Applied Analytics Project

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

Week 12– Save and package your model for deployment

Major: Applied Analytics

Name: Gefan Wang, Chenhe Shi, Tianchen Liu

Date: 04.27.2025

Environment Dependencies

- Operating System: Windows 11
- Python Version: Python 3.11.2
- Python Packages and Versions:
    - numpy==1.26.4
    - pandas==2.2.2
    - scikit-learn==1.4.2
    - XGBoost==2.0.3
    - matplotlib==3.9.0
    - seaborn==0.13.2

Model Deployment Mode

- **Batch Inference:** Selected due to the nature of the data and business context where predictions are generated periodically rather than on-demand. Batch inference is preferable as it reduces computational overhead and aligns well with scheduled business processes, providing scalability and efficiency in handling large volumes of data.

Model Performance

- **Metrics Evaluated:**
    - Accuracy, precision, and recall were primary metrics used for evaluating model performance.
    - Performance was tracked to ensure balanced predictions across diverse demographic groups.

Model Packaging and Deployment Enhancements

In addition to model serialization, we have packaged the complete feature engineering and model pipeline using scikit-learn's *Pipeline* object. This ensures future data inputs are processed consistently during batch inference. Artifacts including model weights, feature

schemas, and input validation logic are saved to enable seamless deployment and rollback if needed.

Monitoring Plan

- **Performance Metrics:**
    - **Accuracy:** Ensures overall prediction reliability.
    - **Precision & Recall:** Balances false positives and false negatives, essential for fairness.
    - **Response Time:** Monitored for user experience.
    - **Prediction Volume:** Assessed for infrastructure scalability.
- **Monitoring Thresholds:**
    - Defined clear thresholds for performance metrics:
        - **Green:** Accuracy >90%, Precision & Recall >85%
        - **Yellow:** Accuracy 80-90%, Precision & Recall 70-85%
        - **Red:** Accuracy <80%, Precision & Recall <70%
- **Monitoring and Drift Detection**:
    - Beyond tracking model performance metrics such as accuracy and precision, we propose monitoring input feature distributions using statistical tests (e.g., PSI, KS test) to detect early signs of data drift. Visualization dashboards will be explored to track model health over time, aiding decision-making for retraining and maintenance.

Risk Mitigation Strategies

- **Green Flag:** Routine checks and system optimization.
- **Yellow Flag:** Enhanced monitoring, root-cause analysis, and preparation for interventions such as retraining.
- **Red Flag:** Immediate model withdrawal, rollback procedures, and urgent corrective measures.

Retraining Strategy

- Monthly scheduled retraining or retraining triggered by metric thresholds. Ensures model adaptation to evolving data.

Data Drift Considerations

- Data drift, seasonal variations, and shifting user behaviors are significant concerns. Continuous monitoring and regular retraining are crucial to mitigate these impacts and maintain model accuracy and fairness.

Future Scalability and Integration

Although batch inference is currently implemented, a basic API-readiness plan has been drafted to support future integration with navigation applications such as Google Maps or Apple Maps. Routine maintenance for our real-time QR code website will ensure continued accessibility for stakeholders.

Real-Time System Integration (Long-Term Vision)

Although we are deploying via batch inference now, a roadmap is in place for future API integration using FastAPI or Flask, allowing for real-time queries and map-based alerts.

As proposed in our symposium poster, the ultimate goal is to integrate our model outputs into navigation systems like Google Maps or Apple Maps, allowing drivers to receive dynamic, real-time accident severity warnings based on live conditions.

QR Code Website Maintenance

- The interactive QR-code based website for real-time severity predictions will be maintained.
- Monthly testing will ensure the QR code link is active, website uptime is stable, and predictions remain accurate.