# 🧪 Mastering EDA Across 10 ML Use Cases

21 Steps. 10 Real Problems. 1 Reusable Framework.

# 🧪 Mastering EDA Across 10 ML Use Cases

### 21 Steps. 10 Real Problems. 1 Reusable Framework.

## 🔍 Why EDA is the Secret Sauce of Machine Learning

In the race to train high-performing machine learning models, one often underestimated step can make or break your solution — \*\*Exploratory Data Analysis (EDA)\*\*.

EDA isn’t just about pretty charts. It's about:

- \*\*Understanding the story\*\* your data is trying to tell

- \*\*Identifying pitfalls\*\* like leakage, imbalance, multicollinearity

- \*\*Engineering features\*\* that make your model shine

But the challenge?

EDA is often ad-hoc, inconsistent, and non-reproducible.

## 📘 Introducing: The MASTER EDA Framework

I developed a \*\*21-step EDA methodology\*\* designed to work across domains and problem types.

This framework categorizes analysis into \*\*structural, statistical, diagnostic, and visual layers\*\*.

| Step | Category | Description |

|------|-----------------|------------------------------------------|

| 1–2 | Data Setup | Load + Overview |

| 3–5 | Visual | Univariate, Bivariate, Multivariate |

| 6–9 | Structural | Missing, Outliers, Skew, Target |

| 10–12| Statistical | Correlation, Class Imbalance, Cardinality|

| 13–17| Diagnostic | Quality, Time, Multicollinearity, etc. |

| 18–21| Advanced | Feature Hints, Clusters, AutoEDA, Stats |

We'll now walk through this framework using \*\*10 ML problems\*\*, from banking to NLP to clustering.

## ✅ Scenario 1: Loan Default Prediction

\*\*ML Task:\*\* Binary Classification

\*\*Target:\*\* `defaulted`

\*\*EDA Angle:\*\* Outliers, skewness, potential leakage

\*\*Dataset:\*\* ~10,000 rows, structured finance data

### 📊 What We Explored

| Step | Insight |

|------|---------|

| \*\*3–4\*\* | Detected right-skewed distribution in `loan\_amount` and `income` |

| \*\*6\*\* | Missing values were concentrated in employment history and credit utilization |

| \*\*7\*\* | Significant outliers in `loan\_to\_income\_ratio` |

| \*\*10\*\* | High correlation between `early\_payoff\_flag` and `defaulted` — flagged for leakage |

| \*\*17\*\* | Created `risk\_score` × `loan\_term` as an interaction feature |

| \*\*21\*\* | ANOVA showed `loan\_amount` and `tenure` had significant predictive power |

### 📈 Favorite Visual

Boxplot showing defaulted loans having heavier tails:

sns.boxplot(x='defaulted', y='loan\_amount', data=df)

### 💡 Takeaway Insight

EDA helped expose a potential \*\*label leakage\*\* through features indirectly tied to `defaulted`. Removing or redefining these features improved generalization.

## 📍 Scenario 2: Customer Churn

\*\*ML Task:\*\* Binary Classification

\*\*Target:\*\* `is\_churned`

\*\*EDA Focus:\*\* Time-based analysis, class imbalance, tenure segmentation

👉 \*(to be expanded in next version)\*

## 📍 Scenario 3: Ad Click Prediction

\*\*ML Task:\*\* Binary Classification

\*\*Target:\*\* `clicked`

\*\*EDA Focus:\*\* High-cardinality features, campaign profiling, CTR imbalance

👉 \*(to be expanded in next version)\*

## 📍 Scenario 4: Employee Attrition

\*\*ML Task:\*\* Binary Classification

\*\*Target:\*\* `attrition`

\*\*EDA Focus:\*\* HR data quality, categorical imbalance, satisfaction scores

👉 \*(to be expanded in next version)\*

## 📍 Scenario 5: House Price Prediction

\*\*ML Task:\*\* Regression

\*\*Target:\*\* `sale\_price`

\*\*EDA Focus:\*\* Skewed numeric, multicollinearity, outliers

👉 \*(to be expanded in next version)\*

## 📍 Scenario 6: Credit Score Segmentation

\*\*ML Task:\*\* Unsupervised (Clustering)

\*\*EDA Focus:\*\* PCA, VIF, outlier trimming, feature engineering

👉 \*(to be expanded in next version)\*

## 📍 Scenario 7: \*(Skipped)\*

\*\*Stock Trend Classification\*\* — intentionally skipped

## 📍 Scenario 8: Disease Diagnosis

\*\*ML Task:\*\* Multiclass Classification

\*\*Target:\*\* `diagnosis\_code`

\*\*EDA Focus:\*\* Text + Categorical, leakage checks, symptom feature engineering

👉 \*(to be expanded in next version)\*

## 📍 Scenario 9: Sentiment Classification (NLP)

\*\*ML Task:\*\* Multiclass (or binary)

\*\*Target:\*\* `sentiment`

\*\*EDA Focus:\*\* Review text, score stats, punctuation signals

👉 \*(to be expanded in next version)\*

## 📍 Scenario 10: Product Recommendation Clustering

\*\*ML Task:\*\* Unsupervised

\*\*EDA Focus:\*\* Behavioral metrics, PCA + KMeans, engineered scoring (value, activity)

👉 \*(to be expanded in next version)\*

## 🧠 Wrapping Up: EDA Isn't Optional

Across all these examples, one pattern held true:

> \*\*The quality of your EDA determines the ceiling of your model.\*\*

From identifying data leakage to generating features to fixing target skew — EDA is how we turn noise into signal.

## 🔗 What’s Next?

- Want the notebooks? ⭐ [Coming Soon on GitHub]

- Want the infographic? 🖼️ [Coming up next!]

- Want to reuse the 21-step template? Copy it, clone it, make it yours.