**Kamal Shaik**

**Senior Data Engineer**

**Professional Summary:**

* **8 years** of experience as Data Engineer with executing data - driven solutions with adept knowledge on Data Analytics, Text Mining, Machine Learning (ML), Predictive Modelling, and Natural Language Processing (NLP) and Python.
* Expertise in Python Machine Learning/Deep Learning Modelling for multiple data science projects of full data science production cycle working with various Python libraries NumPy, Pandas, matplotlib, Sklearn, Beautiful soup, Pyecharts, PySpark, SparkSQL, OpenCV, TensorFlow and Keras.
* Proficient in Data Analysis, Data Migration, Data Profiling, Data Cleansing, Transformation, Integration, Data Import, and Data Export through the use of multiple ETL tools such as a Azure Data Factory, AWS Glue, Informatica power center.
* Skilled in identifying and assessing available machine learning and statistical analysis libraries, including regressors, classifiers, statistical tests, and clustering algorithms, and performed machine learning algorithms using Python.
* Experienced in developing and maintaining CI/CD pipelines for machine learning models using Azure DevOps.
* Expertise in developing end-to-end machine learning models using AWS SageMaker and AWS Lambda for real-time predictions.
* Proficient in working with different data formats including JSON and XML, and experienced in updating Python scripts to match training data with AWS Cloud Search databases for further classification.
* Developed highly scalable classifiers and tools by leveraging machine learning, Apache spark & deep learning.
* Experience in Migration from Teradata On-Prem to Cloud GCP - Snowflake Environment.
* Passionate about gleaning insightful information from massive data assets and developing a culture of sound, data-driven decision making
* Experience with Data Analytics, Data Reporting, Ad-hoc Reporting, Graphs, Scales, PivotTables
* Proficient in in working with various IDE’s such as Jupyter Notebook, PyCharm, Emacs, Spyder, Visual Studio and VSCode.
* Collaborated with cross-functional teams and worked on Agile methodologies, using Jira for project management and version control tools such as Git.
* Robust participation for functioning in fast-paced multi-tasking environment both independently and in the collaborative team.

**Technical Skills:**

|  |  |
| --- | --- |
| **Programming for DS** | Python(Pandas, NumPy, matplotlib, Sklearn, Beautiful soup, TensorFlow, Keras and), SQL, Hive, SparkSQL |
| **Software Methodologies** | SDLC - Waterfall, Agile, SCRUM |
| **Visualization Tools** | Tableau, Python - Matplotlib, Seaborn |
| **Analytical Skills** | Data & Quantitative analysis, Decision Analytics, Predictive Modelling, Data Driven decision making, Data mining and visualization tools, Machine learning (ML) algorithms, SQL Query Interpretations, Forecast, ETL (Extract, Transform, Load) |
| **Querying languages** | SQL, NO SQL, PostgreSQL, MySQL, Microsoft SQL, · Databases: MySQL, Oracle, Teradata, Snowflake, Big Query (GCP), |
| **IDE Tools** | PyCharm, Spyder, Eclipse, Visual Studio |
| **Defect Tracking** | JIRA, Git, and Version One |
| **Operating Systems** | Linux and windows |
| **Cloud Platforms** | Amazon Web Services (AWS), Microsoft Azure |

