**DHANAJAY. Y**

**Sr Data Scientist**

**Email: dyedavilli@gmail.com**

**Phone: 7329077018**

**LinkedIn:** [linkedin.com/in/dhananjay-yedavilli-503374298](https://www.linkedin.com/in/dhananjay-yedavilli-503374298)

**Professional Summary:**

* Around 6 years of experience in software design, development, and implementation across various business applications.
* Specializing in Machine Learning and Deep Learning, I'm proficient in Python, utilizing libraries like NLTK, Scikit-Learn, TensorFlow, and more.
* Successfully built and deployed highly scalable ML and DL models, incorporating ensemble algorithms for robust solutions.
* Possess strong functional knowledge in Health and Data Analytics, applying it effectively in projects.
* Implemented a framework for distributing TensorFlow ML model training, inference, and hyperparameter optimization on Spark clusters.
* Proficient in Pyspark, I've developed ML models on large datasets in distributed environments.
* Applied Geographical Information Systems (GIS) to categorize unstructured data based on location features.
* Contributions include models for Contextual Entity Recognition, aiding in information extraction from large document collections.
* Adeptive at creating and deploying browsable APIs for real-time model inference via REST using Django, including A/B testing features.
* Led the re-classification of a significantly large number of categories in an Enterprise Document Repository.
* Designed and delivered a Restful API that interacts with HBase Database, handling CRUD operations and diverse queries.
* Hands-on experience in H2O (PySparkling), I've implemented machine learning models effectively.
* Trained Data Scientists to directly implement models in PySpark, resulting in significant cost savings.
* Expertise extends to Micro Services and various Machine Learning Topologies.
* Skilled in building Ensemble models with minimal to no reproducibility errors for complex classification tasks.
* Proficient in Flask, Groovy, SQL, Docker, and Kubernetes, I'm well-versed in various tech stacks.
* Worked extensively with ElasticSearch, Kibana, Solr, and other enterprise search engines.
* Known for excellent client interaction skills, demonstrating a dedicated approach to delivering high-quality solutions.
* Experience encompasses AWS (S3, EC2), Apache Spark, Scala, and Principal Component Analysis for propensity models.
* Utilized Azure Databricks as a fast, collaborative spark-based platform on Azure.
* Specialize in Data Engineering, handling Database Development, ETL, and Big Data technologies.
* Skilled in managing ETL workflows with Apache Airflow, I excel in performance tuning SQL queries and have experience with NoSQL databases like HBase.
* Proficient in GCP and AWS, I've built data ingestion pipelines, designed data models, and led end-to-end data migration projects.
* Expertise includes Business Intelligence Solutions with Microsoft SQL Server, covering SSIS, SSRS, and SSAS.
* Successfully implemented Continuous Delivery pipelines using Docker, GitHub, and AWS, and have experience with various data analytics and Big Data services.
* Passionate about staying updated with the latest technologies and advancements in Google Cloud Platform (GCP).

**Technical Skills:**

| **Environment** | **Tools and Technologies** |
| --- | --- |
| Programming Languages: | Python, Machine Learning, R, Natural Language Programming(NLP), NoSQL, MySQL, PySpark. |
| Python and Machine Learning Tools: | Python, NLTK (Natural Language Toolkit), Scikit-Learn, TensorFlow, Pandas, NumPy, Pyspark, H2O (PySparkling), Keras, XGBoost. |
| Distributed Computing and Big Data Tools: | PySpark, Spark, Elasticsearch, Apache Kafka, Apache Flink, HDFS (Hadoop Distributed File System), AWS (Amazon Web Services), EMR (Elastic MapReduce), Apache Zookeeper, Yet Another Resource Negotiator (YARN), Sqoop, Redshift, HBase, Cassandra, MongoDB, Kubernetes, Docker, Django, Snowflake, BigQuery, Dataflow, Cloud Shell, AWS SageMaker, Apache Airflow, Azure Databricks. |
| Data Analysis, Visualization, and Business Intelligence Tools: | Geographical Information Systems (GIS), Django, SQL, Elasticsearch, Kibana, Solr, Flask, Groovy, SSIS (SQL Server Integration Services), SSRS (SQL Server Reporting Services), SSAS (SQL Server Analysis Services), Tableau, Matplotlib. |
| Data Preprocessing and NLP Tools: | CoreNLP, Contract intelligence tool, BERT architecture. |
| Version Control and Collaboration Tools: | GitHub, Slack, Jira, Confluence, Jupyter Notebooks. |
| Cloud Computing Platforms: | Google Cloud Platform (GCP), Amazon Web Services (AWS), Azure. |
| Other Data Science and Analytics Tools: | Principal Component Analysism, Collaborative filtering, Google Optimize, Market basket analysis. |

