Applied Analytics Project

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

Week 6 — Develop First modeling approach

Major: Applied Analytics

Name: Gefan Wang, Chenhe Shi, Tianchen Liu

Date: 03.02.2025

Model Approach

For this week's modeling approach, we implemented multiple variations of classification models to predict accident severity in Massachusetts. These models are well-suited for accident severity prediction due to their diversity in ability to process structured data with multiple variables, capture complex patterns in driver behavior and environmental conditions, and generalize well across different accident scenarios. The models used were:

1. **K-Nearest Neighbors (KNN)**: KNN serves as a simple yet effective baseline model that relies on feature similarity to classify accident severity. However, it becomes computationally expensive as the dataset grows due to the need to compute distances for all data points.
2. **Decision Tree:** A highly interpretable model that splits data based on feature importance, making it useful for understanding accident severity factors. However, decision trees tend to overfit without proper pruning.
3. **Random Forest:** An ensemble of decision trees that reduces overfitting by averaging multiple models, increasing robustness. The trade-off is increased training time and model complexity.
4. **XGBoost:** A powerful gradient boosting model that outperforms other models in structured data tasks by minimizing errors iteratively. It requires careful hyperparameter tuning to avoid overfitting while maximizing performance.
5. **Naive Bayes:** A probabilistic model that is computationally efficient and works well with categorical features, but it assumes feature independence, which is often unrealistic in accident data.
6. **PCA + Logistic Regression:** This approach reduces dimensionality using Principal Component Analysis before applying Logistic Regression, balancing interpretability and computational efficiency while potentially losing some information in the dimensionality reduction process.

These models were selected to compare performance across different algorithm types, including distance-based methods, tree-based approaches, probabilistic models, and dimensionality-reduced regression.

Hyperparameter Evaluation

We evaluated key hyperparameters for each model:

- **KNN:** Number of neighbors (3, 5, 7) and distance metric.
- **Decision Tree:** Max depth (None, 5, 10, 20), and minimum samples split.
- **Random Forest:** Number of estimators (50, 100, 200), max depth, and minimum samples split.
- **XGBoost:** Learning rate, max depth (3, 5, 7), and number of estimators.
- **Naive Bayes:** Variance smoothing parameter.
- **PCA + Logistic Regression:** Number of principal components and regularization strength.

These hyperparameters were chosen to balance model complexity and performance, preventing overfitting while maintaining predictive accuracy

Model Performance Metrics

We assessed model performance using:

- **Accuracy:** Measures overall correctness.
- **Precision & Recall:** Important for imbalanced classes.
- **F1 Score:** Balances precision and recall.
- **AUC-ROC:** Evaluates classification effectiveness across different thresholds.

These metrics were selected to capture different aspects of model performance, ensuring robustness in predicting accident severity levels.

Metrics Calculation (Training & Validation Sets)

We computed the metrics for each model variation using both training and validation datasets. Below is the summary of performance results:

```
Validation Accuracy Scores:
KNN: 0.7112
Decision Tree: 0.7120
Random Forest: 0.7641
XG Boost: 0.7886
Naive Bayes: 0.6955
PCA+LogReg: 0.7048

Winning Model on Validation Set: XG Boost
```

From the results, we discovered that the best model is **XG Boost**, with highest validation accuracy (0.7886), strong F1-score, and recall balance which captures more accurate severity classifications. **Random Forest** is a close second, performing well with slight overfitting. **Decision Tree and KNN** show good results but overfit more than ensemble methods. **Naive Bayes and PCA + Logistic Regression** provide simple baselines but struggle with non-linear relationships.