# 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=100, learning\_rate=0.1, max\_depth=6 | 0.0308 |
| **CatBoost** | iterations=1000, learning\_rate=0.05, depth=4, l2\_leaf\_reg=7 | 0.1913 |
| **Neural Network** | [64, 32], epochs=100, batch\_size=16 | 3.2676 |
| **Stacking Ensemble** | Default base models + CatBoost final estimator | 0.7347 |
| **Extra Trees** | n\_estimators=200 | 0.1572 |
| **Deep Neural Network** | [256, 128, 64], epochs=100, batch\_size=16 | 2.7042 |

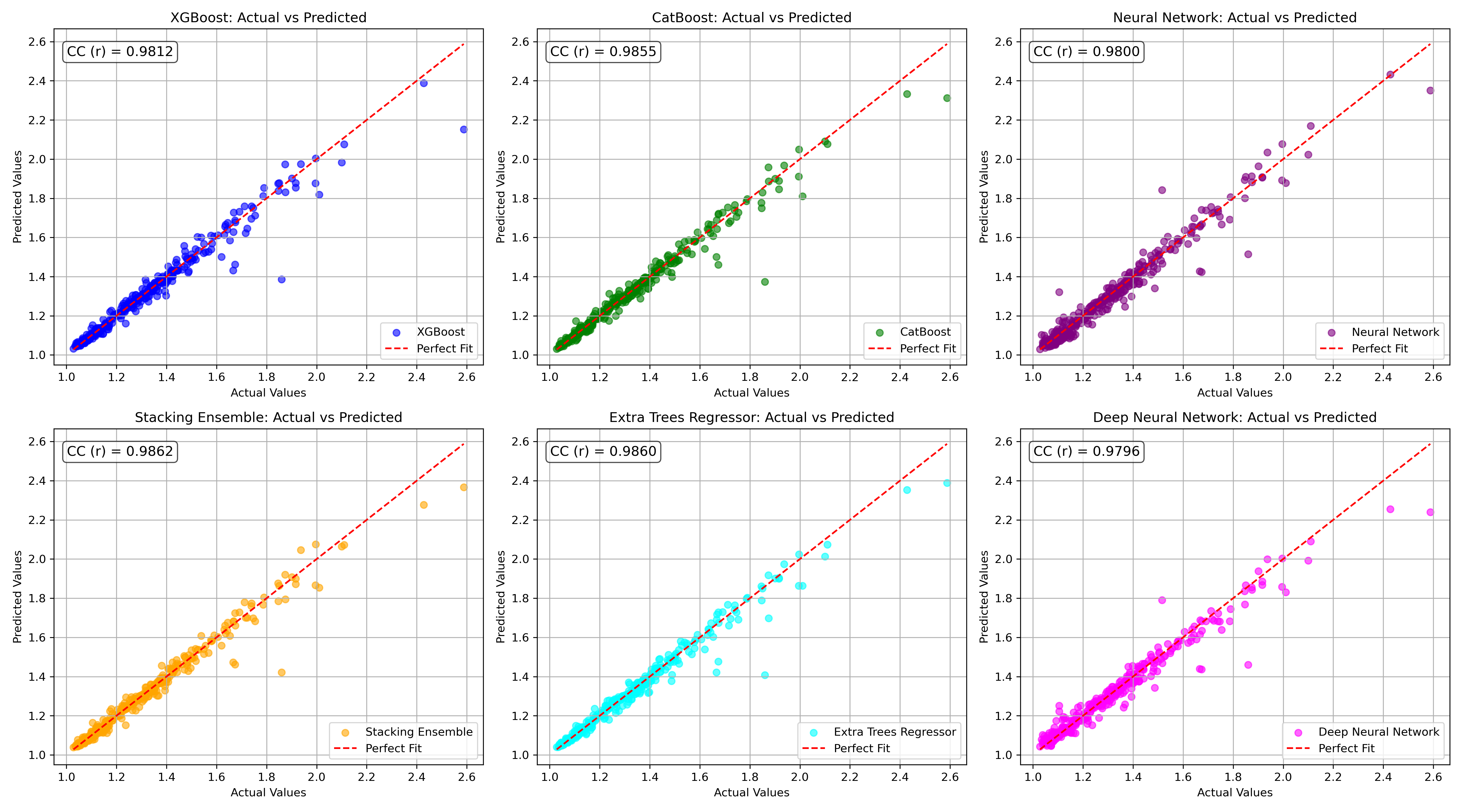
## 4. Evaluation Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | R² Score | Adjusted R² |
| **XGBoost** | 0.0019 | 0.0438 | 0.0200 | 0.9658 | 0.9655 |
| **CatBoost** | 0.0015 | 0.0392 | 0.0199 | 0.9726 | 0.9723 |
| **Neural Network** | 0.0022 | 0.0474 | 0.0288 | 0.9600 | 0.9596 |
| **Stacking Ensemble** | 0.0016 | 0.0399 | 0.0200 | 0.9717 | 0.9714 |
| **Extra Trees** | 0.0017 | 0.0407 | 0.0199 | 0.9705 | 0.9702 |
| **Deep Neural Network** | 0.0019 | 0.0435 | 0.0236 | 0.9663 | 0.9659 |

