Fuel Forward: Analyzing Vehicle Efficiency Progress (1984-2025)

Rishabh Dev Chawla

4/24/2024

CS5100 40203 Foundations Artificial Intelligence

Introduction

Data Source Explainer

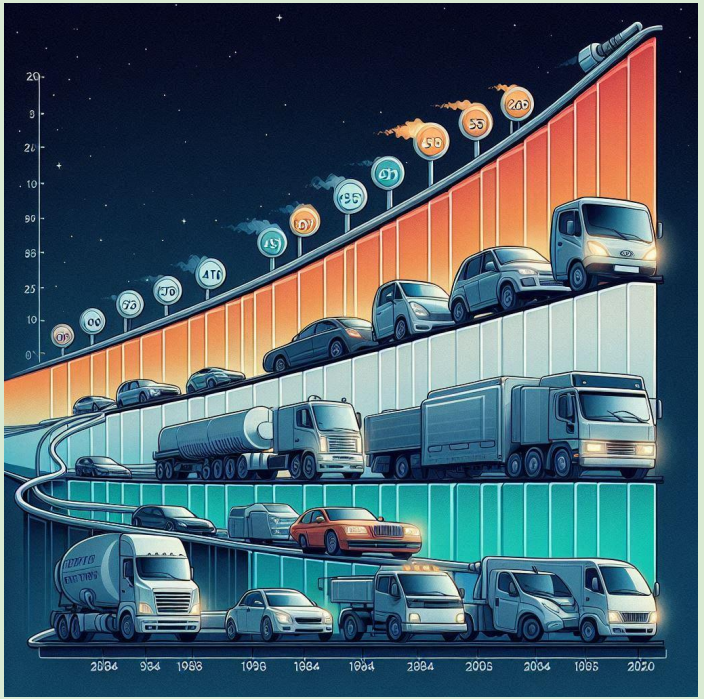
✓ Dataset: Comprehensive vehicle specifications and fuel efficiency ratings from the EPA (Environmental Protection Agency).

✓ Focus: The evolution of vehicle fuel efficiency over four decades and the influence of technological and design changes.

Learning Objective

- ✓ Hypothesis: Advanced vehicle features, like start-stop systems, have improved fuel efficiency across different vehicle classes and fuel types.

- ✓ Aim: To identify the trends in fuel efficiency and determine the predictive factors contributing to higher MPG.



