DATA & AI BOOT-KON EVENT

FraudFix Use Case

Duration: 15 Minutes

Authors: Wissem Khlifi

Use Case Introduction 2
Understand the Structure of the Event 2
About the Dataset 2
The Story: Enhancing Fraud Detection with Machine Learning at FraudFix Technologies 3
Understanding Machine Learning Objectives for Imbalanced Datasets 4
Understanding Good and Bad Values for Metrics 4
Architecture Diagram 6
[Hands-on Lab - 1] Setup your environment: Notebooks & IAM 7
[Hands-on Lab - 2] Data Ingestion with Dataproc , BigLake & PubSub 7
[Hands-on Lab - 3] ELT & Lakehouse : BigQuery, Dataform, LLM 7
<Lunch Time: 45 Minutes> 7
[Hands-on Lab - 4] ML Operations with Vertex AI 7
[Hands-on Lab - 5] Data Governance with Dataplex 7
<Coffee Break: 15 Minutes> 7
[Hands-on Lab - 6] Data Sharing with Analytics Hub 7
[Optional] [Hands-on Lab - 7] Data Canvas & Looker Studio 7
Use Case Closure : Architecture Challenge & Retrospective 7
</Event Closure> 8
Appendices 8

Use Case Introduction

- About the company: FraudFix Technologies is a cutting-edge company focused on making financial transactions safer for Google Cloud enterprise customers across industries (financial institutions, online retailers, etc ...)

- Your role: As a senior data analytics/AI engineer at FraudFix Technologies, you will tackle the challenges of making financial transactions safer using machine learning. Your work will involve analyzing vast amounts of transaction data to detect and prevent fraud, as well as assessing customer sentiment regarding the quality of transaction services. You will leverage a unique synthetic dataset, which includes auto-generated data by Google Gemini and a public European credit card transaction dataset that has been PCA transformed and anonymized. This dataset will be used to train your models, reflecting real-world applications of GCP Data & AI in enhancing financial safety.

Understand the Structure of the Event

During this event, your main focus will be on completing the labs, which are clearly marked with the label "[LAB]" in the instruction manuals. These labs include detailed step-by-step instructions to guide you. In addition to the labs, you'll face several challenges that you'll need to solve on your own or with your group. Groups will be assigned by the event organizers at the start of the event.

About the Dataset

- The datasets contain transactions made by credit cards in September 2013 by European cardholders, but also augmented by Google Gemini.
- This dataset presents transactions that occurred over two days, where there are a few hundred fraudulent transactions out of hundreds of thousands of transactions.
- It is highly unbalanced, with the positive class (frauds) accounting for less than 0.1% of all transactions (subject to testing in your notebooks).
- It contains only numeric input V* variables which are the result of a **PCA** transformation.
- Due to confidentiality issues, the owner of the dataset cannot provide the original features and more background information about the data.

1. Features **V1, V2, ... V28** are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' , 'Feedback' and 'Amount'.
2. Feature **'Time'** contains the seconds elapsed between each transaction and the first transaction in the dataset.
3. Feature **'Amount'** is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning.
4. Feature **'Class'** is the response variable and it takes value 1 in case of fraud and 0 otherwise.
5. Feature **'Feedback'** represents customer selection on service quality after submitting the transaction. This feature has been auto-generated by Google Gemini and added to the original dataset.
6. During your machine learning experimentation using notebooks, one of the notebook cells will add your Google cloud account email address into the prediction dataset for traceability. This email address is treated as PII data and should not be shared externally outside of Fraudfix.
7. The original dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.
8. If you need more details on current and past projects on related topics are available here and here.

The Story: Enhancing Fraud Detection with Machine Learning at FraudFix Technologies

1. **Introduction to the Challenge**

   As a data analytics and AI engineer at FraudFix Technologies, your principal mission is to enhance the detection of credit card fraud. Effective fraud detection is crucial as it protects customer transactions and maintains trust, fundamental for sustaining business integrity and customer relations.

2. **Data Challenges**

   In your role, you face the challenge of working with data in PCA format, which complicates the interpretation of original features and engineering new features. The dataset predominantly consists of normal transactions, with fraudulent ones forming a minor fraction, creating a significant imbalance. Moreover, GDPR compliance necessitates careful adjustment of machine learning outputs to avoid sharing personally identifiable information (PII).

