# 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: | 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 | 0.0434 |
| **CatBoost** | iterations=1000, learning\_rate=0.05, depth=4, l2\_leaf\_reg=7 | 0.1938 |
| **Neural Network** | [128, 64], epochs=100, batch\_size=16 | 3.1911 |
| **Stacking Ensemble** | Default base models + CatBoost final estimator | 0.7844 |
| **Extra Trees** | n\_estimators=200 | 0.1635 |
| **Deep Neural Network** | [256, 128, 64], epochs=100, batch\_size=32 | 2.3129 |

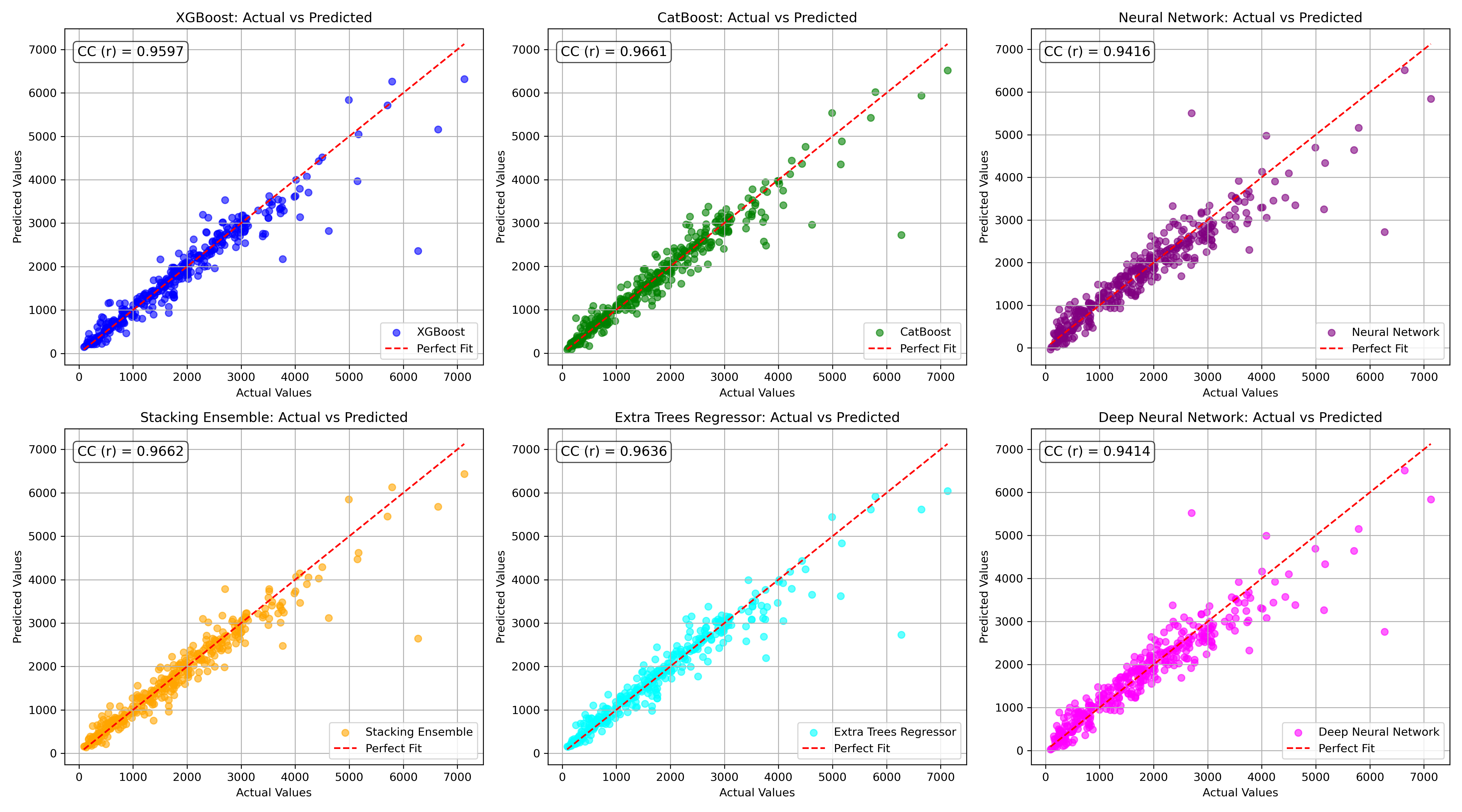
## 4. Evaluation Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | R² Score | Adjusted R² |
| **XGBoost** | 108293.491 | 329.080 | 178.893 | 0.9261 | 0.9253 |
| **CatBoost** | 94122.562 | 306.794 | 177.038 | 0.9358 | 0.9351 |
| **Neural Network** | 170444.040 | 412.849 | 263.698 | 0.8837 | 0.8825 |
| **Stacking Ensemble** | 101285.715 | 318.254 | 178.067 | 0.9309 | 0.9302 |
| **Extra Trees** | 108699.095 | 329.696 | 177.497 | 0.9259 | 0.9250 |
| **Deep Neural Network** | 162918.102 | 403.631 | 264.895 | 0.8889 | 0.8877 |

