Parkinson's Disease Prediction

By: Utkarsh Singh,

Project Overview

- Predicting MDS-UPDRS scores to measure Parkinson's Disease Progression
- Goal: "Develop ML models using protein/peptide data to predict UPDRS score (1-4)"
- Data used from Kaggle dataset competition:
  - https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction

Task

Primary Task: Regression (predict continuous UPDRS score)

Secondary Task: Classification (predict binned UPDRS categories)

Predict all 4 UPDRS scores together:

- UPDRS 1: Mental Behavior & Mood
- UPDRS 2: Daily Activities
- UPDRS 3: Motor Symptoms
- UPDRS 4: Treatment Side Effects

Data

- Original dataset consists of separate datasets: Clinical, Supplemental Clinical, Peptides, Proteins
- Merged into one single dataset
- Size: 4,838 entries, 10 columns
- Features:
  - Categorical: upd23b_clinical_state_on_medication
  - Continuous: PeptideAbundance, NPX, visit_month
- Pre-processing: “Duplicates removed, filled all NaNs with 0 and aggregated means”

Exploratory Data Analysis

- Analyzed data distribution and relationships to determine appropriate binning thresholds
- Preprocessed data by encoding categorical variables and scaling features to enhance model performance

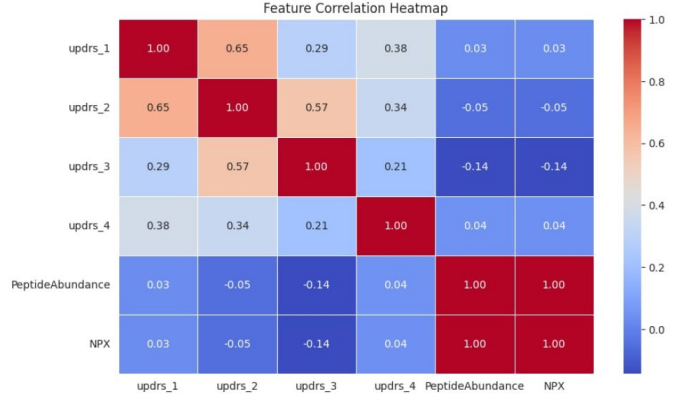


Figure 3. Feature Correlation Heatmap Matrix

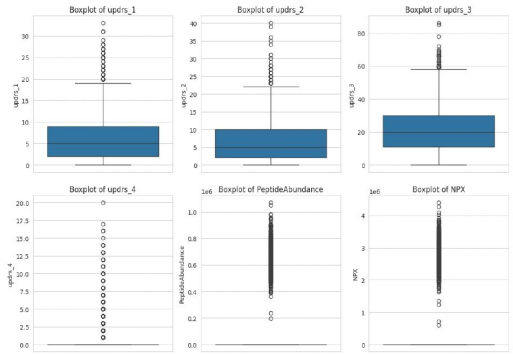
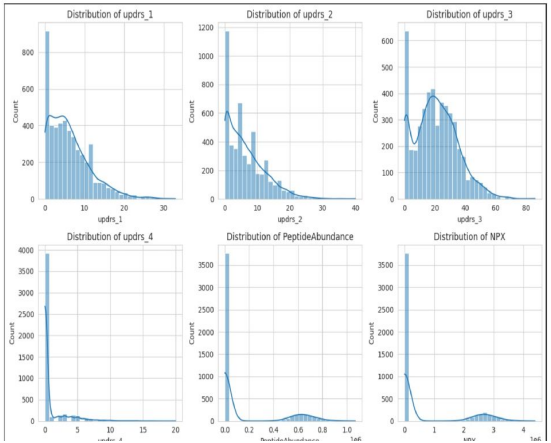


Figure 2. Boxplot the visualise and identify outliers in the data



Evaluation Metrics

Regression Metrics:

- MAE: Averages error magnitude; shows average deviation of prediction from actual UPDRS scores
- MSE/RMSE: Averages squared differences where lower = closer predictions and higher = large errors; RMSE penalises larger errors showing prediction error variance
- SMAPE: Normalizes errors relative to actual MDS-UPDRS
- R^2: Explains how well model explains variance in MDS-UPDRS

Classification Metrics:

