# 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=100, learning\_rate=0.05, max\_depth=4 | 0.0415 |
| **CatBoost** | iterations=1000, learning\_rate=0.01, depth=6, l2\_leaf\_reg=3 | 0.3064 |
| **Neural Network** | [128, 64, 32], epochs=150, batch\_size=32 | 3.1060 |
| **Stacking Ensemble** | Default base models + CatBoost final estimator | 0.8214 |
| **Extra Trees** | n\_estimators=200 | 0.1158 |
| **Deep Neural Network** | [256, 128, 64], epochs=100, batch\_size=16 | 3.7269 |

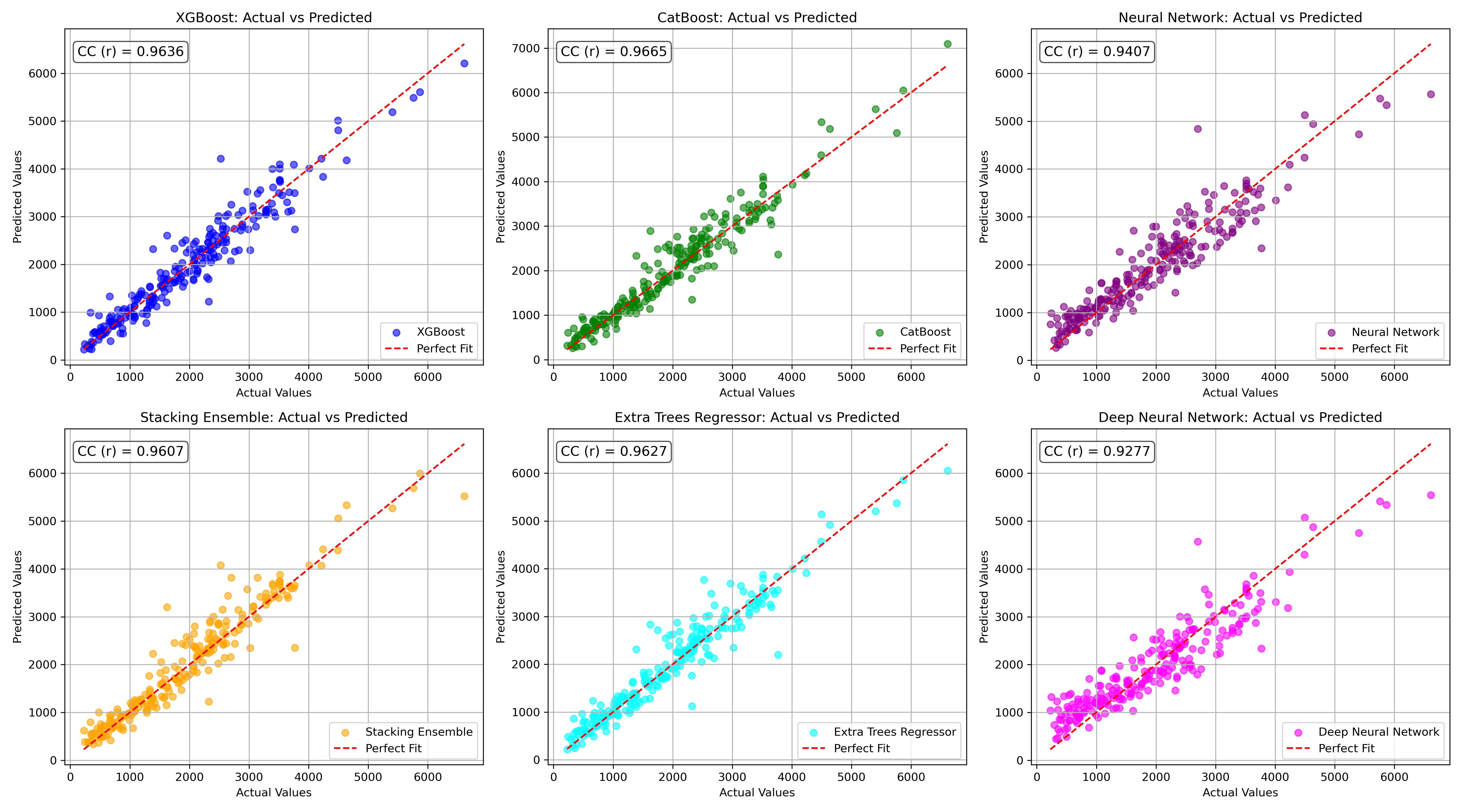
## 4. Evaluation Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | R² Score | Adjusted R² |
| **XGBoost** | 88242.176 | 297.056 | 207.900 | 0.9326 | 0.9315 |
| **CatBoost** | 89964.778 | 299.941 | 214.514 | 0.9313 | 0.9301 |
| **Neural Network** | 152226.257 | 390.162 | 292.070 | 0.8838 | 0.8817 |
| **Stacking Ensemble** | 109131.806 | 330.351 | 217.044 | 0.9167 | 0.9152 |
| **Extra Trees** | 98353.398 | 313.614 | 210.681 | 0.9249 | 0.9236 |
| **Deep Neural Network** | 142519.201 | 377.517 | 290.717 | 0.8912 | 0.8893 |

