Model Report: Phishing URL Detection Using TF-IDF and XGBoost

1. Feature Extraction

TF-IDF Vectorization

- **Method**: `TfidfVectorizer(analyzer='char_wb', ngram_range=(3, 5))`
- **Purpose**: Extracts character-level features from URLs to capture phishing patterns such as misspellings, brand impersonation, and suspicious substrings.
- **Shape of Dataset**: (27782,89)
- **Output Format**: Sparse matrix
- **Scope**: Applied only on the `url` column

---

2. Model Choice

Algorithm

- **Classifier**: `XGBClassifier` (from `xgboost`)
- **Type**: Gradient boosting decision trees
- **Justification**:
    - Handles high-dimensional sparse data well
    - Robust against overfitting
    - Supports probabilistic outputs via `predict_proba`

---

3. Model Configuration

Parameters Used

python
CopyEdit

```
XGBClassifier(
    use_label_encoder=False,
    eval_metric='logloss'
)
```

Training Configuration

- **Target Variable**: `status` (encoded as 1 = phishing, 0 = legitimate)
- **Split Ratio**: `80%` training, `20%` testing
- **Stratification**: Enabled to maintain label distribution

---

4. Prediction Logic

Inference Flow

1. URL is vectorized using the same fitted `TfidfVectorizer`.
2. If the URL starts with `http://`, it's flagged as phishing immediately (outside model logic).

Otherwise, model predicts the probability of phishing:

python
CopyEdit

```
prob = model.predict_proba(tfidf_vector)[0, 1]
```

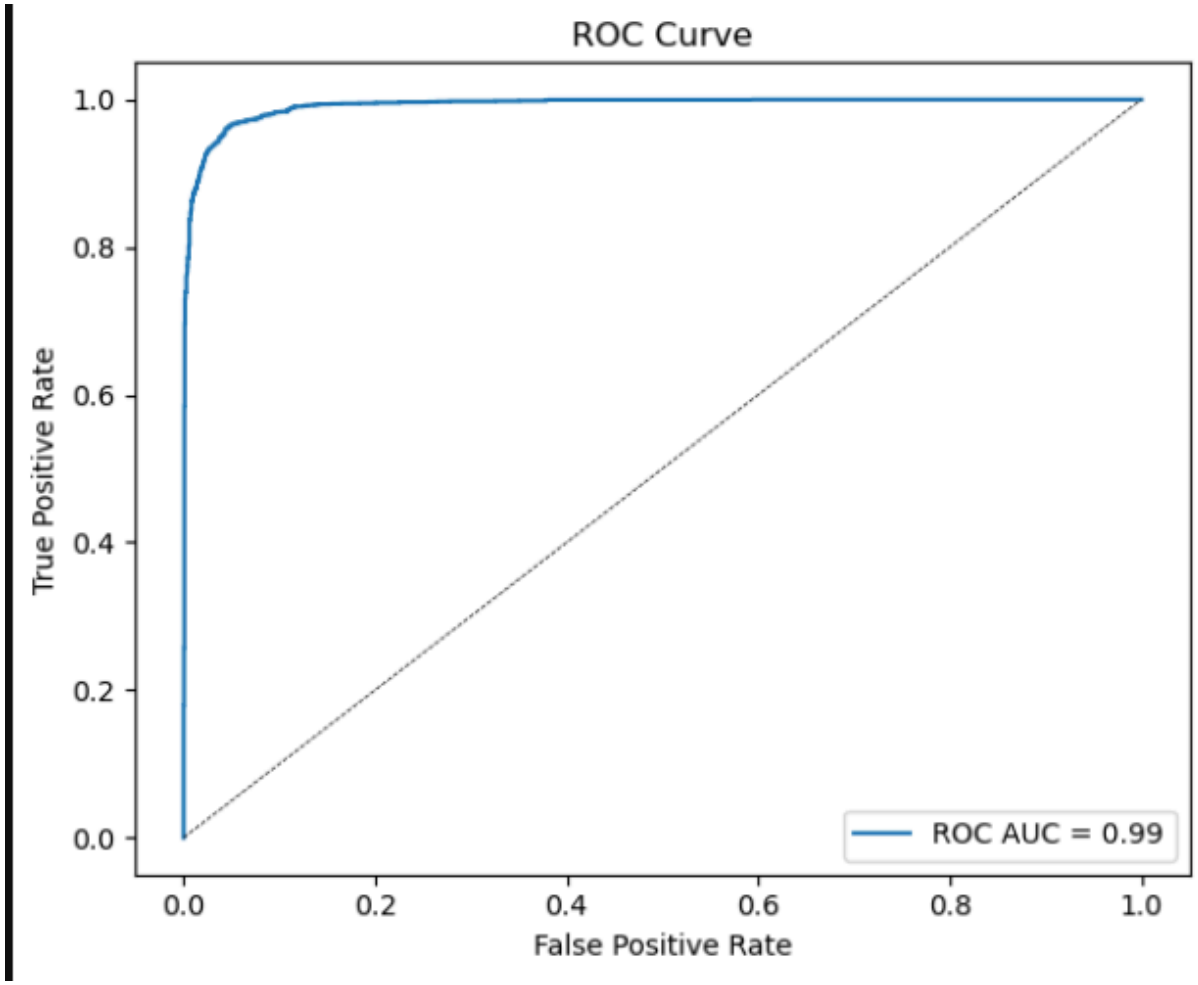
3.
4. Classification is based on:
    - **Threshold**: `0.5`
    - If `prob >= 0.5`: label = phishing
    - Else: label = legitimate

