Passport Classification with Gitcoin

Gitcoin Passport Data Science Take Home Problem

Overview

- Understanding the Challenge
- Data Management & Insights
- Feature Engineering & Model Building
- Advanced Techniques & Reflection
- Forward Thinking & Continuous Learning
- Q&A

Defining the Sybil Challenge: Understanding Business and Product Needs

- Objective: Sybil detection for blockchain-based applications
- Different Use-Cases: Airdrops vs. Grant Programs
- Sybil Detection Metrics: What's an acceptable sybil detection rate? Is there an industry standard or baseline?
- Assumption based on dataset: Predominantly airdrop-like requirements
- Business Implications:
  - High accuracy in human detection crucial
  - Misclassifications can lead to reputation damage and financial losses
  - High volume of airdrop applicants intensifies cost of errors
- Model Approach: Binary classification - Human vs. Sybil

Key Performance Metrics: Selecting What Truly Matters

- Choosing the Right Metrics:
  - Understand stakeholder objectives and constraints.
  - Prioritize simple, observable, attributable metrics.
- Metric Considerations:
  - Macro Average F1 score for avoiding class imbalance pitfalls.
  - Emphasis on precision and recall for nuance.
- Optimizing Performance:
  - Satisficing with secondary metrics.
  - Example: Optimize precision at ≥0.95 recall.
    - Ensures minimal human misses.
    - Maximizes correct sibil detections.

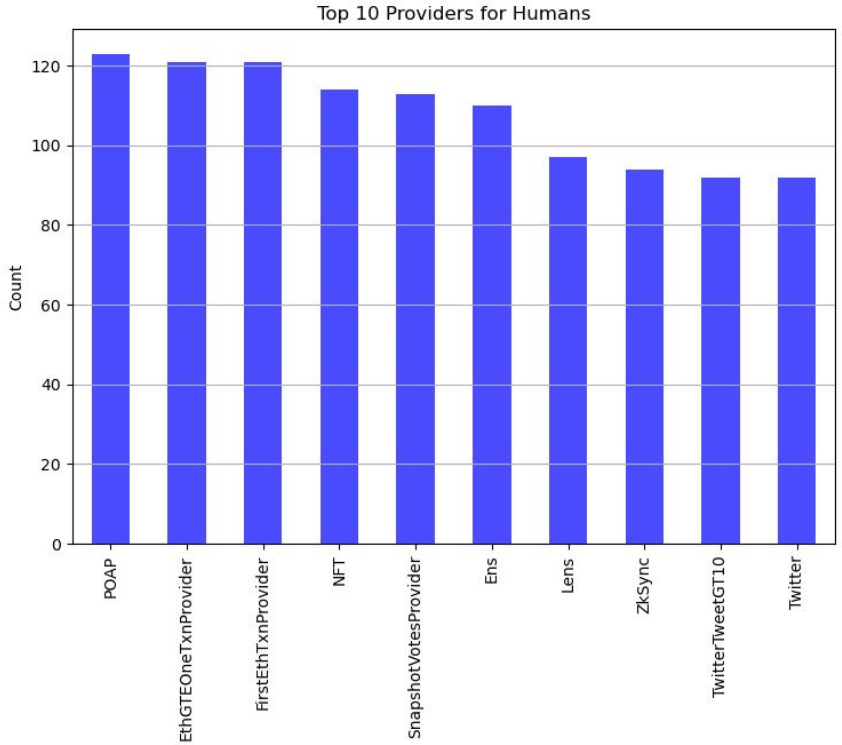
Data Deep Dive: Unlocking the Potential of Our Sources