3. **Initial Challenges with Standard Methods**

   Initially, traditional fraud detection methods such as rule-based systems and basic statistical models (e.g., using averages and standard deviations) were ineffective due to the class imbalance. These methods often misclassified legitimate transactions as fraudulent. This early failure highlighted the need for a better machine learning objective, shifting away from solely maximizing accuracy.

4. **A Major Improvement**

   You pivoted to anomaly detection strategies, treating fraud detection as an outlier detection problem. This shift significantly improved the identification of fraudulent transactions amidst predominantly normal ones, reducing false positives. Consequently, the focus moved towards more appropriate metrics for imbalanced datasets, particularly the AU PRC (Area Under the Precision-Recall Curve), prioritizing it over traditional accuracy.

   To learn more about the meaning and significance of metrics such as accuracy, F1 score, recall, and AU PRC curve in imbalanced dataset contexts, refer to the "***Understanding Machine Learning Objectives for Imbalanced Datasets***" section.

5. **Making Things Better**

   Your initial focus is on ingesting data into your Google Cloud environment, followed by transforming, cleaning, and organizing the data in BigQuery. This ensures smooth data flow and cleanliness for analysis. Tools like Vertex AI , Data Canvas and Looker Studio are employed to better understand transaction patterns and enforce data quality governance. Additionally, sentiment analysis is used to gauge customer perceptions of the transaction service.

6. **Setting Up MLOps and Managing Data Quality**

   Using GCP's Vertex AI, you automated the training and deployment of your models, enhancing efficiency. Google Dataplex is utilized to implement automatic checks to maintain high data quality.

7. **Sharing ML prediction results**

The final step was to provide customers with your model predictions via the Analytics Hub. This empowered them to make well-informed decisions and increased their trust in FraudFix's ability to detect and prevent fraudulent activity. Ultimately, customers decide how to handle these detected transactions, be it manual approval or disapproval.

Understanding Machine Learning Objectives for Imbalanced Datasets

Traditional accuracy metrics can be misleading in datasets where fraudulent transactions are vastly outnumbered by legitimate ones. Instead, we focus on metrics that provide a more accurate evaluation of our model's effectiveness in fraud detection. During the Vertex AI experimentation, you will encounter few optimization metrics such as precision, recall, F1 score and AU PRC. In this section, we will explain what each metric signifies and what are our machine learning objectives.

- **Precision**: Indicates the percentage of transactions correctly identified as fraudulent out of all transactions labeled as fraudulent. In other words, it answers the question: "of all transactions the model labeled as fraudulent, how many were actually fraudulent?". *Formula: Precision = True Positives / (True Positives + False Positives)*
- **Recall**: Measures the proportion of actual fraudulent transactions that the model correctly identified, crucial for capturing as many fraud cases as possible. *Formula: Recall = True Positives / (True Positives + False Negatives)*
- **F1 Score**: A harmonic mean of precision and recall, providing a balance between the two. It is vital for detecting fraud effectively as it considers both false positives and false negatives. *Formula: F1 = 2 x (Precision x Recall / Precision + Recall)*
- **AU PRC** (Area Under the Precision-Recall Curve): Evaluates the model's ability to distinguish between classes under imbalanced conditions. It assesses the trade-off between precision and recall, important in fraud detection to minimize both types of errors.

By focusing on **maximizing the AU PRC metric**, we enhance the model's accuracy in identifying actual fraud while minimizing false alarms, ensuring the system's reliability in operational settings where misclassification has high stakes. This focus is crucial in the financial domain, where inaccuracies can lead to significant financial losses and customer dissatisfaction. The good news is that Google machine learning solutions like **AutoML** will help you achieve these objectives in a more straightforward way. Google Cloud's AutoML is chosen for its ease of use and efficient model development process, enabling rapid deployment of effective fraud detection models. Alternative options within GCP cater to different needs and expertise levels:

- **BigQuery ML:** Enables building and training models directly within BigQuery using SQL, offering a convenient option for those familiar with SQL and seeking quick model development.
- **Vertex AI Custom Training:** Provides greater flexibility and control for experienced data scientists who want to build and train custom models using frameworks like TensorFlow or PyTorch.

Understanding Good and Bad Values for Metrics

While the specific thresholds for "good" and "bad" can vary depending on the context, the general guidelines provided below can serve as a useful reference point for those who encounter these metrics for the first time.

