# Machine Learning Model Tracking Document

## 1. Dataset Information

|  |  |
| --- | --- |
| Dataset Name: | Well 782 |
| Number of Samples: | 782 |
| 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 | 0.0483 |
| **CatBoost** | iterations=1000, learning\_rate=0.05, depth=6, l2\_leaf\_reg=3 | 0.3445 |
| **Neural Network** | [128, 64], epochs=100, batch\_size=16 | 2.5329 |
| **Stacking Ensemble** | Default base models + CatBoost final estimator | 0.7889 |
| **Extra Trees** | n\_estimators=200 | 0.1127 |
| **Deep Neural Network** | [256, 128, 64], epochs=150, batch\_size=16 | 2.3055 |

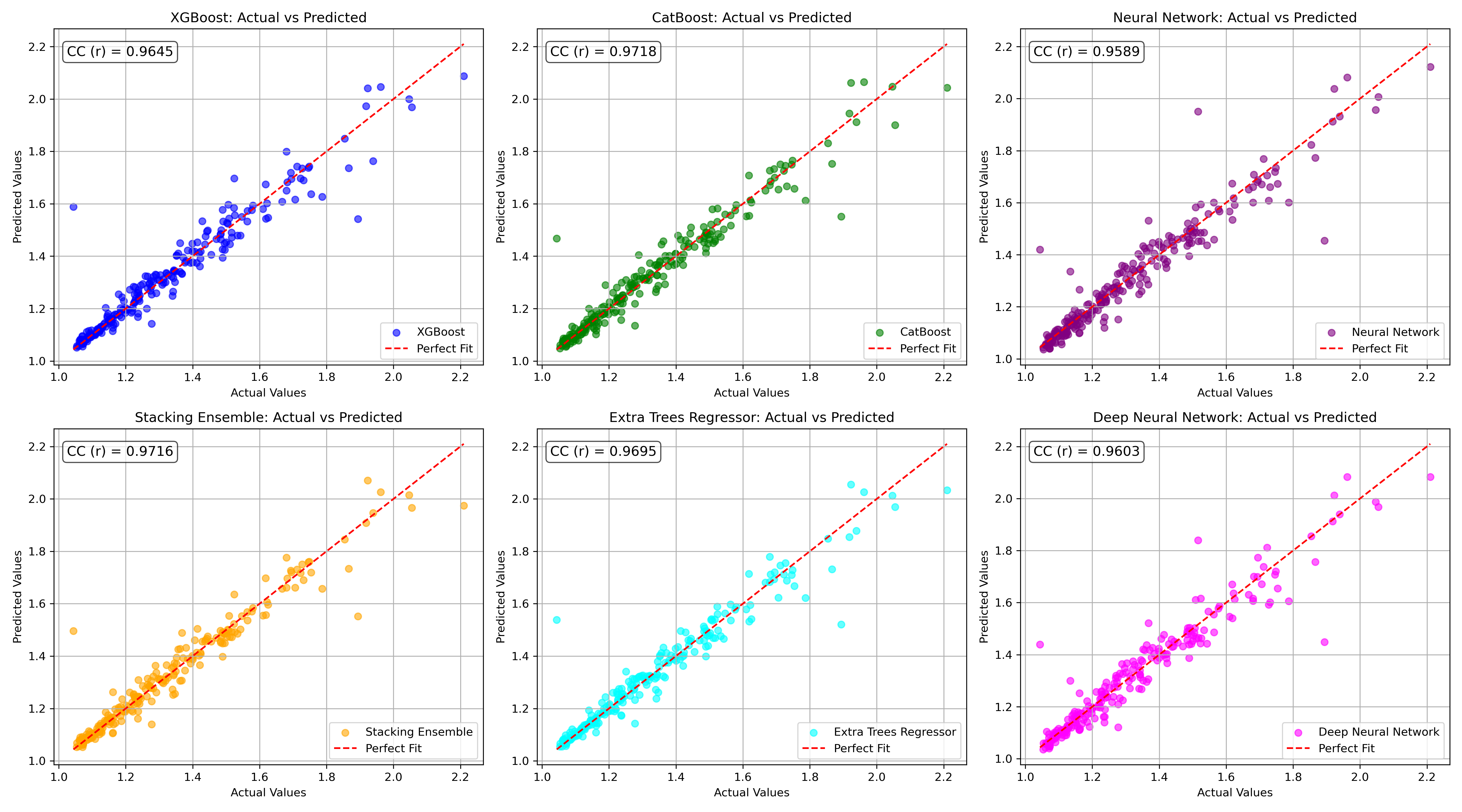
## 4. Evaluation Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | R² Score | Adjusted R² |
| **XGBoost** | 0.0031 | 0.0554 | 0.0311 | 0.9417 | 0.9407 |
| **CatBoost** | 0.0028 | 0.0525 | 0.0295 | 0.9476 | 0.9467 |
| **Neural Network** | 0.0036 | 0.0600 | 0.0382 | 0.9318 | 0.9306 |
| **Stacking Ensemble** | 0.0029 | 0.0543 | 0.0289 | 0.9441 | 0.9431 |
| **Extra Trees** | 0.0032 | 0.0563 | 0.0295 | 0.9398 | 0.9388 |
| **Deep Neural Network** | 0.0037 | 0.0608 | 0.0353 | 0.9299 | 0.9287 |