**Educational Details:**

**Bachelors:** Computer Science Engineering, in 2017 from GITAM University, Hyderabad, India.

**Masters:** Computer Sciences, in 2022 from Case Western Reserve University, Cleveland, OH.

**Professional Experience:**

**Client: C & S Wholesale Grocers, Edison, NJ. Sep 2022 - Till Date**

**Role: Sr. Data Scientist.**

**Responsibilities:**

* Led a cross-functional team in the development of a robust ML Platform, ensuring regular model scoring and automated audience delivery.
* Spearheaded the transformation of existing Auto/Card and Home Lending Models into Distributed Models, showcasing proficiency in managing complex model transitions.
* Collaborated with stakeholders to define project scope and requirements, taking charge of project coordination and ensuring successful implementation of a cloud-based data lake solution using AWS services.
* Collaborated to build the ML Platform, ensuring regular model scoring and automated audience delivery.
* Performed data imputation using the Scikit-learn package in Python. Worked on different data formats such as JSON, and XML to be ingested into ML models in Python.
* Developed Python package useful in distributing Machine Learning model training and inference on Spark cluster.
* Developed new DBT models and pipelines for processing marketing and lending page events data with low latency.
* Developed Airflow DAGs for automated ingestion, delivery of data files for improved data quality, and availability using Python & SQL.
* Created data processing pipelines for impression, and clickstream data in unstructured format into consumable normalized tables.
* Collaborated to build the ML Platform to support the execution of model scoring with regular cadence and automated audience delivery.
* The library helps in distributing Tensorflow, PyTorch, XGBoost, LightGBM, and GLM Models on the Spark cluster.
* Implemented methods to enable Hyper Parameter optimization of the models in the distributed environment which drastically cut down the time and effort needed to find the best choice of hyperparameters.
* Enabling Visualization of hyper-parameters in Tensor Board H-Params Dashboard, using model/trail parallelization is the key to this utility.
* Enabled visualization of the model graphs, training & evaluation metrics, individual worker progress, loss curves, and more useful information during distributed training, in real-time on the Tensor Board dashboard.
* Developed the DistML Framework to be highly scalable and efficient. Performance Benchmarking is done on datasets ranging from 1MM to 1 billion. Linear scaling and efficient model training are observed.
* Converted the existing Auto/Card and Home Lending Models into Distributed Models.
* Implemented a template on how to run H2O models on spark cluster using PySparkling. Productionized the Trained models using H2O MOJO.
* Drive X-sell marketing optimization and measurement of different product conversions across channels and product mix.
* Used MAE, RMSE score, Confusion matrix, F-Score, AUC/ROC, Cross-validation, and A/B testing to evaluate model performance in both the simulated environment and the real world.
* Collaborated with stakeholders to define project scope and gather requirements.
* Designed and developed a cloud-based data lake solution using AWS services (e.g., Apache Spark, Hadoop, Kafka).
* Collaborated with cross-functional teams to design data models, develop data pipelines, and implement data transformations.
* Implemented Apache Spark and Hadoop to build the data lake infrastructure.
* Developed and deployed data integration solutions and advanced analytics using Apache Hadoop.
* Created complex Hive queries for extracting and persisting data from diverse sources into HDFS.
* Specialized in Big Data solutions, with a focus on pattern matching and predictive modeling.
* Utilized Sqoop to import/export data between Oracle and Hadoop in Parquet format.
* Developed Spark and Hive jobs for data summarization and transformation.
* Implemented data reconciliation processes to ensure Elastic Search index document counts matched source records.
* Leveraged Scala and Spark-SQL/Streaming for high-speed data processing.
* Streamlined real-time data processing by integrating Kafka with Spark for dynamic price surging using machine learning algorithms.
* Developed Scala-based Spark applications for data cleansing, event enrichment, aggregation, de-normalization, and preparation for machine learning and reporting teams.
* Ensured error tolerance and performance optimization of Spark applications.
* Proficient in data ingestion to various Azure Services including Azure Data Lake, Azure Storage, Azure SQL, and Azure DW, with subsequent data processing in Azure Databricks.
* Designed scalable, end-to-end solutions on the Azure platform using Azure Data Factory (ADF), Azure Databricks, and Azure SQL Database.
* Ingested real-time streaming data into Azure Databricks using Azure Event Hubs or Azure Event Grid.
* Extensively worked with Sqoop for data import from Oracle databases.
* Operated in AWS cloud environments, including EMR clusters and S3 storage.
* Created batch scripts for data retrieval from AWS S3 and applied transformations using the Spark framework in Scala.
* Designed Hive tables, implemented partitioning, dynamic partitions, and buckets.
* Collaborated closely with data scientists and business analysts to translate requirements into efficient Flink data processing pipelines.
* Utilized SSIS for creating interfaces between front-end applications and SQL Server databases.
* Managed cluster maintenance, Hadoop log files, and AWS cloud infrastructure.
* Conducted performance tuning and optimization of ETL processes, including SQL queries and ODI mappings.
* Documented the solution and provided end-user training.

