Feature Stability and Robustness in Tabular Machine Learning under Natural Distribution Shifts

1. Background and Motivation

Tabular machine learning models are widely used in real-world decision-making systems, including credit risk assessment, healthcare analytics, and insurance underwriting. While such models often achieve strong predictive performance during development, their reliability frequently deteriorates once deployed. A primary reason for this is that real-world data rarely remains stationary: economic conditions change, populations evolve, and institutional practices differ across time and space.

Most tabular learning methods implicitly assume that training and deployment data follow the same distribution. When this assumption is violated, models may rely on correlations that are predictive in historical data but unstable across different environments. As a result, performance degradation under distribution shift remains a major obstacle to the trustworthy deployment of tabular machine learning systems.

Recent advances in invariant and causal learning suggest that predictors based on stable, causally relevant relationships are more robust to such shifts. However, much of this work is theoretical or focused on synthetic settings and non-tabular domains. There is still limited empirical understanding of how feature stability manifests in real-world tabular data, and how it can be practically exploited to improve robustness in deployed systems.

This research is motivated by the need to bridge this gap between theory and practice.

2. Research Gap

Although distribution shift and invariant learning have received increasing attention, several important limitations remain in the current literature:

First, most studies evaluate robustness at the level of models, focusing on comparative performance, rather than examining stability at the level of individual features. Second, empirical investigations using real-world tabular datasets with naturally occurring environments are relatively scarce. Third, existing work rarely addresses the conditions under which stability-aware learning is beneficial, or when enforcing invariance may actually harm performance.

As a result, practitioners lack clear guidance on how to identify stable features, how stability relates to robustness, and when stability-based approaches should be preferred over standard predictive modeling.

3. Research Objectives

The overarching goal of this research is to develop a deeper empirical understanding of feature stability in tabular machine learning and its role in robustness under distribution shift.

Specifically, this study aims to:

- Examine how feature–outcome relationships vary across natural environments in real-world tabular data
- Quantify the relationship between feature stability and predictive robustness
- Evaluate the strengths and limitations of stability-aware learning methods in applied settings
- Provide practical insights for designing more reliable tabular prediction systems

4. Research Questions

The study is guided by the following research questions:

RQ1:
Which features in real-world tabular datasets exhibit stable predictive relationships across natural environments?

RQ2:
How does feature stability influence model robustness under different types of distribution shift, such as temporal or demographic changes?

RQ3:
Under what conditions does enforcing feature stability improve generalization, and when does it negatively affect performance?

RQ4:
Can feature-level stability measures provide practical guidance beyond standard accuracy-based evaluation?

These questions emphasize explanation and understanding, rather than model benchmarking alone.

5. Dataset and Experimental Setting

The primary dataset used in this study will be the **Home Credit Default Risk** dataset, which contains over 300,000 loan applications with rich demographic, financial, and credit-related features. This dataset is well suited for the proposed research, as it reflects realistic, high-stakes decision-making and contains naturally occurring heterogeneity across applicants.

Natural environments will be defined using meaningful data partitions, such as temporal segments, demographic groups, and credit-related strata. These partitions introduce distribution shifts that closely resemble real deployment scenarios, without relying on artificial data manipulation.

Where feasible, the findings will be validated on an additional tabular dataset from a different application domain to assess generalizability.

6. Methodology

The study will begin by establishing strong baseline models commonly used for tabular prediction, including logistic regression and gradient-boosting-based methods. These models will serve as reference points for evaluating robustness under distribution shift.

Stability-aware learning approaches, such as invariant risk minimization and regularization techniques that penalize environment-specific feature effects, will then be applied. Rather than focusing solely on model performance, the analysis will explicitly examine how individual feature contributions vary across environments.

To support this analysis, a feature stability metric will be defined to quantify the variability of feature effects across environments. This enables a systematic comparison between stable and unstable features and facilitates a feature-centric evaluation framework.

Model evaluation will emphasize robustness, measured through performance on unseen environments and performance degradation under increasing distribution shift, rather than single-split accuracy.

7. Expected Contributions

This research is expected to make the following contributions:

- An empirical characterization of feature stability in real-world tabular datasets under natural distribution shifts
- A feature-level evaluation framework that complements existing robustness and invariant learning approaches
- Practical insights into when stability-aware learning improves reliability and when it may be counterproductive
- Actionable guidance for practitioners designing robust tabular prediction systems

8. Significance

By shifting attention from model-level comparisons to feature-level behavior, this study aims to improve understanding of why tabular models succeed or fail under distribution shift. The focus on real-world data and natural environments enhances the practical relevance of the findings, while the emphasis on explanation and conditions for success aligns with the expectations of high-impact journals.

9. Conclusion

This research proposes an empirical investigation into feature stability as a key driver of robustness in tabular machine learning. By integrating causal intuition, invariant learning principles, and applied analysis, the study seeks to generate insights that are both scientifically meaningful and practically useful, contributing to the development of more reliable decision-support systems.