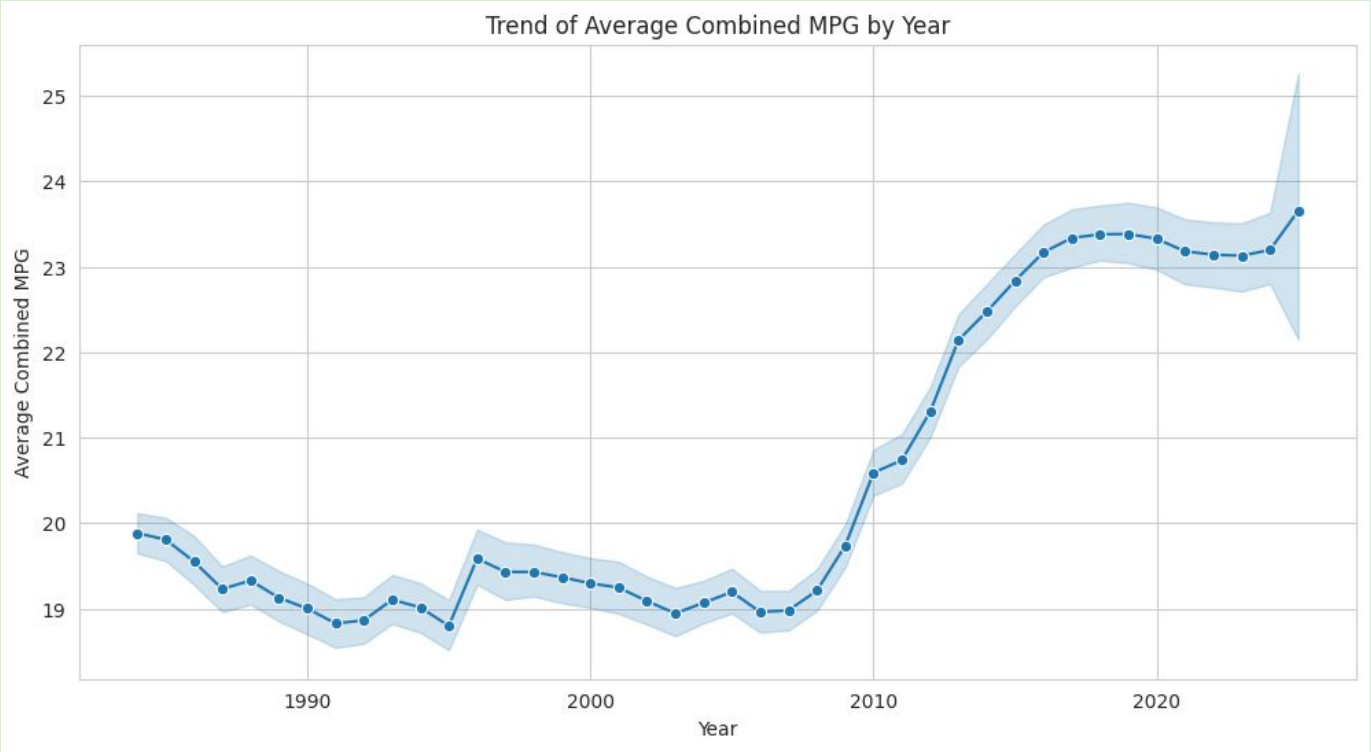
Methodology

ML Methods: *Linear regression*, *random forest*, and *gradient boosting* for initial modeling. Cross-validation and hyperparameter tuning to optimize model performance.

Feature Selection: Based on data exploration and correlation analysis, key