**Business Analytics Capstone**

**BA723**

**Monitoring and Governance Documentation**

**By**

**Aisha Akter Anamika**

**301439588**

**Monitoring and Governance**

In predictive modeling, drift monitoring is essential to ensure that a model continues to perform reliably over time. As new data is collected, the patterns and relationships in the data may change due to factors such as evolving populations, changes in data collection methods, or shifts in the real-world phenomenon being modeled. If these changes are not detected early, the model’s predictions can become inaccurate or biased.

Even if the target distribution remains stable, shifts in feature distributions can reduce model accuracy because the model was trained on a different feature landscape.

For this breast cancer prediction project, monitoring both types of drift ensures that the model continues to make accurate predictions and remains valid for real-world use.

**1. Model Drift Analysis**

In this analysis, the training (build) dataset represents the expected values, while the test dataset represents the new data being evaluated. In the future, as the model is run with different incoming datasets, the expected values will always be based on the original model build, and the actual values will be replaced with the results from each new dataset. This ensures that all future drift checks are consistently compared against the same baseline the model was originally trained on.

**Test Used:** Chi-square test for target distribution shift  
**Result:**

* Chi-square statistic (X²): **0.02566**
* Critical value at 95% confidence: **3.841** (df = 1)
* Conclusion: No significant shift in target distribution between training data and new test data.

**Interpretation:**  
The distribution of benign (Class 0) and malignant (Class 1) cases in the new data is consistent with the original training data. This means the model is still operating in a similar target environment and there’s no evidence of label distribution drift.

**Recommendation**

If the Chi-square statistic is very large with a p-value below 0.01, it signals strong evidence of drift, requiring investigation of the root cause and potentially retraining or recalibrating the model. In a medical context like breast cancer detection, even moderate drift warrants prompt attention to ensure recall does not decline, especially if the affected features are among the top predictors.

**Variable Drift Analysis**

**Test Used:** Population Stability Index (PSI) for the five predictors in the stepwise regression model.  
PSI measures the stability of feature distributions over time. Values close to **0** indicate no drift, while higher values indicate changes in the distribution that could affect model performance.

**Feature-by-Feature Results:**

1. **concave\_points\_worst — PSI = 0.0835**
   * This variable measures the severity of the concave points in the worst (largest) tumor area.
   * PSI below 0.10 means the distribution has remained stable between the training and new data.
   * No action is needed — the model is likely still interpreting this feature in the same way as during training.
2. **radius\_se — PSI = 0.1046**
   * Represents the standard error of the mean radius of the cell nuclei.
   * PSI slightly above 0.10 suggests a small distribution shift.
   * Although this is not yet critical, the change should be monitored more frequently to ensure it doesn’t grow into a significant drift that could affect model accuracy.
3. **texture\_mean — PSI = 0.1102**
   * Indicates the average variation in surface texture of cell nuclei.
   * PSI slightly above 0.10 indicates a minor change in distribution over time.
   * This may be due to differences in imaging quality or patient sample characteristics in the new data.
   * Recommend closer monitoring to detect if the change is increasing over time.
4. **concave\_points\_se — PSI = 0.0599**
   * Measures the standard error of the number of concave points in the tumor boundary.
   * PSI is well below 0.10, indicating no drift.
   * This feature remains consistent with the training data, so no immediate action is required.
5. **radius\_mean — PSI = 0.0718**
   * Represents the average radius of the cell nuclei.
   * PSI below 0.10 suggests the distribution is stable.
   * This feature’s predictive contribution is likely unaffected by data changes.

Summary Interpretation:

* Three features (concave\_points\_worst, concave\_points\_se, radius\_mean) show stable distributions (PSI < 0.1).
* Two features (radius\_se and texture\_mean) show minor drift (0.1 < PSI < 0.2) and should be monitored more closely to ensure the drift does not increase.
* No features are in the critical drift zone (PSI ≥ 0.25).

Key Assessment:

* Model drift result shows the target distribution is unchanged, meaning the model is still predicting in a familiar context.
* Variable drift result shows most features remain stable, but two features have slight shifts that could become problematic if trends continue.
* Since input distributions are mostly the same, there’s no immediate threat to model reliability — but these early shifts should be watched.

**4. Recommendations**

1. **Maintain Stability:**
   * Keep PSI values as close to **0** as possible to ensure feature stability over time.
   * For the two variables with PSI > 0.10 (radius\_se, texture\_mean), increase monitoring frequency to monthly instead of quarterly.
2. **Threshold-based Action Plan:**
   * **PSI < 0.1:** No action required.
   * **0.1 ≤ PSI < 0.2:** Monitor more frequently and investigate potential causes.
   * **PSI ≥ 0.25:** Retrain or recalibrate the model, as this indicates significant drift.