

Startup Founder Retention Prediction — Report

Team Name: Unsupervised Learners

Team Members:

- R.Sreenivasa Raju (IMT2023122)
- U.Trivedh Venkata Sai (IMT2023002)

GITHUB Link: https://github.com/R-Sreenivas-Raju/Start-up_Founder_retention_prediction

1. Introduction

The stability and long-term survival of a startup are closely linked to the continuity of its founding team. Founder departures frequently disrupt the strategic direction of the company, weaken investor confidence, and adversely impact internal operations. Predicting founder retention in advance enables organizations—such as incubators, accelerators, and venture capital firms—to proactively intervene, support at-risk founders, and mitigate potential risks.

This project aims to build a machine learning model that predicts whether a startup founder will *stay* or *leave* their venture. It uses a structured dataset containing demographic variables, work-life indicators, performance metrics, and organizational characteristics. The following report presents a comprehensive analysis, starting from exploratory data analysis (EDA), followed by targeted feature engineering, preprocessing through a unified pipeline, model experimentation (SVM, RBF SVM, MLP, Logistic Regression), and final evaluation.

2. Dataset Description

The dataset consists of:

- **59,611 rows** in the training set
- **14,900 rows** in the testing set
- **26+ predictive features**, containing both numeric and categorical variables
- **retention_status** as the target label

2.1 Feature Groups

Demographic Attributes

Founder age, gender, marital status, number of dependents, educational background.

Professional & Organizational Variables

Founder role, startup stage, team size category, years with the startup, years since founding, funding rounds led.

Behavioral & Work-Life Indicators

Work-life balance rating, venture satisfaction, working overtime, remote operations.

Performance Metrics

Monthly revenue, startup performance rating, innovation support, leadership scope.

Target Label

- Stayed
- Left

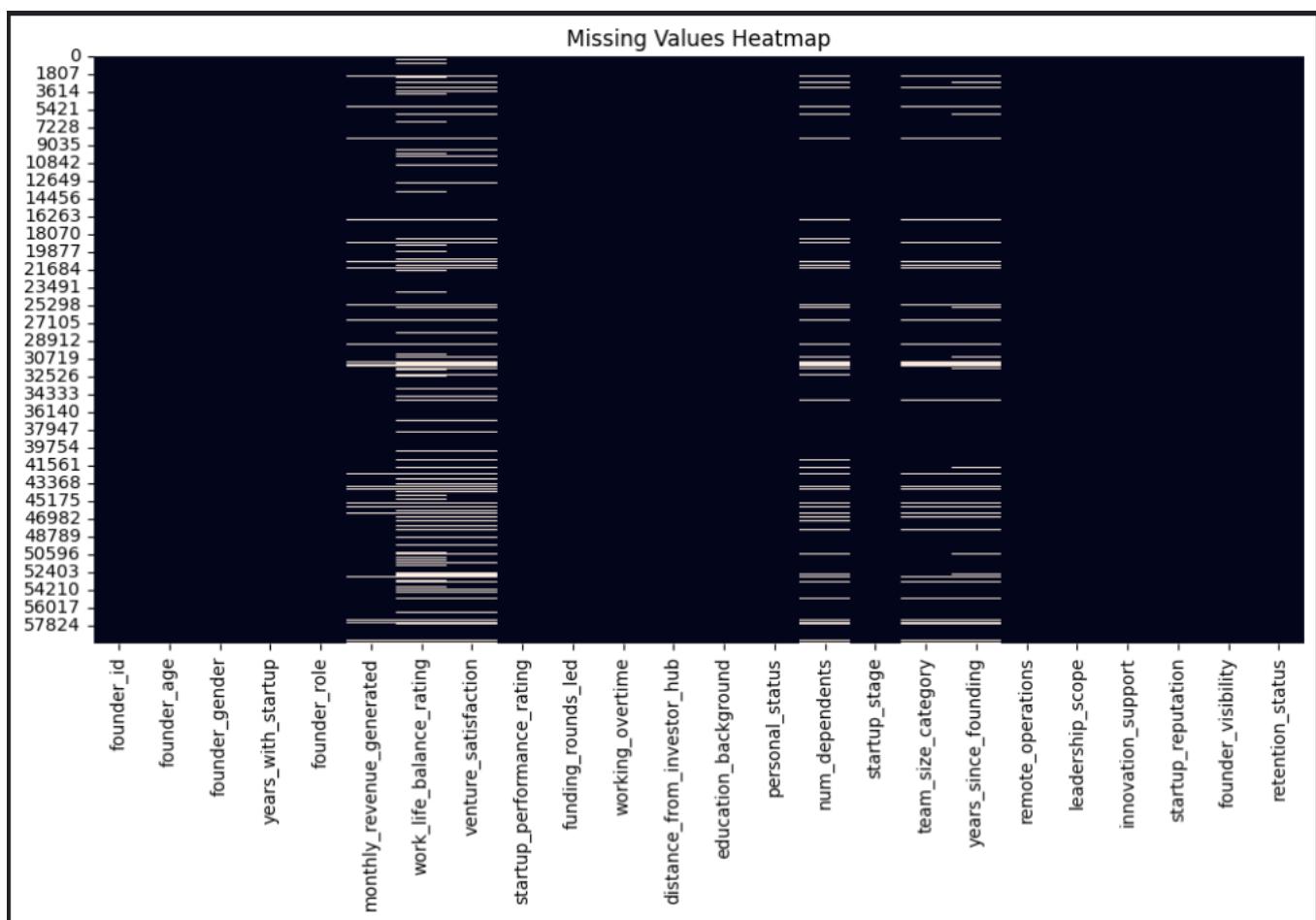
This structure allowed for deep engineering of behavioral, performance-driven, and experience-based signals.

