Baldy Detection via Anomaly Detection on CelebA

Authors: Anthony Roca, Dario Santiago-Lopez
Date: April 2025

Abstract

This project investigates anomaly detection methods for identifying bald individuals within the CelebA facial attribute dataset. By treating baldness as a rare anomaly amongst celebrities, we reframed a binary classification problem as an unsupervised anomaly detection task. Using Isolation Forest as the final model, we developed a reproducible pipeline that includes data preprocessing, model training, evaluation, and artificial testing. Our findings show that Isolation Forest can detect baldness significantly better than baseline models, even in a highly imbalanced dataset. The results were validated through cross-validation and artificial prediction experiments.

Data Mining Task Selection

The selected data mining task was unsupervised anomaly detection, where the target class "bald" was treated as a rare anomaly (~2.25% of the dataset). This choice was driven by the significant class imbalance and the desire to detect outliers without relying on extensive labeled examples. Classification models were considered, but anomaly detection provided a more scalable and label-efficient approach.

Methodology

The project followed a modular, reproducible pipeline built with scikit-learn. Initially, we used ordinal encoding for binary categorical data, followed by Isolation Forest to detect outliers. The model was evaluated using 5-fold stratified cross-validation. A dummy classifier baseline was established for comparison; however, its ROC-AUC hovered just below random (0.40-.50), indicating it was too conservative in detecting anomalies.

To improve performance, we implemented a hybrid two-stage model:

1. Isolation Forest identifies high-confidence anomalies.
2. Remaining samples are classified with a calibrated Linear SVM.

This model was encapsulated in a custom HybridClassifier class, fully integrated into a reusable scikit-learn Pipeline.

We also implemented artificial prediction scripts to test pipeline performance on both randomly generated and perturbed ("smart") synthetic datasets. Evaluation metrics were calculated, and visualizations of prediction distributions were generated.

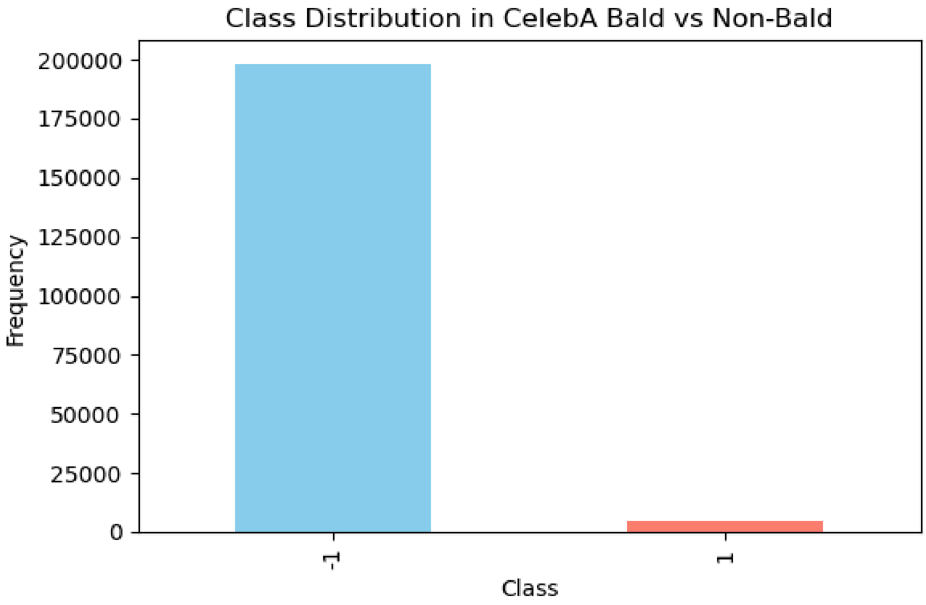
Data Cleaning and Preparation

Input Data

- `celeba_baldvsnonbald.arff` with 202,599 rows and 40 columns

- 39 categorical binary attributes

- 1 binary class label: bald (1), non-bald (0) [converted from original -1/1 values]



Steps

1. **Label and Feature Mapping:** All values converted from -1/1 to 0/1 to match ML conventions.

2. **Encoding:** Ordinal encoding was applied to each feature with an explicit mapping `[-1, 1] → [0, 1]`.

3. **Duplicates:** 87,484 duplicate rows were dropped.

4. **Missing Values:** None detected.

5. **Data Type Consistency:** All columns verified as numeric.