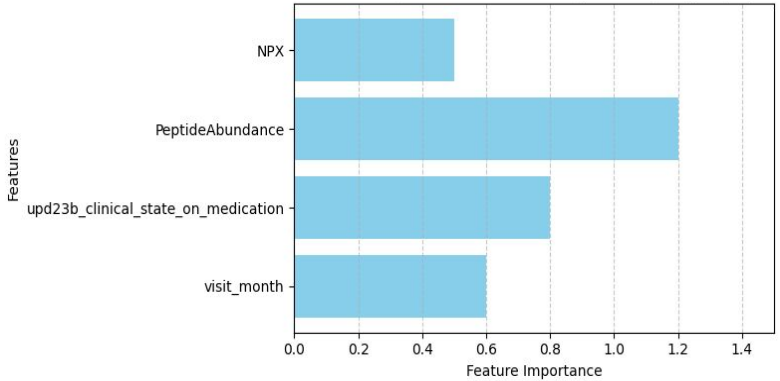
- Precision, Recall, F1-score

Initial Baseline - Linear Regression

- Multi output linear regression
- Features: PeptideAbundance, NPX, upd23b_clinical_state_on_medication
- Results (Single Train Test Split):

Evaluation Metrics for Linear Regression (Single Train-Test Split):

Metric	Value
MAE	5.16
MSE	61.46
RMSE	7.84
R^2	0.056
SMAPE (%)	99.61



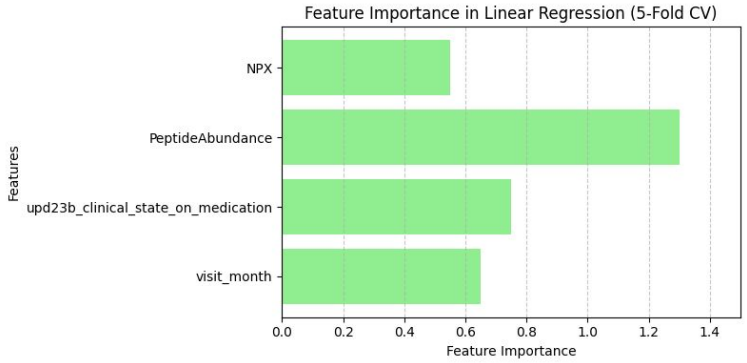
Feature-Engineered Linear Regression

- Enhanced Linear Regression with Feature Engineering
- New Features:
  - Interaction: PeptideAbundance * NPX
  - Polynomial: $PeptideAbundance^2$, NPX^2
  - Time: visit_month, $visit\_month^2$, log(visit_month + 1)
  - Scaling: Standardized features
- Results: 5 fold CV

```
Evaluation Metrics for Linear Regression (5-Fold CV):

+-----------+--------+
|  Metric   | Value  |
+-----------+--------+
|    MAE    |  5.0   |
|    MSE    |  58.1  |
|   RMSE    |  7.62  |
|    R²     |  0.1   |
| SMAPE (%) | 100.08 |
+-----------+--------+

Best Fold: Fold 4 (Global SMAPE: 97.89%)
```



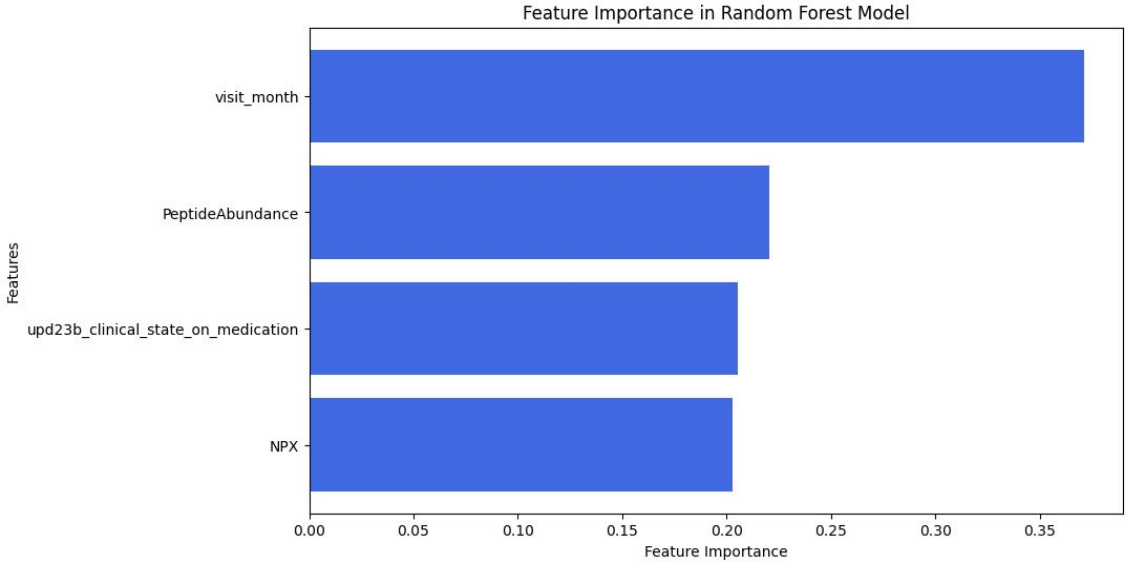
Random Forest Regressor

Evaluation Metrics:

