**Main Challenges**

* **Severe class imbalance**
  + Bankruptcies rare → accuracy alone misleading
  + Used class weights in LR, RF, XGBoost to penalize missed bankruptcies
  + Applied SMOTE on the training set to oversample the minority class
* **Highly skewed numeric features**
  + Some ratios had extreme outliers/long tails
  + Tree models fine, but LR/SVM needed StandardScaler + selective log1p transforms
* **Multicollinearity**
  + Many ratios overlapped in meaning
  + Removed one variable from pairs with correlation |ρ| > 0.95
  + Ran VIF to drop unstable variables
* **Sampling bias between train/test**
  + Used stratified split
  + Checked PSI to confirm similar distributions
* **Interpretability vs. accuracy**
  + Gradient boosting → best scores
  + Logistic Regression → most transparent for stakeholders

**How Lab 4 Shaped Lab 5**

* **Model shortlist**: LR (baseline), RF/XGBoost (non-linear), dropped CatBoost (runtime)
* **Preprocessing**: Scaling only for LR/SVM, median imputation, consistent categorical encoding
* **Imbalance handling**: Combined SMOTE + class weights as tested in Lab 4
* **Feature engineering restraint**: Avoided excessive synthetic ratios for clarity
* **Multicollinearity checks**: Correlation filter + VIF before selection
* **Feature selection**: Correlation pruning + XGBoost importance; kept features agreed by multiple methods
* **Hyperparameter tuning**: RandomizedSearchCV first, then narrowed search for top models
* **Cross-validation**: StratifiedKFold for balanced folds
* **Metrics**: ROC-AUC (primary), PR-AUC/F1 (secondary), calibrated probabilities saved
* **Drift monitoring**: PSI run between train/test, flagged >0.25
* **Interpretability tools**: SHAP for trees, coefficients for LR

**Recommended Model for Deployment**

* **Primary choice**: Tuned XGBoost
  + ROC-AUC ~0.96, PR-AUC high, minimal train–test gap (<0.02)
  + Isotonic calibration → lowest Brier score (probabilities reliable)
  + SHAP shows liquidity/leverage ratios as top stable features
  + Fast training/inference, fits pipeline, supports drift alerts
* **Fallback**: Tuned Logistic Regression
  + ROC-AUC ~0.92
  + Fully transparent → good for regulatory/high-drift cases
* **Deployment plan**:
  + XGBoost in production with SHAP dashboards + PSI drift checks
  + Logistic Regression in shadow mode for explainability audit trail