- **Precision**
  - *Good Precision:* High precision indicates that most of the transactions labeled as fraudulent are indeed fraudulent. This is important to minimize the number of false positives (legitimate transactions incorrectly labeled as fraudulent).
    - *Example of Good Precision:* Precision > 0.9 (90%) might be considered good in many contexts.
  - *Bad Precision:* Low precision means that many of the transactions labeled as fraudulent are actually legitimate, which can be problematic for customer experience and trust.
    - *Example of Bad Precision:* Precision < 0.5 (50%) indicates that more than half of the transactions labeled as fraudulent are false alarms.

- **Recall**
  - *Good Recall*: High recall indicates that most of the actual fraudulent transactions are correctly identified. This is crucial in fraud detection to capture as many fraud cases as possible.
    - *Example of Good Recall:* Recall > 0.8 (80%) is often considered good because it means that the model is catching most fraud cases.
  - *Bad Recall*: Low recall means that many fraudulent transactions are missed, which is a major issue in fraud detection.
    - *Example of Bad Recall*: Recall < 0.5 (50%) suggests that the model is missing a significant number of fraud cases.

- **F1 Score**

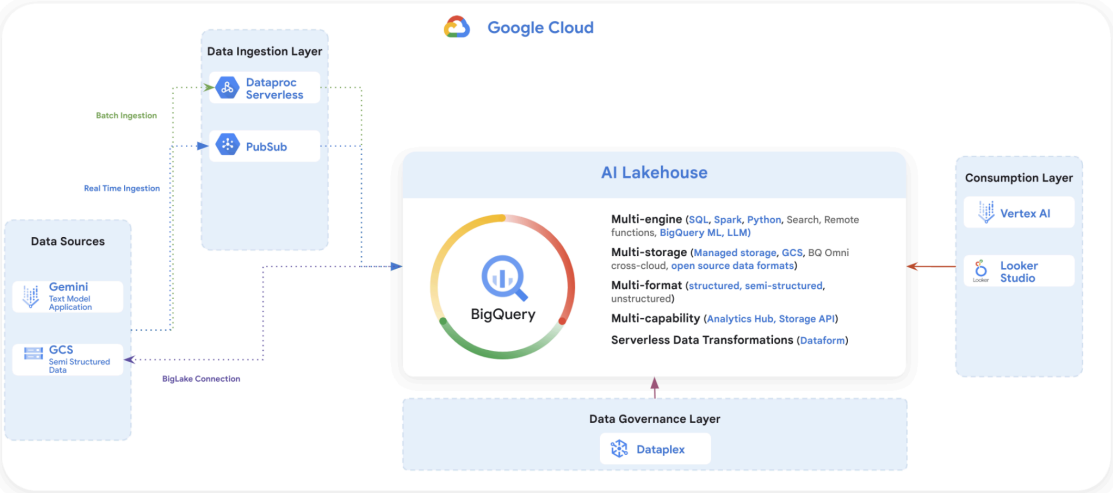
  - *Good F1 Score*: A high F1 score indicates a good balance between precision and recall. It's particularly useful when you need to balance the cost of false positives and false negatives.
    - *Example of Good F1 Score:* F1 Score > 0.8 (80%) is usually considered good in many contexts.
  - *Bad F1 Score*: A low F1 score indicates poor performance in both precision and recall.
    - *Example of Bad F1 Score*: F1 Score < 0.5 (50%) suggests the model is neither accurately identifying fraudulent transactions nor capturing enough of them.

- **AU PRC (Area Under the Precision-Recall Curve)**

  - *Good AU PRC*: A high AU PRC indicates that the model performs well across different thresholds, effectively balancing precision and recall.
    - *Example of Good AU PRC*: AU PRC >~ 0.8 (80%) is often considered good.
  - *Bad AU PRC*: A low AU PRC indicates poor performance in distinguishing between the classes, especially under imbalanced conditions.
    - *Example of Bad AU PRC*: AU PRC < 0.5 (50%) suggests that the model is performing poorly, barely better than random guessing.

- **Context Matters:** It's important to note that these values are relative and depend on the specific use case and the acceptable trade-offs between false positives and false negatives. For example, in some high-stakes environments, even higher precision and recall might be required, while in others, slightly lower values might be acceptable if the cost of false positives or negatives is lower.

Architecture Diagram



Data Sources

- You'll start by working with raw data that comes in different formats (csv , parquets).
- Those data files are stored in a github repository *(link to github is shared afterwards)*
- Your first task is to store the raw data into your Google Cloud Storage (GCS) bucket.

