# Machine Learning Model Tracking Document

## 1. Dataset Information

|  |  |
| --- | --- |
| Dataset Name: | Well 160 |
| Number of Samples: | 160 |
| Number of Features: | 4 (Tf, Rs, Gg, Api) |
| Target Variable: | Bob |
| Outlier Handling: | None |
| Feature Engineering Applied: | None |
| Scaling/Normalization Applied: | Only for Neural Networks |
| Encoding Applied: | None |

## 2. Preprocessing Steps

|  |  |
| --- | --- |
| Step | Description |
| Train-Test Split | 70% - 30% |
| Shuffling | Yes, using random\_state=42 |
| Handling Missing Data | None |
| Feature Scaling | Only for Neural Networks |
| Feature Selection | None |

## 3. Models Used & Hyperparameters

|  |  |  |
| --- | --- | --- |
| Model | Hyperparameters | Training Time |
| **XGBoost** | n\_estimators=200, learning\_rate=0.05, max\_depth=4 | 0.0553 |
| **CatBoost** | iterations=1000, learning\_rate=0.05, depth=4, l2\_leaf\_reg=5 | 0.1069 |
| **Neural Network** | [128, 64], epochs=100, batch\_size=16 | 2.5446 |
| **Stacking Ensemble** | Default base models + CatBoost final estimator | 0.4885 |
| **Extra Trees** | n\_estimators=200 | 0.0532 |
| **Deep Neural Network** | [256, 128, 64], epochs=150, batch\_size=32 | 2.3703 |

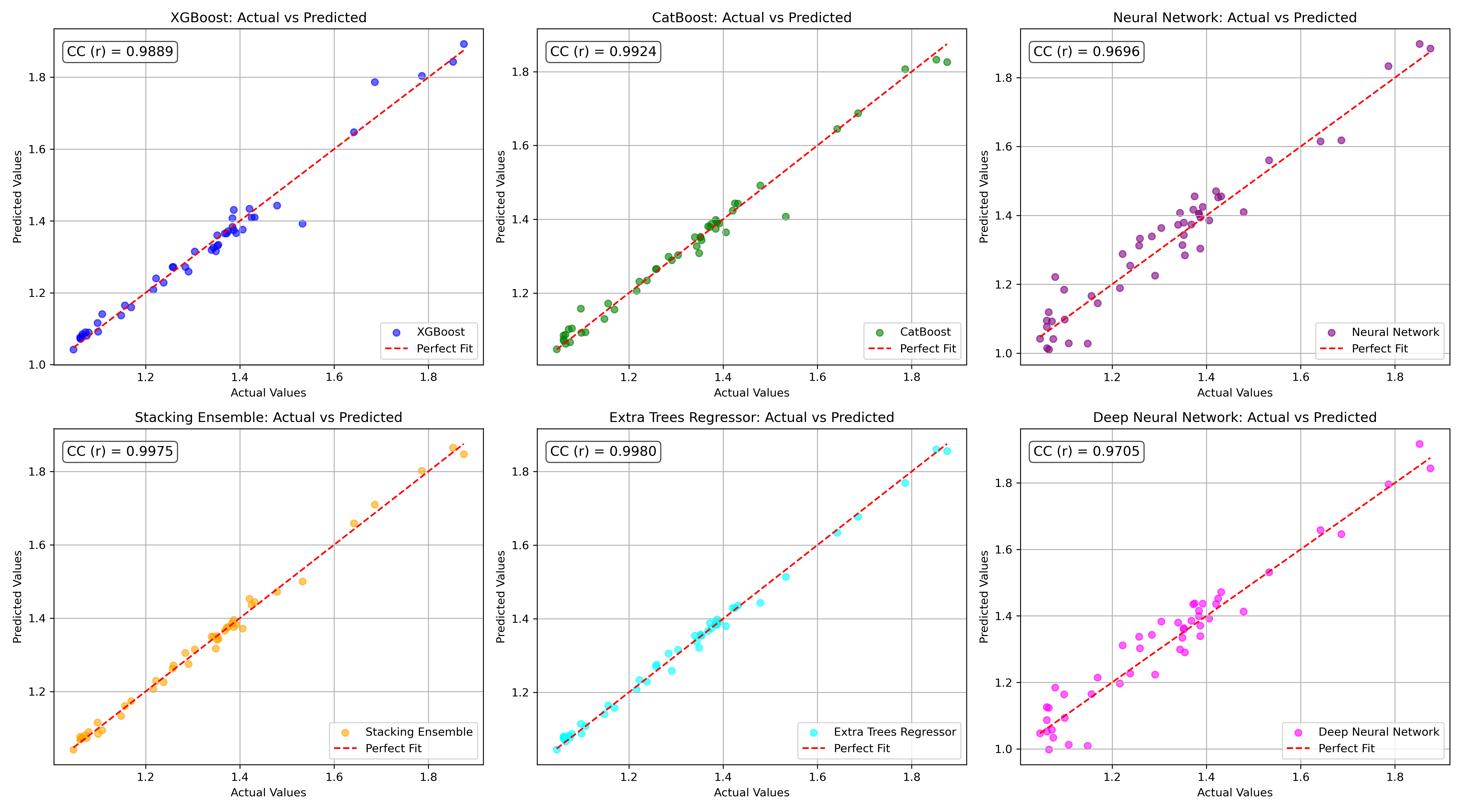
## 4. Evaluation Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | R² Score | Adjusted R² |
| **XGBoost** | 0.0006 | 0.0251 | 0.0159 | 0.9851 | 0.9838 |
| **CatBoost** | 0.0003 | 0.0181 | 0.0123 | 0.9922 | 0.9915 |
| **Neural Network** | 0.0022 | 0.0473 | 0.0373 | 0.9472 | 0.9423 |
| **Stacking Ensemble** | 0.0002 | 0.0147 | 0.0120 | 0.9949 | 0.9944 |
| **Extra Trees** | 0.0002 | 0.0138 | 0.0110 | 0.9955 | 0.9951 |
| **Deep Neural Network** | 0.0027 | 0.0517 | 0.0418 | 0.9369 | 0.9311 |

