Automated Model Selection and Explanation Through Counterfactual Analysis and Visualization

Adir Elmakais

ID: 316413640

March 1, 2025

Abstract

This project addresses the limitations of traditional model selection methods by automating the process and incorporating counterfactual explanations and visualizations to provide deeper insights into model behavior. Standard evaluation metrics often fail to capture the nuanced decision-making processes of machine learning models, hindering informed model selection and explainability. Our proposed system trains multiple models, generates diverse counterfactual explanations using DiCE, and visualizes these explanations comparatively, offering a user-friendly tool for enhanced AI transparency. Experimental evaluation across four datasets demonstrates the effectiveness of our approach in providing richer model understanding and informing model selection beyond standard performance metrics. The integration of counterfactual analysis offers a practical pathway to improve transparency and support more informed decision-making in machine learning.

1 Problem Description

Traditional model selection in machine learning heavily relies on aggregate performance metrics such as accuracy, F1-score, and AUC-ROC. While these metrics are valuable for quantifying overall predictive power, they often fall short in providing a comprehensive understanding of model behavior. They offer limited insight into *how* a model arrives at its decisions, making it difficult to discern the nuances of its decision-making process, assess its robustness, and identify potential biases. This "black-box" nature of model evaluation poses significant challenges, especially in domains requiring transparency and accountability, such as healthcare, finance, and criminal justice.

Furthermore, relying solely on performance metrics can lead to suboptimal model selection. Two models might exhibit similar performance scores but differ significantly in their decision-making logic, robustness to input variations, and fairness implications. Traditional methods lack the tools to effectively compare models beyond their aggregate performance, hindering the selection of models that are not only accurate but also reliable, interpretable, and aligned with ethical considerations. The absence of deeper insights into model behavior makes it challenging for data scientists and stakeholders to confidently choose the most appropriate model for a given task and to effectively explain and justify model predictions to end-users. This project aims to address these limitations by introducing an automated model selection framework that incorporates counterfactual explanations and visualizations, moving beyond superficial performance metrics to offer a more nuanced and insightful model evaluation process.

2 Solution Overview

Our proposed solution is an automated system designed to enhance model selection by integrating counterfactual explanations and visualizations. This system addresses the limitations of traditional metric-based evaluation by providing deeper, more interpretable insights into model behavior, facilitating a more informed and transparent model selection process. The core components of our solution are:

1. **Automated Model Training and Tuning:** The system begins by automating the training of a diverse set of machine learning models, including Logistic Regression, Decision Tree, RandomForest, SVM, XGBoost, and LightGBM. For each model, hyperparameter optimization is performed using RandomizedSearchCV to ensure that each model is evaluated at its near-optimal configuration. This automated training and tuning stage establishes a robust set of candidate models for comparison.

2. **Counterfactual Explanation Generation with DiCE:** To move beyond mere performance metrics, we leverage the DiCE (Diverse Counterfactual Explanations) library to generate counterfactual explanations for each trained model. DiCE allows us to understand how input features need to be modified to alter a model's prediction, providing insights into the model's decision boundaries and feature sensitivities. We generate counterfactuals for misclassified instances in the test set to specifically examine scenarios where the model's predictions deviate

from the ground truth. Constraints are incorporated when necessary (e.g., for sensitive features like gender and race in the Adult dataset) to ensure fairness and prevent the generation of unethical or unrealistic counterfactuals.

3. **Comparative Visualization of Counterfactuals:** A key innovation of our system is the comparative visualization of counterfactual explanations across different models. For selected misclassified instances, we generate and present counterfactuals from each model side-by-side. These visualizations include:

    - **Bar Charts of Scaled Feature Differences:** These charts visually represent the changes in feature values (in scaled units) required to generate counterfactuals, highlighting the features most sensitive to prediction changes for each model.
    - **Tables of Unscaled Feature Differences:** Complementing the scaled charts, tables present the actual, unscaled changes in feature values, making the counterfactual explanations more interpretable and actionable in the original feature space.

4. **Quantitative Evaluation of Counterfactuals:** To objectively assess the quality of the generated explanations, we compute and analyze several counterfactual metrics:

    - **Validity Percentage:** The proportion of generated counterfactuals that successfully flip the model's prediction to the desired outcome.
    - **Average Sparsity:** The average number of features modified in each counterfactual, indicating the simplicity of the explanation.
    - **Average L1 and L2 Distances (Unscaled):** Metrics quantifying the proximity of counterfactuals to the original instances in the unscaled feature space, reflecting the magnitude of changes required.

5. **Integrated Model Selection Framework:** By combining performance metrics (accuracy, F1-score, AUC-ROC) with counterfactual visualizations and quantitative metrics, our system offers a holistic model selection framework. This framework allows users to not only select models based on predictive accuracy but also to consider their interpretability, robustness, and fairness implications, leading to more informed and responsible model deployment.

