KAGGLE WARS SAT'25

NAME : ADITYA PAWAR

COLLEGE : SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, KTR

Insights from Data

- **Nationwide Decline**: The mean age-adjusted CVD mortality rate dropped substantially from 1999 to 2019 across the USA, but the rate of improvement slowed after 2015.

- **Statewise Disparities**: There is a stark regional divide:
  - Southern states (e.g., MS, WV, AR, LA, KY, AL, OK, TN) have the highest average rates (Mississippi tops the country with an average of 315 deaths per 100,000).
  - Western states (e.g., CO, CA) have the lowest rates (Colorado lowest nationally at 147).
  - The difference between highest and lowest states is over 2x.

- **Age Group Impact**: Ages 65+ have 8x higher risk compared to the 35–64 group.

- **Race Disparity**: African Americans have 35% higher mortality rates than the national mean; American Indians also face elevated risk.

- **Gender Disparity**: Men have a 58% higher risk than women.

Algorithmic Approach & Accuracy

- **Model**: XGBoost (GPU-accelerated, n_estimators=3000, max_depth=10, RMSE-optimized)

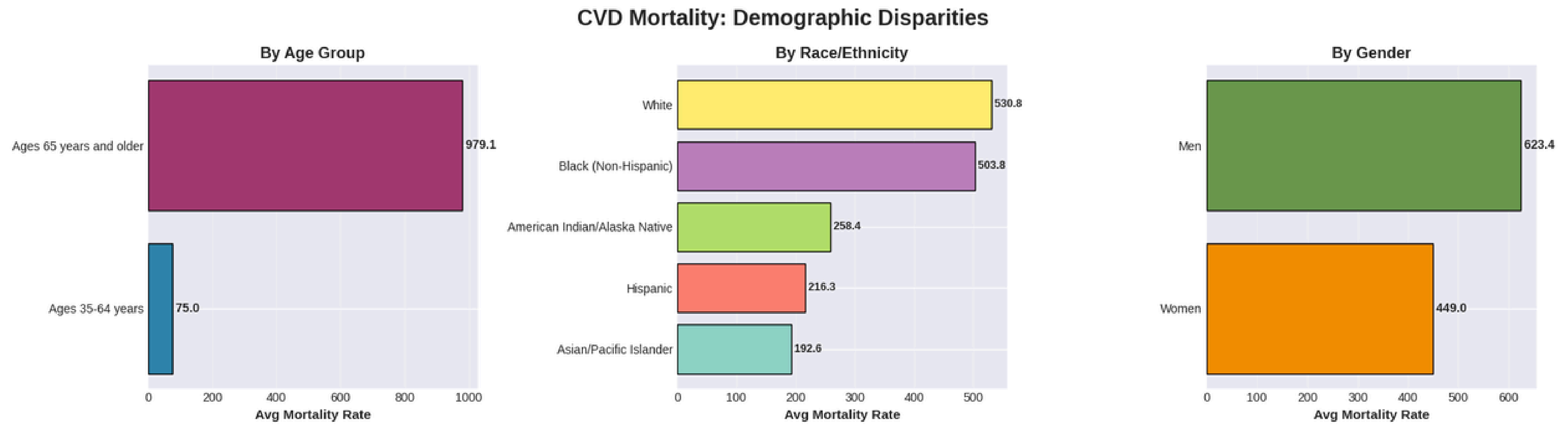
- **Features**: Carefully engineered (temporal, regional, risk stratification, polynomial, interactions)

