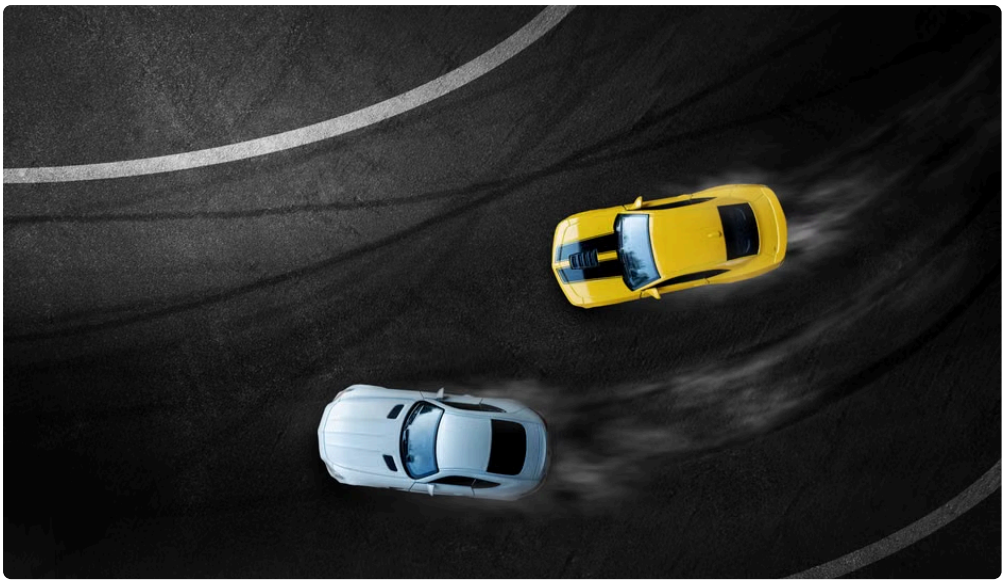
In the realm of risk-scored decision systems—particularly in lending and healthcare—the choice and treatment of features can make or break model performance and ethical compliance. One common pitfall is the use of proxy variables like zip codes, which often introduce contradictions, bias, and fairness issues. This blog post dives deep into why zip code proxy features create confusing model behaviors, how tools like **disagreement rate** and **predictive entropy** help diagnose issues, and strategies to mitigate risks via feature exclusion and better loss alignment.



Understanding the Problem: Proxy Variables and Model Contradictions

I've seen this play out countless times: wished they had known this beforehand.. The term proxy variables refers to features that indirectly represent sensitive https://reportz.io/ai/when-models-disagree-what-contradictions-reveal-that-a-single-ai-would-miss/ attributes such as race, income, or socio-economic status. Zip codes are a classic example—they can encode demographic or economic disparities that a model might use as shortcuts rather than true causal predictors.

Models trained with zip codes often produce *contradictory predictions*, undermining both fairness and accuracy. These contradictions manifest when the model outputs a positive risk classification for a borrower in one zip code but a negative risk classification for a highly similar borrower in a nearby zip code. Such inconsistencies are more than just technical quirks—they hint at deeper *distribution shifts* and *data gaps* embedded in the training data's geography-dependent sampling.

What Happens on the Worst Day in Prod?

Let me push my favorite question here: **what happens on the worst day in production?** If your model relies heavily on zip code proxies, worst-case scenarios include unintended disparate impact on protected subgroups, regulatory non-compliance under fair lending laws, and dramatic model degradation when geographic patterns change due to economic or demographic shifts.

High-Signal Risk Indicators: Disagreement Rate and Predictive Entropy

Two advanced diagnostic metrics help us understand and uncover model contradictions stemming from zip code proxies:

- **Disagreement Rate:** This measures the frequency at which multiple sub-models or ensemble members disagree on the prediction for a given instance. High disagreement on data points associated with particular zip codes signals uncertainty and inconsistency in the model's understanding.
- **Predictive Entropy:** An uncertainty metric derived from the predicted class probabilities. High entropy means the model has low confidence in its prediction, often suggesting conflicting or sparse data information.