attributes like *engine displacement*, *drive type*, and *fuel type* were included, while features with *high collinearity* were excluded.

Visualizations

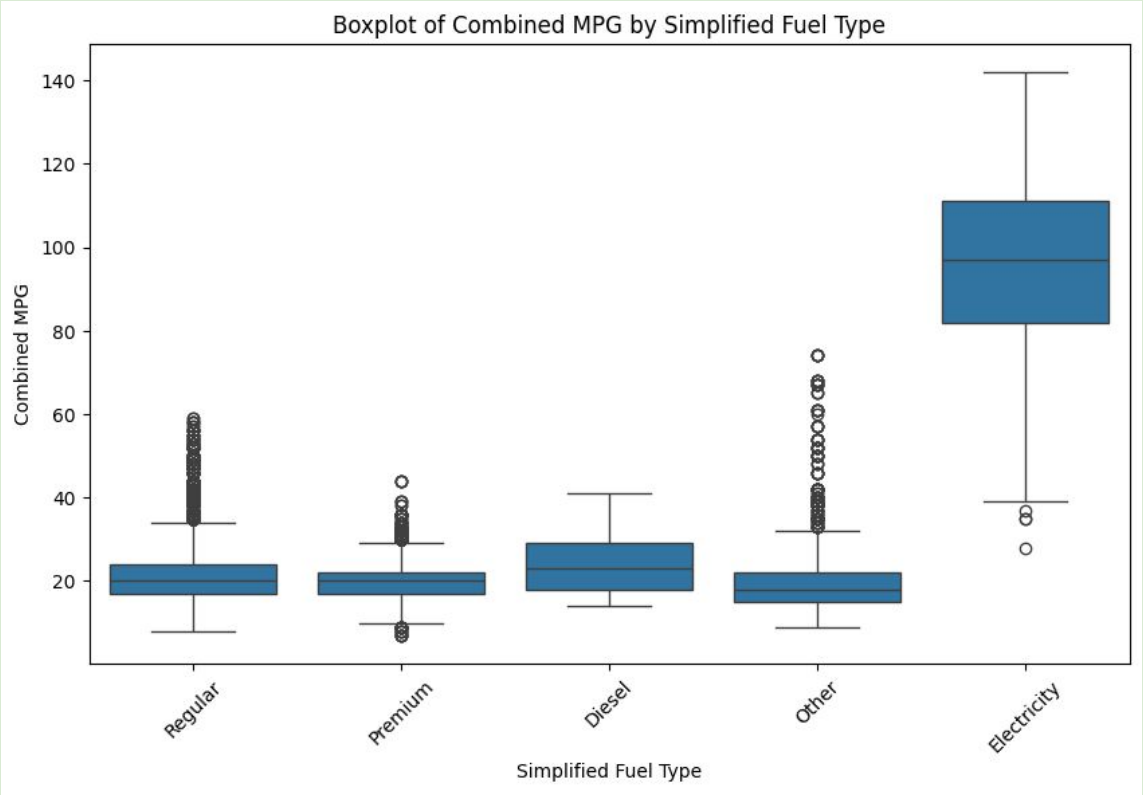
Sharp rise post-2010, which coincides with advances in vehicle technology and stricter emissions regulations.