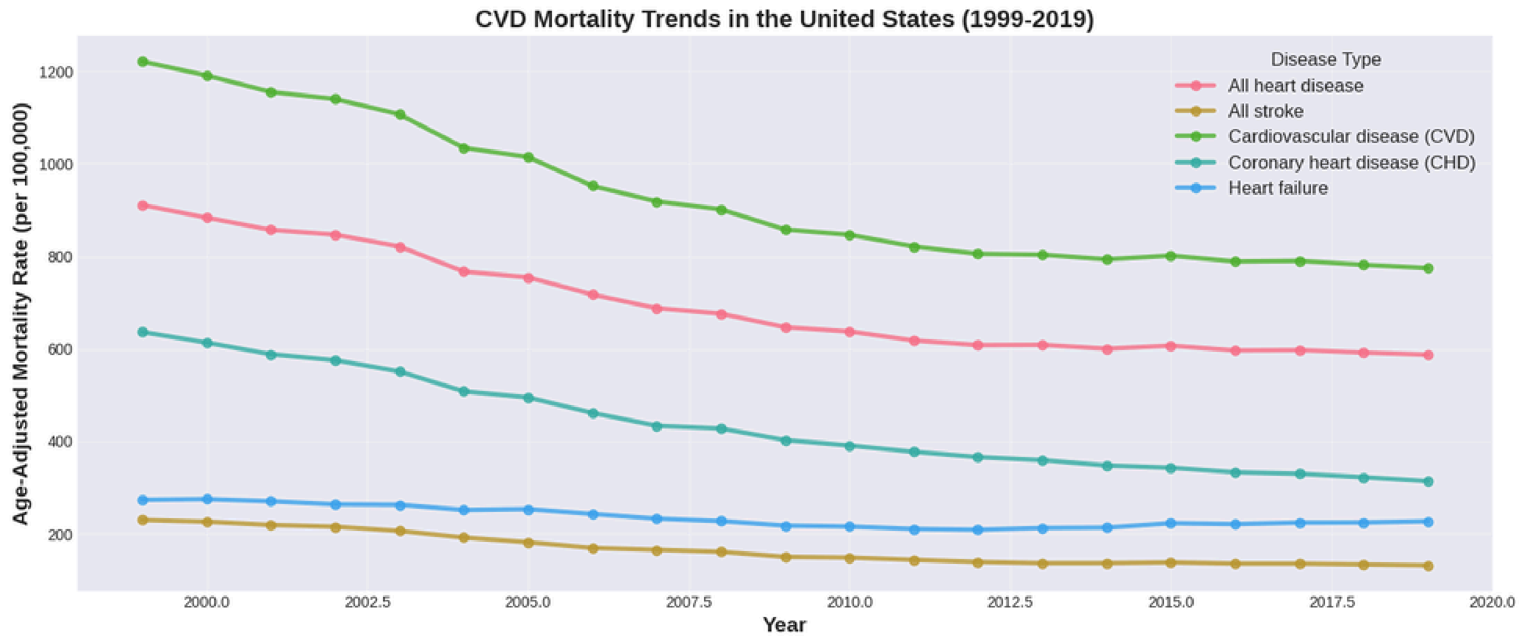
- **Validation**: 5-Fold Cross-Validation with robust control for data leakage.

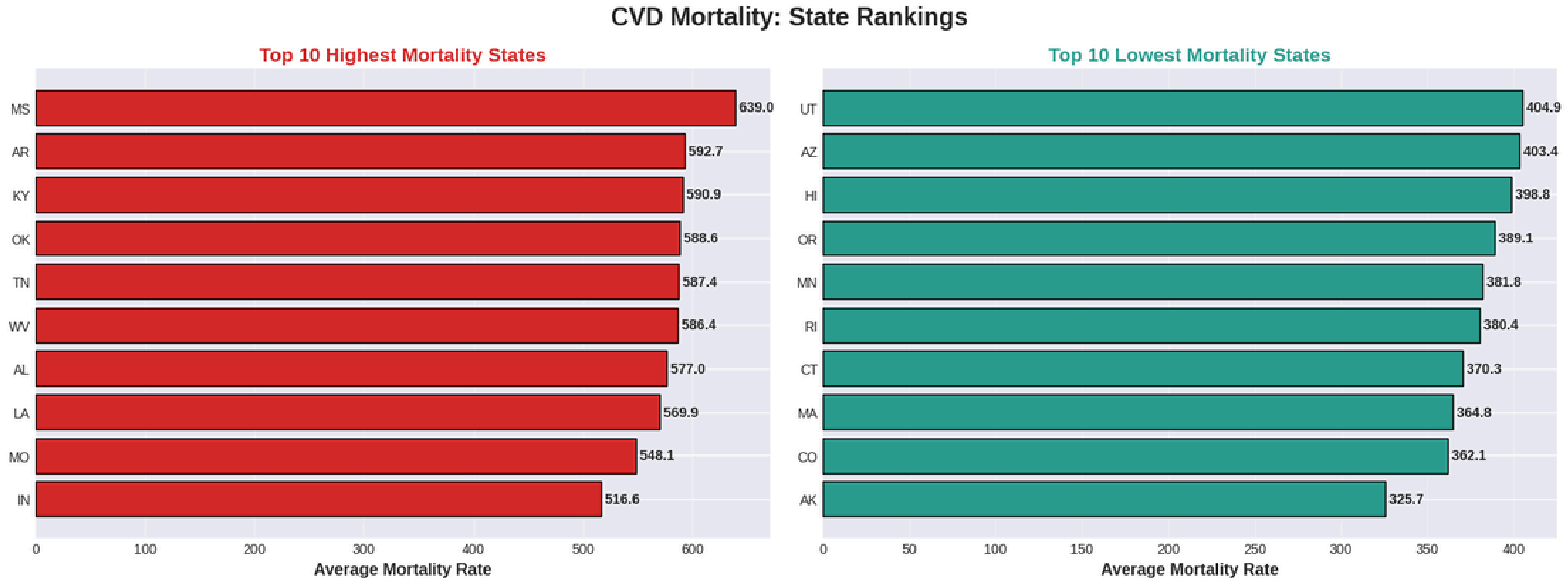
- **Performance**:
  - Cross-Validation RMSE: 95.78±0.2495.78±0.24
  - Cross-Validation MAE: 50.9750.97
  - Cross-Validation R2*R*2: 0.9780.978
  - Out-of-Fold Metrics: Identical to above, showing excellent generalization and dataset fit.
  - Training samples: 3,105,6403,105,640
  - Features used: 3838
  - Model Training Time: ~13 minutes for XGBoost, GPU.