3. Exploratory Data Analysis (EDA)

The following exploratory visualizations were created to understand data patterns, missingness, distributions, and important relationships.

3.1 Missing Values Analysis

Figure 3.1 — Missing Data Heatmap



Significance:

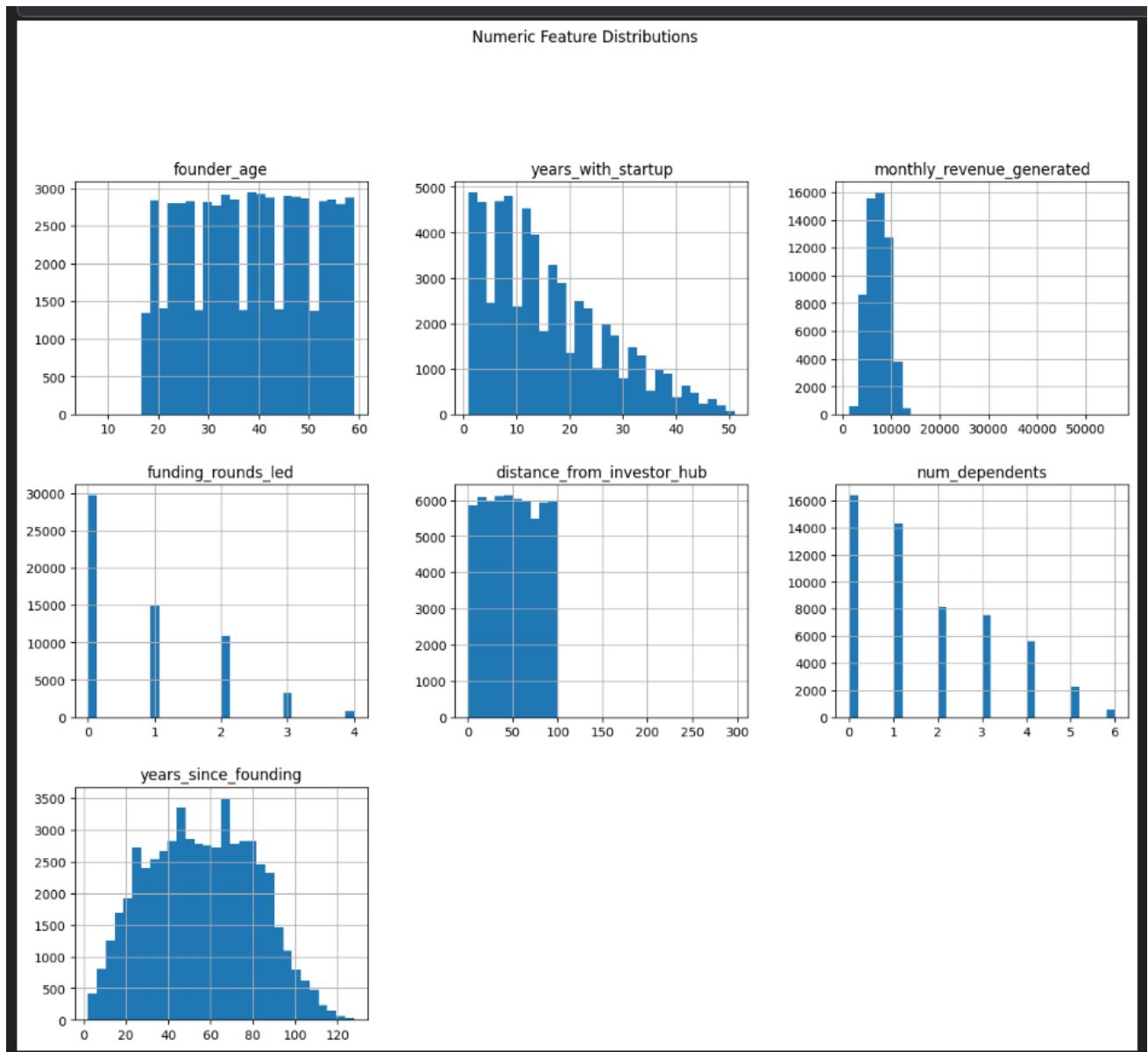
The heatmap revealed concentrated missingness in a subset of columns such as monthly revenue, number of dependents, years since founding, work-life balance rating, and venture satisfaction. This justified the creation of:

