Evaluating Machine Learning Algorithms on Google BigQuery ML

Ayush Sharma (0774583), Jasmeet Singh (0758054)

INTRODUCTION

1. The project evaluates **Logistic Regression** and **K-nearest Neighbours (KNN)** in Google BigQuery to understand BigQuery ML's capabilities in training, evaluating, and comparing models.
2. Leveraging the human resources dataset with 10,000 and 1 million entries, the aim is to scrutinize Big Query's handling of machine learning tasks.
3. Deriving insights into BigQuery ML's efficiency in model training, evaluation, and comparative analysis is the primary ambition.
4. The project aims to delineate BigQuery ML's strengths, weaknesses, and optimal use cases in various machine learning scenarios.
5. Logistic Regression achieved high precision targeting gender (0.978) but lower precision for salaries above $100,000 (0.619). KNN displayed a Davies Bouldin Index of 0.684 for the 10,000-entry dataset.
6. Larger datasets showed similar trends with longer execution times, emphasizing nuances in BigQuery ML's performance with varying dataset scales.
7. Logistic Regression's varied predictive accuracy and KNN's consistent performance offer insights into BigQuery ML's suitability for diverse machine learning tasks.
8. These results highlight BigQuery's potential in handling different data scales and nuances of machine learning tasks.

AIM

The project aims to assess the performance of Logistic Regression and K-nearest Neighbours(KNN) algorithms using Google BigQuery ML. Focused on understanding BigQuery ML's functionalities, the evaluation seeks insights into training, evaluating, and comparing models.

Leveraging a human resources dataset with subsets of 10,000 and 1 million entries, the project scrutinizes BigQuery's efficiency in machine learning processes. The ambition is to derive valuable observations delineating BigQuery ML's strengths, weaknesses, and optimal use cases across various machine learning scenarios. The assessment emphasizes resource utilization, model evaluation metrics, and the platform's potential in handling diverse data scales.

METHOD

1. Experiment design aimed at comprehensive evaluation of BigQuery ML using **Logistic Regression** and **K-means** on datasets of 10,000 and 1 million entries.
2. Dataset underwent selection ensuring representation across scales; preprocessing handled missing values, categorical variables, and feature engineering.
3. Implementation involved configuring Logistic Regression for Gender and Salary (> $100,000) using multiple metrics. K-means was assessed via Davies Bouldin Index and Mean Squared Distance.
4. Leveraged BigQuery Studio's SQL for data manipulation, ML for model creation, and Looker Studio for visualization (see **figure 1**).

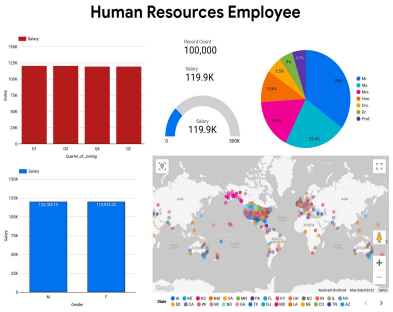

*Fig.1. Screenshots show the **interactive dashboards** created in the Google Looker Studio.*

5. Jupyter Notebook facilitated flexible visualization, exploratory analysis, and experiment documentation.
6. Performance evaluation focused on predictive capabilities and resource consumption within BigQuery, allowing iterative refinement.
7. Methodical documentation in Jupyter Notebooks enabled thorough analysis, comparisons between algorithms, and insights across dataset scales.

RESULTS

1. **Logistic Regression** displayed varied performance based on target variables; **precision/recall/accuracy** for gender: **0.978/0.746/0.863**, but for >$100,000 salaries: **0.619 precision** with full recall.
2. **KNN,** with specific parameters, exhibited metrics like Davies **Bouldin Index** (0.684) and **Mean Squared Distance** (0.174) providing insights into performance and resource consumption (see **figure 2**).

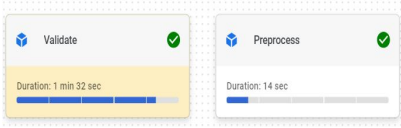

*Fig.2. Screenshot of the execution graph while **creating** a K-means model in BigQuery studio.*

3. Analysis of the one million-entry dataset showed Logistic Regression maintaining **0.624 precision/recall** for >$100,000 salaries, requiring increased execution time and resources (see **figure 3**).
4. KNN, optimized for clustering, offered consistent metrics with varying execution times and substantial resource usage, notably in bytes shuffled.

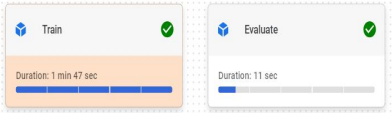

*Fig.3. Screenshot shows the execution graph while **training** and evaluating K-means model in BigQuery studio.*

5. Observations highlight algorithm performance variance concerning dataset sizes and target variables, emphasizing resource demands in BigQuery ML.
6. Findings provide crucial insights into BigQuery ML's applicability across diverse machine learning scenarios, highlighting trade-offs between execution efficiency and performance in data-rich contexts.

CONCLUSIONS

Structured Insights: The research project culminated in nuanced insights into the efficacy of BigQuery ML and associated tools. The structured experiment design facilitated a comprehensive understanding of the platform's capabilities.

Integration Strengths:

1. **Seamless Integration:** BigQuery ML seamlessly integrates within the BigQuery environment, simplifying workflows for organizations heavily reliant on Google Cloud services.
2. **Unified Ecosystem:** Conducting machine learning tasks directly within a robust data warehousing platform provides a unified ecosystem for data analysis and machine learning.

Scalability and Considerations:

1. Efficiency with Large Datasets: BigQuery ML demonstrated efficiency in handling large-scale datasets, providing valuable insights into resource consumption and execution times.
2. Resource Trade-offs: Organizations should weigh computational resources, execution times, and expenses, especially with substantial data volumes.

Algorithm Performance:

1. Logistic Regression: Exhibited varying performance based on target variables, with high precision but fluctuations in effectiveness.
2. K-means Clustering: Showcased consistency in clustering tasks, presenting a robust option for pattern identification within datasets.

Tools Synergy:

The combination of Looker Studio for visualization and Jupyter Notebook for exploratory analysis enriched the experimentation process, facilitating iterative refinement of models.

Adoption Considerations:

1. For Google Cloud Reliance: BigQuery ML is optimal for businesses invested in Google Cloud, streamlining machine learning tasks within a familiar ecosystem.
2. Cost Implications: For smaller-scale operations, careful consideration of potential cost implications due to resource consumption is crucial.

TRENT UNIVERSITY