Crashspot – Week 6 Starter Report

Focus: Model Enhancement + Class Imbalance Handling + Gradient Boosting

Objectives

- Address class imbalance in dataset (Logistic Regression & Random Forest with class_weight='balanced').
- Apply SMOTE oversampling (when available).
- Add Gradient Boosting as third model.
- Compare models using validation F1-score.
- Evaluate best model on test set.
- Save figures, dataset snapshot, and trained model.

Workflow

1. Load engineered dataset (week5_features.csv).
2. Split into train/validation/test (70/15/15).
3. Train models with class balancing (LogReg_bal, RF_bal, GB).
4. Apply SMOTE oversampling and retrain (if available).
5. Select best model based on validation F1-score.
6. Evaluate on test set → classification report, confusion matrix.
7. Generate ROC and Precision–Recall curves.
8. Save model and dataset snapshot for future weeks.

Outputs

- Notebook: Crashspot_Week6_Starter.ipynb
- Figures: docs/figures/week6_roc.png (ROC curve), docs/figures/week6_pr.png (Precision–Recall curve).
- Dataset: data_clean/week6_features.csv (engineered dataset snapshot).
- Model: models/week6_best_model.pkl (serialized best model).

Insights

- Random Forest with class weighting performed best in validation.
- SMOTE oversampling improved class balance but results were dataset-dependent.
- Test set showed near-perfect accuracy (small dataset caveat, possible overfitting).
- Outputs provide stronger foundation for Week 7 hyperparameter tuning and robust evaluation.