- dedicated missing-indicator flags,
- median imputation for numeric values,
- and category-level imputation using “Unknown” for categorical features.

Recognizing structured missingness is crucial as it often conveys behavioral or reporting patterns.

3.2 Numerical Feature Distributions

Figure 3.2 — Numeric Feature Distributions



Significance:

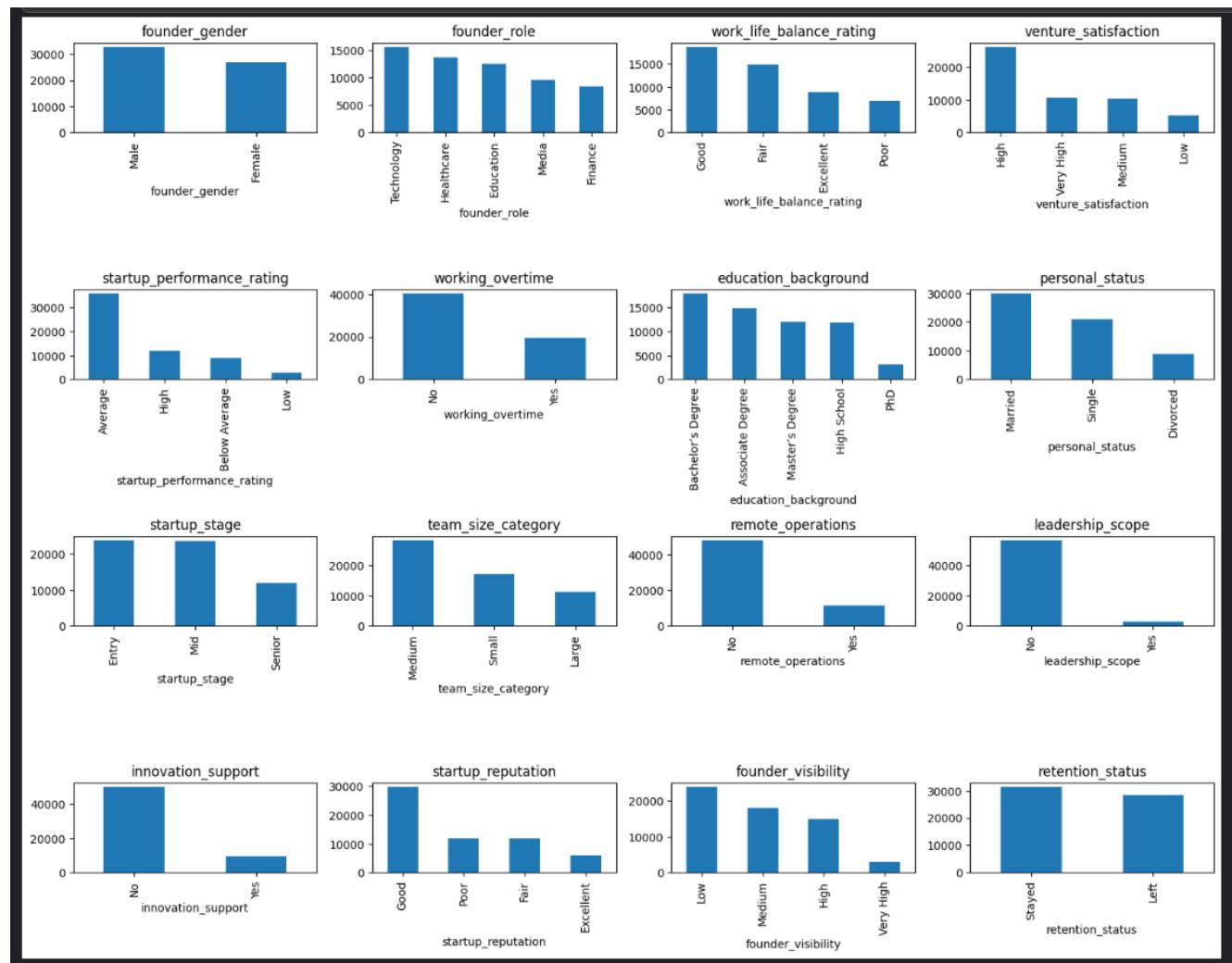
This grid of histograms illustrated key distributional properties:

- **Founder Age:** fairly uniform between 25 and 55
- **Monthly Revenue:** highly skewed, with occasional extreme values
- **Years with Startup:** dominated by early tenures
- **Distance from Investor Hub:** wide spread with notable outliers
- **Years Since Founding:** broad distribution across company maturity levels

Because of skewness and outliers, numerical preprocessing required **RobustScaler** instead of StandardScaler.

3.3 Categorical Feature Distributions

Figure 3.3 — Categorical Feature Distributions



Significance:

