Abhijeet Singh
Halifax, NS

Career Objective:

Aim to contribute to an innovative, data-driven organization by leveraging my expertise in Machine Learning (AI) and data engineering to address complex projects and propel business growth. Committed to delivering impactful solutions while continuously enhancing my technical skills and professional development.

Skills and Abilities:

- Machine Learning
- Python (pandas, NumPy, matplotlib, scikit-learn)
- MLOps
- Airflow
- Big data Management
- Dbt
- GCP
- AWS
- Spark
- Data Mining and Data Warehousing
- BigQuery
- Tableau
- Looker
- Alteryx
- NLP
- Javascript
- LangChain
- DevOps
- GIT
- Requirement Analysis
- Prompt Engineering
- MySQL
- Oracle
- Postgres

Education:

- **Master, Computer Science (ECommerce), Machine Learning:** Dalhousie University, Halifax
  Duration: September 2019 – May 2021
- **Bachelor of Technology: Computer Science and Engineering:** AKTU University, India, Duration: 2014 – 2018

Personal Projects:

Quora Question pair similarity - The objective of the project was to identify whether a given pair of Quora questions were semantically duplicate

- Developed an NLP-powered binary classification system using Pandas and Scikit-learn to identify duplicate Quora question pairs from a dataset of over 400,000 entries.
- Achieved significant performance improvements (log loss reduction from 0.88 to 0.34) by engineering features like TF-IDF and Word2Vec, and through hyperparameter-tuned models, with XGBoost reaching 84% precision and 90% recall in duplicate detection
- Evaluated model performance using log loss and confusion matrix, ensuring a robust classification system with no overfitting (train/test log loss: 0.34), enhancing Quora's duplicate detection efficiency.

Sentiment Classification - The Objective of this Machine Learning project was to perform the sentiment classification on the online reviews (Yelp Dataset)

LinkedIn | GitHub |
Phone: 782-446-6007
Email: abhijeetsingh0727@gmail.com

Profile:

Results-driven Data professional with 5+ years of experience designing and delivering scalable, cloud-native data solutions. Skilled in building end-to-end **machine learning** pipelines, **advanced analytics** models, and **data-driven architectures** on cloud platforms like GCP & AWS. Expertise includes data science workflows, data modeling, data warehousing, and ETL/ELT pipeline development for large-scale, high-volume environments. Proven ability to collaborate with cross-functional teams to enable predictive analytics, enhance business intelligence (BI) capabilities, and drive data-driven decision-making across organizations:

- Designed and implemented end to end **machine learning** solution
- Led the design and deployment of **data warehouse solutions**, including data integration, transformation, and exploration, ensuring robust data architectures that meet business needs.
- Developed **scalable data pipelines** and engineered cloud-based solutions to support advanced analytics, using technologies like **Apache Beam** on GCP Dataflow, Data Fusion**,** Cloud Functions, **SQL**, **Python** and **ETL tools**.
- Conducted **exploratory data analysis (EDA)** and **feature engineering** to optimize machine learning model performance, improving forecast accuracy and reducing model drift.
- Championed best practices in **ML model retraining pipelines**, **CI/CD for data pipelines**
- Collaborated closely with data scientists, engineers, and business stakeholders to translate complex analytical requirements into scalable, production-grade data solutions.

Professional Experience:

Bell Canada, Halifax, NS

Data Engineer II

Jan 2024- Present

Telecom Hardware upgrade fraud detection:

Collaborated closely with architects to design, develop, and implement a machine learning–based fraud detection model to address telecom Hardware Upgrade (HUG) fraud. Engaged with domain experts and business stakeholders to gain in-depth understanding of critical features, operational workflows, and fraud patterns, ensuring the solution was highly aligned with real-world scenarios and business needs:

- Processed over 15 million web session records using Google Cloud Dataproc and PySpark for distributed stream processing, performing advanced ETL workflows, data cleaning, and feature engineering. Engineered 50+ predictive features (e.g., user behavior, device metadata) using Python (pandas, NumPy, scikit-learn) and stored optimized datasets in BigQuery, enabling real-time analytics for fraud detection.
- Developed and optimized supervised machine learning models using Vertex AI and Python frameworks (scikit-learn, TensorFlow, XGBoost) to classify telecom hardware upgrades as fraudulent or authentic, achieving an F1 score of 0.92. Conducted correlation analysis and model comparisons (gradient-boosted trees vs. neural networks) using precision, recall, and AUC, ensuring high-performance quantitative models for real-time insights.
- Automated model deployment and batch-to-real-time prediction pipelines using Vertex AI and Cloud Composer, integrating Cloud Functions for SHAP-based interpretability visualizations stored in Cloud Storage. Built scalable frameworks to transition from batch processing to real-time fraud detection, collaborating with cross-functional teams to align technical solutions with business objectives
- Delivered over $1 million in annual savings by detecting fraudulent activities, with model performance monitored via Vertex AI Model Registry. Communicated actionable insights to stakeholders using feature importance visualizations, ensuring alignment on deliverables and enhancing analytics capabilities through GCP-native, high-quality data engineering solutions.