**Environment:** Scikit-learn, Python, Spark, DBT, Airflow, JSON, XML, Tensorflow, PyTorch, XGBoost, LightGBM, Azure Databricks, GLM, Tensor Board, H2O MOJO, AWS (Amazon Web Services), Apache Spark, Databricks, Hadoop, Kafka, EMR (Elastic MapReduce), S3 (Simple Storage Service), Hive, Sqoop, Scala, Kafka, Flink, SSIS (SQL Server Integration Services), ODI (Oracle Data Integrator).

**Client: Elavon Inc., Atlanta, GA. Jan 2021 – Dec 2021**

**Role: Sr. Data Scientist.**

**Responsibilities:**

* Collaborated on building an ML Platform for model scoring and automated audience delivery.
* Conducted data imputation using Scikit-learn in Python, handling JSON and XML formats for ML model ingestion.
* Created a Python package for distributing ML model training and inference on a Spark cluster.
* Developed DBT models and pipelines for processing marketing and lending page events data with low latency.
* Implemented Airflow DAGs for automated data ingestion and delivery, enhancing data quality and availability.
* Established data processing pipelines for unstructured impression and clickstream data, transforming them into normalized tables.
* Designed a utility library for distributing Tensorflow, PyTorch, XGBoost, LightGBM, and GLM Models on a Spark cluster.
* Implemented methods for Hyper Parameter optimization in a distributed environment, improving efficiency in model training.
* Enabled visualization of hyper-parameters and model metrics in Tensor Board H-Params Dashboard.
* Developed the DistML Framework for highly scalable and efficient model training, benchmarked on datasets of varying sizes.
* Converted existing Auto/Card and Home Lending Models into Distributed Models.
* Created a template for running H2O models on a Spark cluster using PySparkling and productionized trained models using H2O MOJO.
* Led X-sell marketing optimization efforts and measured product conversions across channels and product mix.
* Utilized various evaluation metrics such as MAE, RMSE, Confusion matrix, F-Score, AUC/ROC, Cross validation, and A/B testing to assess model performance.
* Worked as a Hadoop Developer, responsible for managing clusters and building scalable distributed data solutions.
* Developed custom RDDs in PySpark for data transformations and actions on RDDs.
* Used Spark API over Cloudera Hadoop YARN for analytics on data in Hive.
* Conducted performance tuning of Spark jobs, leveraging caching and cluster advantages.
* Wrote Spark scripts using Scala Shell commands, as needed.
* Scheduled Oozie workflows to execute multiple Hive and Pig jobs.
* Worked with various file formats including Text, Sequence files, Avro, ORC, and Parquet.
* Converted Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Leveraged Spark API over Hadoop YARN for data analytics using Hive.