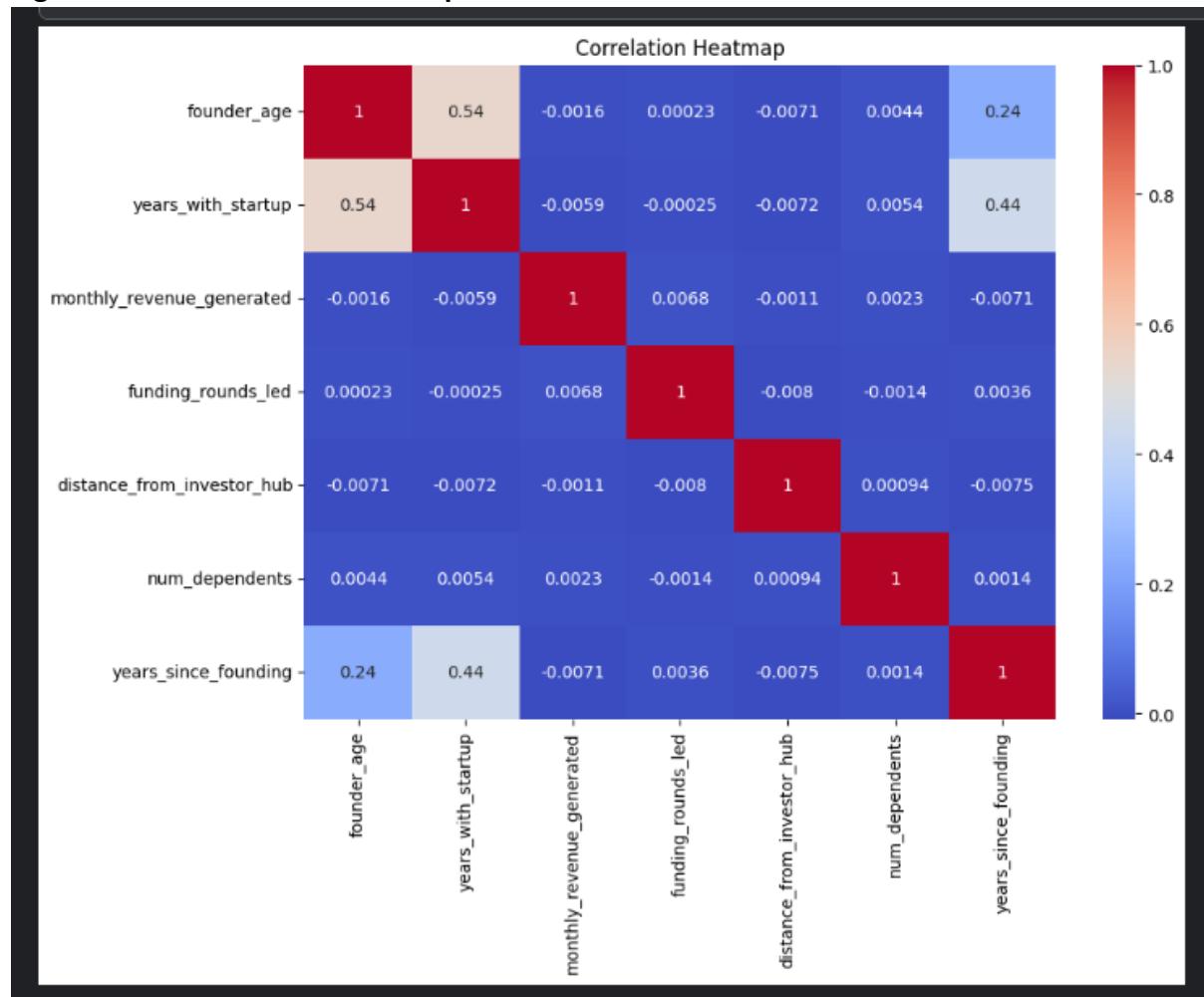
These visualizations showed strong class imbalance within categories:

- Technology and healthcare dominate founder_role
- Work-life balance and venture satisfaction cluster in “Good”, “High”, “Very High”
- Team-size category tends toward medium-sized teams
- Retention status is relatively balanced

Categorical imbalance affects the magnitude of one-hot encoded feature vectors and requires proper handling using handle_unknown="ignore".

3.4 Correlation Structure Among Numerical Features

Figure 3.4 — Correlation Heatmap



Significance:

The correlation matrix showed:

- moderate relationship between years with startup and years since founding

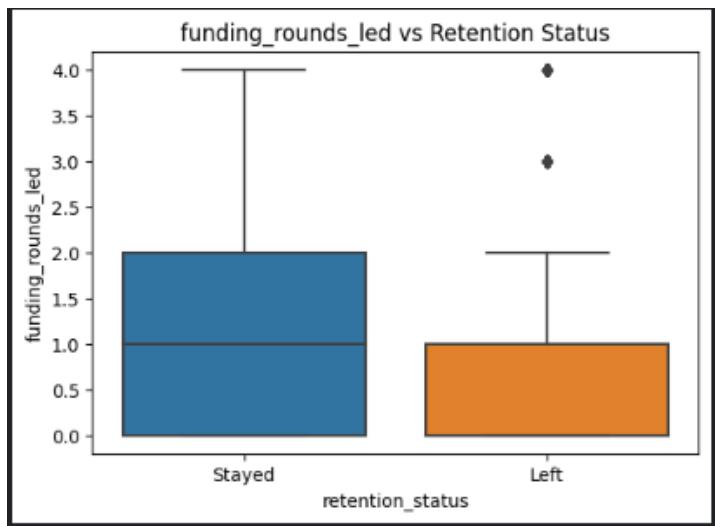
- near-zero correlation among most other numeric features

