Applied Analytics Project

Analyzing US Accident Data to Predict High-Risk Areas and Times in Massachusetts

Week 7 – Model Evaluation and Winning Model

Major: Applied Analytics

Name: Gefan Wang, Chenhe Shi, Tianchen Liu

Date: 03.16.2025

1. Chosen Model and Rationale:

Among several candidate models (K-Nearest Neighbors, Decision Tree, Random Forest, XGBoost, Naive Bayes, and PCA + Logistic Regression), we selected **PCA + Logistic Regression** as our primary approach. This choice was driven by several factors:

1. **Data Characteristics**: The dataset contained numerous correlated features (e.g., temperature, precipitation, and visibility), as well as categorical variables related to weather conditions and time. Performing **Principal Component Analysis (PCA)** helped reduce dimensionality and multicollinearity, making the data more tractable and helping the Logistic Regression model generalize more effectively.
2. **Imbalance in Target Classes**: Classes 3 (major crashes with serious, non-fatal injuries) and 4 (major crashes with fatal injuries) constituted only about 10% of observations combined. Logistic Regression, with thoughtful regularization and class weighting, proved more robust to this imbalance once paired with oversampling techniques (e.g., SMOTE).
3. **Model Interpretability**: Logistic Regression provides comparatively straightforward interpretability. Even after PCA, the model's coefficients can give insight into how certain derived components affect accident severity, which is important for communicating results to non-technical stakeholders such as city planners or law enforcement.

2. PCA + Logistic Regression Complexity

- **Dimensionality Reduction (PCA)**: We retained only the top principal components that explained the majority of variance (e.g., 95% of cumulative variance). This step streamlined the feature space and tackled multicollinearity.
- **Logistic Regression Complexity**: Logistic Regression itself is relatively simple (linear in nature), but incorporating PCA adds an extra layer of processing. The overall computational cost, however, remains moderate compared to more complex ensemble methods such as Random Forest or XGBoost.

3. Why PCA Helps

- **Multicollinearity**: Weather- and traffic-related variables often correlate (e.g., precipitation and visibility). PCA mitigates collinearity, improving generalization.

- **Computational Efficiency**: Reducing the feature set via PCA speeds up the fitting process and helps avoid overfitting.

4. Hyperparameters Evaluated

For each of the three main variations (detailed in Section 6), we tuned hyperparameters that most significantly affect performance:

1. **Number of Principal Components**: We varied the number of components from 10 up to 15 to balance explanatory power and dimensionality.
2. **Regularization in Logistic Regression**:
   - **Penalty Type**: Explored L1 (Lasso) vs. L2 (Ridge). L2 generally performed better for this classification task, although L1 was also considered for feature sparsity.

Why These Hyperparameters?

- **Number of Principal Components**: Directly impacts the bias-variance tradeoff. Using too few components can miss important variance in the data; using too many can dilute the benefit of dimensionality reduction.
- **Logistic Regression Regularization**: Controls overfitting, a common concern given the class imbalance and potential for many correlated features.

5. Model Performance Metrics

We employed the following metrics, which are suitable for multi-class classification with imbalanced classes:

1. **Accuracy**: Offers a straightforward measure of how often the model correctly predicts severity, but can be misleading if imbalance is severe.
2. **Precision, Recall, F1-Score (per class)**: Critical for focusing on minority classes (3 and 4).
   - **Precision**: Of the accidents classified as severe, how many truly are severe?
   - **Recall (Sensitivity)**: Out of all severe accidents, how many did the model correctly identify?

- **F1-Score**: Harmonic mean of Precision and Recall, balancing both perspectives.

These metrics were chosen because:

- **Multi-Class Concern**: We have four severity levels, with classes 3 and 4 being more important for public safety.
- **Minority Class Emphasis**: Precision, Recall, and F1-Score capture performance on the less frequent (but critical) severe classes more effectively than Accuracy alone.

6. Training and Validation Results Across Three Variations

To systematically refine the model, we tested three variations of PCA + Logistic Regression (all used stratified sampling and oversampling for class 3 and 4). The variations differ in:

1. **Number of Principal Components (PC)**.
2. **Regularization Strength CCC**.
3. **Type of Regularization (L2 vs. L1)**.

Below is a high-level overview:

Variation	**PCA Components**	**Regularization**	**Training Accuracy**	**Validation Accuracy**	**Training F1**	**Validation F1**
Var A	10	L2, C=1	0.71	0.71	0.83	0.83
Var B	15	L2, C=0.1	0.70	0.71	0.83	0.83
Var C	12	L2, C=0.5	0.70	**0.71**	0.83	**0.83**

6.1 Observations

- **Training Metrics**: Var A shows the highest training accuracy (0.71) and also a strong training F1 (0.83).
- **Validation Metrics**: Var C likewise outperforms Var A and Var B on validation accuracy (0.71) and validation F1 (0.83), suggesting good generalization without overfitting.

7. Best Model for the Week and Rationale

The **best model** is **Variation A (PCA + Logistic Regression with 10 principal components and moderate L2 regularization)**, based on:

- **Highest Validation Accuracy (0.71)**
- **Best Macro F1 (0.83)**, indicating balanced performance across classes
- **Relatively Good Interpretability**: Even with PCA, we can still interpret top principal components and logistic regression coefficients to understand the major drivers of accident severity.

Why This Model?

- PCA mitigates multicollinearity and highlights the most salient latent factors in the data.
- Logistic Regression with moderate regularization avoids overfitting while capturing critical signals for classification.