46. Exploring Statistical Models for Predicting Planetary Habitability A Comparative Analysis of Logistic Regression, Naive Bayes, and Quadratic Discriminant Analysis

**Abstract**

This research paper investigates various statistical methods to predict planetary habitability based on features such as Solar Radiation, Atmospheric Composition, and Distance from Star. Using several statistical models, including Quasi-Poisson, Binomial Generalized Linear Models (GLM), Logistic Regression, Naive Bayes, and Quadratic Discriminant Analysis (QDA), we aim to uncover the underlying relationships between planetary attributes and their habitability status. The findings highlight both the potential and limitations of these traditional models, suggesting the need for more sophisticated approaches to capture complex planetary dynamics.

**1. Data Overview and Statistical Modeling**

The datasets provided contain information on various planetary features such as Solar Radiation, Atmospheric Composition, Distance from Star, and their corresponding Habitability (0 or 1, indicating non-habitable and habitable, respectively). The analysis involves utilizing several statistical models to predict and understand the underlying relationships between these planetary attributes and their habitability status.

**Dataset Summary**

* **Planetary Features**:
  + **Solar Radiation**: Measure of the solar energy received by each planet.
  + **Atmospheric Composition**: The makeup of the atmosphere, which could range from essential gases to harmful elements.
  + **Distance from Star**: The distance of the planet from its star, impacting temperature and potential habitability.
  + **Habitability**: Binary classification indicating if a planet is habitable (1) or not (0).

**2. Model Interpretations**

**A. Logistic Regression Model for Planetary Habitability**

* **Dataset Summary**: This dataset has 200 observations across 5 features. We employed a binomial GLM to model habitability based on Solar Radiation, Atmospheric Composition, and Distance from Star.
* **Model Summary**:
  + **Coefficients**:
    - Intercept: -0.1455750, indicating the baseline log-odds of habitability when all features are zero.
    - Solar Radiation: Coefficient of -0.0003044, suggesting that increasing solar radiation slightly decreases the odds of habitability, but this result is not statistically significant (p = 0.569).
    - Atmospheric Composition: Coefficient of 0.0039738, indicating a positive but non-significant relationship (p = 0.417).
    - Distance from Star: Coefficient of 0.0005712, also not statistically significant (p = 0.918).
  + **Model Fit**:
    - Residual deviance = 275.90 on 196 degrees of freedom, indicating the model's goodness of fit.
    - AIC = 283.9 suggests a relatively moderate model fit.
  + **Accuracy**: The test set accuracy is approximately 51.67%, indicating the model's ability to predict planetary habitability is only slightly better than random guessing.

**B. Decision Boundary and Probability Predictions**

* **Logistic Regression Decision Boundary**:
  + The plot shows that the decision boundary is linear, separating habitable and non-habitable planets based on solar radiation and atmospheric composition. However, there is substantial overlap between the two classes, indicating a low discrimination capability of the logistic regression model.
* **Predicted Probability Plot**:
  + The scattered nature of predicted probabilities across different levels of solar radiation suggests no clear trend, confirming the low predictive power.

**C. Naive Bayes and QDA Models for Planetary Classification**

* **QDA Decision Boundaries**:
  + QDA allows for non-linear decision boundaries, which are evident in the complex curves shown in the planetary classification plot. This flexibility enables the model to capture more intricate relationships among planetary features.
  + However, the decision regions are highly interwoven, suggesting a model that may be overfitting to the training data.
* **Naive Bayes Decision Boundaries**:
  + The Naive Bayes classification boundaries show smoother contours compared to QDA, reflecting the assumption of feature independence. The model identifies regions of habitability with varying levels of probability density.
  + While visually distinct, Naive Bayes has limitations in accounting for correlated planetary features, which may explain some misclassification.

**D. Confusion Matrix Analysis**

The confusion matrix provides insights into the classification accuracy of the models. Both the QDA and Naive Bayes models demonstrate varying levels of accuracy across different planets, with several misclassifications. For example, some planets such as Gigglex and Guffaw-9 are frequently misclassified, indicating potential limitations in feature representation or model assumptions.

**3. Discussion of Findings**

**Trends and Insights**

* **Solar Radiation Impact**:
  + Across the models, solar radiation consistently shows a weak association with habitability, potentially indicating that the planets in this dataset have a diverse range of tolerances to solar exposure.
* **Atmospheric Composition**:
  + The variability in atmospheric composition, particularly in the Naive Bayes decision boundaries, suggests that atmospheric elements play a critical role in determining habitability. However, the independence assumption may oversimplify these relationships.
* **Distance from Star**:
  + The impact of the distance from the star is also not consistently significant across models, implying that other unobserved factors may be influencing habitability.

**Opportunities and Threats**

* **Opportunities**:
  + Refining feature engineering by considering additional planetary characteristics (e.g., gravitational force, surface temperature) could enhance model accuracy.
  + Exploring more complex models such as ensemble methods or neural networks might capture non-linearities and interactions more effectively than the current approaches.
* **Threats**:
  + Overfitting, as seen in the QDA model, remains a concern with high model complexity.
  + The assumption of feature independence in Naive Bayes may not hold true, potentially skewing results.

**4. Conclusion and Future Directions**

This research explores various statistical methods to predict planetary habitability, revealing both the potential and limitations of traditional models such as logistic regression, Naive Bayes, and QDA. The findings indicate that while simple models provide basic insights, they lack the sophistication required to capture complex planetary dynamics. Future studies could leverage richer datasets, advanced feature engineering, and state-of-the-art machine learning algorithms to improve predictive performance and deepen our understanding of exoplanet habitability.

**Proposed Actions**

1. **Data Augmentation**: Collect and integrate more granular planetary data to enrich feature sets.
2. **Model Experimentation**: Test alternative models like Random Forests, Gradient Boosting Machines, or Deep Neural Networks for improved accuracy.
3. **Validation Techniques**: Implement cross-validation techniques to assess model generalizability and mitigate overfitting risks.