Comparative Analysis of Rule-Based and ML-Based Startup Health Scoring Models

This document presents a detailed comparison between two methods used for evaluating startup health: a **manual, rule-based scoring model** and an **enhanced machine learning (ML)-based model**. Both methods rely on core startup features, but differ significantly in methodology, capabilities, and applicability.

Summary

The first notebook (Task_1.ipynb) implements a rule-based approach, where startups are scored using a weighted sum of normalized features such as team experience, market size, monthly active users, burn rate, funds raised, and valuation. The weights are predefined and applied directly, making this method simple, interpretable, and effective for quick analysis. However, it lacks adaptability and does not account for non-linear relationships or feature interactions.

The second notebook (Task1_ML.ipynb) builds on the same scoring logic but integrates machine learning models — specifically, **Random Forest Regressor** and **XGBoost Regressor** — to predict scores and validate the reliability of the manual formula. The ML model achieves strong predictive performance with R^2 values of **0.8277** (Random Forest) and **0.8196** (XGBoost). Additionally, it introduces **feature importance plots** and **KMeans clustering** to segment startups into categories such as *High Potential*, *High Burn Risk*, and *Undervalued but Growing*. These enhancements provide more strategic and actionable insights, making the ML-based version more powerful and scalable.

Table 1: Methodology Comparison

Aspect	Rule-Based Model (Task_1.ipynb)	ML-Based Model (Task1_ML.ipynb)
Approach	Manual weighted formula	Scoring + Regression + Clustering
Techniques Used	Min-Max scaling + custom weights	Min-Max scaling + ML + KMeans
Score Calculation	Predefined linear formula	Same formula + ML regression output

Interpretability	High	High with added feature importance
Adaptability	Low (static weights)	High (learns patterns from data)

Table 2: Capabilities and Outputs

Feature	Rule-Based Model	ML-Based Model
Prediction Capability	Not available	Predicts scores for new startups
Clustering	No	KMeans clustering with custom labels
Feature Importance	Not available	Provided via XGBoost
Validation Metrics	Not applicable	RMSE & R^2 used for model assessment
Visualizations	Histogram, bar chart, heatmap	Cluster plot, feature importance, basic charts
Output	Score, rank	Score, rank, cluster label, prediction metrics

Table 3: Use Case and Suitability

Criterion	Rule-Based Model	ML-Based Model
Use Case	Simple evaluations, prototyping	Scalable applications, production systems
Complexity	Low	Moderate to high
Speed of Implementation	Fast	Slower (training and tuning required)
Best For	Quick insights, academic demos	Real-world analysis, strategic decision-making

Scalability	Limited	High

Conclusion

The **ML-based scoring model** offers substantial advantages over the rule-based version. By incorporating supervised learning, clustering, and feature importance analysis, it allows not only score computation but also strategic segmentation, trend discovery, and prediction. The rule-based model remains useful for rapid prototyping or scenarios with limited resources, but for robust, explainable, and scalable solutions, the ML-based model is the more effective and future-ready choice.