# 1. Problem Statement

Customer retention is a key driver of sustainable growth, yet many organizations make churn-related decisions without clear insights into the factors driving customer attrition. This lack of visibility makes it difficult to identify which customer behaviors and operational signals should be prioritized for effective retention strategies. Given that acquiring new customers typically costs 5x more than retaining existing ones, understanding how to engage current customers is critical in any industry or organization.

This project aims to answer one question: What customer behavior and operational factors have the greatest impact on churn rates?

# 2. About the Data

The original dataset consisted of 10,000 observations with 32 features (full list included in python notebook), including a binary churn indicator ( 0 indicates retention/ 1 indicates the customer churned) and a mix of behavioral, financial, satisfaction, and demographic metrics. The data is taken from Kaggle and is synthetic. The experiment's target variable is churn.

## Preprocessing steps:

- Dropped observations with missing values
- Removed identifier columns not relevant to prediction
- Separated features and target variable
- Applied a stratified train/test split to preserve churn distribution
- Standardized numeric features
- Encoded categorical variables

| Category | Features |
|---|---|
| Customer Profile | Age, gender, location, tenure, contract type |
| Product Usage | Logins, session duration, feature usage, activity trends |
| Billing & Payment | Subscription fees, revenue, payment failures, discount |
| Customer Support | Tickets, resolution time, CSAT, complaints |

After preprocessing, the final dataset used for the project consisted of 7,955 observations with 32 features.

# 3. Methodology

## Logistic Regression

Logistic regression was used as a baseline model given the binary target variable and its coefficient-based interpretability. Coefficient analysis revealed the strongest drivers of churn:

| Feature | Description | Coefficient |
|---|---|---|
| csat_score | Customer satisfaction | -0.54 |
| tenure_months | Months with company | -0.42 |
| monthly_logins | Account activity | -0.38 |
| payment_failures | Failed payments | 0.37 |

The Logistic regression showed that churn is primarily driven by financial behavior and engagement. Payment failures increase churn risk, while higher satisfaction, longer tenure, and more monthly logins reduced it.

**Model Performance:**
The logistic regression model achieved an accuracy of 89.6% and a ROC-AUC of 0.716, indicating the model's reasonable predictive power to distinguish between churned and retained customers.

- Accuracy: 0.896
- Precision: 0.316
- Recall: 0.019
- F1-score: 0.035
- ROC-AUC: 0.716

## Random Forest

After running the regression, I decided to run a random forest model to capture any potential nonlinear relationships and feature interactions that logistic regression might miss. The model achieved an accuracy of 89.8% and a ROC-AUC of 0.712, which is similar to the logistic regression model. While the Random Forest did not improve predictive performance, it confirms that the relationships between the predictors and churn are largely linear in the dataset.

# 5. Conclusion

This analysis demonstrates that customer churn in this data is mainly influenced by financial behavior and engagement, with payment failures increasing churn risk and higher satisfaction, longer tenure, and more monthly logins reducing it. Logistic regression provided a clear, interpretable baseline, while Random Forest confirmed that nonlinear relationships did not substantially improve prediction, suggesting that the key drivers are largely linear.