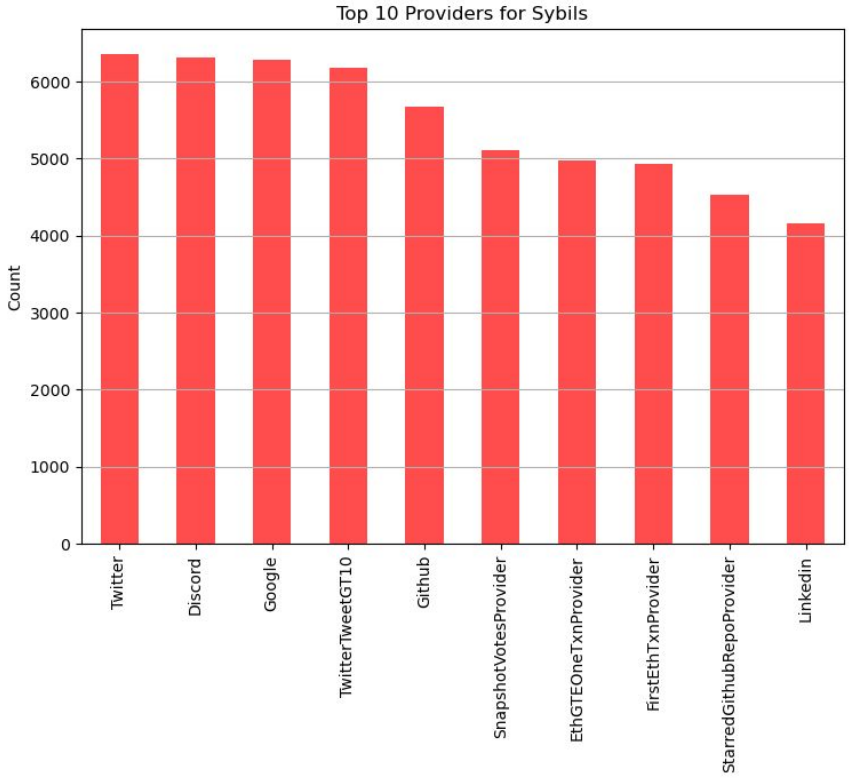
- Quality Matters: A model's performance is tied to its data. "Garbage in, garbage out."
- Beyond Fixed Datasets: Real-world scenarios offer flexibility in data choices.
- Potential Data Enhancements:
  - Leverage on-chain wallet data.
  - Request additional user data during verification.
  - Scrape relevant data from online platforms.
  - Explore data augmentation techniques.

Exploratory Data Analysis: Visualizing and Interpreting the Data Landscape

- Datasets Loaded:
  - Human dataset with 127 entries and 3 columns: publicaddress, Confidence Score, and passport.
  - Sybil dataset with 6,936 entries and 2 columns: publicaddress and passport.
- Passport Data Structure:
  - type: Credential type.
  - proof: Details about the proof.
  - issuer: Credential issuer.
  - @context, issuanceDate, expirationDate: Credential metadata.
  - credentialSubject: Contains subject data such as id, hash, and provider.

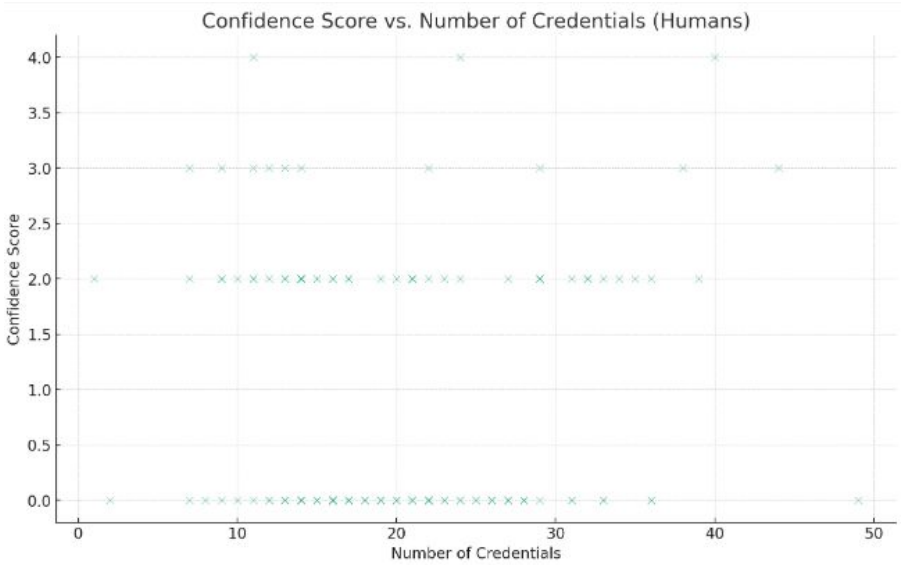
Exploratory Data Analysis: Visualizing and Interpreting the Data Landscape





Data Integrity: Cleaning and Preparing for Modelling

- Key Observations from EDA:
  - The passport column is rich in JSON-encoded data detailing various credentials.
  - Preliminary analysis on the relationship between Confidence Score and Number of Credentials shows no clear correlation.
  - Decision: Drop the Confidence Score for model simplicity.



