# Machine Learning Model Tracking Document

## 1. Dataset Information

|  |  |
| --- | --- |
| Dataset Name: | Well 283 |
| Number of Samples: | 283 |
| Number of Features: | 4 (Tf, Rs, Gg, Api) |
| Target Variable: | Pb |
| 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=0.1, max\_depth=4 | 0.0454 |
| **CatBoost** | iterations=1000, learning\_rate=0.1, depth=4, l2\_leaf\_reg=7 | 0.1244 |
| **Neural Network** | [128, 64, 32], epochs=150, batch\_size=32 | 3.7586 |
| **Stacking Ensemble** | Default base models + CatBoost final estimator | 0.6145 |
| **Extra Trees** | n\_estimators=200 | 0.0635 |
| **Deep Neural Network** | [256, 128, 64], epochs=150, batch\_size=16 | 4.2367 |

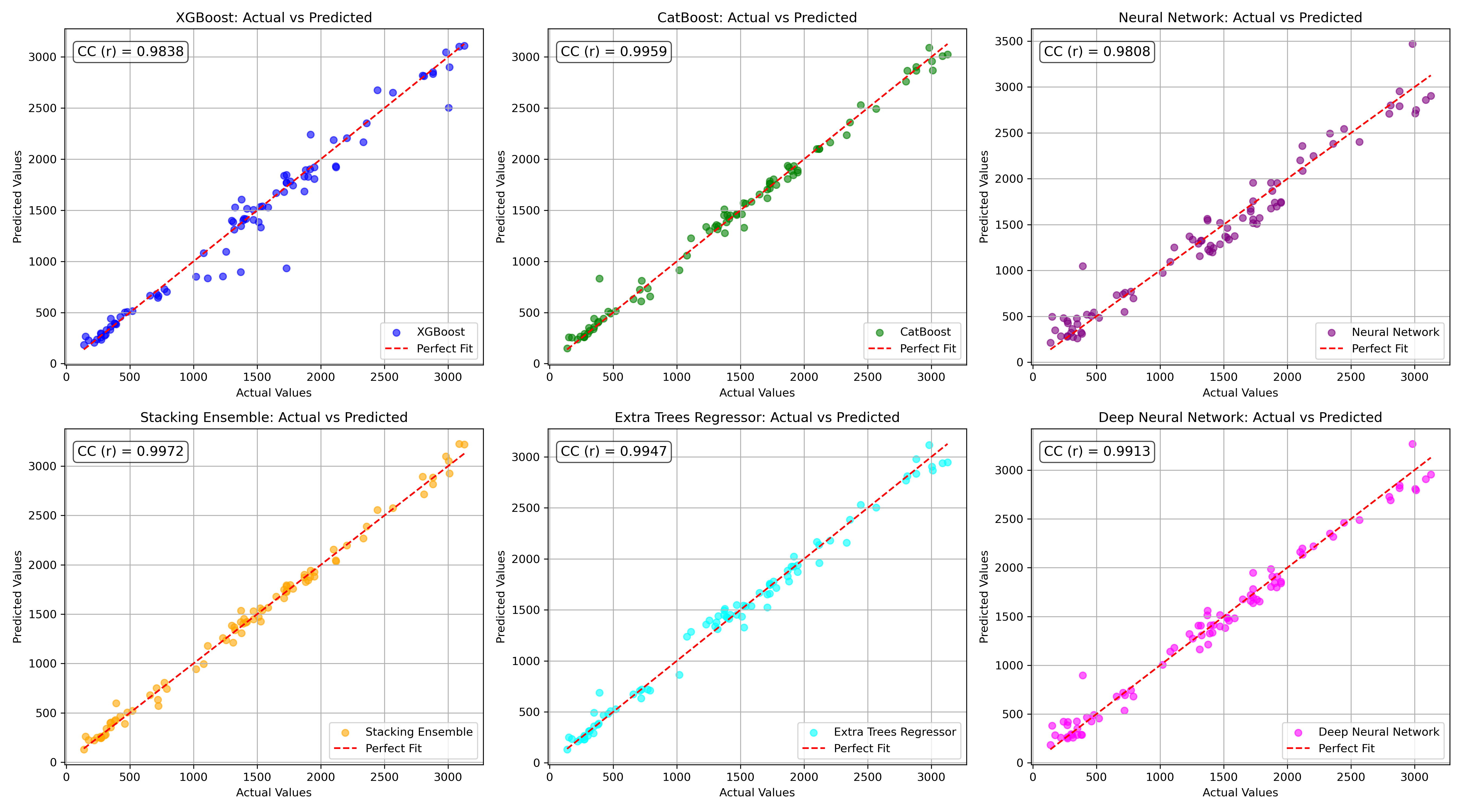
## 4. Evaluation Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | R² Score | Adjusted R² |
| **XGBoost** | 13935.197 | 118.047 | 70.132 | 0.9804 | 0.9794 |
| **CatBoost** | 3668.776 | 60.570 | 42.216 | 0.9948 | 0.9946 |
| **Neural Network** | 27735.372 | 166.539 | 128.185 | 0.9610 | 0.9590 |
| **Stacking Ensemble** | 4023.883 | 63.434 | 49.558 | 0.9943 | 0.9941 |
| **Extra Trees** | 7648.432 | 87.455 | 63.572 | 0.9892 | 0.9887 |
| **Deep Neural Network** | 8428.792 | 91.809 | 69.474 | 0.9881 | 0.9875 |

