VERITAS: Verifiable Evaluation and Reporting In Transparent AI Systems

Author:

Gerald Enrique Nelson Mc Kenzie

Abstract:

In this work, I present VERITAS, a self-introspective engine designed to autonomously generate comprehensive model cards for machine learning systems. VERITAS integrates performance monitoring, explainable AI, bias auditing, and compliance logging into a unified framework. By enabling models to generate their own “model cards” VERITAS enhances transparency, accountability, and trustworthiness in AI deployments. The Iris dataset is used for validation, using a RandomForestClassifier, and demonstrates that VERITAS produces verifiable, detailed reports that capture the model’s capabilities, limitations, and operational metrics.

1. Introduction

The growing complexity of machine learning models has prompted the development of tools to document their inner workings, performance metrics, and potential biases. Traditional model cards require manual or semi-automated generation, which can lead to outdated or incomplete documentation. VERITAS (Verifiable Evaluation and Reporting In Transparent AI Systems) is designed to address these shortcomings by enabling AI models to autonomously generate self-reflective reports. This paper details the architecture, methodology, and experimental evaluation of VERITAS, underscoring its potential to improve model governance and foster trust in AI systems.

2. Related Work

Recent efforts such as Model Cards [Mitchell et al., 2019] and NVIDIA's ModelCard++ have advanced the cause of transparent AI. However, these methods still rely on human intervention for model documentation. Research in explainable AI (e.g., SHAP [Lundberg & Lee, 2017] and LIME [Ribeiro et al., 2016]) has contributed valuable techniques for model introspection, yet fully automated, self-reporting systems remain underexplored. VERITAS bridges this gap by combining automated performance evaluation, explainability, bias auditing, and governance logging into one coherent framework.

3. Methodology

3.1. System Architecture

VERITAS is structured into four main modules:

1. **Self-Monitoring Module:**

2. Collects performance metrics (e.g., accuracy, confusion matrix) during model evaluation.

3. **Explainability Engine:**

Uses techniques such as SHAP to generate explanations for model predictions. For demonstration purposes, the mean absolute SHAP values are computed for individual features.

4. **Bias and Fairness Auditor:**

Implements preliminary checks for bias, reporting on fairness metrics and known limitations.

5. **Reporting Module:**

Integrates the outputs from the other modules to generate a comprehensive model card in JSON format. Additional compliance and governance logs are embedded to support regulatory audits.

3.2. Implementation Details

The implementation is provided as an open-source code repository (listed in the appendix). In the prototype, the Iris dataset from scikit-learn is used, which is small and well-suited for visualization and demonstration. The following pseudo-code outlines the primary operations:

veritas_monitor.py

class VeritasMonitor:

def **init**(self, model, data, feature_names, compliance_info=None): # Initialize with a trained model and dataset ...

```
def evaluate_performance(self):
    # Predict on test data and calculate accuracy and confusion
matrix
    ...

def generate_explanation(self, sample_index=0):
    # Compute SHAP values for a sample to explain predictions
    ...

def audit_bias(self):
    # bias detection logic
    ...

def generate_model_card(self):
    # Compile model information, performance metrics,
explainability data,
    # bias audit results, and compliance logs into a report
    ...
```

```
    def save_model_card(self, filename="veritas_model_card.json"):
        # Save the model card as a JSON file
        ...
```

Example usage with the Iris dataset

if __name__ == "__main__":

    # Load Iris data, train a RandomForestClassifier,

    # instantiate VERITAS, and generate a model card

...

A complete version of this code, along with detailed documentation, is available in the repository.

3.3. Experimental Setup

Experiments were conducted using the Iris dataset and a RandomForestClassifier. The small dataset size facilitates rapid execution and easy visualization of performance metrics and SHAP explanations. All experiments were run on standard hardware without requiring significant computational resources.

4. Results

4.1. Model Performance

The VERITAS module successfully evaluated the model's performance, generating a model card that includes:

- **Accuracy:** A measured accuracy score on the test set.
- **Confusion Matrix:** Detailed confusion matrix outputs in JSON format.

4.2. Explainability

A SHAP-based explanation was generated for a selected test sample. The report includes:

- **Mean Absolute SHAP Values:** Quantitative importance of each feature.
- **Feature Names:** A mapping of SHAP values to the corresponding input features.

4.3. Bias and Fairness Audit

Preliminary bias audits reported no significant bias in the evaluated model. The audit summary is included in the final model card.

4.4. Governance and Compliance

Audit logs and compliance information (e.g., GDPR compliance notes) were automatically embedded into the report, ensuring that all aspects of the model's operation are traceable.

5. Discussion

VERITAS demonstrates the feasibility of a self-generative model card system, automating the introspection and documentation process. While the prototype uses simplified bias auditing and compliance logging, future work will integrate more advanced statistical tests and secure logging mechanisms. The modular architecture of VERITAS facilitates its extension to other models and datasets, paving the way for broader adoption in production environments.

6. Conclusion

I presented VERITAS, a novel framework for automated, self-generated model cards that enhances transparency and accountability in AI systems. The initial experiments on the Iris dataset confirm that VERITAS can successfully capture essential model characteristics, generate meaningful explanations, and log compliance-related

information. Future research will focus on expanding the bias auditing capabilities and integrating advanced security measures.

References

- Mitchell, M., et al. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*.
- Lundberg, S.M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Appendix: Code Availability

The full implementation of VERITAS, including detailed instructions and examples, is available in the GitHub repository:

https://github.com/lordxmen2k/VERITAS