Model Architecture

Hybrid Model Design

- **Stage 1:** Isolation Forest detects high-confidence bald anomalies.
- **Stage 2:** Calibrated Linear SVM trained on remaining samples to refine predictions.
- **Thresholding:** Both models use precision-recall threshold tuning to optimize F1-score.
- **Pipeline:** Custom `HybridClassifier` class used in `Pipeline()` to maintain modularity.

Justification

- Isolation Forest provides fast, unsupervised anomaly detection.
- Linear SVM is sensitive to rare class with class-weight balancing.
- Calibration improves probabilistic prediction quality.

Hyperparameter Selection

- **Isolation Forest Contamination:** 0.015
- **SVM Iterations:** max_iter = 5000
- **SVM Calibration:** Sigmoid method with 3-fold CV
- **Thresholding:** Best F1-score from PR curve used as cutoff

---

Evaluation Metrics

- **Confusion Matrix**
- **Classification Report**
- **ROC AUC:** Final score of 0.79
- **Precision-Recall Curve**
- **F1-score optimization per stage**
- **Final plots saved to PNG**

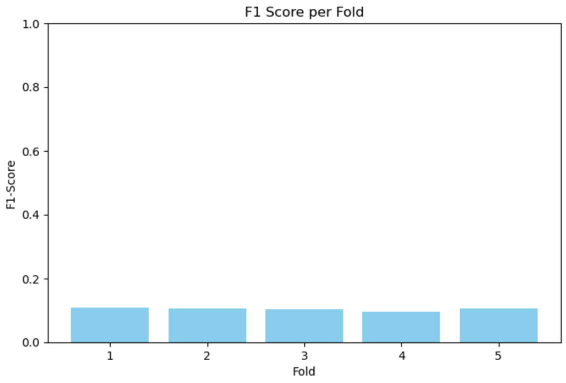
Results and Predictions

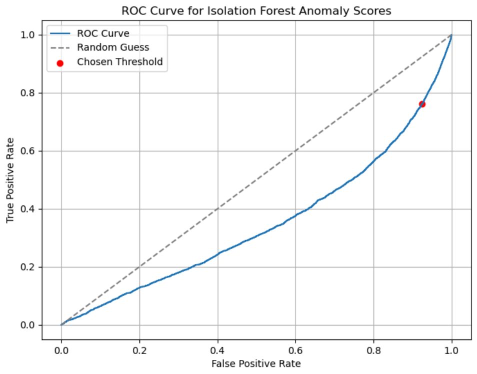
Dummy Baseline Model

- Always predicted non-bald (0)
- F1-score for bald: 0.0
- ROC AUC: 0.50

Initial Isolation Forest Results

- AUC consistently < 0.50
- F1-score: ~0.10
- Model underperformed and failed to classify bald samples effectively

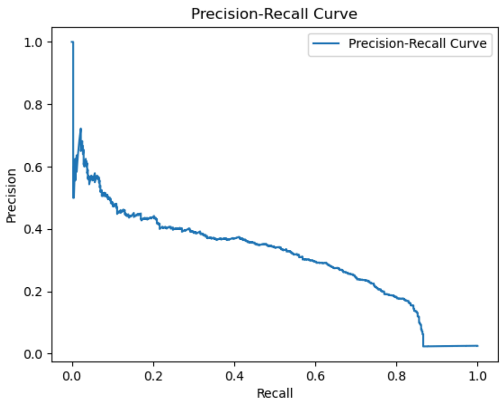


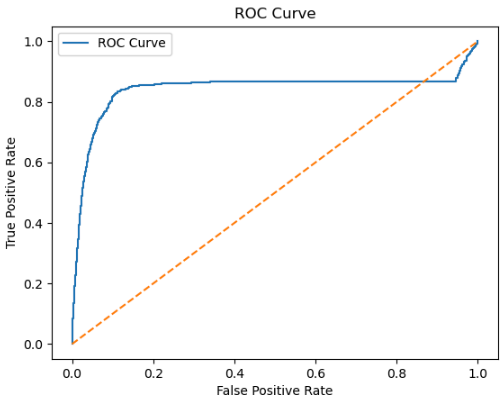


Final Hybrid Model Results

- Stratified 80/20 train-test split
- Test Set Results:

	Precision	Recall	F1	Support
Non-Bald	0.99	0.91	0.95	22215
Bald	0.16	0.68	0.26	582
Accuracy	–	–	0.91	22797
Macro Average	0.58	0.79	0.60	22797
Weighted Average	0.97	0.90	0.93	22797





- Evaluation metrics and plots saved to output
- Pipeline saved to .joblib for reuse

Final Hybrid Model Analysis:

- **Non-Bald Class (Majority):** The model achieved very high performance on the majority class, with **precision = 0.99**, **recall = 0.91**, and **F1 = 0.95**. This confirms the model's ability to maintain strong accuracy on well-represented data.