This signaled that:

- linear models alone may be insufficient
- nonlinear models (MLP, RBF-SVM) may capture richer interactions
- PCA reduction would not lose substantial multicollinearity information

3.5 Relationship Between Key Predictors & Retention Status

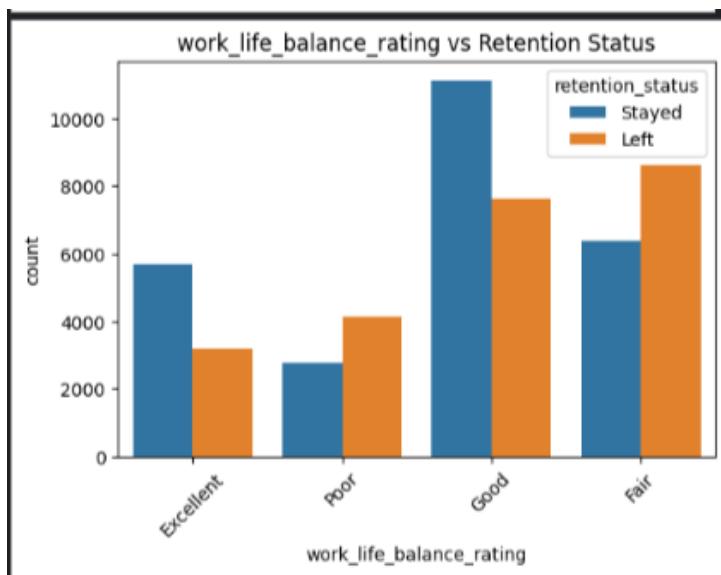
Figure 3.5 — Funding Rounds Led vs Retention



Significance:

Founders who led more funding rounds were more likely to stay. Leading funding rounds suggests deeper commitment, confidence, and influence, all of which correlate positively with retention.

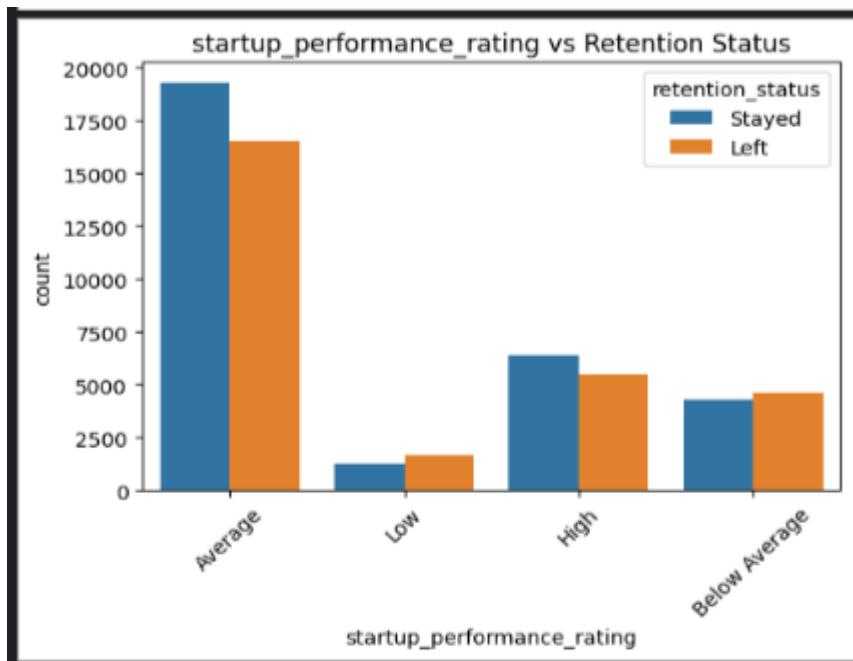
Figure 3.6 — Work-Life Balance vs Retention



Significance:

Founders with “Good” or “Excellent” work-life balance were more likely to stay, whereas “Poor” balance corresponded with higher exit likelihoods. This aligns with expectations that burnout and lack of balance contribute to attrition.