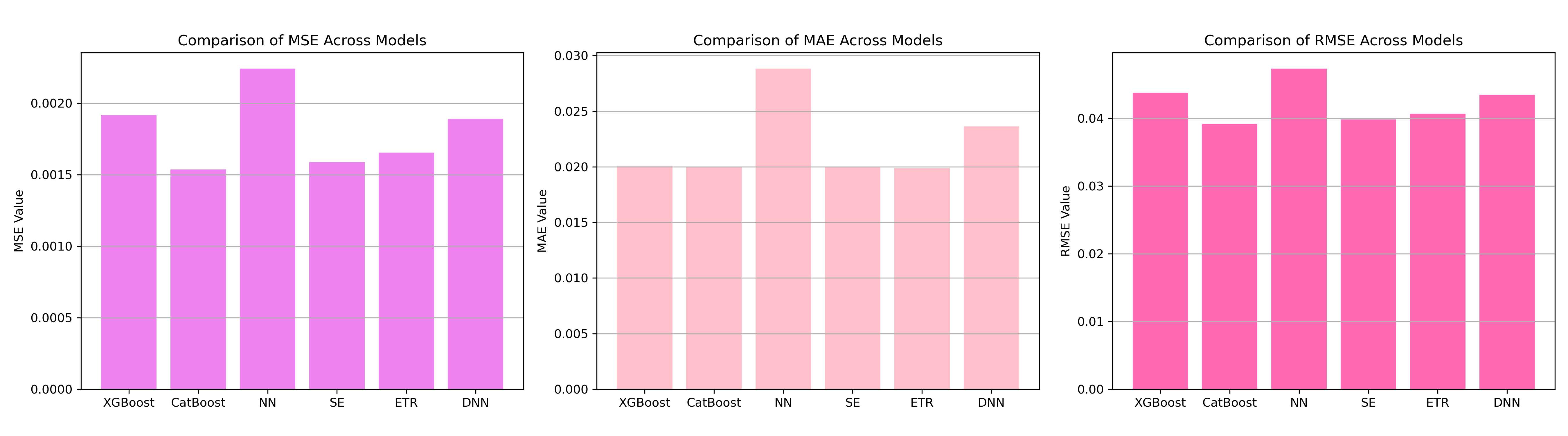
## 5. Cross-Validation Summary (5-Fold)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | RMSE Mean | RMSE Std | MAE Mean | MAE Std | R² Mean | R² Std |
| **Stacking Ensemble** | 0.0448 | 0.0083 | 0.0215 | 0.0023 | 0.9654 | 0.0118 |
| **Extra Trees Regressor** | 0.0460 | 0.0099 | 0.0209 | 0.0021 | 0.9636 | 0.0128 |
| **CatBoost** | 0.0466 | 0.0103 | 0.0227 | 0.0025 | 0.9614 | 0.0181 |
| **XGBoost** | 0.0481 | 0.0085 | 0.0228 | 0.0017 | 0.9610 | 0.0095 |
| **Deep Neural Network** | 0.1586 | 0.0108 | 0.1173 | 0.0248 | 0.5588 | 0.1217 |
| **Neural Network** | 0.1877 | 0.0502 | 0.1408 | 0.0187 | 0.3865 | 0.2732 |