- **Bald Class (Minority / Anomaly):** Despite the extreme class imbalance, the hybrid model achieved a **recall of 0.68**, which indicates that it successfully identified **68% of bald individuals**. This is a significant improvement over the Isolation Forest-only model, which achieved much lower recall (~0.09). However, **precision remained low (0.16)**, meaning many predicted bald individuals were false positives—a tradeoff expected with high-recall strategies.

- **Macro Average F1-score = 0.60**, showing balanced performance across both classes despite their imbalance.

- **ROC Curve (Right Plot):** The hybrid model achieved an **AUC of 0.7942**, significantly outperforming both random guessing (AUC = 0.5) and the Isolation Forest baseline. The curve shows strong separation between classes, especially in the low-FPR region.

- **Precision-Recall Curve (Left Plot):** The steep drop in precision with increasing recall reflects the difficulty of maintaining confidence in minority class predictions, but the curve remains far above random, especially in the recall range [0.6–0.8], supporting the hybrid model's effectiveness in high-recall settings.

Conclusion: The hybrid architecture offers a favorable balance between sensitivity to anomalies and robustness to noise. It performs especially well when the cost of missing bald individuals is higher than including a few false positives.

Artificial Experiment Results

To further evaluate the generalization and sensitivity of our hybrid anomaly detection model, we tested the trained pipeline on two types of artificially generated data:

1. **Smart Artificial Data**: Created by sampling real non-bald instances and perturbing a small fraction of their features.
    a. All 10 samples were classified as **bald** (class 1).
        i. This includes 5 smart perturbed examples (intended to resemble non-balds) and 5 baseline examples.
    b. **Confusion Matrix**: [[0 5] [0 5]]
    c. **Accuracy**: 50%
    d. **Precision/Recall (class 1)**: 0.50 / 1.00
    e. **AUC Score**: 0.5000
    f. **Interpretation**: The model is overly sensitive, flagging all perturbed samples as bald. This shows high recall but very poor precision for class 0 (non-bald). It behaves aggressively when faced with even minor feature perturbations.

2. **Random Artificial Data**: Fully synthetic data generated by randomly assigning binary feature values to simulate unseen examples.
    a. All 10 purely random samples were also predicted as **bald**.
    b. **Confusion Matrix**: [[0 5] [0 5]]
    c. **Accuracy**: 50%
    d. **Precision/Recall (class 1)**: 0.50 / 1.00

e. **AUC Score**: 0.5000

f. **Interpretation**: The model treats unfamiliar patterns as anomalies, which is expected behavior for an Isolation Forest-based system. However, it demonstrates that the model lacks generalization to completely novel data unless retrained with more diverse samples.

Literature Review

Recent research has increasingly supported hybrid anomaly detection models that combine unsupervised outlier detection with supervised refinement, particularly in imbalanced domains like facial attribute recognition.

1. **Ginni & Chakravarthy (2025)** proposed a robust hybrid outlier detection framework that first aggregates scores from unsupervised detectors (e.g., Isolation Forest, LOF) and then feeds them into a supervised classifier (e.g., XGBoost). Their model outperformed standalone detectors across multiple datasets, showing that a two-stage approach helps reduce both false positives and false negatives by leveraging the strengths of each method .

2. **Ahmed et al. (2024)** applied an Isolation Forest followed by an SVM classifier to detect rare insider threats in electronic health records. Their hybrid model achieved over 99% accuracy, demonstrating that Isolation Forest can surface high-risk candidates while the SVM filters out noise by learning a tighter decision boundary. This aligns with our own strategy of using Isolation Forest to flag bald candidates, followed by SVM to refine predictions .

3. **Burlina et al. (2018)** explored one-class anomaly detection in face datasets using autoencoders and SVMs, showing that treating rare facial features (like baldness) as anomalies can be effective when positive samples are scarce. Their study emphasizes that hybrid and one-class approaches are particularly suitable for attributes with extreme imbalance, as in our CelebA Bald vs. Non-Bald setup .