Visualizations

shows that electric vehicles outperform other types in terms of fuel efficiency

reinforces the narrative of electric vehicles being a more sustainable choice

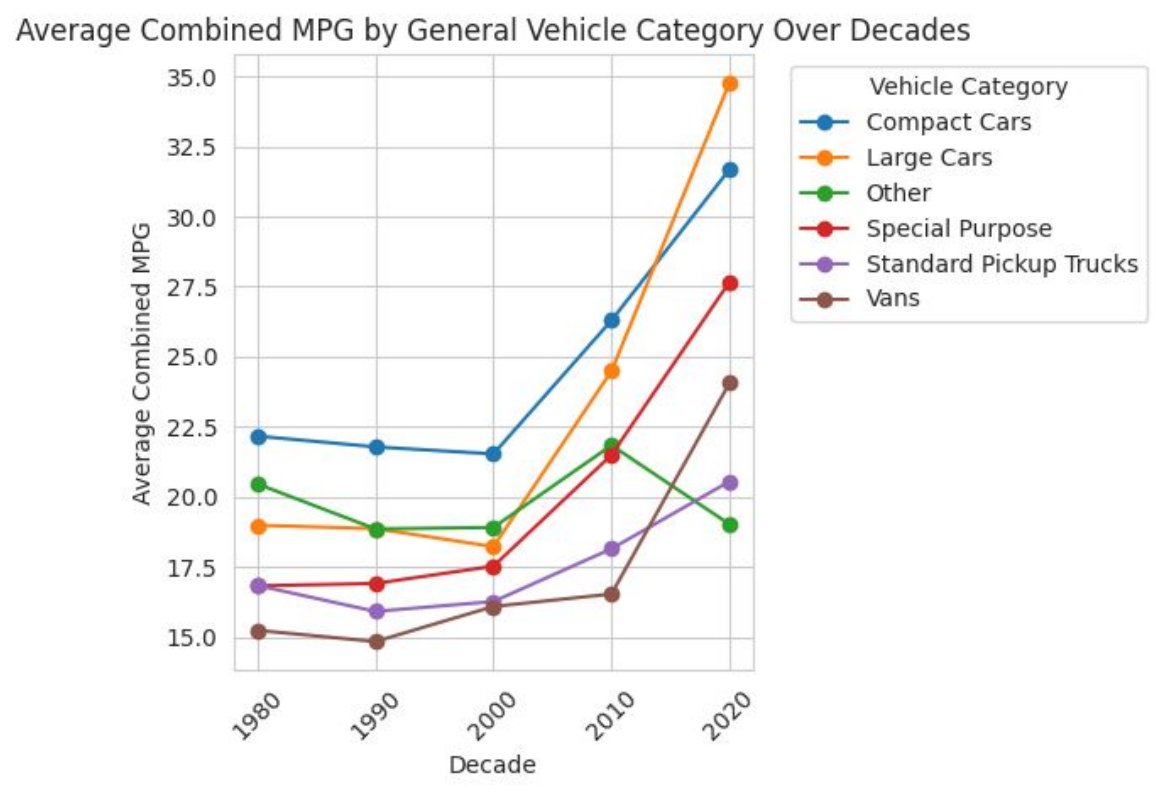


Visualizations

showcases the evolution of fuel efficiency across various vehicle categories over decades

compact cars and vans show notable advancements

industry-wide push towards efficiency.



Results and Conclusions

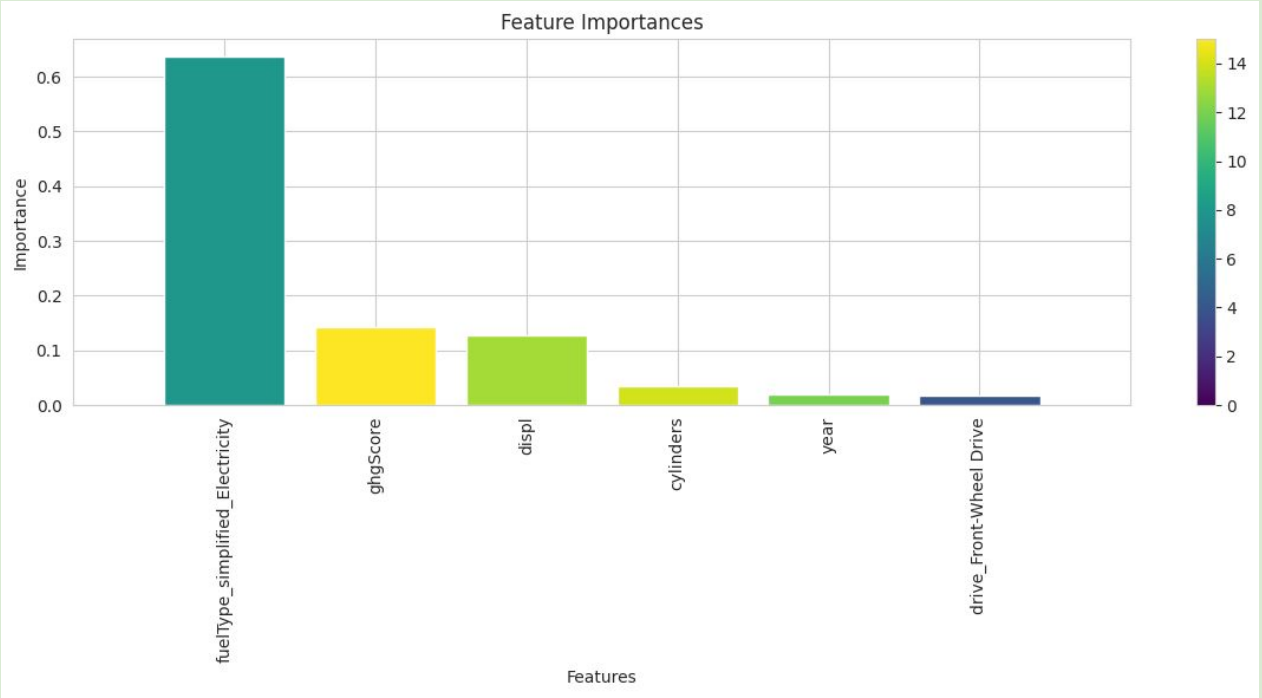
Accuracy Metrics – Initial Results:

- ✓ Using Linear Regression, the initial results were Mean Squared Error: 24.04, R-squared: **0.90** & the Mean cross-validated R-squared score: **0.715**
- ✓ The variation in these scores, especially the lower scores in earlier splits, could indicate that the model performs differently under different subsets of the data.