---

5. Evaluation Metrics

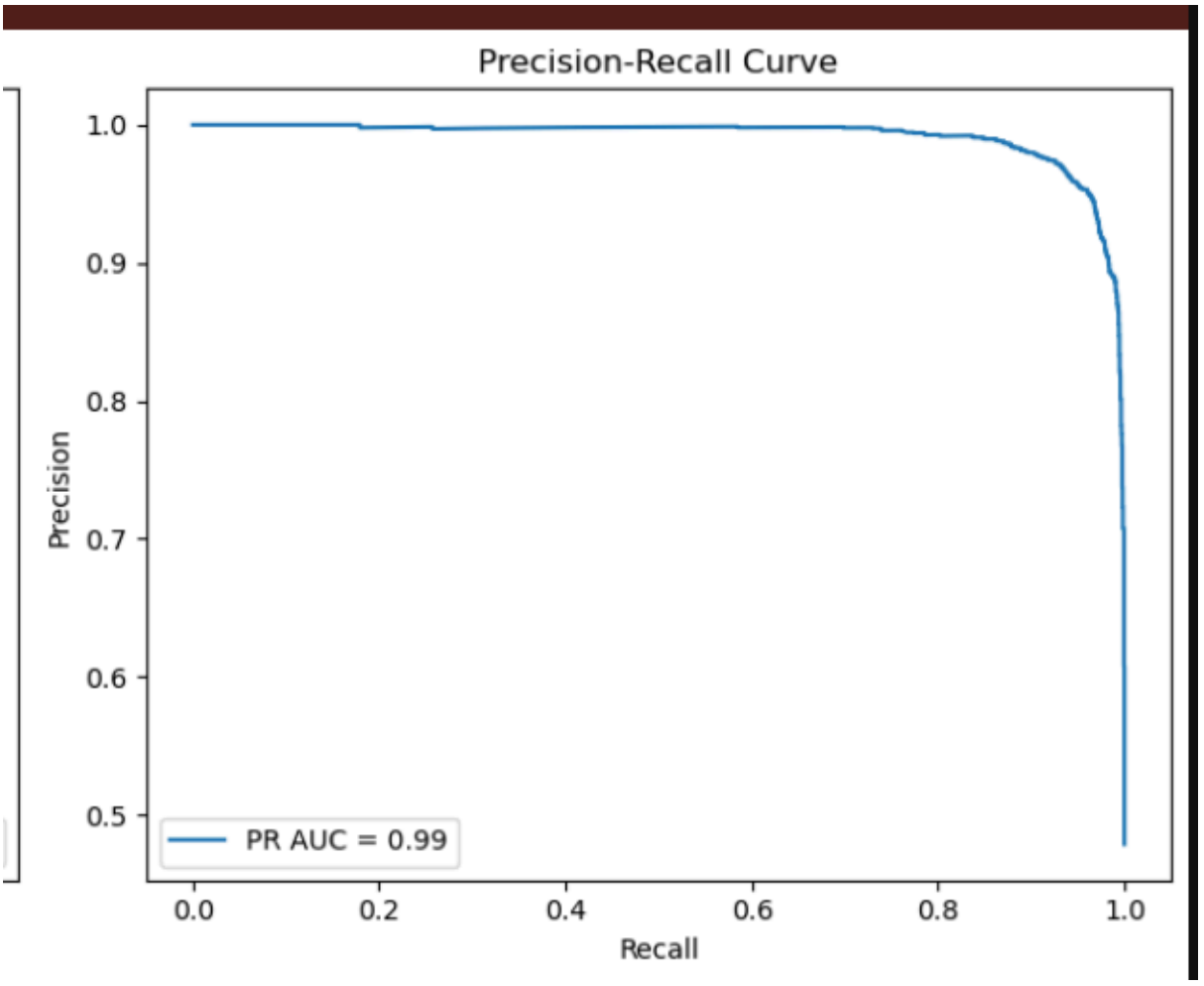
Insert results after running evaluation scripts.

