Modeling-III

Team Members:
Chandra Sekhar Katipalli
Sindura Reddy Challa
Sanjana Reddy Soma

Context Recap

- **Phase I**: Traditional ML models (Linear Regression, Decision Trees)

  ✅ Gave baseline performance

  ❌ Limited in capturing complex temporal/spatial patterns

- **Phase II**: Ensemble Models (Random Forest, XGBoost, Gradient Boosting)

  ✅ Improvement in performance on many routes

Phase II Ensemble Model (Before Tuning)

	Models	Train MSE	Test MSE	Train R^2	Test R^2
0	Random Forest	113.95	311.25	0.87	0.65
1	Gradient Boosting	239.62	358.01	0.72	0.59
2	Voting Regressor	174.60	325.05	0.80	0.63
3	XGBoost	205.30	354.72	0.76	0.60

Observation: Models struggled on dataset with large variance between routes.

Current Progress: Ensemble Models After Tuning & Route-Wise Modeling

Route-wise training & evaluation

- Custom model fit for each ROUTE_ID

Significant **Test R^2 improvement** on high-traffic & clean routes

Advanced modeling with **route-level tuning** and **hyperparameter optimization** improved generalization.

Bias- Variance Tradeoff

ROUTE_ID	Model	Train MSE	Test MSE	Train R2	Test R2	Train Size	Test Size
IN0000100000	Gradient Boosting	95.31055419	155.2640871	0.914073598	0.861504585	980	285
IN0000100000	Voting Regressor	59.96074957	191.2753859	0.945942907	0.829382541	980	285
IN0000200000	Gradient Boosting	65.16132541	174.7990556	0.894610381	0.744837085	523	153
IN0000200000	Voting Regressor	44.69868763	177.3464134	0.927705926	0.74111858	523	153
IN0000220000	Voting Regressor	77.00933733	282.2153559	0.875910208	0.489643953	401	121
IN0000220000	Gradient Boosting	116.1610751	324.2314621	0.812822651	0.413662355	401	121
IN0000590000	Voting Regressor	83.80391786	481.7638142	0.92636408	0.674882144	789	214
IN0000590000	Gradient Boosting	133.2645836	513.4753012	0.882904518	0.653481677	789	214

- Gradient Boosting and Voting Regressor consistently performed best across most routes.
- Routes like IN0000100000 showed strong predictive power with Test $R^2 > 0.85$.
- Some routes exhibited lower Test R^2 due to limited data or high variance.
- Indicates that route-wise modeling is effective, but results vary depending on data quality and volume.

Hyperparameter Tuning Techniques Applied

- **Manual Tuning**

  Adjusted: n_estimators, learning_rate, max_depth, subsample, min_samples_split, min_samples_leaf, max_features

  ✅ Gave the most consistent improvements

- **RandomizedSearchCV**

  Searched random combinations over parameter grid

- **GridSearchCV**

  Exhaustive search over selected parameter values

We manually combined the best-performing parameters from RandomizedSearchCV, GridSearchCV, and domain knowledge leading to the most consistent improvements in model performance.

Advanced Models

Need for More Powerful Models

Issues Observed:

Some routes still show high error / low R^2

Ensemble models struggle with sequential dependencies or route-wise irregularities

Underfitting on small or noisy sections

Use neural network:

Capture sequential trends (yearly degradation/improvement)

Better model nonlinear and interaction effects

Next Steps

We will convert our data to sequential data
Advanced Models:

- Recurrent Neural Network (RNN)
- Long Short-Term Memory (LSTM)

we plan to explore **deep neural networks**, including **RNNs** and **LSTMs**, to capture temporal and nonlinear patterns in our transportation data.

We believe these models can complement our ensemble approaches.

Research Questions

1. How does the fusion of HPMS and FAF datasets (2013–2022) enhance the predictive performance of highway deterioration models in estimating IRI, compared to traditional statistical and machine learning approaches using single-source data?

2. What are the most influential predictive features—such as traffic volume, freight load, and pavement condition indices—derived from the integrated datasets, and how do their contributions vary across different machine learning models?

3. How effectively can the proposed predictive model, leveraging data fusion and advanced machine learning techniques, minimize forecasting errors and improve the optimization of maintenance scheduling to reduce unplanned highway repairs?

4. What is the optimal approach to forecasting IRI at different levels of granularity—both for entire highway routes (RouteID level) and for specific highway sections (0.1-mile segments)—to support more precise maintenance planning?

5. What might be the most effective method for visualizing and presenting findings to highway maintenance teams, like using geospatial mapping to find the roughest sections along a highway and their projected deterioration over time?

UMBC

THANK YOU