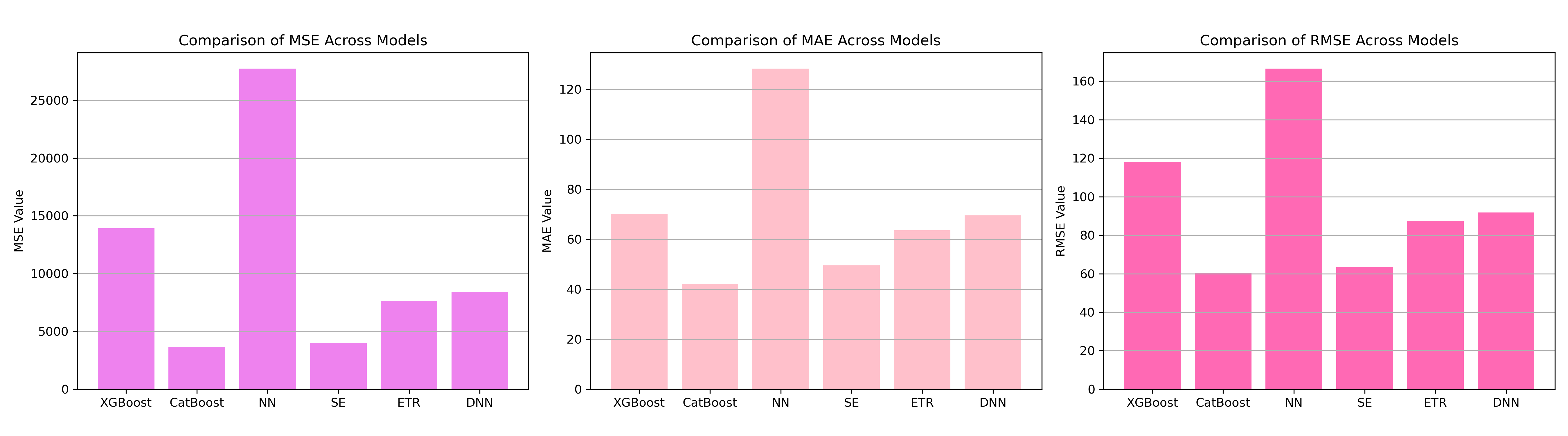
## 5. Cross-Validation Summary (5-Fold)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | RMSE Mean | RMSE Std | MAE Mean | MAE Std | R² Mean | R² Std |
| **CatBoost** | 74.068 | 28.261 | 46.221 | 15.143 | 0.9916 | 0.0058 |
| **Stacking Ensemble** | 81.157 | 18.042 | 54.367 | 8.836 | 0.9906 | 0.0038 |
| **Extra Trees Regressor** | 94.514 | 23.095 | 62.729 | 13.049 | 0.9872 | 0.0057 |
| **XGBoost** | 98.519 | 25.548 | 61.706 | 14.022 | 0.9860 | 0.0071 |
| **Deep Neural Network** | 405.133 | 64.305 | 282.550 | 44.065 | 0.7700 | 0.0643 |
| **Neural Network** | 457.439 | 79.216 | 328.972 | 61.720 | 0.7043 | 0.1021 |