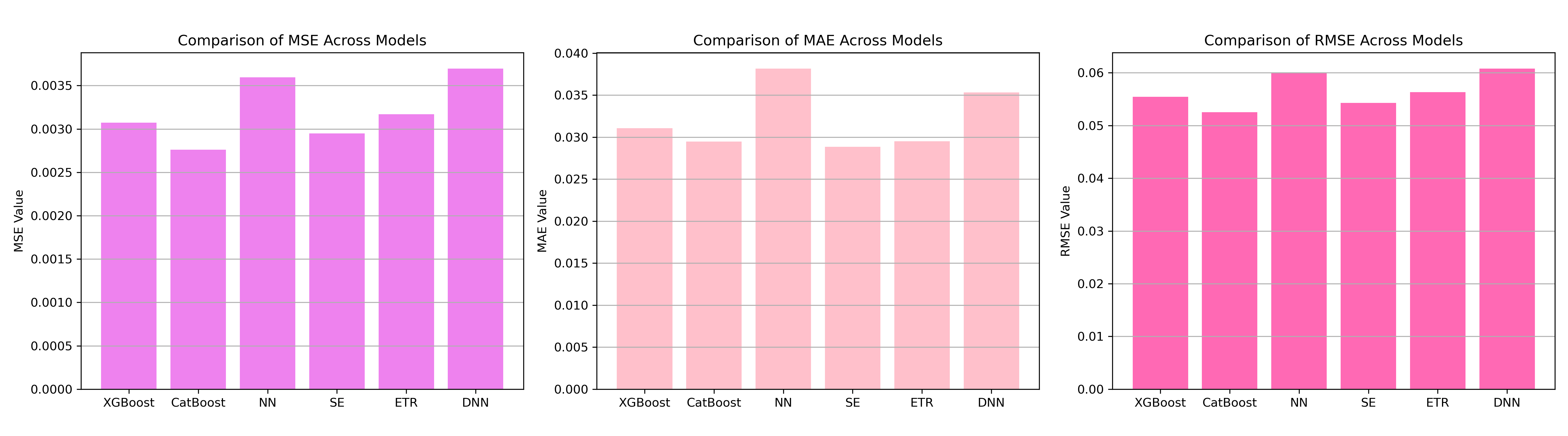
## 5. Cross-Validation Summary (5-Fold)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | RMSE Mean | RMSE Std | MAE Mean | MAE Std | R² Mean | R² Std |
| **Stacking Ensemble** | 0.0668 | 0.0213 | 0.0311 | 0.0059 | 0.9381 | 0.0189 |
| **CatBoost** | 0.0668 | 0.0253 | 0.0315 | 0.0077 | 0.9381 | 0.0256 |
| **Extra Trees Regressor** | 0.0684 | 0.0212 | 0.0317 | 0.0064 | 0.9349 | 0.0196 |
| **XGBoost** | 0.0724 | 0.0293 | 0.0339 | 0.0086 | 0.9276 | 0.0337 |
| **Neural Network** | 0.2229 | 0.0316 | 0.1641 | 0.0266 | 0.2816 | 0.1071 |
| **Deep Neural Network** | 0.2230 | 0.0235 | 0.1692 | 0.0173 | 0.2400 | 0.2437 |