**Professional Experience:**

**Client: CVS – Providence, RI May 2021 - Till date**

**Role: Senior Data Engineer**

**Responsibilities:**

* Applied supervised machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest for predictive modelling on AWS, using libraries including pandas, NumPy, Seaborn, and scikit-learn.
* Implemented AWS Glue for data integration and ETL jobs to move data between different sources such as S3, DynamoDB, and RDS.
* Used AWS Comprehend service for natural language processing tasks such as sentiment analysis and entity recognition.
* Leveraged AWS Kinesis to collect, process, and analyze real-time streaming data.
* Developed and deployed end-to-end machine learning models using AWS SageMaker and AWS Lambda for real-time predictions.
* Implemented feature engineering techniques using Databricks to extract relevant features from raw data to improve model performance and conducted exploratory data analysis to ensure data quality and accuracy.
* Identified and assessed available machine learning and statistical analysis libraries (including regressors, classifiers, statistical tests, and clustering algorithms).
* Performed data cleaning and feature selection using MLLib package in Apache Spark and worked with deep learning frameworks such as TensorFlow.
* New user/service accounts set up through IAM, Dataset creations & granting access through Terraform in GCP.
* Designed and developed AWS Step Functions workflows to automate machine learning pipelines and orchestrate various services such as AWS Lambda, AWS Glue, and AWS SageMaker.
* Optimized machine learning models using AWS Elastic Inference to accelerate inference performance.
* Managed datasets and version control using AWS S3 and AWS CodeCommit.
* Ensured data security and compliance by implementing AWS Identity and Access Management (IAM) policies and encryption mechanisms.
* Used AWS EMR to perform data processing at scale and optimized performance by configuring cluster settings and tuning Spark configurations.
* Implemented end-to-end CI/CD pipelines for machine learning models using AWS CodePipeline, CodeBuild, and CodeDeploy.
* Used visualization tools like Amazon QuickSight and Tableau for data visualization and reporting.
* Configured Airflow connections to integrate with various data sources such as AWS S3, DynamoDB, and MySQL databases.
* Worked on different data formats such as JSON, XML and performed machine learning algorithms in Python.
* Updated Python scripts to match training data with our database stored in AWS Cloud Search, to be able to assign each document a response label for further classification.

**Environment:** Python (pandas, NumPy, SciPy, matplotlib, sci-kit-learn, NLTK), Tableau, Shell Scripting, AWS (SageMaker, Lambda, Glue, EMR, S3, Kinesis, Comprehend, Elastic Inference, QuickSight, CodePipeline, CodeBuild, CodeDeploy, CloudWatch, IAM), Airflow, GIT, Jira, Agile, Apache Spark, PySpark, Linux.

**Client: Wells Fargo – Charlotte, NC Dec 2019 – May 2021**

**Role: Data Engineer**

**Responsibilities:**

* Used Python to transform data from nested JSON, and various formats into usable data.
* Derived data from relational databases to perform complex data manipulations and conducted extensive data checks to ensure data quality.
* Performed data modelling operations using Power BI, Pandas, and SQL.
* Researched extensively on the nature of the customers and designed multiple models to perfectly fit the necessity of the client and Performed Extensive Behavioral modelling and Customer Segmentation to discover behavior patterns of customers by using K-means Clustering.
* Participated in features engineering such as feature generating, PCA, Feature normalization and label encoding with Scikit-learn pre-processing.
* Designed and implemented machine learning workflows using Azure Machine Learning, including data preparation, model training, and model deployment.
* Managed and tracked machine learning experiments using MLflow, an open-source platform for managing the end-to-end machine learning lifecycle.
* Configured and automated the deployment of machine learning models to Azure Kubernetes Service (AKS) using Azure DevOps Release pipelines.
* Used Azure DevOps to manage project artifacts, including source code, models, and datasets.
* Implemented and maintained CI/CD pipeline using Azure DevOps to ensure the quality and timely deployment of machine learning models.
* Used Azure services like Azure Synapse Analytics (formerly SQL Data Warehouse) for data warehousing, Azure Event Hubs for data streaming, and Azure Databricks for data processing and analysis.
* Used Azure Functions to build serverless data processing pipelines and Azure Key Vault for data security and encryption.
* Worked with Azure services such as Azure Blob Storage, Azure SQL Database, and Azure Cosmos DB to store and manage machine learning data and metadata.
* Provided technical guidance and support to team members on Azure DevOps, MLflow, and MLOps practices and tools.
* Worked on developing **ETL** processes (Data Stage Open Studio) to load data from multiple data sources to **HDFS** using **FLUME** and **SQOOP**, and **Big Query (GCP)** performed structural modifications using **Map Reduce**, **HIVE**.
* Installed and configured Hadoop Map Reduce jobs, HDFS and developed multiple MapReduce jobs and used different UDF's for data cleaning and processing.

