| **Aspect** | **Impact** | **Example** |
| --- | --- | --- |
| **Biasing the model** | Duplicates can give **more weight** to some observations than others, leading to a biased model. | If a "loyal customer" record is duplicated 5 times, the model thinks such customers are more important than they actually are. |
| **Overfitting** | The model might **overfit** to the duplicated records instead of learning general patterns. | The model memorizes the duplicate examples and cannot generalize to new data. |
| **Inflated performance metrics** | Validation metrics (accuracy, F1 score, etc.) might be **artificially high** because duplicated records make predictions easier. | Test dataset may have near-identical samples, causing misleading high accuracy. |
| **Wrong feature importance** | Features related to duplicated records might **appear more important** than they truly are. | Age or region could falsely seem important if duplicated in the dataset. |
| **Resource Wastage** | Duplicate records **increase memory and computation** cost unnecessarily. | Training time and model size become much bigger without real gain. |
| **Skewed probability outputs** | If using probabilistic models (e.g., Logistic Regression), duplicates **push the probability distributions** unfairly. |  |
| **Bad customer segmentation** | In clustering (K-Means etc.), duplicates **distort cluster centroids**. | Wrong customer segmentation results. |

**🧠 How Duplicate Records Affect Model Outcomes**