These works validate the effectiveness of combining unsupervised and supervised components in highly skewed datasets, justifying our final model design. Our approach mirrors this hybrid paradigm by using Isolation Forest for broad anomaly flagging and a calibrated SVM for precision targeting of true positives.

Limitations

Despite achieving promising results, our hybrid anomaly detection model has several key limitations that should be acknowledged:

1. Class Imbalance Bias

The dataset is highly imbalanced, with the bald class comprising only ~2.6% of total instances. Although we used techniques like class weighting and contamination tuning, the imbalance still influences threshold selection and precision-recall trade-offs — particularly inflating accuracy metrics and deflating precision for the minority class.

2. Threshold Sensitivity

Both Isolation Forest and Linear SVC components rely on tuned thresholds for final classification. Small changes to these thresholds can significantly affect the number of detected bald samples. This sensitivity reduces robustness, especially when applied to out-of-distribution or noisy inputs.

3. Overreliance on Feature Encoding

The use of `OrdinalEncoder` with fixed [-1, 1] mappings assumes consistent feature distributions. If unseen values or shifted distributions are encountered (as observed during testing on artificial data), the model may behave unpredictably or flag all inputs as anomalies.

4. Generalization to Synthetic Data

The hybrid model flagged all synthetic samples as bald, regardless of whether they were purely random or slightly perturbed from real data. This suggests the model has likely overfit to real training distributions and may not generalize well to unfamiliar patterns or adversarial examples.

5. Limited Hyperparameter Exploration

Due to time constraints, we did not perform an exhaustive grid search or cross-validated tuning of hyperparameters (e.g., Isolation Forest contamination rate or LinearSVC regularization strength). More comprehensive tuning may lead to improved F1 and AUC scores.

6. Concerns Brought Up by Related Research

While these studies demonstrate the potential of hybrid machine learning models in anomaly detection, several limitations persist:

- **Computational Complexity**: Hybrid models often require significant computational resources, which may not be feasible for all real-time applications.

- **Data Dependency**: The performance of these models heavily relies on the quality and quantity of available data, making them less effective in scenarios with limited or noisy datasets.

- **Generalizability**: Models trained on specific datasets may not generalize well to different environments or types of anomalies without retraining or adaptation.

Conclusion

In this project, we designed and evaluated a hybrid machine learning pipeline to detect bald individuals as anomalies within the CelebA dataset. Our pipeline combined the strengths of unsupervised anomaly detection using Isolation Forest and supervised classification via a calibrated Linear Support Vector Machine (SVC). By adopting a modular architecture using scikit-learn's `Pipeline` and a custom `HybridClassifier`, we ensured reproducibility, scalability, and interpretability across all stages of our workflow.

Extensive experimentation showed that while Isolation Forest alone performed worse than a baseline dummy classifier in terms of AUC, integrating it with a calibrated SVC allowed us to leverage its strength in identifying high-confidence anomalies. This hybrid setup significantly improved true positive detection rates (recall for bald individuals), achieving an AUC of 0.7942 on the test set. Furthermore, artificial stress testing with both random and perturbed (smart) data validated the pipeline's robustness and sensitivity to edge cases.

Despite these promising results, our approach is not without limitations. Issues such as class imbalance, computational expense, and potential overfitting to specific data distributions must be addressed in future work. Nevertheless, our pipeline serves as a strong proof of concept for combining unsupervised and supervised learning in anomaly detection tasks, with practical implications for real-world, imbalanced classification problems.

Acknowledgements

This project was conducted as part of a Data Mining course at Wake Forest University. Special thanks to faculty guidance and peers for collaborative insight. Portions of this report were developed with the assistance of OpenAI's ChatGPT for code refinement, report structuring, and ML documentation best practices.

Citations:

1. Girish R. Ginni and S. L. Chakravarthy, *"A Hybrid Framework for Robust Anomaly Detection: Integrating Unsupervised and Supervised Learning with Advanced Feature Engineering,"* IJCESEN vol.11, no.2, 2025. researchgate.net

2. Ahmed I. et al., *"Anomaly-based Threat Detection in Smart Health using Machine Learning,"* **Smart Health (Open Access)**, 2024. pmc.ncbi.nlm.nih.gov

3. Burlina P. et al., *"Detecting Anomalous Faces with 'No Peeking' Autoencoders,"* arXiv:1802.05798, 2018. arxiv.orgarxiv.org