Modeling Process

1. **Feature/Target Setup**
    - Input features (X) excluded `driver_id` and `enhanced_risk_score`.
    - Target (y) was the `enhanced_risk_score`, which had been engineered as a composite measure of driving risk.
2. **Categorical & Numeric Handling**
    - `vehicle_type` was treated as categorical and one-hot encoded.
    - All other numeric features were passed through directly.
    - This hybrid preprocessing ensured the models could leverage both **continuous driver behavior metrics** and **vehicle-type effects**.
3. **Models Tested**
    - **CatBoost Regressor**: Excellent at handling categorical data natively, baseline for comparison.
    - **Random Forest Regressor**: Robust, interpretable tree-based method.
    - **XGBoost Regressor**: Gradient-boosted trees, strong performance on structured data.
    - **Gradient Boosting Regressor**: Another boosting method for risk scoring.
    - **Stacking Ensemble (final choice)**: Combined RF, XGB, and GBR as base models, with a **Ridge regression meta-model**. This allowed the ensemble to learn strengths of each model and gave the best generalization.
4. **Validation Strategy**
    - **5-fold Cross Validation** (MAE, RMSE, R²) ensured consistency across splits.
    - **Hold-out test set** was used for unbiased evaluation.
5. **Model Performance** (example test results)

  - CatBoost: MAE ~ 5.99, R^2 ~ 0.63
  - RandomForest: MAE ~ 6.26, R^2 ~ 0.55
  - XGBoost / GradientBoosting: Weaker (R^2 ~ 0.13–0.17)
  - **StackingEnsemble: MAE ~ 1.85, RMSE ~ 2.28, R^2 ~ 0.95 → best performer**

Pricing Engine Design

1. **Base Premium**
   - Set at **$2,285/year**, consistent with the U.S. average for full coverage (ensures industry realism).
2. **Risk Normalization**
   - Risk scores were clipped between `0–100` and then normalized to `0–1`.
3. **Scaling Factor**
   - Designed to scale premiums **up to 50% higher** for the riskiest drivers.
4. **Outputs**
   - Both **annual premium** and **monthly premium** were generated.
   - Example output included predicted risk score, premium annual, and premium monthly per driver.

Why This Approach

- **Interpretability:** Premiums are directly linked to a risk score that is both machine-learned and human-readable.
- **Industry Alignment:** Ties to realistic base premiums avoids outputs that would feel disconnected from real insurance markets.

- **Scalability:** By keeping the pricing engine modular, different scaling factors or base premiums can be applied for various geographies, policies, or risk tolerances.