

How a Data Scientist Chooses Between Checking Assumptions vs. Directly Applying Complex Models

The decision depends on **project goals**, **data nature**, **time constraints**, and **stakeholder expectations**.

1. When to Check Assumptions First

You check assumptions if:

- **Interpretability matters** → You want to **explain why** predictions are made.
- **Regulatory requirements** → In healthcare, finance, or legal projects, models **must be explainable**.
- **Small datasets** → With limited data, assumptions ensure you don't **overfit**.
- **Statistical inference is important** → You care about **coefficients**, **p-values**, **confidence intervals**.

Example:

- Predicting how **study hours** affect **student performance** → you want to **explain** relationships, not just predict scores.
 - Linear Regression or Logistic Regression is used → **assumptions must be checked**.
-

2. When to Skip Assumptions and Use Complex Models

You can directly apply non-linear models like XGBoost, Random Forest, LightGBM, Neural Networks if:

- **Prediction accuracy is the only goal**.
- The dataset is **large and complex**.
- You suspect **non-linear relationships** or **high-dimensional interactions**.
- Stakeholders don't need **explanations**, only **results**.

Example:

- Predicting **customer churn** for a telecom company with millions of records.

- Using **XGBoost** may outperform Linear Regression, and you don't need to justify individual coefficients.
-

3. Hybrid / Practical Industry Approach

Most data scientists **combine both approaches**:

Step 1 — Start Simple

- Try Linear or Logistic Regression.
- Check **basic assumptions** like **multicollinearity, outliers, residuals**.



Step 2 — Compare Models

- If Linear Regression performs well → **keep it** for interpretability.
- If accuracy is poor → **move to complex models** like XGBoost.

Step 3 — Validate with Cross-Validation

- Always validate predictions, whether using simple or complex models.
-

Decision Framework

Factor	Check Assumptions 	Directly Use Complex Models 
Goal	Explain relationships	Maximize prediction accuracy
Data Size	Small / medium	Large, high-dimensional
Complexity	Simple, linear relationships	Non-linear interactions
Time	Plenty of time for EDA	Limited time, fast delivery

Factor	Check Assumptions 🔍	Directly Use Complex Models ⚡
Stakeholder Needs	Interpretability required	Only final predictions matter
Industry	Healthcare, finance, research	E-commerce, recommendation, ads

Key Takeaway ✅

- If you need interpretability → check assumptions and start simple.
- If you need accuracy and data is complex → use advanced models directly.
- Best practice → start simple, validate, then move to complex.