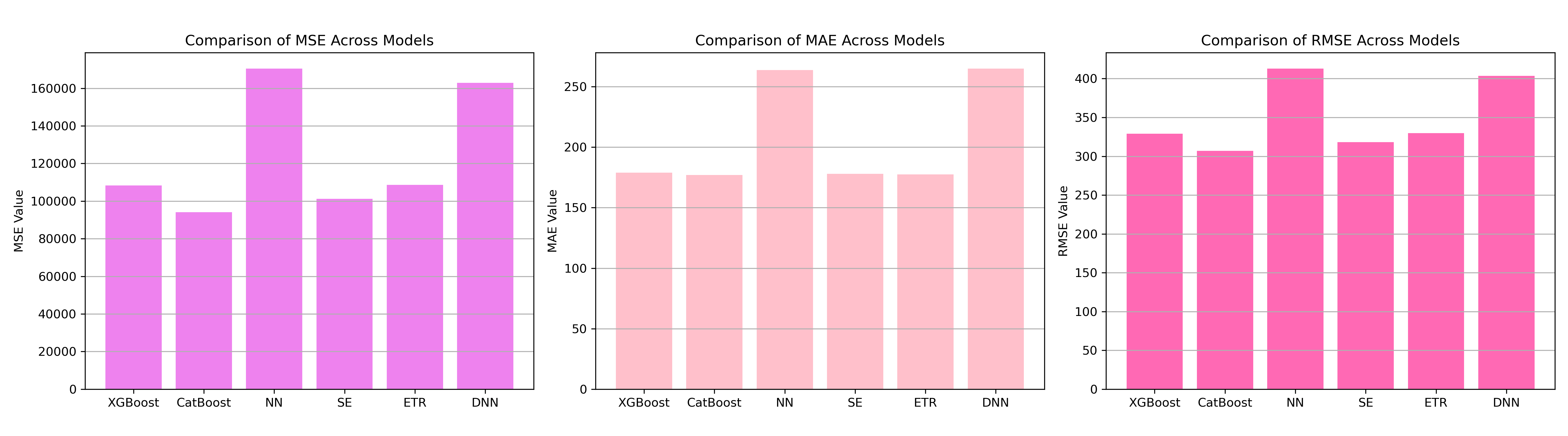
## 5. Cross-Validation Summary (5-Fold)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | RMSE Mean | RMSE Std | MAE Mean | MAE Std | R² Mean | R² Std |
| **Extra Trees Regressor** | 290.750 | 33.520 | 170.724 | 9.566 | 0.9407 | 0.0140 |
| **XGBoost** | 292.065 | 53.057 | 178.745 | 18.818 | 0.9389 | 0.0217 |
| **Stacking Ensemble** | 294.349 | 49.304 | 181.307 | 15.466 | 0.9381 | 0.0209 |
| **CatBoost** | 296.746 | 42.848 | 181.322 | 17.429 | 0.9380 | 0.0162 |
| **Deep Neural Network** | 496.948 | 26.444 | 358.732 | 13.780 | 0.8294 | 0.0132 |
| **Neural Network** | 499.334 | 25.269 | 362.252 | 12.390 | 0.8277 | 0.0132 |