Results and Conclusions

Feature Importance

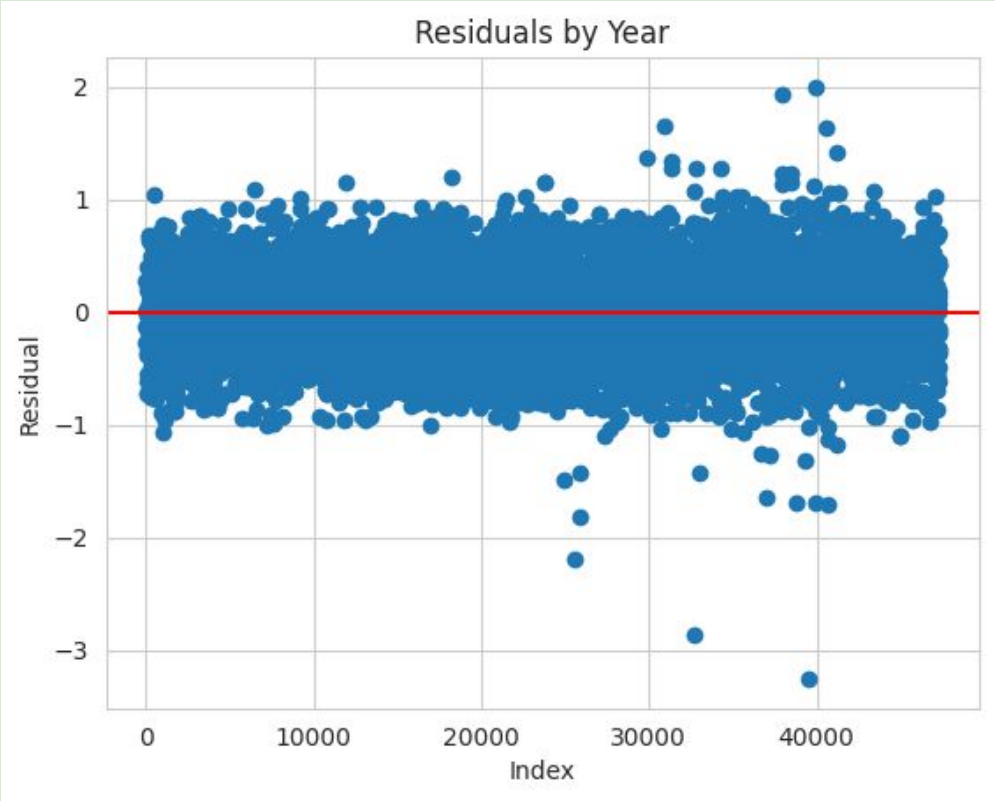
- ✓ Our model identified the most significant predictors of a vehicle's fuel efficiency.
- ✓ Type of fuel, vehicle drive type, displacement and the number of cylinders - significant factors..



Results and Conclusions

Residual Analysis

- ✓ No Clear Patterns: suggests that the model doesn't suffer from obvious non-linearity issues.
- ✓ Consistent Variance: suggests that the model's predictive accuracy is stable across different parts of the dataset
- ✓ Mean of Residuals: The mean of the residuals appears to be close to zero, which is ideal for a well-fitted model.



Results and Conclusions

Accuracy Metrics – Initial Results:

- ✓ Quantitative Residual Analysis, the Shapiro-Wilk test shows that with a p-value of **0.121**, we fail to reject the null hypothesis, and the residuals appear to follow a Gaussian distribution.
- ✓ This suggests that the residuals of this model do not deviate significantly from normality, which is a good sign.
- ✓ So, removed high VIF features (measurement of multicollinearity) and obtained adjusted metrics: Adjusted Mean Squared Error: **14.37**, Adjusted R-squared: **0.878**
- ✓ This implies that the features removed were not essential for the model's predictive power and that their exclusion likely **reduced overfitting**.