**Environment:** Scikit-learn, Python, JSON, XML, Spark, DBT (Data Build Tool), Airflow, Tensorflow, PyTorch, XGBoost, LightGBM, GLM (Generalized Linear Models), H2O, H2O MOJO, Tensor Board, DistML Framework, Hadoop, PySpark, Cloudera Hadoop YARN, Scala Shell, Oozie, Text, Sequence files, Avro, ORC, Parquet, Apache Hive, Apache Pig.

**Client: Express Scripts, St. Louis, MO. Feb 2020 – Dec 2020**

**Role: Data Scientist.**

**Responsibilities:**

* Created a robust rule hierarchy for extracting relevant information from diverse types of documents.
* Implemented automated validation using advanced NLP string matching algorithms, ensuring the accuracy of the extracted data.
* Developed a module for cleaning data and generating training datasets. This involved employing Scikit-learn's preprocessing techniques to enhance data quality.
* Implemented end-to-end pipelines, making extensive use of Amazon Web Services (AWS) Elastic MapReduce (EMR) for efficient data processing.
* Deployed and managed files in AWS Simple Storage Service (S3) buckets via Amazon Elastic Compute Cloud (EC2) clusters, optimizing data handling.
* Actively contributed to the production data pipelines, enhancing the decision-making process and overall solution effectiveness.
* Leveraged a suite of cutting-edge technologies including Apache Spark, Hadoop Distributed File System (HDFS), Apache Hive, Apache Zookeeper, and Yet Another Resource Negotiator (YARN) for developing enterprise solutions.
* Conducted data ingestion into AWS S3 using Sqoop and performed crucial transformations with Hive, streamlining the data processing workflow.
* Successfully migrated on-premises MS SQL databases to AWS Redshift using AWS Database Migration Service. Additionally, utilized SQL and Python for creating ETL scripts.
* Designed a comprehensive security framework to safeguard S3 access, employing AWS Lambda functions and DynamoDB for fine-grained access control.
* Utilized Apache Airflow and Databricks Jobs for efficient task scheduling and orchestration. Also, implemented data quality checks and validations with Databricks.
* Implemented real-time data processing solutions with Apache Flink, enhancing the analysis of streaming data.
* Ventured into Azure Databricks, exploring data frame creation and business analysis with Apache Spark. Imported and integrated data into various Azure services.
* Developed ETL pipelines using tools like SQL Server Integration Services (SSIS) and Data Transformation Services (DTS) Packages, ensuring seamless data flow.
* Designed and implemented security frameworks for precise access control to S3 objects, using AWS Lambda and DynamoDB.
* Managed streaming systems processing over 100 million data packets daily, utilizing tools like Streamsets, Kinesis streams, and API gateways.
* Worked extensively with various AWS databases, including Elastic Cache, HBase, Cassandra, and MongoDB, for efficient data modeling and optimization.
* Employed technologies like Apache Spark, EMR, and Hive for large dataset transfers and computations, ensuring optimal performance.
* Used Kubernetes for deploying, scaling, and managing Docker containers, streamlining the deployment process.
* Orchestrated seamless data copying between AWS S3 and Snowflake, interfacing with SQL Workbench for effective data operations.
* Demonstrated proficiency in utilizing Google Cloud Platform (GCP) services, including BigQuery, Dataflow, and Cloud Shell, for various tasks.
* Developed an unsupervised multi-stage image classification model for categorizing document images, significantly improving document classification efficiency and searchability.
* Implemented a module for enhancing image quality through super resolution, contributing to the accuracy of the image classification process.
* Developed an advanced text parser capable of automatically detecting and extracting text from over 50 different file types, streamlining the document processing workflow.
* Designed a RESTful API for performing CRUD operations on an HBase database integrated with the Elastic Search engine, providing an intuitive interface for data retrieval and manipulation.