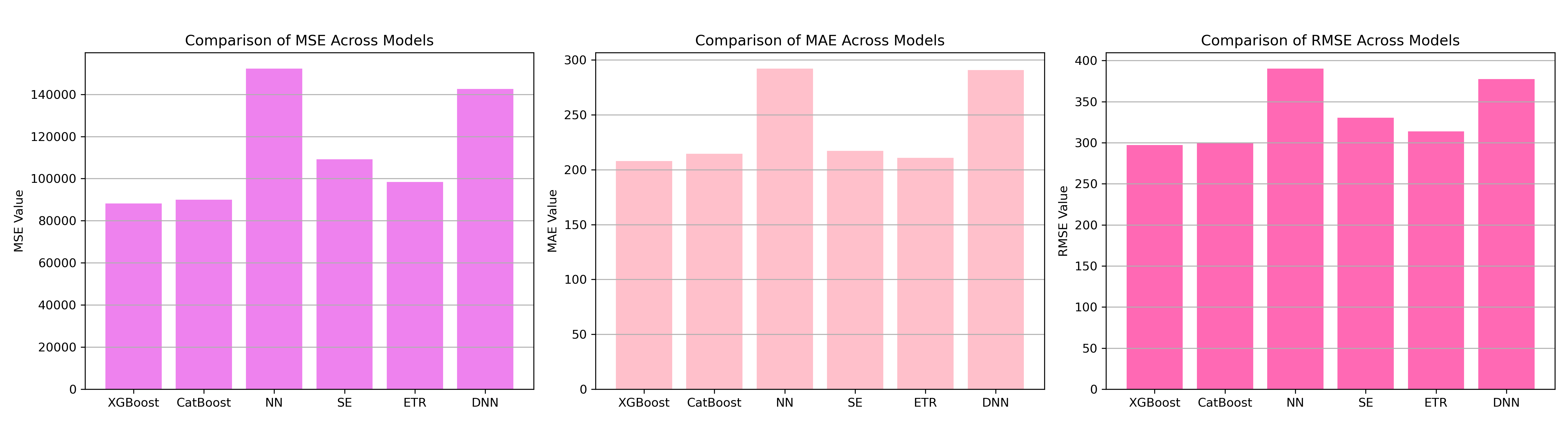
## 5. Cross-Validation Summary (5-Fold)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | RMSE Mean | RMSE Std | MAE Mean | MAE Std | R² Mean | R² Std |
| **CatBoost** | 348.176 | 70.255 | 234.746 | 25.258 | 0.9249 | 0.0186 |
| **Extra Trees Regressor** | 351.781 | 75.700 | 227.877 | 32.070 | 0.9235 | 0.0204 |
| **Stacking Ensemble** | 352.655 | 67.819 | 232.352 | 27.937 | 0.9216 | 0.0218 |
| **XGBoost** | 366.470 | 80.905 | 246.418 | 37.622 | 0.9165 | 0.0243 |
| **Deep Neural Network** | 528.849 | 50.928 | 375.089 | 24.855 | 0.8269 | 0.0157 |
| **Neural Network** | 531.050 | 52.696 | 381.294 | 20.825 | 0.8258 | 0.0135 |

