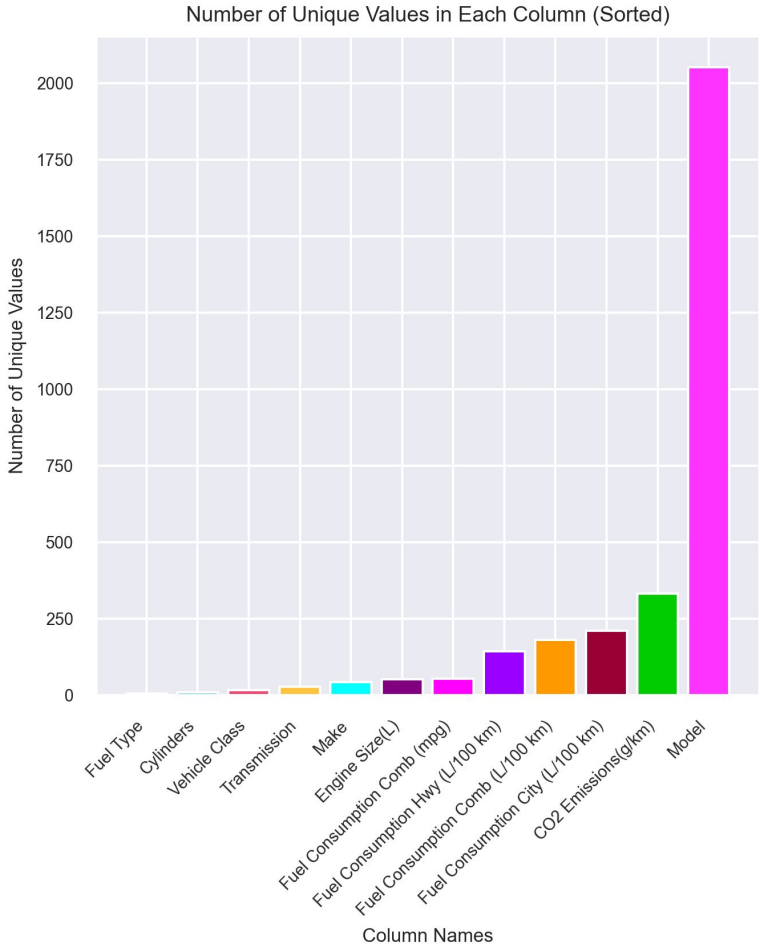
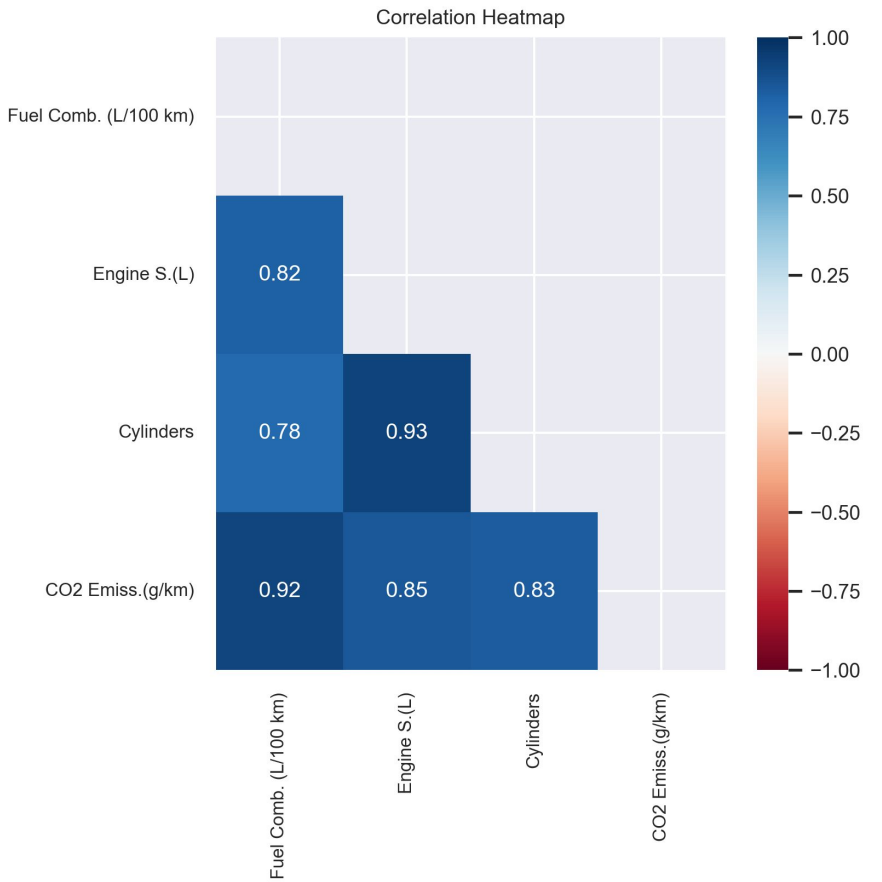