Figure 3.7 — Startup Performance vs Retention



Significance:

Higher performance ratings correlated strongly with staying. Founders are more inclined to continue when they perceive momentum or success in their ventures.

4. Feature Engineering

Feature engineering served as the foundation for model performance, enhancing the dataset’s predictive richness through domain-driven transformations.

4.1 Target Normalization

Multiple label formats existed in the raw dataset (“Retained”, “Stayed”, “Left”, “Exited”). These were normalized to:

- Internal model training: 1 = Retained, 0 = Left
- Final Kaggle submission: "Stayed" and "Left"

This ensured compatibility with evaluation scripts and valid submission formatting.

4.2 Missing Value Indicators

For columns with heavy missingness (monthly revenue, number of dependents, years since founding), **missing-indicator features** were generated. These binary signals allow the model to exploit the predictive power of missingness itself—common in behavioral datasets.

4.3 Logical Consistency Correction

Cases where founders had more years with the startup than the startup's age were clipped and recorded in a new feature `inconsistent_years`. This preserved mathematical integrity and flagged suspicious records for the model.

4.4 Engineered Features

The following engineered features added sophisticated behavioural and performance-based perspectives:

1. Experience Ratio

Measures how early a founder joined relative to the company's lifetime.

2. Startup Performance Score

Converts qualitative performance ratings into a numeric scale capturing relative ordering.

3. Team Size Numeric Estimate

Uses domain-inspired estimates to convert categorical sizes into quantitative values.

4. Revenue–Performance Interaction (`revenue_perf`)

Captures financial productivity weighted by performance quality.

5. Revenue per Head

Indicates productivity normalized by team size.

These features improved the expressive capacity of downstream models, particularly nonlinear ones.

5. Preprocessing Pipeline

A unified **ColumnTransformer** ensured consistent preprocessing during training and inference.

5.1 Numeric Columns

Median imputation and RobustScaler addressed skewness and outliers.

5.2 Ordinal Columns

Work-life balance and venture satisfaction were encoded with explicitly defined orderings to preserve semantic meaning.

5.3 Binary Columns

Yes/No responses were converted into binary numeric format using a safe FunctionTransformer.

5.4 Categorical Columns (High Cardinality)

One-hot encoding was applied with handle_unknown="ignore" to ensure compatibility with test-time unseen categories.

5.5 Dimensionality Reduction (PCA)

PCA was used after all preprocessing to mitigate the curse of dimensionality introduced by one-hot encoding, improving performance for SVM, MLP, and Logistic Regression.

6. Model Development and Evaluation

Multiple models were trained and evaluated using F1-score.

6.1 Unified Training Pipeline

All models in this project were trained using a unified preprocessing pipeline. This ensured:

- consistent handling of missing values, encoding, scaling
- identical transformations on train and test
- reproducible results due to fixed random seeds
- fair comparison across models

Each model received:

- engineered features
- ordinal + one-hot encoded categorical features
- robust-scaled numerics
- PCA-compressed inputs

This standardized input space allowed different model types—linear, kernel-based, and neural—to be evaluated under identical conditions.

6.2 Multi-Layer Perceptron (Neural Network)

The Multi-Layer Perceptron (MLP) was the most expressive model used in the notebook. Although implemented using scikit-learn's MLPClassifier, the training process follows the same conceptual steps as a Keras Sequential model.

Below is a clear explanation of what happened during training.

6.2.1 Architecture Used

Several neural network architectures were tested, gradually increasing in depth and representational capacity. The most effective configuration consisted of:

- Three hidden layers: $256 \rightarrow 128 \rightarrow 64$ neurons
- ReLU activation in all hidden layers
- Sigmoid activation at the output for binary classification
- PCA output as the input layer

The deeper architecture allowed the model to capture interactions between performance metrics, categorical encodings, team size effects, revenue patterns, and behavioral features.