Crafting Features: Tailoring Data for Optimal Performance

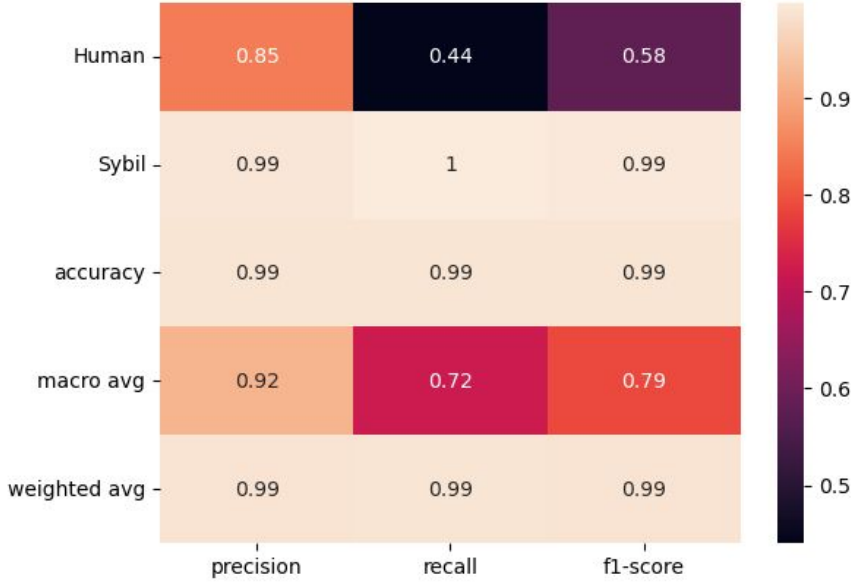
- Feature Engineering Highlights:
  - Extracted the number of credentials in each passport.
  - One-hot encoded providers, turning each unique provider into a separate feature.

Choosing the Right Model: Selection Criteria and Rationale

- Factors to Consider in Model Selection:
  - Speed: Both in training and prediction.
  - Budget: Cost implications of deploying and running the model.
  - Data Volume & Dimensionality: Size and complexity of the dataset.
  - Feature Types: Handling categorical vs. numerical features.
  - Explainability: Ability to interpret and explain model decisions.
- Reasons for Choosing Random Forest:
  - Out-of-the-Box Efficiency: Typically performs well without extensive tuning.
  - Some Explainability: Offers insights into feature importances.
  - Avoids Assumptions: Doesn't require confirming assumptions like linear relationships, as in logistic regression.
  - Time-Efficiency: For the sake of time, a straightforward model like Random Forest was optimal.

Training Insights: Evaluating Model Performance and Outcomes

- Model & Training:
  - Training: 5,650 samples (80%); Test: 1,413 samples (20%).
- Evaluation Metrics:
  - Accuracy: 98.87%; ROC-AUC: 0.97.
  - Macro F1-Score: 0.72.
  - Human Recall: 44%; Sybil Precision: 99%.

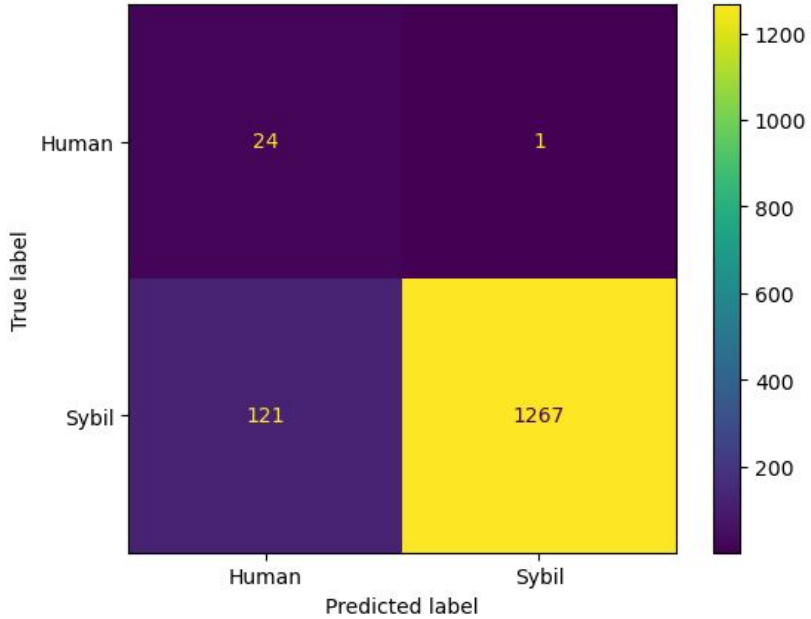


Training Insights: Evaluating Model Performance and Outcomes

- Challenges:
  - Low recall for Human class.
  - Imbalanced data.
- Recommendation:
  - Oversampling or synthetic data methods.
- Business Implications:
  - Aim: Reduce false negatives (Humans misclassified).
  - High applicant volume intensifies impact of errors.
  - Align model with business priorities for reputation safeguarding.

Boosting Human Recall: Sampling Strategies Explored

- Techniques Used:
  - Oversampling with SMOTE.
  - Undersampling.
  - No resampling (baseline).
- Results:
  - Oversampling (SMOTE):
    - F1: 74%.
    - Human Recall: Improved, but misclassified 14 as Sybil.
    - Sybil Precision: 99%.
  - Undersampling:
    - F1 (on undersampled test set): 92%.
    - F1 (on original test set): 94%.
    - Human Recall: Good; only 1 misclassified.
    - Sybil Precision: Decreased with 121 misclassified.