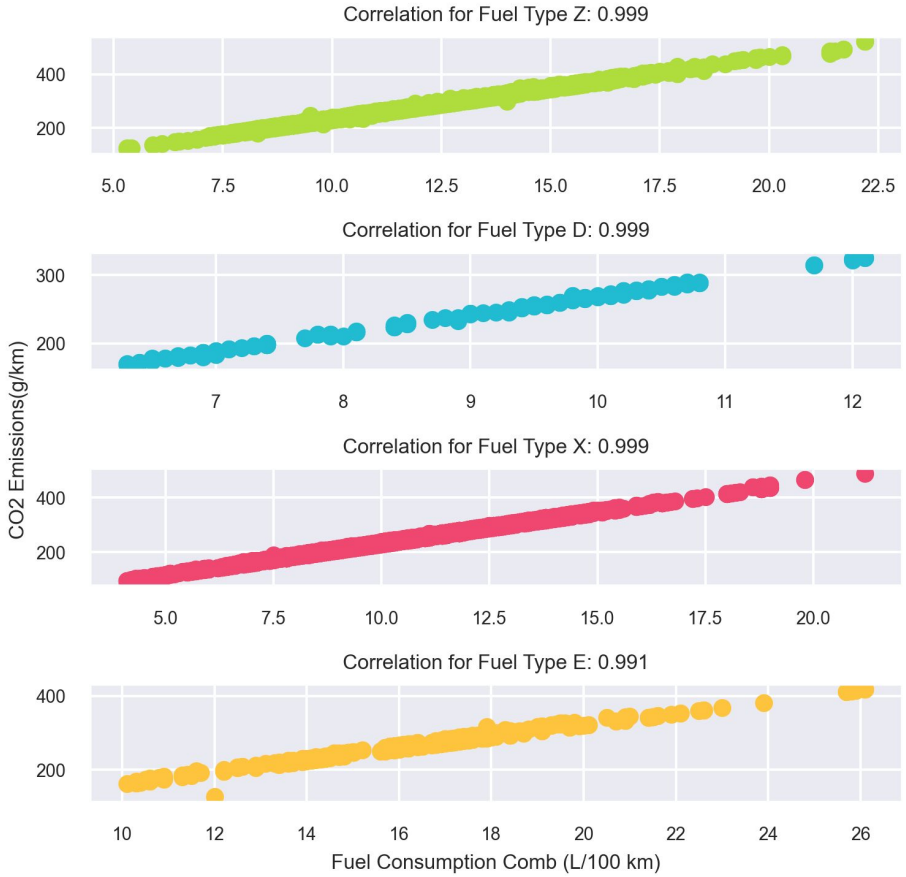
Cars' CO2 Emissions Prediction

