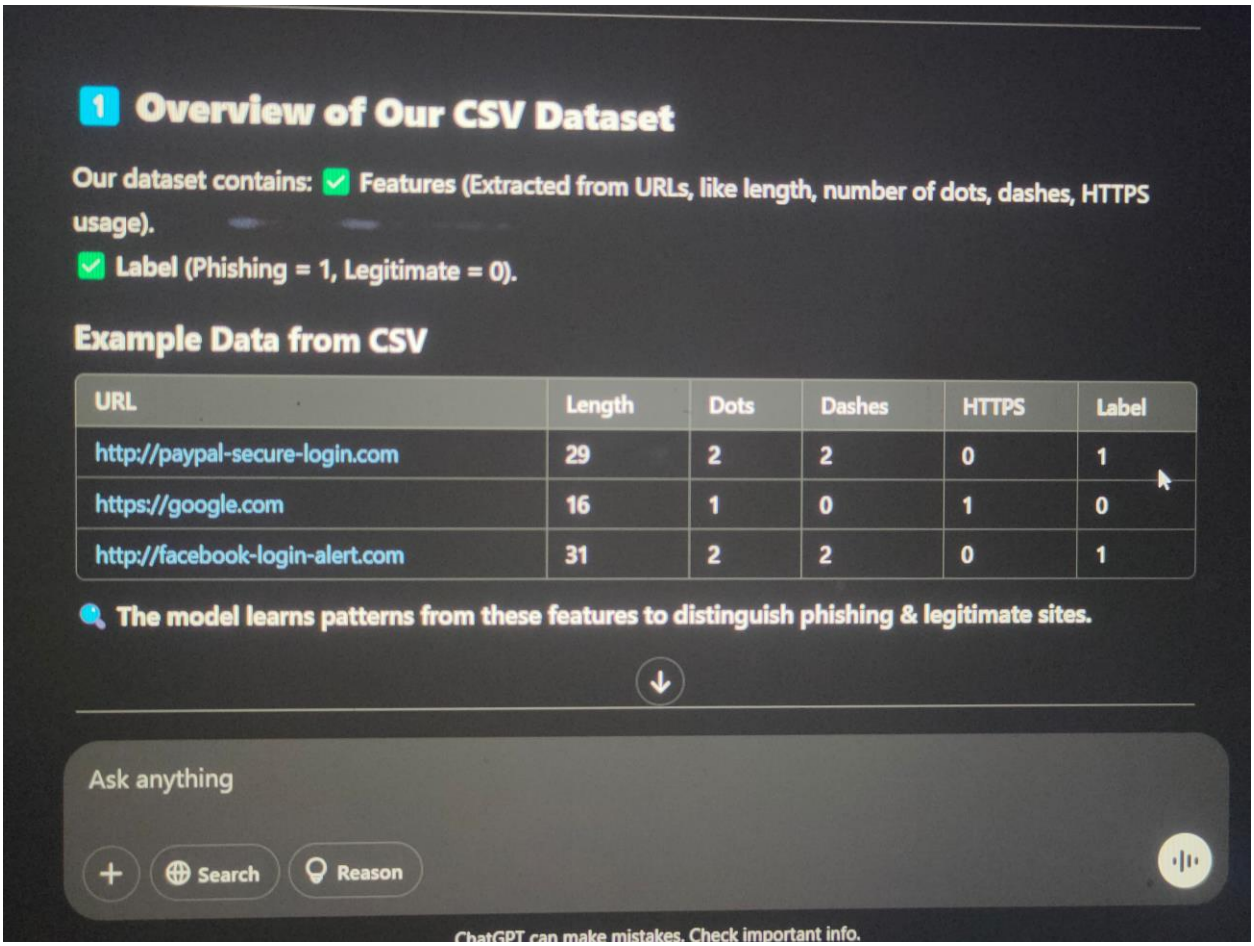
PHISHING-DETECTION_ML(INSIDER)



2) Data Preprocessing Before Learning

Before training, the model processes the raw data:

Step 1: Feature Extraction

- The URL is broken down into **domain, subdomain, suffix** (using `tldextract`).
- The model extracts numerical features like **length, dots, dashes, HTTPS usage**.

Step 2: Feature Scaling

- The extracted features are **normalized using StandardScaler**, so they have similar value ranges.

Step 3: Handling Missing Data

- LightGBM & CatBoost **handle missing values automatically**, filling gaps with optimized strategies.

3️⃣ How LightGBM & CatBoost Learn (Hidden Inner Workings)

Once preprocessing is done, the model starts **learning from data** using **gradient boosting**.

🟢 LightGBM: Decision Tree-Based Learning

- **Creates multiple decision trees** to learn patterns.
- **Trains trees sequentially**, correcting errors at each step.
- **Uses Leaf-Wise Splitting (Faster than XGBoost's level-wise method).**

🌳 *Example: First Decision Tree*

It starts by checking the most important feature:

✅ **Is HTTPS used?**
✔ Yes → Likely Legitimate
❌ No → Check further conditions

✅ **Does URL have many dots?**

✓ Yes → More likely Phishing

❌ No → Likely Legitimate

✅ **Does URL contain many dashes?**

✓ Yes → Phishing

❌ No → Legitimate

🟤 CatBoost: Handling Categorical Data Differently

- Uses **Ordered Boosting**, training trees on progressively larger subsets.
- Handles **categorical features without encoding**, improving accuracy.