**Environment:** Apache Spark, Hadoop (HDFS), Hive, Apache Zookeeper, YARN, Sqoop, EMR, Amazon Simple Storage Service (S3), Amazon Elastic Compute Cloud (EC2), AWS Database Migration Service, SQL, Python, Scikit-learn, NLP (Natural Language Processing), AWS Lambda, DynamoDB, Apache Airflow, Databricks, Apache Flink, Azure Databricks, Microsoft SQL Server, SQL Server Integration Services (SSIS), Data Transformation Services (DTS) Packages, Elastic Cache, HBase, Cassandra, MongoDB, Kubernetes, Docker.

**Client: Bank Of America, Charlotte, NC. Sep 2018 – Jan 2020**

**Role: Data Scientist.**

**Responsibilities:**

* Developed and deployed a model based on the BERT architecture for contextual entity recognition with their semantic relations intact.
* This model enabled me to extract relational data from contract agreements related to the LIBOR transition.
* Using BERT allowed me to recognize entities within the context of the agreements, such as names, dates, and locations.
* By preserving the semantic relationships between the entities, I was able to extract useful data related to the LIBOR transition from the agreements.
* Built a data source by converting completely unstructured data from documents into structured data using a contract intelligence tool.
* Intelligence tool helped to extract information from unstructured documents such as contract agreements and turn it into structured data that could be easily analyzed.
* Using the Intelligence tool to structure the data, I was able to gain valuable insights that would have been difficult to obtain from the unstructured documents.
* This process allowed me to better understand the data contained in the documents and make more informed decisions based on that information.
* Ingested data from the DataLake into several downstream applications, which allowed the data to be used in multiple ways throughout the organization.
* The data from the DataLake played a crucial role in enabling another noteworthy project - an enterprise search engine - by providing a rich source of data that could be indexed and searched.
* Built an attention-based encoder-decoder model for a question-answering system to extract answers and classify contracts tied to LIBOR.
* Able to achieve test accuracy scores of up to 95%, which was a significant improvement over previous methods used to classify these contracts.
* Implemented streamline the process of identifying contracts tied to LIBOR and allowed us to more efficiently manage the transition away from that benchmark.
* Leading to the effort to classify and tag approximately 50 million documents as LIBOR Tied Agreements, understood how criticalty to our overall strategy for managing the transition away from the LIBOR benchmark.
* Identifying and properly categorized to identify potential risks and opportunities related to agreements.
* Led to proactive approach to managing the transition and to mitigate potential risks by identifying opportunities for growth and optimization.

**Environment:** Keras, Scikit Learn, Python, NLTK, CoreNLP, Pandas, Numpy, BERT architecture, Contract intelligence tool, Attention-based encoder-decoder model, DataLake, Downstream applications.

**Client: Amazon, HYD- India. May 2017 – Aug 2018**

**Role: Data Scientist.**

1. **Project Name: Amazon Pay**

**Description:** Aimed to enhance personalized marketing strategies, optimize product recommendations, and implement fraud detection within a large financial dataset.