- **Goal:** develop a model to **predict CO2 emissions** using cars' characteristics data.
- **Tools and Techniques**: different models, Grid Search, Cross Validation.
- **Dataset: ~7K** data points.
- **Features: 6** numerical, **5** categorical.
- **Data Source**: **Canadian Government** Portal (compiled on Kaggle).

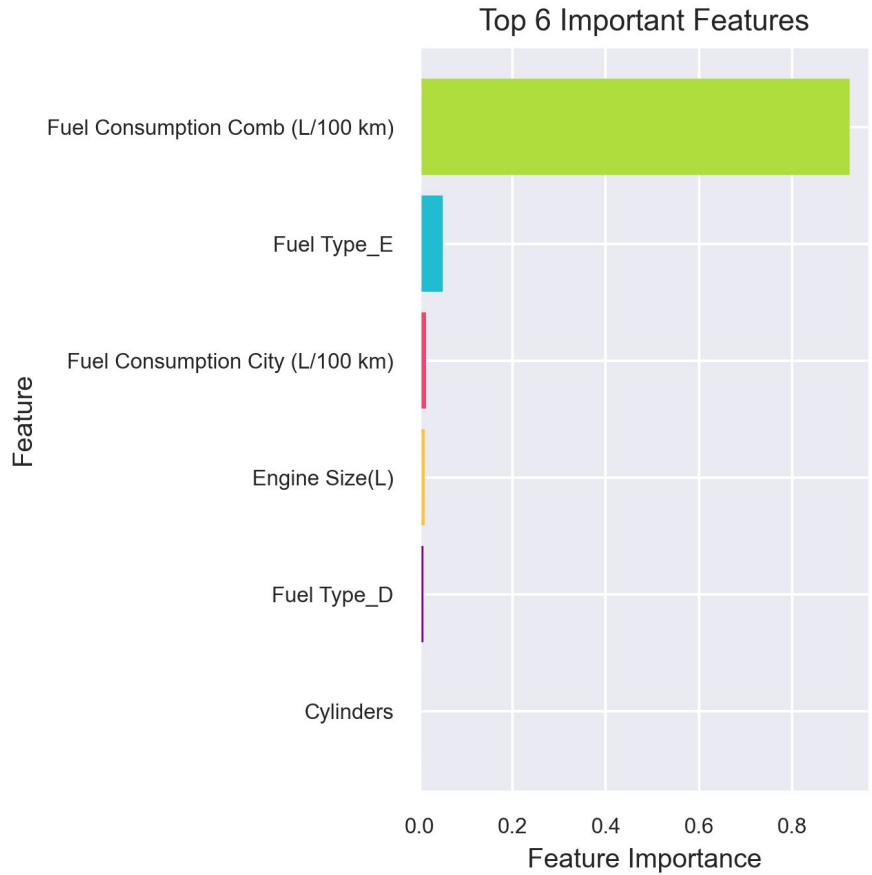


Correlations





RF Model with Fuel Consumption or Formula for Best Result



Initial state:

- ❏ **5908** Data Points
- ❏ **2148** Features

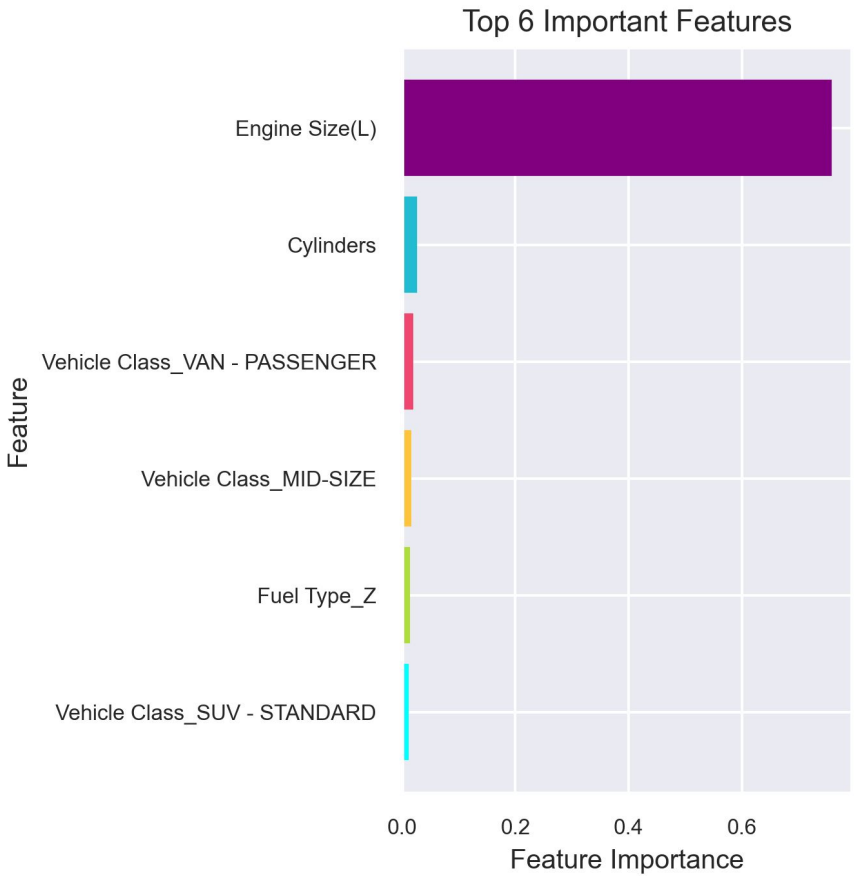
With Weighted Importance:

- ❏ **5908** Data Points
- ❏ **5** Features

Test Score:

- ❏ R2: **0.994**
- ❏ RMSE: **4.5**

RF Model without Fuel Consumption



Initial state:

- **5908** Data Points
- **2145** Features

With Weighted Importance:

- **5908** Data Points
- **6** Features

Test Score:

- R2: **0.959**
- RMSE: **11.8**

Cat Boost Models (Gradient Boosting on Decision Trees)

- **With** Car Model Column
- **No** Fuel Consumption

- **5908** Data Points
- **2145** Features

Test Score:

- R2: **0.96**
- RMSE: **10.3**

- **No** Car Model Column
- **No** Fuel Consumption

- **5908** Data Points
- **92** Features

Test Score:

- R2: **0.95**
- RMSE: **12.8**

Main Findings:

- **Use formula**, when **Fuel Consumption** is present, it yields the best prediction of CO2.
- **Engine Size:** provides good prediction (**R2 = 0.959**), even without Fuel Consumption.
- **Categorical columns** don't significantly improve performance (**R2 = 0.96 - 0.95**).
- **Better prediction - better cars' optimization.**