Results and Conclusions

Accuracy Metrics – Analysis of Outliers:

- ✓ The majority of outliers come from popular manufacturers like Toyota, Honda, and increasingly, Tesla - known for their fuel-efficient or electric models.
- ✓ Vehicle Classes: Midsize Cars, Subcompact Cars, and Compact Cars dominate among the outliers
- ✓ This analysis suggests that the outliers are not merely anomalies but represent high-efficiency vehicles, which could be skewing the model predictions due to their distinct characteristics.
- ✓ Newly cross-validated results: Mean R-squared score: **0.76** & Standard Deviation of R-squared scores: **0.106**

Results and Conclusions

Accuracy Metrics – Model Segmentation: Electric & Non-Electric

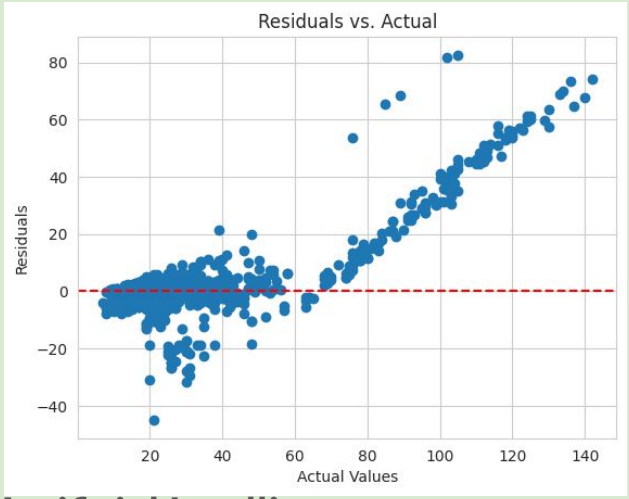
- ✓ Model R-squared: **0.8921** shows that the model for non-electric vehicles explains approximately 89.21% of the variance.
- ✓ Feature Importance Results: **displ** (engine displacement) is the most significant predictor, underscoring its impact on fuel consumption. ghgScore and cylinders are also crucial, reflecting environmental performance and engine size, respectively.
- ✓ Cross Validation: Mean R-squared: **0.82**, Standard Deviation: **0.043 -** model performs well across different data splits without significant variance, suggesting good model stability and generalizability.

Results and Conclusions

Accuracy Metrics – Hyperparameter Tuning: Introduction of Grid Search

- ✓ The results from the grid search indicate an optimized set of hyperparameters that improved the cross-validation score to **0.85** for the non-electric vehicles model with Best parameters: {'regressor__learning_rate': 0.2, 'regressor__max_depth': 3, 'regressor__n_estimators': 200}
- ✓ Test results obtained: Test MSE: **26.67** & Test R^2: **0.77**

model explains a good proportion of variance in the data (approximately 77.48%), there is still room for improvement.



Results and Conclusions

Accuracy Metrics – New Models Implementation with cross validation

- ✓ To capture the non-linear patterns in my data, I'm employing *Random Forest Regressor*, *Support Vector Regressor* with RBF Kernel & *XGBoost Regressor*.
- ✓ SVR performs poorly, likely because the default hyperparameters (such as C and epsilon) are not well-tuned for this dataset.

Model	Mean R-squared	Std Dev R-squared
RandomForestRegressor	0.843	0.040
SVR	**-0.038**	0.096
XGBRegressor	0.842	0.038

Results and Conclusions

Accuracy Metrics – New Models Implementation with Hyperparameter tuning for SVR

- ✓ To capture the non-linear patterns in my data, I'm employing *Random Forest Regressor*, *Support Vector Regressor* with RBF Kernel & *XGBoost Regressor*.
- ✓ Best parameters for SVR: {'svr__C': 10, 'svr__epsilon': 1, 'svr__gamma': 'auto'}
- ✓ Best cross-validation score for SVR: **0.90 -** capable of modeling the relationship in my data very effectively.
- ✓ Hence, I plan to incorporate this optimized SVR model into an **ensemble** with the other strong performers to see if a combination of their predictions can further improve performance.