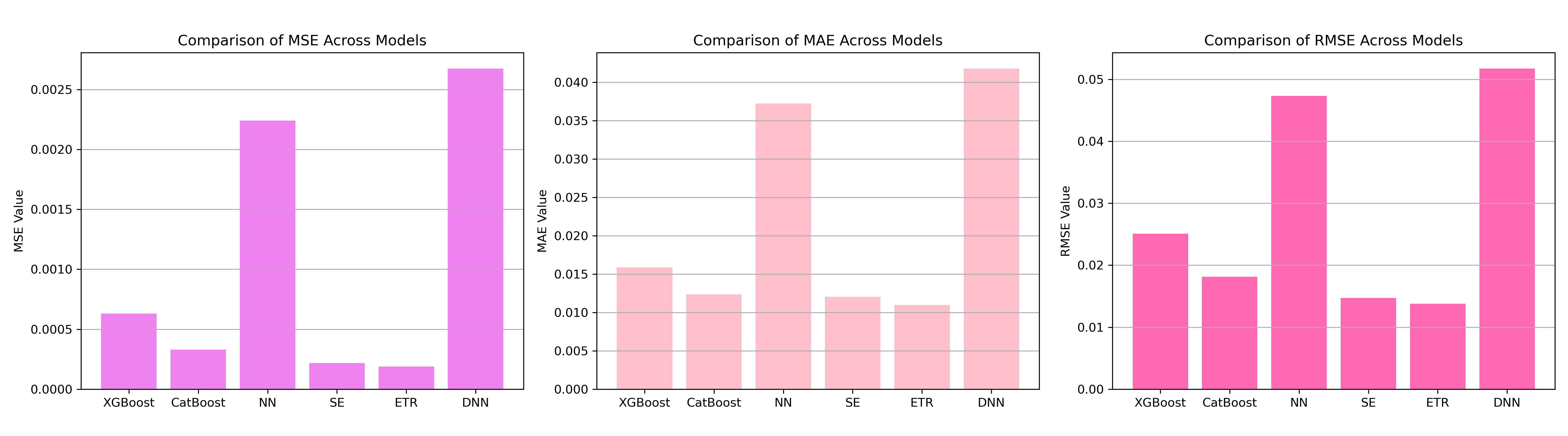
## 5. Cross-Validation Summary (5-Fold)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | RMSE Mean | RMSE Std | MAE Mean | MAE Std | R² Mean | R² Std |
| **Stacking Ensemble** | 0.0159 | 0.0036 | 0.0110 | 0.0015 | 0.9937 | 0.0019 |
| **Extra Trees Regressor** | 0.0162 | 0.0017 | 0.0120 | 0.0010 | 0.9932 | 0.0022 |
| **XGBoost** | 0.0201 | 0.0033 | 0.0138 | 0.0012 | 0.9894 | 0.0039 |
| **CatBoost** | 0.0237 | 0.0106 | 0.0138 | 0.0032 | 0.9842 | 0.0145 |
| **Neural Network** | 0.1867 | 0.0467 | 0.1444 | 0.0334 | 0.0315 | 0.5853 |
| **Deep Neural Network** | 0.2188 | 0.0545 | 0.1770 | 0.0433 | -0.3119 | 0.6285 |