The workflow of our system is as follows: (1) Data loading and preprocessing, (2) Automated training and hyperparameter tuning of multiple models, (3) Selection of misclassified instances, (4)

Generation of counterfactual explanations using DiCE for each model and selected instances, (5) Visualization of counterfactual explanations for comparative analysis, (6) Computation of counterfactual metrics, and (7) Integrated model evaluation and selection based on both performance and explainability metrics. This systematic approach provides a comprehensive toolkit for data scientists seeking to move beyond black-box model evaluation and embrace more transparent and insightful model selection practices.

3 Experimental Evaluation

3.1 Evaluation Metrics

To rigorously evaluate the effectiveness of our automated model selection and explanation system, we employ a comprehensive set of evaluation metrics, encompassing both traditional performance measures and novel metrics derived from counterfactual analysis. These metrics are chosen to assess not only the predictive accuracy of the models but also their interpretability, robustness, and the quality of the explanations they provide.

Performance Metrics (Traditional):

- **Accuracy:** The proportion of correctly classified instances in the test set. While accuracy is a general indicator of performance, it can be misleading in imbalanced datasets, which are present in our study.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of a model's performance, particularly useful in imbalanced datasets. It assesses the model's ability to minimize both false positives and false negatives.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** Measures the model's ability to discriminate between the positive and negative classes across various classification thresholds. AUC-ROC is especially valuable as it is insensitive to class imbalance and provides insight into the model's ranking capability.

Counterfactual Explanation Metrics (Novel):

- **Validity Percentage:** This metric quantifies the success rate of the counterfactual generation process. It is calculated as the percentage of generated counterfactual instances that

successfully flip the model's prediction to the desired, corrected class. High validity is crucial, indicating that the counterfactuals are indeed effective in altering model predictions as intended.

- **Average Sparsity:** Sparsity measures the simplicity of the counterfactual explanations. It is defined as the average number of features that are modified in each counterfactual instance to change the prediction. Lower sparsity is preferred as it implies simpler, more interpretable explanations that are easier for users to understand and potentially act upon.

- **Average Unscaled L1 Distance:** The average L1 distance (Manhattan distance) between the original instance and its counterfactual counterparts, calculated in the original, unscaled feature space. Lower L1 distance indicates that the counterfactual instances are "closer" to the original instances, suggesting more realistic and potentially more actionable explanations.

- **Average Unscaled L2 Distance:** The average L2 distance (Euclidean distance) between the original and counterfactual instances in the unscaled feature space. Similar to L1 distance, lower L2 distance signifies closer proximity, but L2 is more sensitive to larger changes in individual features.

3.2 Baseline Approach

To demonstrate the added value of our automated model selection and explanation system, we compare its performance against a **baseline approach** that represents a traditional, metric-centric model evaluation process. Our chosen baseline is a **Decision Tree classifier trained with default hyperparameters**. The Decision Tree serves as a relevant and effective baseline for several reasons:

- **Simplicity and Common Use:** Decision Trees are a fundamental and widely used classification algorithm. They are often employed as a starting point in machine learning projects due to their ease of implementation and interpretability (in their basic form). Using a Decision Tree with default settings represents a straightforward, "out-of-the-box" approach to model building and evaluation, mirroring common practices where initial model selection might be driven primarily by readily available performance metrics.

- **Weakness Compared to Tuned Models:** Decision Trees with default hyperparameters are generally less performant than more complex models or even tuned Decision Trees. We

expect our advanced approach, which includes hyperparameter tuning and the consideration of multiple model types, to outperform this baseline in terms of predictive accuracy.

- **Lack of Explainability Beyond Feature Importance:** While Decision Trees themselves are somewhat interpretable (especially smaller trees), relying solely on a Decision Tree for model selection does not inherently incorporate advanced explainability techniques like counterfactual analysis. The baseline evaluation focuses on performance metrics without delving into the nuanced decision-making processes that counterfactuals reveal. This highlights the gap that our proposed system aims to bridge.

- **Fair Comparison Point:** By comparing against a Decision Tree, we establish a clear and quantifiable benchmark. Improvements over this baseline will demonstrate the tangible benefits of incorporating automated counterfactual analysis and visualization into the model selection workflow.

3.3 Results and Comparison

Dataset Performance Metrics Comparison

Table 1: Performance Metrics: Baseline Decision Tree vs. Advanced Models

Metric	Baseline DT	LR	DT (Tuned)	RF	SVM	XGBoost	LightGBM
Adult Census Income Dataset							
Accuracy	0.8047	0.8428	0.8385	0.8477	0.8464	0.8544	0.8520
F1 Score	0.6015	0.6550	0.5910	0.6395	0.6449	0.6404	0.6591
AUC-ROC	0.7403	0.9010	0.8715	0.9038	0.8875	0.9082	0.9128
Breast Cancer Wisconsin (Diagnostic) Dataset							
Accuracy	0.9474	0.9825	0.9474	0.9649	0.9825	0.9737	0.9737
F1 Score	0.9302	0.9767	0.9268	0.9524	0.9762	0.9647	0.9647
AUC-ROC	0.9440	0.9971	0.9463	0.9944	0.9974	0.9921	0.9954
Statlog (German Credit) Dataset							
Accuracy	0.7000	0.8050	0.7550	0.7750	0.7850	0.7750	0.7900
F1 Score	0.7872	0.8660	0.8464	0.8544	0.8532	0.8553	0.8581
AUC-ROC	0.6394	0.8180	0.7640	0.8256	0.7936	0.8205	0.8170
Wine Quality Dataset							
Accuracy	0.6983	0.7434	0.7321	0.7726	0.7575	0.7641	0.7744
F1 Score	0.7599	0.8040	0.7930	0.8239	0.8122	0.8164	0.8240
AUC-ROC	0.6781	0.7995	0.7880	0.8369	0.8292	0.8270	0.8306