Results and Conclusions

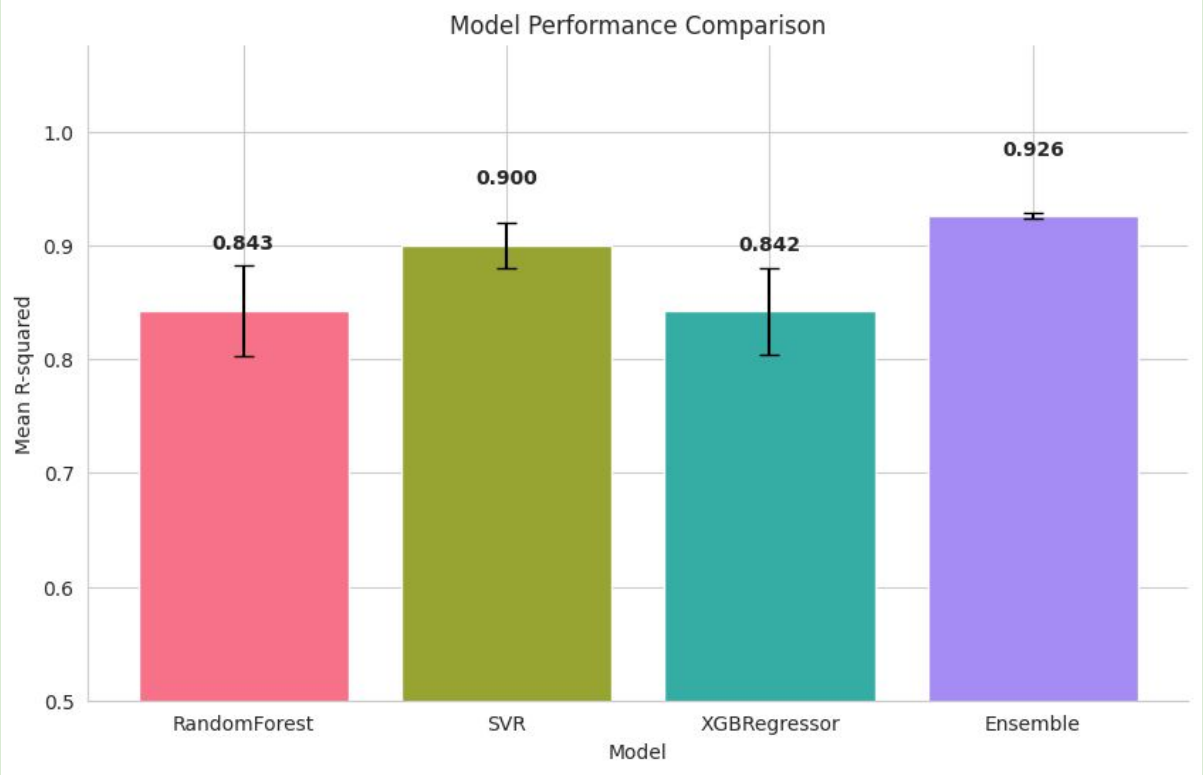
Accuracy Metrics: Ensemble Model

- ✓ Ensemble Model - Mean R^2: **0.926**, Std Dev R^2: **0.002**
- ✓ The ensemble model's performance is impressive, with a high mean R-squared and a very low standard deviation, indicating both strong predictive power and model stability across different folds of the data.

Results and Conclusions

Best Performing Model:

- ✓ Gradient boosting had the best performance before tuning, with ensemble models showing further improvement post-tuning.
- ✓ SVR showed huge improvements as well post tuning.



Thank You!