Why These Metrics Matter

Both disagreement rate and predictive entropy serve as **early-warning signals** for:

- **Edge Cases:** Individual profiles not well represented in the training set, often in less populous or economically diverse zip codes.
- **Distribution Shifts:** Changing socio-economic patterns that break the original association between zip code and risk.
- **Proxy Confounds:** When zip code proxies encode multiple correlated factors, leading to contradictory signals when those factors uncouple.

By monitoring these signals in prod, teams can flag high-risk predictions where accuracy hides inconsistencies and fairness risks abound.

Data Gaps and Subgroup Coverage

One root cause of these contradictions is **data gaps**. Zip codes with sparse or skewed data samples cause the model to extrapolate in misleading ways. This leads to poor subgroup coverage, especially for minority or low-frequency populations within certain geographic boundaries.

Table 1 below illustrates typical zip-code-based coverage issues and their implications:

Zip Code Type Data Density Typical Model Behavior Fair Lending Risk High-density urban High Stable predictions, low entropy Lower risk Sparse rural Low High disagreement, unstable scores Higher risk for bias and disparity Economically diverse suburbs Medium Moderate disagreement, sensitive to shifts Moderate risk

Objective Mismatch and Loss Function Tradeoffs

The core ML optimization goal often doesn't align perfectly with fair lending objectives. For example, using a standard cross-entropy loss to maximize accuracy may inadvertently push the model to exploit zip code proxies that correlate highly with defaults but also with protected groups.

This **objective mismatch** leads to:

- Models overconfidently assigning risk scores based on proxy signals rather than causal factors.
- Inadequate representation of uncertainty or disagreement, camouflaging unfair treatment.
- Limited calibration of probability outputs, affecting threshold decisions tied to costs.

Instead, strategic tradeoffs in loss design can reduce these effects, such as incorporating:

- **Fairness-aware constraints** that penalize proxy reliance.
- **Uncertainty-based regularization** to encourage calibrated probability outputs reflecting predictive entropy.
- **Subgroup-aware metrics** directly embedded into the objective to minimize disparate impact.

How to Fix Zip Code Proxy Issues

1. **Evaluate Feature Importance with Caution:** Use explainability tools to identify zip code proxy effects beyond pure accuracy contributions.
2. **Monitor Disagreement and Entropy in Production:** Build monitoring dashboards focused on these metrics grouped by zip code and subgroups to identify emerging risk concentrations.
3. **Feature Exclusion or Transformation:**
    - Remove zip code features entirely when they do not meet fairness constraints.
    - Aggregate or anonymize zip code data into broader geographic or socio-economic indicators to reduce proxy effects.
    - Use adversarial debiasing techniques to remove zip code information from learned representations.
4. **Rebalance Training Data:** Collect more data or reweight samples in underrepresented zip codes and subgroups to close data gaps.
5. **Align Loss Functions with Fair Lending Goals:** Implement objectives that trade off accuracy and fairness, incorporating uncertainty regularization.
6. **Threshold Calibration Based on Costs:** Prioritize setting classification thresholds that respect economic and fairness costs instead of arbitrary accuracy-driven cutoffs.

Bonus: Common Quirks and Things Accuracy Hides

- Accuracy can improve by leaning on zip code proxies but may simultaneously increase subgroup disparity.
- Calibrated probabilities often reveal poor confidence, warning of overtrust in predictions from geographic proxies.
- Systems that only report test-set accuracy miss worst-day-in-prod scenarios caused by zip code shifts.

Conclusion

Zip code features are a double-edged sword—while they can boost predictive power, they also mask contradictions, biases, and calibration issues that jeopardize fair lending compliance and operational robustness. By systematically diagnosing model uncertainty through disagreement rate and predictive entropy, understanding data gaps and distribution shifts, and aligning loss functions with fairness goals, practitioners can mitigate these risks.

Ultimately, feature selection, monitoring, and loss engineering together form the foundation for safe, trustworthy decision systems that don't merely deliver high accuracy but also uphold fair and consistent treatment across geographies and subpopulations.

Remember: never trust overconfident probabilities blindly, always ask what happens on the worst day, and prioritize thresholds driven by cost-sensitive fairness tradeoffs—not vibes.