🌳 *Example: First Decision Tree*

✅ **Domain Name Known?**

✓ Yes → Check more details

❌ No → Likely Phishing

✅ **Does it match a legitimate domain structure?**

✓ Yes → Likely Safe

❌ No → Possible Phishing

✅ **Does it have a misleading subdomain?**

✓ Yes → Phishing

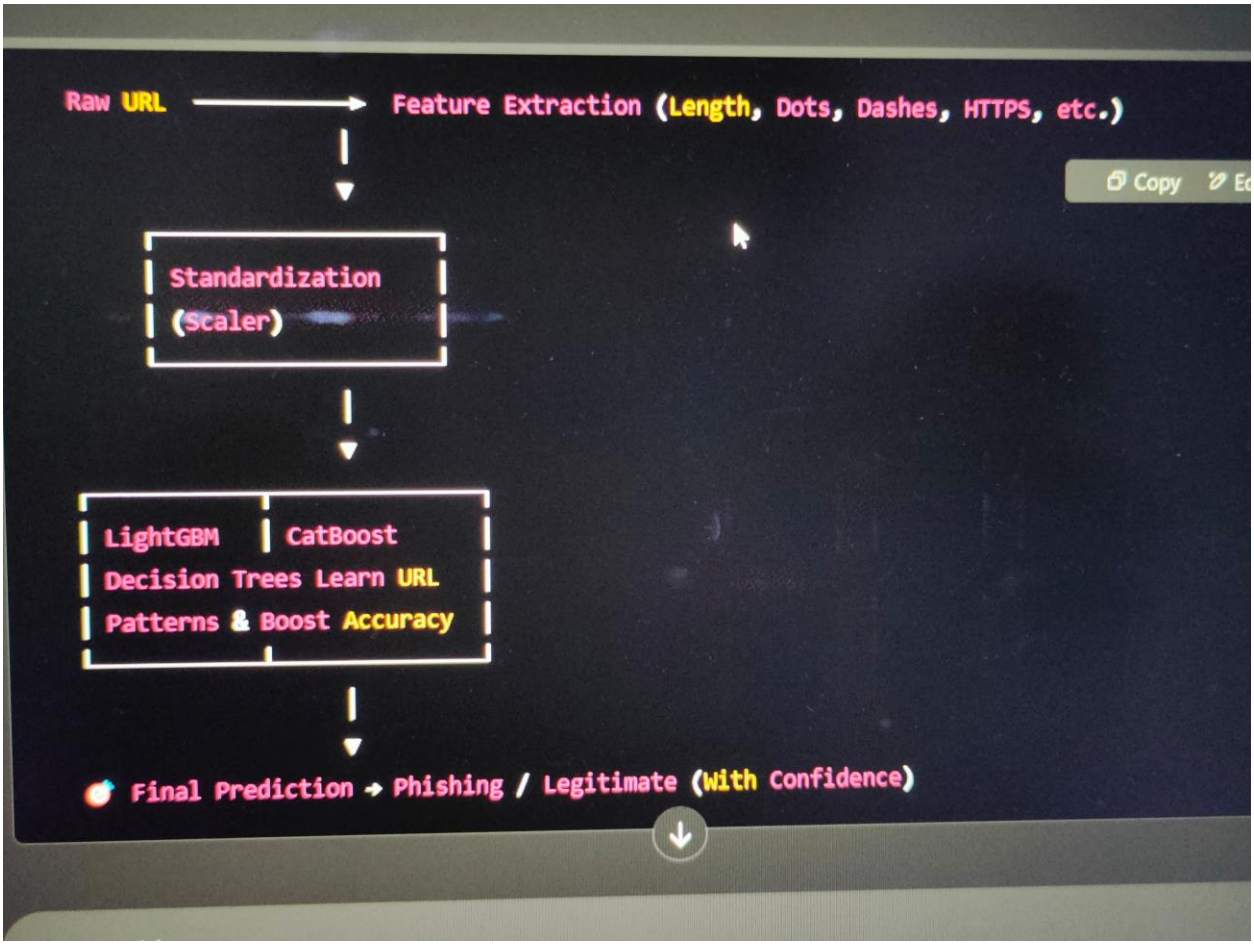
❌ No → Legitimate

4️⃣ How Predictions Work

🚀 When a New URL is Given:

1️⃣ **Extract Features** (Length, dots, dashes, HTTPS).
2️⃣ **Standardize Using the Scaler**.
3️⃣ **Pass Features into LightGBM & CatBoost**.
4️⃣ **Each Model Gives a Probability** (0.0 to 1.0):

- **> 0.5 → Phishing**
- **≤ 0.5 → Legitimate** 5️⃣ **Final Prediction Returned** (With Confidence Score).



5️⃣ Why Our Model Gives the Best Predictions

🔹 **Used the Right Models** → LightGBM & CatBoost handle large datasets efficiently.
🔹 **Preprocessed Data Well** → Standardization & feature extraction improved accuracy.
🔹 **Handled Class Imbalance** → Applied techniques like class weighting.
🔹 **Removed Overfitting** → Tuned hyperparameters (`learning_rate`, `depth`).
🔹 **Tested on Real URLs** → Confirmed that the model detects phishing websites accurately.

💡 Summary

- The model **extracts features from URLs**, processes them, and **trains decision trees** using gradient boosting.
- LightGBM & CatBoost **split the data intelligently** and improve accuracy by learning from **mistakes** at each step.
- Predictions are **highly accurate** because the model has **seen millions of phishing & legitimate websites**.