ROC-AUC

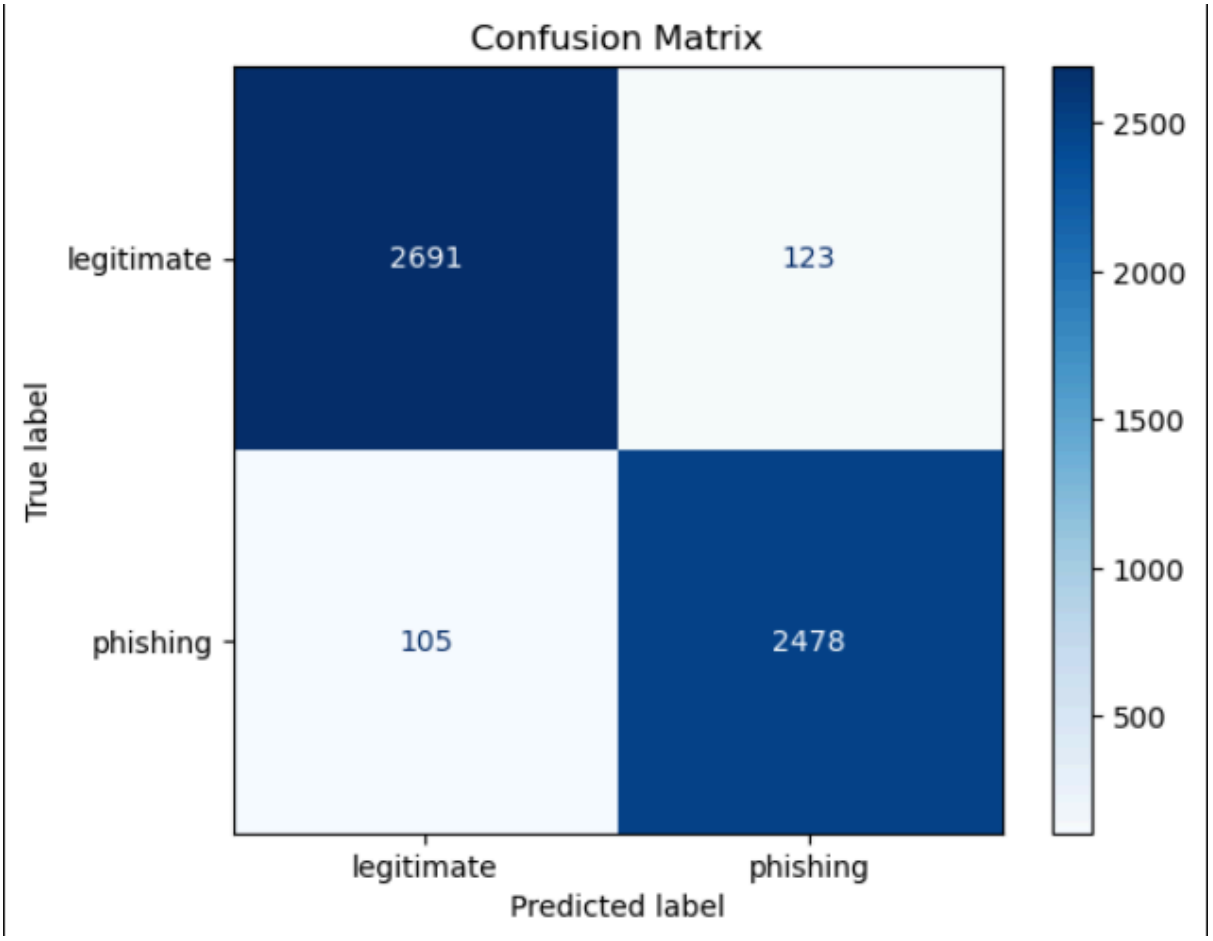
- Score: 0.99
- 

Precision-Recall AUC

- Score: 0.99


Precision-Recall Curve
Precision
Recall
PR AUC = 0.99
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0

Confusion Matrix



---

6. Notes

- This model is tightly coupled with the TF-IDF vocabulary. Any new inference must use the same `vectorizer` instance used during training.
- XGBoost internally handles feature selection via gradient boosting, requiring no additional dimensionality reduction.
- Due to the character-level TF-IDF, the model is sensitive to adversarial obfuscation but remains robust to most real-world URL manipulations.