**Responsibilities:**

* Analyzed extensive customer data for personalized marketing strategies.
* Extracted data using SQL from databases.
* Cleaned and prepared data with Python (pandas, NumPy) and R (dplyr).
* Conducted exploratory data analysis and visualized segments with Tableau.
* Built customer segmentation models with Python (scikit-learn, TensorFlow).
* Developed a recommendation system using collaborative filtering (TensorFlow).
* Conducted A/B tests using Google Optimize.
* Communicated findings through Slack, Jira, and Confluence.
* Collaborated closely with product managers to align with business requirements.
* Documented process and models in Jupyter Notebooks.
* Ensured GDPR compliance for data privacy.
* Stayed updated with trends through research papers, courses, and forums.
* Developed a fraud detection model for a large financial dataset.
* Cleaned data using Python (pandas, NumPy) to handle missing values and outliers.
* Optimized models with Python (scikit-learn, XGBoost).
* Deployed the model using AWS SageMaker for real-time detection.
* Implemented automation with Apache Airflow for data pipeline and retraining.
* Collaborated with software engineers and DevOps teams using Slack and Jira.
* Documented the entire process and findings in Jupyter Notebooks.
* Ensured ethical compliance with AI guidelines and privacy impact assessments.
* Stayed updated with fraud detection techniques through conferences and workshops.
* Identified purchase behavior patterns for optimized product recommendations.
* Preprocessed and cleaned transaction data using Python (pandas) and Excel.
* Conducted advanced statistical analysis and hypothesis testing with R and Python (statsmodels, SciPy).
* Created visualizations using Tableau and Matplotlib.
* Deployed association rule mining with Apache Spark MLlib for scalability.
* Collaborated closely with business analysts and product managers to understand requirements.
* Documented the entire analysis process and insights using Jupyter Notebooks and Google Docs.
* Engaged in online courses and webinars to stay updated with market basket analysis techniques.

**Environment:** SQL, Python, R, Pandas, NumPy, Scikit-learn, TensorFlow, XGBoost, R(dplyr), Excel, Tableau, Matplotlib, Collaborative Filtering, Google Optimize, AWS SageMaker, Apache Airflow, Slack, Jira, Confluence, Jupyter Notebooks, Google Docs, GDPR Compliance, Apache Spark MLlib.

**Client: Amazon, HYD- India.**

**Role: Data Scientist.**

1. **Project Name: Rosetta**

**Description:** Machine Learning Language Translation tool used as an internal LLM tool, designed and deployed for Amazon customer Service.

**Responsibilities:**

* Tasks included training the LLM with prompt engineering and testing the Rosetta tool, sending feedback, observations, and analyzed data to the backend team.
* Used MySQL for querying and Tableau for Visualizations.
* Used PySpark to build ETL pipelines.
* Developed text preprocessing pipelines to handle tasks such as tokenization, stop-word removal, and stemming/lemmatization, improving the quality of NLP model.
* Experience with LLMs, including training/fine-tuning, prompt engineering, RLHF, performance evaluation and cost analysis.
* Experience with NLP techniques such as named entity recognition, relationship extraction, document classification, document summarization, topic modeling, machine translation, sentiment analysis, dialogue systems.
* Analyze model performance, and explore zero-shot/few-shot label generation to augment or supersede iterating with manual labeling resources to curate and refine training sets to improve model performance
* Collaborate with ML Ops and Data Science Engineers to deploy datasets, truthsets, models, pipelines, training and inference code to cloud based model registry
* Effectively communicate and present deliverables and solutions to teams, stakeholders, and executives
* Improved model efficiency by 80% by the end of one year while the tool was being deployed live to Amazon Customer Service to translate from multiple languages in multiple countries.
* Familiarity with LLMs and GenAI models such as, LLaMA, Falcon, GPT\*, BLIP, CLIP, etc.
* Handled customer issues, created tickets for issues, bugs & features regarding Prime Video, Kindle, Alexa & Fire TV Stick on osTicket .
* Analytics experience with Price forecasting and Retail Analysis
* Visualization experience with Tableau for Visualization
* Used Advanced Excel for Excel Visualizations and Data augmentation with pivot tables.

**Environment:** Python, SQL (MySQL), Amazon Redshift, Advanced Excel, PySpark, Tableau, LLMs, Prompt engineering for LLMs, RLHF, Named Entity Recognition, Relationship Extraction, Document Classification and Summarization, Topic Modeling, Machine Translation, Sentiment Analysis, Dialogue Systems, ML Ops, datasets, truth sets, models, osTicket, LLaMA, Falcon, GPT\*, BLIP, CLIP, Other GenAI models.