Data Ingestion Layer

- You will bring this data into your BigQuery AI Lakehouse environment.
- For batch data, you'll use Dataproc Serverless and BigLake.
- For near real-time data, you'll use Pub/Sub to handle data as it comes in.
- Because we want to simulate data ingestion at scale, we will be using the raw data that you have stored in GCS to simulate both batch and real time ingestion.
- These tools help you get the data ready for processing and analysis.

BigQuery AI Lakehouse

Think of this as the main camp where all your data hangs out. It's a GCP product called BigQuery, and it's designed to work with different types of data, whether it's structured neatly in tables or unstructured like a pile of text documents. Here, you can run different data operations without moving data around.

Data Governance Layer

This is where you ensure that your data is clean, secure, and used properly. Using Dataplex, you'll set rules and checks to maintain data quality and governance.

Consumption Layer

- Once you have your insights, you'll use tools like Vertex AI for machine learning tasks and Looker Studio for creating reports and dashboards. This is where you turn data into something valuable, like detecting fraud or understanding customer sentiment.
- Your goal is to share the results of your data predictions to your customers in a secure and private way. You will be using Analytics Hub for data sharing.

Throughout the event, you'll be moving through these layers, using each tool to prepare, analyze, and draw insights from the data. You'll see how they all connect to make a complete data analytics workflow on the cloud.

Cost

- If you are using your GCP environment, running all labs will cost you around 200$ / month.
- We will decommission the environments provided to you by Google a few hours after the event.

[Hands-on Lab - 1] Setup your environment: Notebooks & IAM

Follow Step by Step Instructions here

[Hands-on Lab - 2] Data Ingestion with Dataproc , BigLake & PubSub

Follow Step by Step Instructions here

- [LAB] Biglake Object Tables
- [LAB] Batch data ingestion into BigQuery using Dataproc
- [LAB] (Near-)Real time data ingestion into BigQuery via PUSUB

[Hands-on Lab - 3] ELT & Lakehouse : BigQuery, Dataform, LLM

Follow Step by Step Instructions here

- **[TODO] Challenge:** Setup a daily frequency execution of the workflow we have just created.

<Lunch Time: 45 Minutes>

[Hands-on Lab - 4] ML Operations with Vertex AI

Note: You can start Hands-on Lab 5 while the Hands-on Lab 4 training jobs in Notebooks 2 & 3 are still running.

Follow Step by Step Instructions here

- [LAB]Data exploration and preparation using Vertex AI workbench
- [LAB]MLOPS: Operationalize Machine learning

[Hands-on Lab - 5] Data Governance with Dataplex

Follow Step by Step Instructions here

- Data Catalog Exploration
- Automatic Data Lineage: Supported Data Sources & Targets
- Setup Data Quality Jobs

<Coffee Break: 15 Minutes>

[Hands-on Lab - 6] Data Sharing with Analytics Hub

Follow Step by Step Instructions here

- Data Clean Room

[Optional] [Hands-on Lab - 7] Data Canvas & Looker Studio

Follow Step by Step Instructions here

Use Case Closure : Architecture Challenge & Retrospective

Follow Step by Step Instructions here

</Event Closure>

Appendices

- *Google Cloud Data Analytics in 10 minutes*
  https://www.youtube.com/watch?v=g-f_mWXK9sU
- *Dataproc in a minute*
  https://www.youtube.com/watch?v=Jj6mp7Sam10
- *Run Spark and Hadoop faster with Dataproc Duration : Duration 16 minutes*
  https://www.youtube.com/watch?v=shzKmZ6Yqtk
- *What is Cloud IAM? Duration 10 minutes*
  https://www.youtube.com/watch?v=xQCIVtAECdg
- *Virtual Private Cloud in a minute*
  https://www.youtube.com/watch?v=hS_uvz4ohbo
- *What is Vertex AI ? Duration: 7 minutes*
  https://www.youtube.com/watch?v=gT4qqHMiEpA
- *BigQuery Spotlight Playlist: Duration 90 minutes*
  https://youtube.com/playlist?list=PLIivdWyY5sqLAbldmcMwsxWg-w8Px34MS&si=WlbLukGRZk88tOjx
- *Principal Component Analysis (PCA) in Machine Learining*
  https://towardsdatascience.com/a-complete-guide-to-principal-component-analysis-pca-in-machine-learning-664f34fc3e5a
- *Principal Component Analysis (PCA) Video Explanation: https://www.youtube.com/watch?v=aSfvqTQQkcs*
- *PCA Visualization : https://setosa.io/ev/principal-component-analysis/*