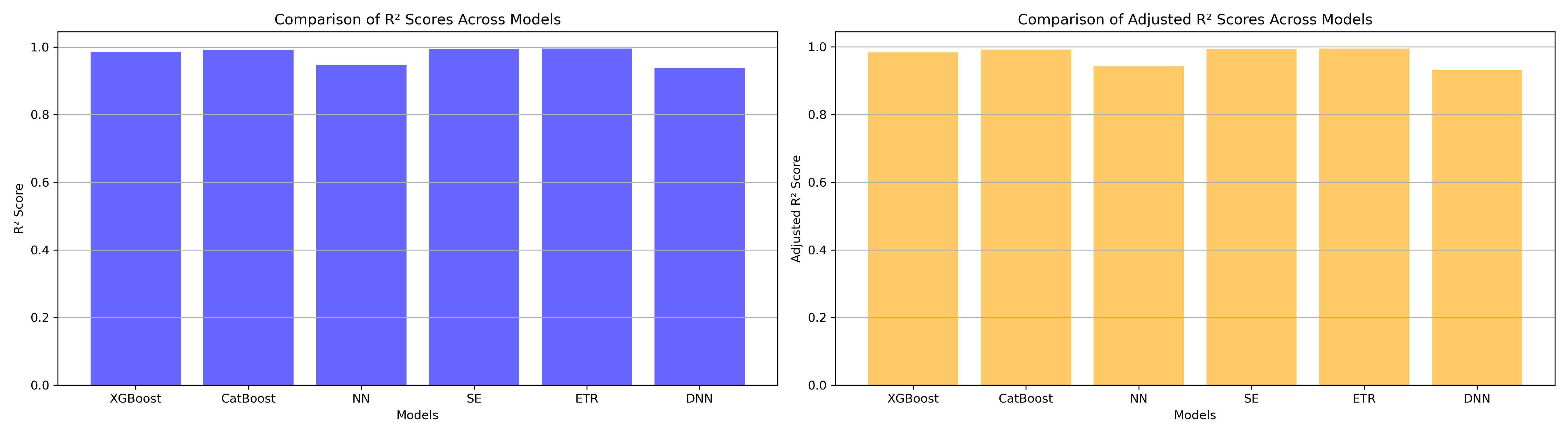
## 6. Visualizations



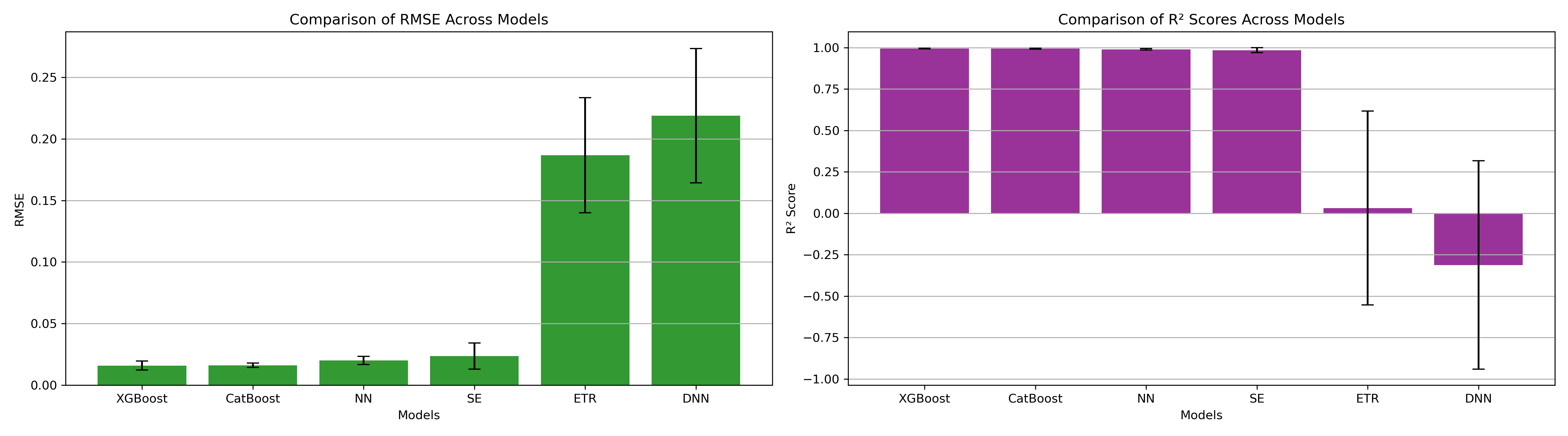
*Figure 1: Actual vs. Predicted Values for Bob (Well 160)*



*Figure 2: Bar Charts of MSE, RMSE, MAE for Bob (Well 160)*



*Figure 3: Bar Charts of R² and Adjusted R² for Bob (Well 160)*



*Figure 4: Error Bars for RMSE and R² from CV for Bob (Well 160)*

### 7. Observations & Next Steps

* **Best Performing Model:** **Extra Trees Regressor** (RMSE: 0.0138, R²: 0.9955)
  + Stacking Ensemble was very close, with RMSE 0.0147 and R² 0.9949.
  + CatBoost and XGBoost also performed well but slightly behind ensemble methods.
  + Neural-based models (**NN** and **DNN**) **performed poorly** compared to tree-based models.
    - **Deep Neural Network had a negative mean R² during cross-validation (-0.3119)**.
    - Indicating generalization problems on small datasets.

### 8. Code Access

The complete source code for data preprocessing, model training, evaluation, and visualization is [available here.](http://github.com/BoushBoo/pvt-prediction-ml-/tree/main) The repository includes organized Jupyter notebooks for each phase, dataset, and target, as well as requirements for reproducibility.