithms for Association Rule Mining

Description:

This project involved developing custom implementations of the Apriori and Brute Force algorithms for association rule mining without external libraries. Through meticulous design and implementation in Python, we aimed to gain deeper insights into the principles and challenges of association rule mining, culminating in thorough testing and validation using synthetic and real-world datasets.

Responsibilities:

- **Algorithm Design**: Conceptualized and designed custom implementations of the Apriori and Brute Force algorithms using **Python 3.9**.
- **Data Preprocessing**: Developed data preprocessing pipelines utilizing the latest versions of **Pandas (1.3.3)** and **NumPy (1.21.2)** for efficient parsing, cleaning, and structuring of transactional data.
- **Algorithm Implementation**: Implemented the Apriori algorithm with optimized data structures and algorithms, leveraging **TensorFlow (2.7.0)** for parallelization and efficient computation.
- **Brute Force Approach**: Designed and implemented a brute force approach using the latest versions of Python libraries, including **Pandas (1.3.3)** and **NumPy (1.21.2)**, for exploring all possible combinations of itemsets.
- **Performance Optimization**: Optimized algorithm performance with **Apache Spark (3.1)** for scalability and efficiency, employing techniques such as pruning and parallelization to reduce computational complexity and runtime.
- **Testing and Validation**: Conducted rigorous testing and validation of algorithms with the latest versions of Python libraries, ensuring correctness and accuracy of association rule mining results.
- **Documentation and Reporting**: Documented the design, implementation, and testing processes comprehensively, providing clear explanations and code comments using **Jupyter Notebook** for knowledge sharing and future maintenance.

Environment: In addition to Python 3.9, the project harnessed the power of **Pandas (version: 1.3.3)**, **NumPy (version: 1.21.2)**, **TensorFlow (version: 2.7.0)**, **Jupyter Notebook**, and **Excel** for algorithmic tasks,whileleveragingbig data technologies such as **Apache Spark (version: 3.1)** to handle large-scale data processing and analysis, ensuring robustness and scalability throughout the project lifecycle.

Client Name: Amazon, Pune, India **June 2018 -- September 2019**
Job Title: Data Engineer
Project: Natural Language Processing (NLP) Applications at Amazon
Project Description:
This project focuses on leveraging Natural Language Processing (NLP) techniques at Amazon, encompassing sentiment analysis of customer reviews and the implementation of chatbots for customer support, along with voice recognition in devices like Amazon Echo, aiming to enhance user experiences and optimize operational efficiency. Through the application of NLP algorithms, the project aims to extract valuable insights from textual and voice data, enabling personalized interactions, efficient support services, and seamless user interactions with Amazon's platforms and devices.

Responsibilities:

- In the capacity of leading the development of state-of-the-art NLP models for sentiment analysis, named entity recognition, and text summarization, leveraging cutting-edge techniques such as deep learning, transformers (e.g., BERT), and attention mechanisms, successful extraction of invaluable insights from extensive unstructured textual data was achieved.
- Moreover, taking charge of designing and optimizing chatbot systems for customer support ensured seamless interactions across Amazon's platforms. Integration of advanced NLP algorithms powered by libraries like NLTK (version: **3.6.2**), spaCy (version: **3.1.3**), and TensorFlow (version: **2.7.0**) for intent detection, dialogue management, and sentiment-aware responses significantly enhanced user experiences.
- Additionally, playing a pivotal role in enhancing voice recognition algorithms for devices like Amazon Echo involved utilizing NLP techniques such as automatic speech recognition (ASR) and natural language understanding (NLU), supported by libraries like PyTorch (version: **1.10.0**) and OpenNLP (version: **1.9.3**), thereby substantially improving accuracy, language understanding, and response generation, enriching user interactions.
- Furthermore, meticulous design and implementation of robust data pipelines for processing and analyzing large volumes of textual and voice data ensured seamless data preparation for NLP model training and inference.

Employment of technologies like Apache Spark (version: **3.1.2**) and Apache Kafka (version: **2.8.0**), alongside Python's Pandas (version: **1.3.3**) and NumPy (version: **1.21.2**), were pivotal in this endeavor.

- Moreover, comprehensive evaluation and optimization of NLP models, leveraging tools like scikit-learn (version: **0.24.2**) and Keras (version: **2.6.0**), and employing metrics such as precision, recall, and F1-score, consistently enhanced model accuracy and generalization.

Environment: The project environment featured **Python 3.9** along with cutting-edge technologies including **NLTK (version: 3.6.2)**, **spaCy (version: 3.1.3)**, **TensorFlow (version: 2.7.0)**, **PyTorch (version: 1.10.0)**, **Apache Spark (version: 3.1.2)**, **Apache Kafka (version: 2.8.0)**, **Pandas (version: 1.3.3)**, **NumPy (version: 1.21.2)**, **scikit-learn (version: 0.24.2)**, and **Keras (version: 2.6.0)**, facilitating advanced NLP model development and big data processing.

EDUCATION:

- **Master’s in Data Science**
- **Bachelor of Engineering in Computer Engineering**