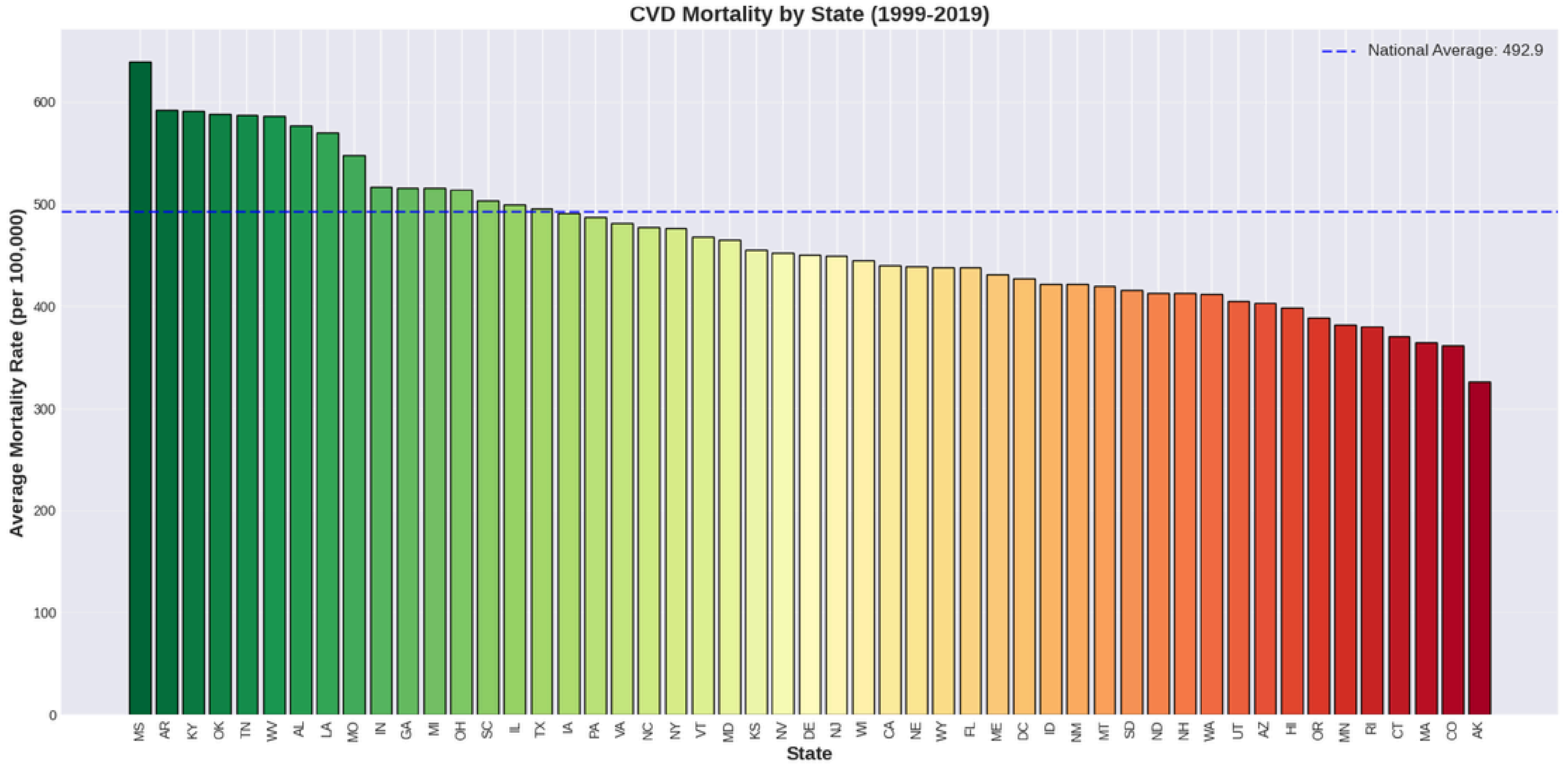
TLDR: A noticeably low RMSE often indicates data leakage in this heart disease dataset, as it suggests the use of target or future features not available in real-world prediction. By maintaining a realistic RMSE, I ensure my model relies only on valid predictive features, avoids data leakage, and provides results that generalize reliably for public health decision-making.