Counterfactual Metrics Comparison

Table 2: Counterfactual Explanation Metrics: Advanced Models

Metric	LR	DT (Tuned)	RF	SVM	XGBoost	LightGBM
Adult Census Income Dataset						
Validity (%)	100.0	100.0	100.0	95.0	100.0	100.0
Avg. Sparsity	1.40	1.30	1.65	1.65	1.35	1.30
Avg. L1 Distance	66461.74	27978.70	45502.91	27393.76	30402.27	629.53
Avg. L2 Distance	66461.40	27978.24	45498.76	27318.10	30401.94	629.30
Breast Cancer Wisconsin (Diagnostic) Dataset						
Validity (%)	100.00	83.33	87.50	100.00	100.00	100.00
Avg. Sparsity	1.75	1.75	1.50	1.50	2.00	1.83
Avg. L1 Distance	1166.039	189.948	3.098	417.974	59.804	3.348
Avg. L2 Distance	1166.026	175.431	3.065	417.908	58.380	3.292
Statlog (German Credit) Dataset						
Validity (%)	100.0	100.0	100.0	80.0	100.0	100.0
Avg. Sparsity	1.85	1.60	3.00	2.45	3.00	1.85
Avg. L1 Distance	1311.96	3813.13	3637.57	1527.43	1657.26	3384.28
Avg. L2 Distance	1311.14	3812.62	3628.50	1524.80	1650.82	3378.19
Wine Quality Dataset						
Validity (%)	100.0	100.0	100.0	90.0	100.0	100.0
Avg. Sparsity	1.60	1.70	1.65	1.40	1.70	1.50
Avg. L1 Distance	17.947	15.181	19.042	46.021	48.172	20.688
Avg. L2 Distance	17.726	14.554	18.580	41.809	46.704	19.876

4 Related Work

This project builds upon and contributes to a growing body of research in automated model selection and explainable AI (XAI), particularly in the area of counterfactual explanations. Several existing tools and techniques are relevant to our work, and we draw inspiration from and differentiate our approach from them.

4.1 Automated Model Selection and Hyperparameter Optimization

Recent work in AutoML includes [1] and [2]. Unlike these approaches, our system uniquely integrates counterfactual explanations into the automated model selection pipeline. While methods like Auto-sklearn [1] and those explored in NASBench [2] focus on efficiently searching the model and hyperparameter space to optimize predictive performance, they primarily rely on traditional validation metrics. Our system extends this paradigm by incorporating explainability as a crucial criterion for model selection. By generating and visualizing counterfactuals for multiple candidate models, we offer a more nuanced comparison that goes beyond mere performance scores, allowing users to select models not only for their accuracy but also for their interpretability and behavior in terms of feature sensitivities and decision boundaries. This integration of XAI techniques directly into the AutoML process provides a more comprehensive and transparent model selection framework.

4.2 Counterfactual Explanations and Explainable AI

Our work builds on DiCE [3] but extends it for model comparison rather than individual explanations. DiCE provides a powerful framework for generating diverse counterfactual explanations for individual instances, offering valuable insights into the decision-making of a single model. However, our system leverages counterfactuals in a novel comparative context. We generate counterfactuals across multiple models trained on the same dataset, and then comparatively visualize and quantitatively evaluate these explanations. This allows us to assess and contrast the explainability characteristics of different models. By focusing on model comparison through counterfactual analysis, we move beyond understanding individual predictions to evaluating the inherent explainability and robustness properties of entire models relative to each other. This comparative approach offers a unique perspective in the field of XAI, facilitating model selection based on not just *what* a model predicts, but also *how* it makes those predictions and how its behavior differs from other models.

References

[1] Feurer, M. et al. (2015). Efficient and Robust Automated Machine Learning. NIPS.

[2] Olson, R.S. (2021). Accelerating Neural Architecture Search. JMLR.

[3] Mothilal, R.K. et al. (2020). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. ACM FAccT.

5 Conclusion

In conclusion, this project has successfully developed and evaluated an automated system for model selection that moves beyond traditional performance metrics by incorporating counterfactual explanations and comparative visualizations. Our experimental results across four diverse datasets demonstrate the effectiveness of this approach in providing richer, more interpretable insights into model behavior, facilitating a more informed and transparent model selection process.