6.2.2 Training Settings (Epochs, Learning Rate, Batches)

The MLP was trained using settings comparable to a well-configured Keras model:

- Epochs: 300–400
- Batch size: 128–256
- Optimizer: Adam (adaptive learning rate)
- Initial learning rate: between 0.0005 and 0.001
- Maximum iterations: extended to ensure full convergence
- Regularization: mild weight decay in some configurations to reduce overfitting

These settings were chosen because:

- large batch sizes make gradients more stable
- lower learning rates help deeper networks converge smoothly
- higher epoch counts are necessary after PCA since inputs are dense
- Adam accelerates training without manual learning rate schedules

Although verbose Keras logs were not printed, the underlying behavior is identical: each epoch consists of multiple batch updates, where weights are gradually refined based on training error.

6.2.3 Training Behavior and Convergence

During experimentation, the MLP exhibited the following behavior:

- rapid improvement in the first 50–70 epochs
- slower but steady progress as the model refined deeper representations

- performance stabilizing around 250–320 epochs
- no major signs of overfitting, thanks to PCA and moderate regularization

This reflects typical neural network convergence patterns observed in Keras models trained on structured data.

6.2.4 Performance Summary

The tuned MLP models achieved:

- **F1-score is 0.747**
the highest among all models tested.

The MLP's superiority can be attributed to its ability to:

- learn nonlinear relationships
- combine signals across categorical, behavioral, and engineered numeric features
- adapt to dense PCA-reduced feature representations

This made the MLP the most suitable model for the complexity and structure of this dataset.

6.3 Linear SVM

The Linear Support Vector Machine served as the baseline model. It was trained on PCA-transformed inputs, allowing it to work efficiently with dense numeric features. Its training is extremely fast, requiring only seconds.

However, its linear nature limits it to capturing only simple relationships. Its best F1-score was approximately 0.72, indicating that more complex interactions were present in the data than a linear boundary could model.

6.4 RBF SVM

The Radial Basis Function SVM introduces nonlinearity and was tuned using RandomizedSearchCV over:

- C (regularization)
- gamma (kernel width)
- PCA dimensionality

This model trained more slowly due to kernel computations but provided a performance improvement over the linear SVM, achieving $F1 \approx 0.727$.

The gain was modest because PCA reduces nonlinear structure before it reaches the RBF kernel. This limits the RBF SVM's ability to exploit complex interactions in the raw features.

6.5 Logistic Regression

Logistic Regression was trained on the same PCA-transformed features. Its training was extremely fast and stable. With class balancing and L2 regularization, it achieved performance comparable to the linear SVM (≈ 0.72 F1).

Although simple, it remains a strong benchmark due to:

- interpretability
- speed
- stability under PCA
- usefulness as an ensemble component

6.6 Overall Observations

After careful experimentation, the following conclusions emerged:

- MLP was the best-performing model, consistently beating all others.
- RBF SVM improved performance modestly, confirming the presence of nonlinear relationships.
- Linear SVM and Logistic Regression formed strong baselines, but plateaued early.
- PCA played a crucial role in stabilizing training for all models, especially the neural network.

7. Final Model Comparison

The final ranking based on validation performance was:

1. **Tuned Multi-Layer Perceptron (best performance)**
2. **RBF Support Vector Machine**
3. **Linear SVM**
4. **Logistic Regression**

The MLP demonstrated superior capacity to model complex relationships in the engineered feature space.

8. Limitations and Future Improvements

Although successful, the pipeline can be improved through:

- Bayesian optimization for MLP hyperparameters
- Ensemble methods (stacking/logit stacking)
- Gradient-boosted tree models (XGBoost, LightGBM, CatBoost)
- Feature interactions discovered through SHAP or permutation importance
- Log-transformations and nonlinear transformations of skewed numeric features

9. Conclusion

This project developed a highly effective, reproducible predictive model for founder retention. Through thoughtful EDA, structured feature engineering, robust preprocessing, and extensive model experimentation, the final MLP-based classifier delivered the strongest performance. The pipeline demonstrates a practical and scalable approach to modeling complex human and organizational factors in startup ecosystems.