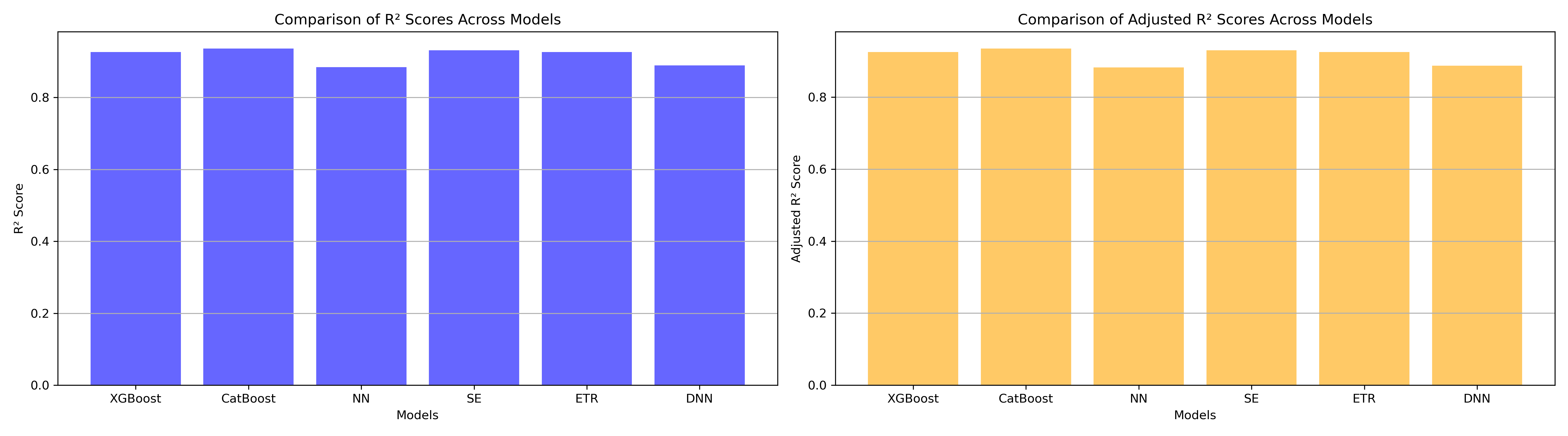
## 6. Visualizations



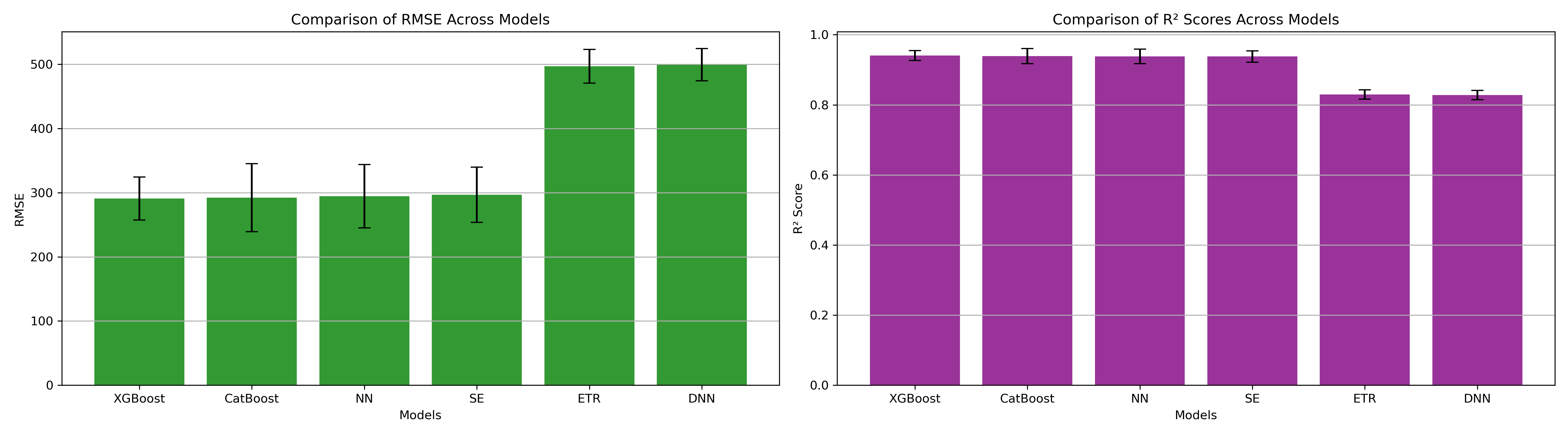
*Figure 1: Actual vs. Predicted Values for Pb (Well 782)*



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



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



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

### 7. Observations & Next Steps

**Best Performing Models:** **Extra Trees Regressor** and **XGBoost** with RMSE around ~291–293 and R² above 0.94.

* Stacking and CatBoost closely followed in performance.
* Neural Networks (both NN and DNN) were significantly worse:
  + RMSE values exceeded 400.
  + R² values dropped below 0.89.
* Overall, ensemble-based tree models proved more reliable for larger structured datasets like 1225 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.