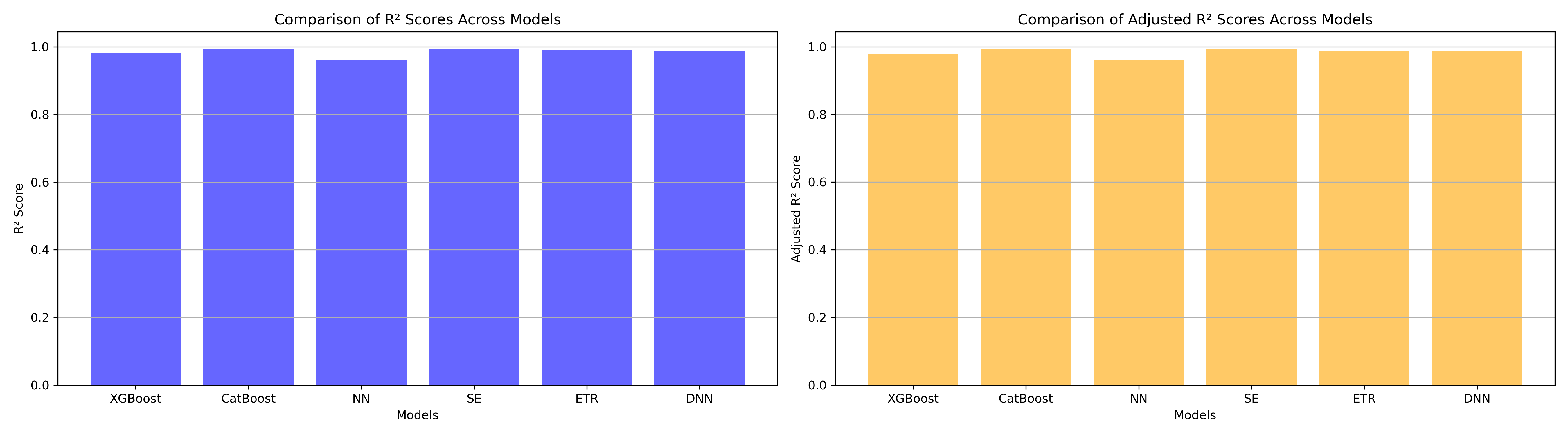
## 6. Visualizations



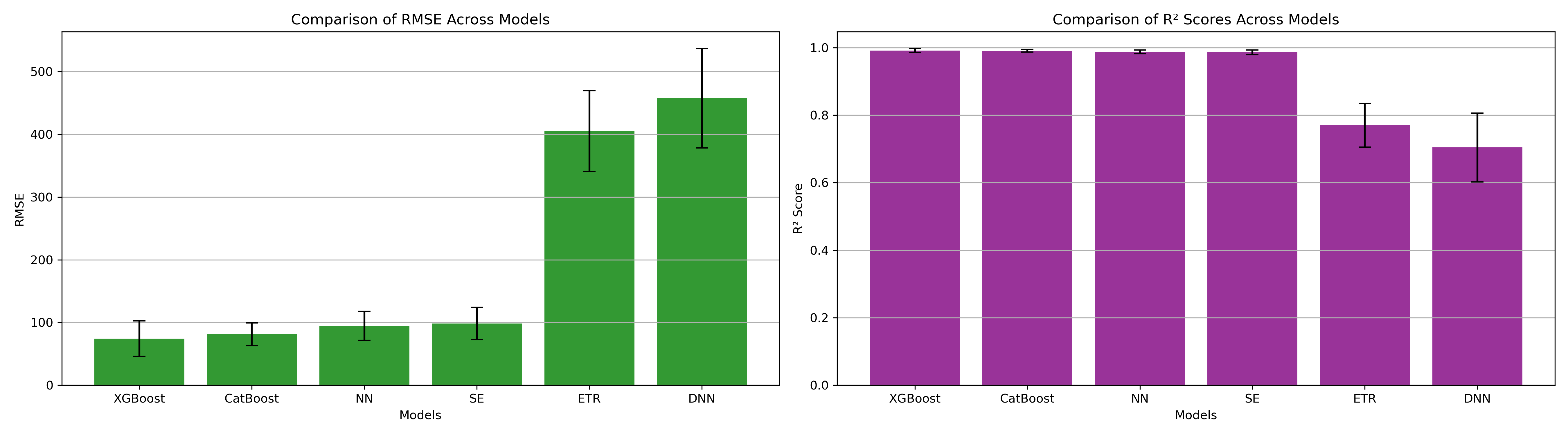
*Figure 1: Actual vs. Predicted Values for Pb (Well 283)*



*Figure 2: Bar Charts of MSE, RMSE, MAE for Pb (Well 283)*



*Figure 3: Bar Charts of R² and Adjusted R² for Pb (Well 283)*



*Figure 4: Error Bars for RMSE and R² from CV for Pb (Well 283)*

### 7. Observations & Next Steps

**Best Performing Model:** **CatBoost** (RMSE: 60.570, R²: 0.9948) closely followed by **Stacking Ensemble**.

* Extra Trees and XGBoost performed well but slightly lower in R².
* **Neural Networks** had significantly lower performance, with CV R² values around 0.70 (NN) and 0.77 (DNN).

### 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.