Boosting Human Recall: Sampling Strategies Explored

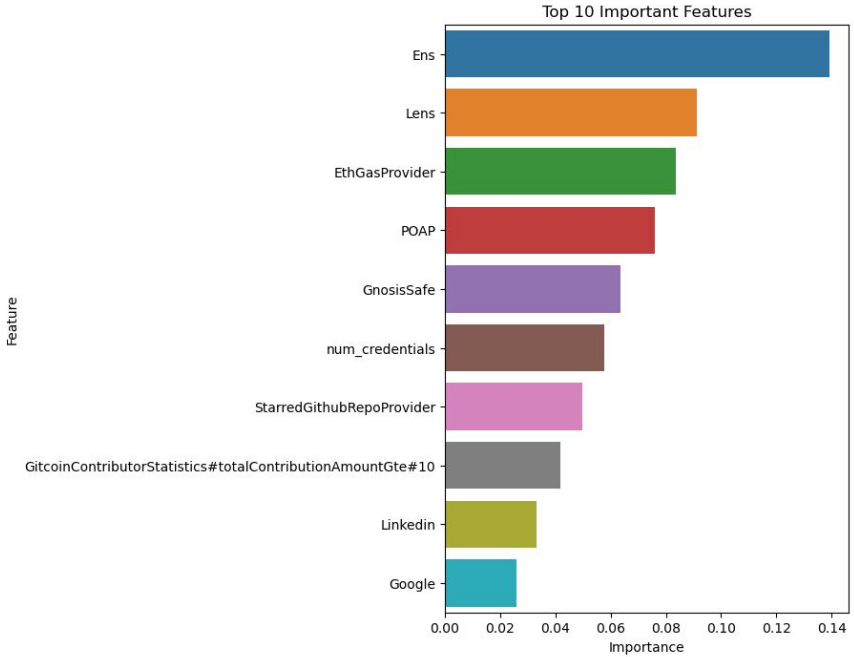
- Undersampling Observations:
  - Best performance among techniques.
  - However, not utilizing all available Sybil data might be a missed opportunity for model learning.
- Sybil Misclassifications:
  - Increase in Sybil misclassification rate with undersampling.
  - Important to align with client's tolerance level: Is it acceptable?
- Recommendation:
  - Consider a two-model approach:
    - One model focused on accurately classifying Humans.
    - Another model focused on Sybils.
  - This allows for specialized models that can optimize for the unique characteristics and importance of each class.

Key Takeaways: Reflecting on Our Discoveries

- Data Insights:
  - Imbalanced class distribution.
  - Rich yet complex passport data.
- Feature Observations:
  - No clear correlation between Confidence Score and number of credentials.
  - passport features like number of credentials and provider specific information proved valuable.
- Model Insights:
  - Random Forest showcased robust performance.
- Sampling Techniques:
  - Undersampling improved Human recall but with trade-offs.
- Business Considerations:
  - Model performance directly impacts user experience.
  - Alignment with business goals is paramount.

Key Takeaways: Reflecting on Our Discoveries

- Top 5 Features: All are on-chain event providers.
  - Signifies their critical role in defining users.
- Implications:
  - These features should have higher weightage in the Passport Score.
  - Can be key differentiators between Humans and Sybils.
- Product Recommendations:
  - Integrate more of such on-chain event providers in Gitcoin Passport.
  - Examples for future integration:
    - More staking integrations.
    - Aave supplier/borrower integration.



Continuous Learning: Iteration and Model Refinement

- Continuous Evolution:
  - Recognize that deploying a model is not the conclusion of the ML journey.
  - Models often require periodic updates and refinements.
- Shifting Business Objectives:
  - As company goals evolve, the criteria for model success may shift.
  - Stay aligned with business priorities to ensure the model remains relevant.
- Evaluation Metrics:
  - Periodically reassess the model using updated metrics to maintain optimal performance.
  - Consider using tools like A/B testing to compare the efficacy of old vs. new models.
- Data Dynamics:
  - Data sources and features can change over time.
  - Regularly check for data drift or anomalies that might impact model performance.

Horizon Scanning: A Glimpse into Future Improvements

- Model Robustness:
  - Implement a Train-Test-Validation split for better evaluation.
  - Utilize cross-validation for more robust model assessments.
- Model Exploration:
  - Iterate on different models to evaluate performance variations.
  - Consider a two-model approach:
    - One focusing on accurate Human classification.
    - Another emphasizing Sybil detection.
- Complexity:
  - Experiment with more intricate models like Gradient Boosted Trees, Neural Networks, or Ensemble methods.
  - Anomaly Detection: Instead of supervised learning, consider unsupervised anomaly detection techniques to identify unusual or suspicious behavior.
- Customization for Users:
  - Develop adjustable models:
    - Allow clients to prioritize recall or precision based on their needs.
    - Offer the capability to select from a suite of models based on specific requirements.

Questions & Answers