ML-Based: Startup Health Scoring Model

Objective

To improve the rule-based startup evaluation system using **machine learning techniques** for better insights, validation, and clustering. This project enhances interpretability and allows prediction of startup health scores using regression models.

Dataset Overview

The dataset contains details about 100 startups with the following columns:

- team_experience
- market_size_million_usd
- monthly_active_users
- monthly_burn_rate_inr
- funds_raised_inr
- valuation_inr

Data Preprocessing

- **Min-Max normalization** applied to all features.
- **Burn rate** inversely scaled since high burn is considered risky.

Composite Score Calculation

A manual scoring formula was designed using the following weights:

Feature	Weight
team_experience	15%
market_size_million_usd	20%

monthly_active_users	25%
monthly_burn_rate_inr	10%
funds_raised_inr	15%
valuation_inr	15%

The final score was scaled to 100 and startups were ranked accordingly.

Machine Learning Enhancements

Regression Models

Two regression models were used to predict the score:

Model	RMSE	R^2 Score
Random Forest	6.0895	0.8277
XGBoost	6.2320	0.8196

These results confirm that the score is **predictable** and follows a meaningful trend across features.

Feature Importance

The XGBoost model revealed the top predictors of score:

- monthly_active_users had the highest impact
- funds_raised_inr and team_experience also contributed significantly
- burn_rate and valuation had the least impact

KMeans Clustering

KMeans (k=3) was used to group startups into behavioral segments:

Cluster Label	Description

High Potential	Strong users, decent funding, low burn
Undervalued but Growing	High users but low valuation
High Burn Risk	High expenses and low users

Visualization:

This helped reveal startups that might not score well but have hidden growth signals.

Deliverables

- Task1_ML.ipynb – ML-based notebook
- ranked_startups.csv – Scores + Ranks + Cluster Labels
- Graphs:
    - XGBoost Feature Importance
    - KMeans Cluster Plot
- This PDF Report

Insights

- ML regression supported the integrity of the scoring formula
- Feature importance suggests startups with more users are consistently strong
- Clustering allowed us to segment and highlight startups that standard scoring might miss

Conclusion

This project demonstrates a hybrid approach where **human logic meets machine learning**. The scoring engine was validated, clustered, and made more robust through ML techniques, laying a strong foundation for real-world startup health assessment.