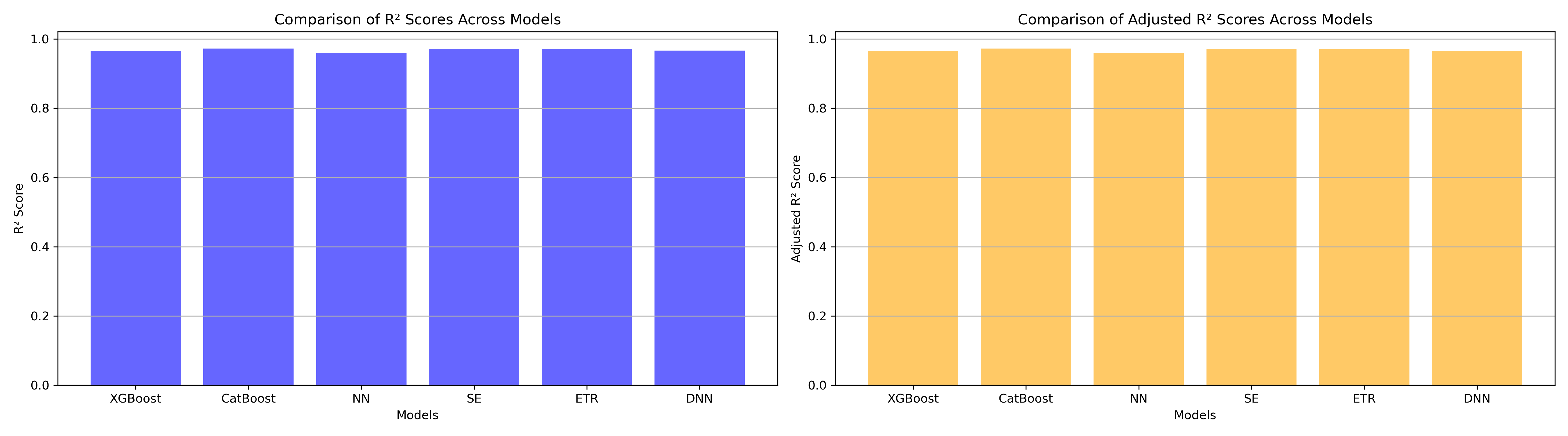
## 6. Visualizations



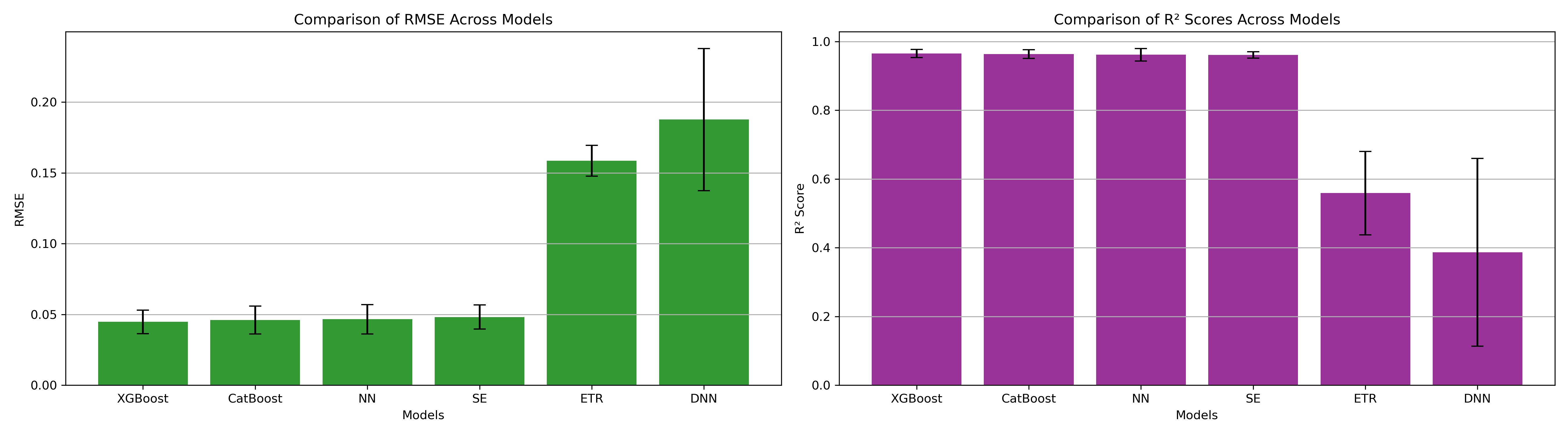
*Figure 1: Actual vs. Predicted Values for Bob (Well 782)*



*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 Model:** **Stacking Ensemble** (RMSE: 0.0399, R²: 0.9717)

* CatBoost and Extra Trees also performed very well, with RMSEs around 0.040–0.041 and R² > 0.97.
* XGBoost trailed slightly behind the ensemble and CatBoost.
* Neural Networks (both **NN** and **DNN**) performed **significantly worse**:
  + **Cross-validation R² for Deep NN:** ~0.56
  + **Cross-validation R² for NN:** ~0.39
* Again, this confirms that for **small to moderate datasets**, **tree-based ensembles are better** than deep learning models.

### 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.