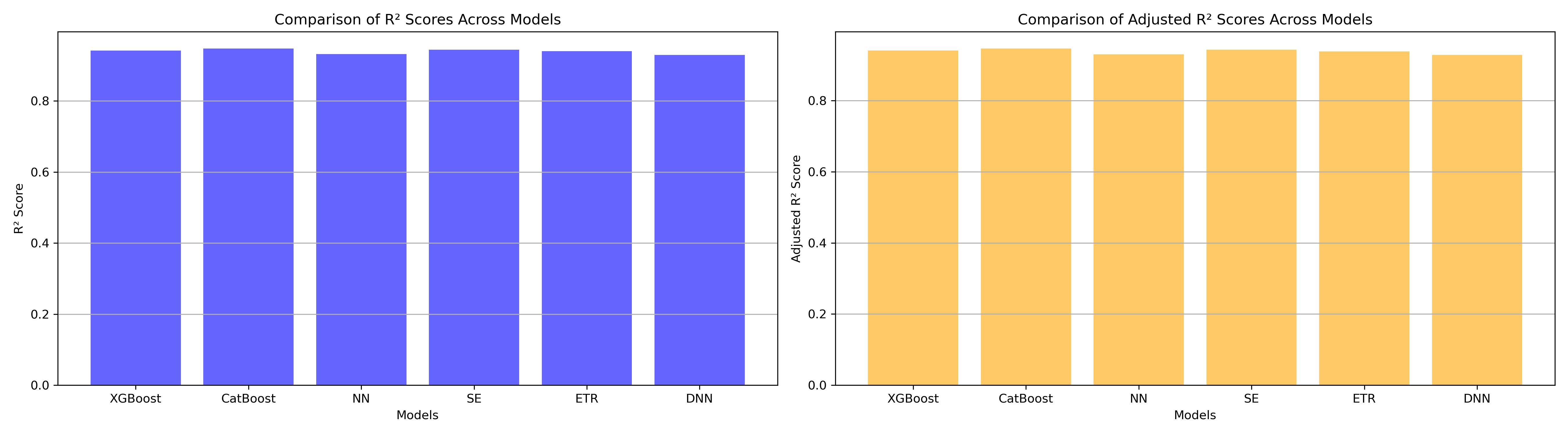
## 6. Visualizations



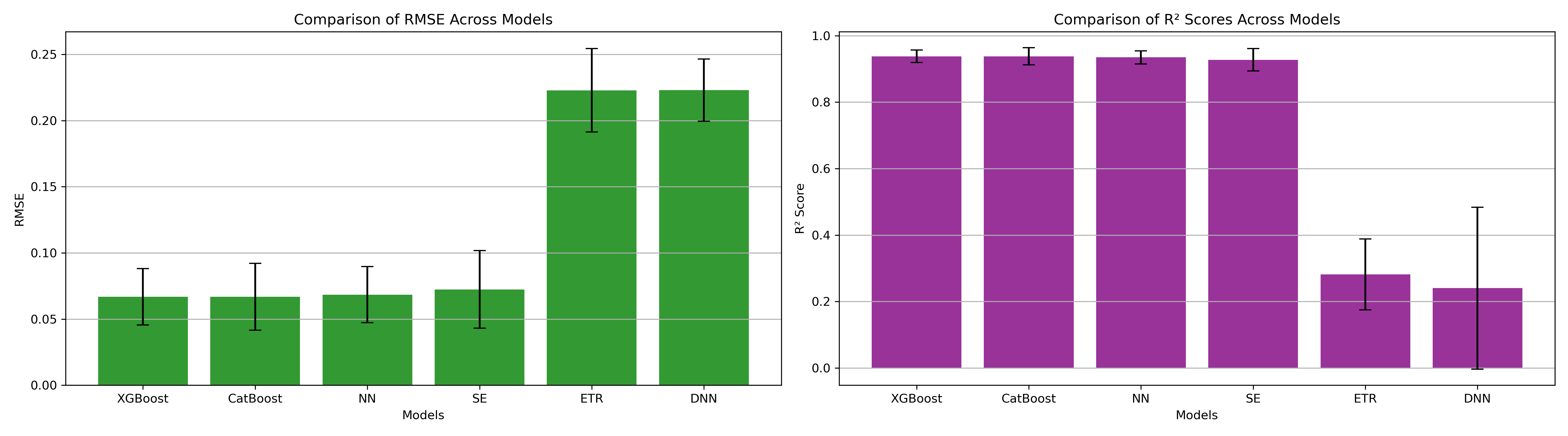
*Figure 1: Actual vs. Predicted Values for Bob (Well 782)*



*Figure 2: Bar Charts of MSE, RMSE, MAE for Bob (Well 782)*



*Figure 3: Bar Charts of R² and Adjusted R² for Bob (Well 782)*



*Figure 4: Error Bars for RMSE and R² from CV for Bob (Well 782)*

### 7. Observations & Next Steps

**Best Performing Models:** **CatBoost** and **Stacking Ensemble** — both with RMSE around 0.052–0.054 and R² above 0.944.

* Extra Trees and XGBoost also performed competitively but slightly behind.
* **Neural Networks (NN and DNN)** underperformed heavily:
  + **NN R² during CV:** ~0.28
  + **DNN R² during CV:** ~0.24
* **Reason:** Deep learning requires larger datasets and more complex feature interactions, while boosting and ensemble trees generalize better on medium datasets like 782 samples.

### 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.