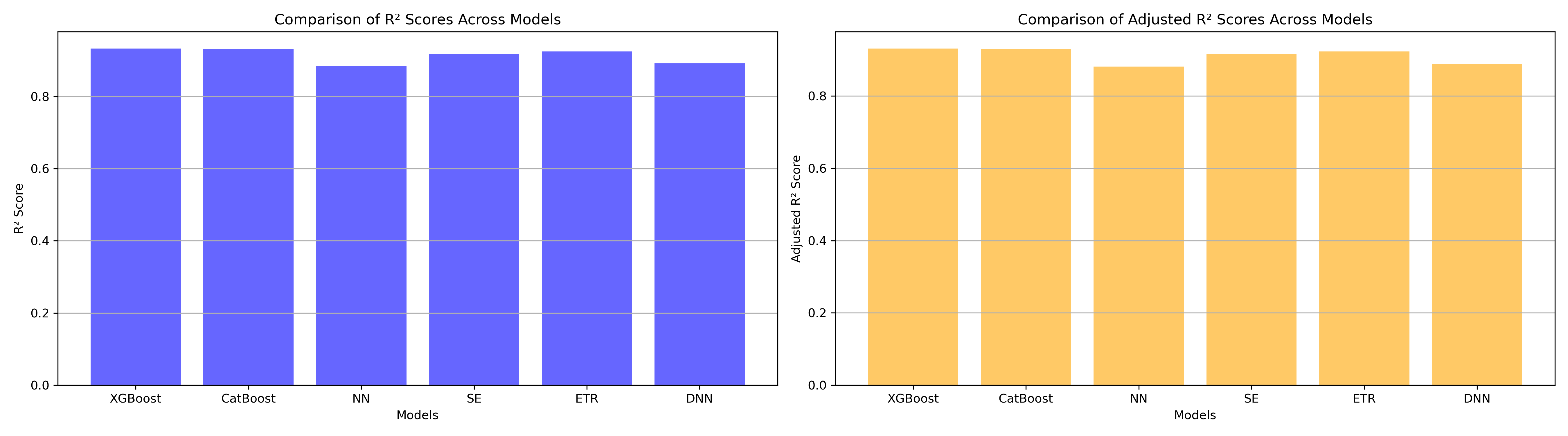
## 6. Visualizations



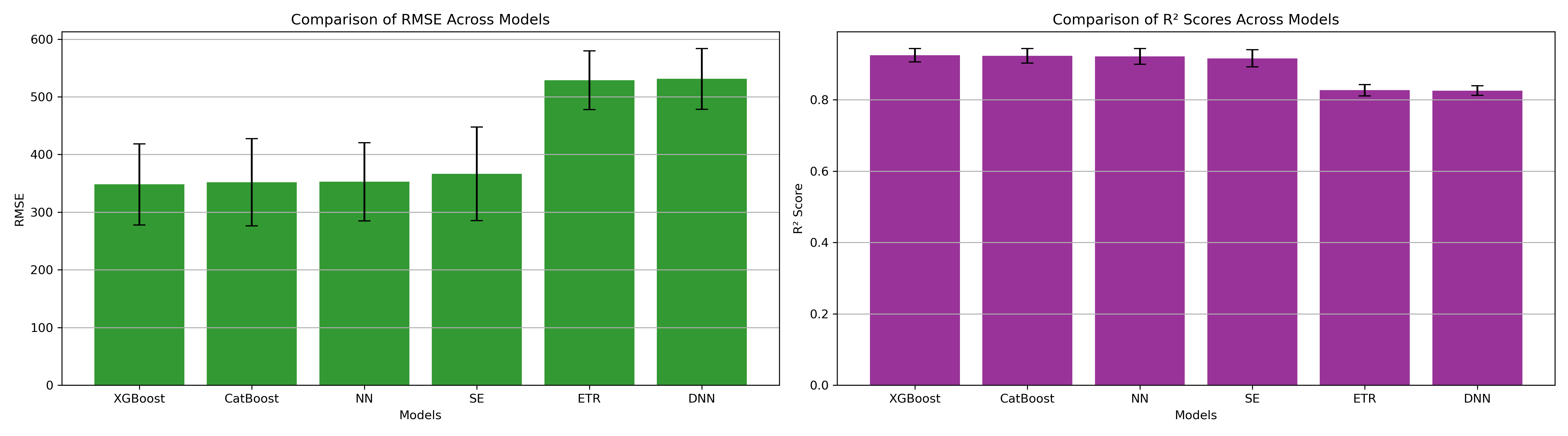
*Figure 1: Actual vs. Predicted Values for Pb (Well 782)*



*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 Model:** **CatBoost** (RMSE: ~299, R²: ~0.9313)

* Extra Trees and Stacking Ensemble models also showed good performance but slightly lower R².
* Neural networks (both NN and DNN) performed poorly with R² around ~0.83 during cross-validation.
* Ensemble-based models clearly outperform deep learning architectures on this dataset.

### 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.