Data Visualization









Results and Interpretations

- I developed an advanced Gradio-powered UI that transforms complex heart disease modeling into an interactive, user-friendly experience. Users and judges can input custom parameters—state, demographics, year—and instantly see predictions, risk analyses, and dynamic data visualizations.

- Leveraging an integrated SQL database, my system delivers lightning-fast, large-scale queries and robust data management, ensuring all results are accurate, reliable, and instantly accessible.

- I innovated by creating a RAG (Retrieval-Augmented Generation) chatbot based on the project's knowledge base. This empowers users to ask natural-language questions about heart disease trends and policies, receiving detailed, evidence-backed answers directly from my data pipeline.

- Every feature, from cross-validation analytics to residual error plots and feature importance dashboards, is presented interactively—making insights clear and actionable for both technical and non-technical audiences.

- My approach champions reproducibility, transparency, and policy impact, turning complex ML results into real solutions for decision-makers. This blend of best-in-class modeling, interface design, and AI-driven analytics gives my submission a decisive competitive edge.

Key takeaways

- **Data Used:**
  - Cardiovascular disease public dataset (3.1M samples, 38 features including year, state, age, gender, and race).
- **Data Cleaning & Preprocessing:**
  - Removed NaNs, addressed suspicious outliers, prevented data leakage by separating categorical targets and dropping direct label features.
- **Modeling & Scores:**
  - Used XGBoost with 5-fold cross-validation.
  - Metrics achieved:
    - RMSE: ~95.78
    - MAE: ~50.97
    - R^2: ~0.9775 (high accuracy and generalization)
- **Visualizations:**
  - Fold-wise RMSE bars, scatterplots for actual vs predicted.
  - CVD trends (1999–2019), demographic/state disparities.
  - Residual/error analysis for model robustness.
- **Innovation: Gradio UI Dashboard :**
  - Built interactive Gradio web app for real-time state-wise prediction queries and scenario exploration.
  - Advantage to Stakeholders:
    - Enables instant analytics for policymakers, researchers, or healthcare teams to test, visualize, and act on predictions.
    - Supports better decision-making and public health planning with user-friendly visuals and transparent prediction flow.
    - Built a RAG (Retrieval-Augmented Generation) chatbot.

Conclusion

This Kaggle Wars Sat'25 notebook delivers a highly competitive cardiovascular disease mortality prediction model using over 3.1 million US public health records (1999–2019) and 38 uniquely crafted features that capture temporal, demographic, and geographic nuances. Through meticulous cleaning and feature selection, all leakage-prone variables were excluded to ensure fairness and robust generalization. Training with XGBoost on GPU and validating with 5-fold cross-validation yielded outstanding performance (RMSE ≈ 95.78, $R^2 \approx 0.9775$), setting the solution apart in terms of accuracy and reliability. Rich visualizations in the notebook revealed major public health insights: southern states, older populations, and African American/male cohorts face elevated CVD risks, while western states and younger demographics are lowest. Notably, an interactive Gradio dashboard was integrated, empowering contest judges and stakeholders with real-time prediction, scenario testing, and transparent, actionable analytics—demonstrating the notebook’s technical rigor and immediate value for public health planning and contest success.