Anshum Banga
www.linkedin.com/in/anshumbanga

ML Algorithms & Outliers: A Complete Guide

Why Outliers Matter

Outliers can:

- Skew statistical measures (mean, variance)
- Distort model performance (e.g., regression coefficients, cluster centroids)
- Reduce generalization by fitting noise instead of true patterns

Algorithms Highly Sensitive to Outliers

Worst choices for noisy or uncleaned data

Algorithm	**Vulnerability**	**Mitigation Strategy**
Linear Regression	Squared error amplifies outliers	Use Huber loss or RANSAC
Logistic Regression	Extreme features shift probabilities	Apply L1 regularization (Lasso)
KNN	Distance metrics misclassify outliers	Use Mahalanobis distance
K-Means	Centroids drift toward outliers	Try K-Medoids or DBSCAN
PCA	Outliers dominate variance directions	Use Robust PCA

Robust Algorithms

Better choices when outliers exist

Algorithm	**Why It Works**	**Pro Tip**
Decision Trees	Splits ignore magnitude, only rank	Limit max_depth to prevent overfit
Random Forest	Bootstrap aggregation dilutes outliers	Use for anomaly detection
XGBoost/LightGBM	Focuses on residuals	Tune min_child_weight
SVM (RBF Kernel)	Maximizes margin; isolates outliers	Increase gamma

Anshum Banga
www.linkedin.com/in/anshumbanga

Advanced Techniques

1. Preprocessing

- **Winsorization**: Cap values at 5th/95th percentiles
- **Quantile Transformation**: Non-linear mapping
- **Robust Scaling**: (X - median) / IQR

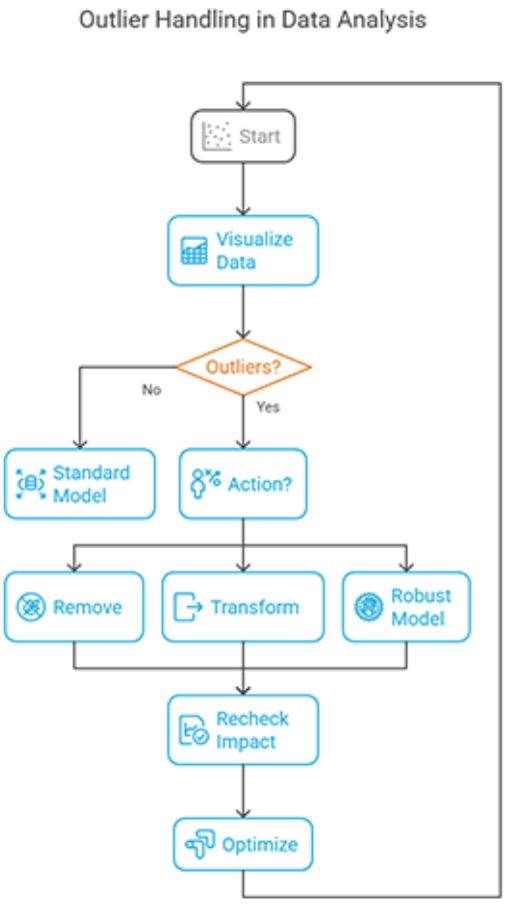
2. Anomaly Detection

- **Isolation Forest**: Effective for high-dimensional data
- **DBSCAN**: Density-based clustering with noise exclusion

3. Modeling Adjustments

- **Huber Regression**: Hybrid MSE/MAE loss function
- **Quantile Regression**: Predicts conditional median or quantile

Outlier Handling Flowchart



Anshum Banga
www.linkedin.com/in/anshumbanga

Practical Checklist

- Visualize first: Boxplots, scatterplots
- Scale robustly: Avoid StandardScaler
- Test models: Compare Random Forest vs. Linear Regression
- Iterate: Remove → Train → Validate

Real-World Example

Problem: Predicting house prices with luxury home outliers
Solution:

- Log-transform prices
- Detect outliers with Isolation Forest
- Train XGBoost with quantile loss

Further Reading

- Scikit-learn's Robust Scaling Guide
- XGBoost documentation on regularization and loss functions
- Research papers on Robust PCA and Isolation Forest