**Environment:** Python, Microsoft Azure, CI/CD, DevOps, MLflow, MLOps, Hadoop MapReduce, NumPy, Pandas, Databricks, Scikit-learn, Power BI, SQL, Agile, Windows.

**Client: First Data Corporation – Atlanta, GA Jan 2017 – Nov 2019**

**Role: Data Engineer**

**Responsibilities:**

* Implemented application of various machine learning algorithms and statistical modelling like Decision Tree, Naive Bayes, Logistic Regression and Linear Regression using Python to determine the accuracy rate of each model.
* Developed collaborative filtering-based recommendation engines using Python and R to recommend retail products.
* Analysed data trends using packages like NumPy, Pandas and Matplotlib in Python.
* Created sentiment analysis model and complex query model data using Hadoop ecosystem, HiveQL, Impala, and regular expression.
* Building prediction models using Linear and Ridge Regression, for predicting future customers based on historical data.
* Developed the model with 3 million data points from historical data and evaluated the model with F-score and adjusted R-squared measure.
* Design Develop and test ETL Processes in **AWS Glue Teradata, GCP-**  to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Performed Exploratory Data Analysis and Data Visualizations.
* Text analytics on review data machine learning technique in python using NLTK.
* Generated ETL mappings, sessions and workflows based on business user requirements to stack data from source files, RDBMS tables to target tables.
* Used Meta data tool for importing metadata from repository, new job categories and creating new data elements.
* Used text-mining process of reviews to determine customers' concentrations.
* Designed and implemented a probabilistic churn prediction model with 80k customer data to predict the probability of customer churn out using Logistic Regression in Python.
* Data wrangling to clean, transform and reshape the data utilizing NumPy and Pandas library.
* Designed Tableau bar graphs, scattered plots, and geographical maps to create detailed level summary reports and dashboards.

**Environment**: Python, NumPy, Pandas, Matplotlib, NLTK, Tableau, HiveQL, Agile, GITHUB, Windows.

**IBM – India August 2015 – December 2016**

**Role**: **Data Engineer**

**Responsibilities**:

* Involved in daily Scrum meetings and follows agile methodology.
* Wrote Python routines to log into the websites and fetch data for selected options. Used Python modules such as requests, urllib, urllib2 for web crawling.
* Used other packages such as beautiful soup for data parsing.
* Involved in developing views and templates with Python and Django view controller and templating language to create a user-friendly website interface.
* Worked on writing and as well as read data from csv and excel file formats.
* Worked on analysing SQL scripts and designed the solution for the process using PySpark.
* Responsible for search engine optimization (SEO) to improve the visibility of the website.
* Developed automated testing framework for command-line based tests on Linux using Objected Oriented Perl and for selenium-based tests using Python.
* Created database using MySQL, wrote several queries to extract data from database.
* Setup automated cron jobs to upload data into database, generate graphs, bar charts, upload these charts to wiki, and backup the database.
* Wrote scripts in Python for extracting data from HTML file.
* Used Git as version control tool, and Maven as Build tool.
* Responsible for debugging and troubleshooting the web application using JIRA Bug Tracking Tool.
* Extracting data from multiple source systems using Oracle SQL.
* Used SQL for Data Analysis and to understand the behaviour of data.
* Performed Data Analysis and presented findings and issues in Microsoft Excel.
* Developed process flow documents, questionnaires and training documents for the team.
* Involved in Index operations (rebuild & reorganize) and database maintenance.
* Written SQL Scripts for Test Cases and Test Scenarios.
* Created various SQL Scripts for verification of the required functionality.

**Environment**: Python, PySpark, SQL, JIRA, Windows