Scalable GCP Data Pipelines for Legacy Migration:

Responsible for designing, developing, and managing end-to-end data pipelines for customer segmentation, ensuring seamless integration with finance billing workflows

- The process involved significant data preprocessing, including removing noise like contractions and stop words, followed by exploratory data analysis to understand data characteristics.
- Advanced feature engineering using SpaCy was employed to convert text into meaningful vectors. Recognizing a class imbalance in the sentiment labels, an oversampling technique was implemented to ensure the model was trained on a balanced representation of positive and negative reviews.
- The Support Vector Machine (SVM) emerged as the best-performing model, achieving a strong F1 Score of 90.08%, indicating its effectiveness in accurately classifying the sentiment of the Yelp reviews.
- Evaluated models with multi-class log loss and confusion matrices, ensuring no overfitting (CV log loss: 0.9771)

Built a Python tool utilizing Claude LLMs for text summarization from files/URLs via prompt engineering, ensuring concise outputs

- Integrated Claude API to generate 100–200-word summaries from text files or URLs, demonstrating proficiency in LLM-based natural language processing
- Optimized prompts for context-aware summaries across varied inputs using advanced engineering techniques
- Implemented text extraction and preprocessing using BeautifulSoup and requests

Developed an interactive chatbot web application using LangChain, OpenAI's GPT, and Streamlit, enabling users to query through a user-friendly interface, leveraging prompt engineering and API integration for accurate and context-aware responses.

- Built a chatbot web app using LangChain and GPT API for real-time, context-aware responses, showcasing LLM integration and NLP skills.
- Designed a prompt template with LangChain's ChatPromptTemplate to optimize clear, relevant chatbot responses using prompt engineering
- Developed a Streamlit web interface for seamless user input and response display, demonstrating front-end development skills
- Integrated LangSmith to track and debug chatbot performance, ensuring reliable API interactions and response consistency
- Configured secure API key management with dotenv, adhering to best practices for scalable development

Github Events Analytics:

- Engineered a scalable data pipeline using PySpark and Cloud Composer, extracting 5M daily GitHub events, transforming them with dbt, and storing as Parquet in GCS for efficient, real-time processing, aligning with cross-functional analytics needs
- Loaded transformed data into BigQuery, enabling advanced analytics for stakeholders and supporting real-time insights with optimized storage and query performance, leveraging Python for pipeline orchestration and data quality checks

Certificates and Awards:

- AWS Cloud Practitioner
- Bell Excellence and Innovation Award 2022
- Bell Excellence and Innovation Award 2024
- Received Honours in Bachelors

Focused on optimizing data flow, enhancing processing efficiency, and delivering accurate, real-time insights to support billing and financial operations:

- Migrated a legacy data processing system to a serverless GCP architecture, ingesting 5M daily records from 7 sources using Cloud Pub/Sub and Apache Beam on Dataflow. Built ETL pipelines with Python (pandas, NumPy) to enable real-time data integration into BigQuery, enhancing analytics scalability.
- Optimized data transformation and relational modeling in BigQuery and Cloud SQL, reducing processing time by 30% and improving data accuracy. Developed robust data models for transactional processing, supporting high-quality inputs for quantitative analytics.
- Automated data workflows with Cloud Composer, achieving 99.9% uptime and seamless data flow across systems. Built scalable stream processing frameworks with Dataflow, enabling real-time insights and aligning with cross-functional team goals.
- Implemented Python-based data validation checks with BigQuery SQL, reducing discrepancies. Communicated performance metrics to stakeholders, ensuring data governance and delivering actionable insights for business decisions.

Revenue Forecasting Using Machine Learning:

Built an automated machine learning pipeline for telecom revenue forecasting, reducing forecast error (MAPE) by 20% and manual effort by 40%. The solution improved financial planning accuracy, enhanced operational efficiency, and enabled faster, data-driven decision-making:

- Processed large-scale historical telecom revenue data and engineered 15+ features with Python (pandas, NumPy, scikit-learn), enabling scalable, real-time analytics for forecasting.
- Applied machine learning regressors (e.g., XGBoost) for revenue forecasting and used ARIMA and SARIMA models for trend analysis and statistical benchmarking.
- Conducted extensive hyperparameter tuning using Grid Search and Cross-Validation, successfully reducing Mean Absolute Percentage Error (MAPE) by 20%, significantly improving forecasting accuracy and reliability for quarterly business planning.

Bell Canada, Halifax, NS

Data Engineer I

January 2021 – December 2023

SingleBan Billing System Data Processing:

Developed an optimized billing pipeline for finance stakeholders, leveraging serverless cloud architecture to execute complex transformations, enabling efficient and insightful reporting:

- Designed a scalable BigQuery data model to manage 5M customer billing records, with automated GCS-to-BigQuery pipelines processing 500K records daily, ensuring scalability and data quality.
- Optimized query performance by 40% and reduced costs by 25% using clustering and partitioning, integrating data into Looker for 50 stakeholders to drive data-driven decisions

Labour Timesheet Reconciliation Report:

Developed an advanced data pipeline leveraging AWS cloud-native tools, executing sophisticated transformations to streamline timesheet reconciliation and deliver actionable insights for stakeholders:

- Engineered an automated data pipeline using AWS Glue and PySpark, processing 200k+ daily timesheet records from source systems, storing in Amazon S3, and loading into Redshift, reducing reconciliation errors by 30% and cutting processing time by 25%.
- Optimized Redshift query performance by 40% using daily-refreshed materialized views, enabling real-time analytics for stakeholders and saving 20% in retrieval costs, supporting data-driven decisions with scalable Python-driven ETL workflows

Employee Attrition Model:

Developed a machine learning model to predict and mitigate voluntary employee attrition, optimizing for precision and recall addressing data imbalance:

- Processed employee dataset through cleaning, transformation, and feature engineering, addressing missing values and outliers, while conducting EDA with Pandas to uncover attrition patterns like tenure and job satisfaction trends.
- Built and optimized Logistic Regression, Random Forest, and XGBoost models using SMOTE and class weighting to handle class imbalance, achieving a 15% recall improvement and providing insights to reduce voluntary attrition.