- Total MAE: 4.7
- Total MSE: 53.54
- Total RMSE: 6.20
- Total R^2 Score: 0.13

Features:

- Visit Month
- Peptide Abundance
- Clinical State on Medication
- NPX



K-Nearest Neighbors (KNN) Regressor

- Approach:
  - Standardized features
  - Grid search for optimal k = 20
- Features:
  - visit_month
  - upd23b_clinical_state_on_medication (encoded numerically)
  - PeptideAbundance
  - NPX

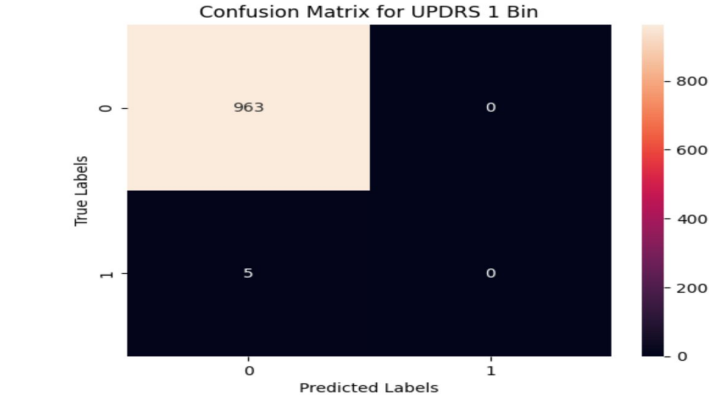
Evaluation Metrics:

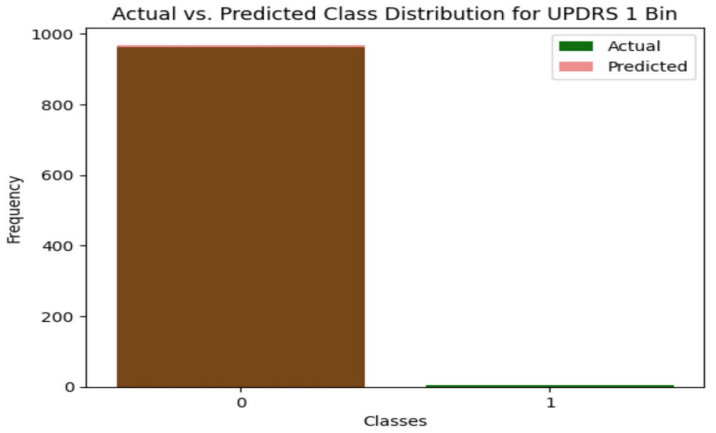
- Best k: 20
- MAE: 4.732760847107438
- MSE: 52.78826510847104
- RMSE: 7.265553324315433
- R^2 Score: 0.13034528382740965
- SMAPE (%): 89.0646

KNN is not effective for predicting UPDRS scores in this dataset.

Likely due to complex and non-linear relationships in the data.

Classification





- Implemented a RandomForestClassifier to categorize UPDRS scores into severity levels: Mild, Moderate, Severe, based on clinical thresholds
- This approach aids in quick assessment and decision-making in clinical settings
- Our classification model complements the regression models by providing quick, categorical insights into disease severity, enhancing clinical utility

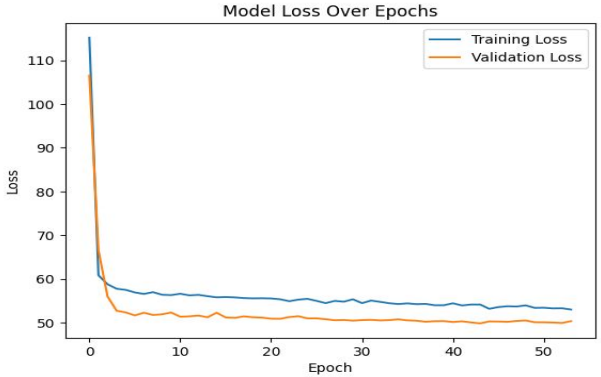
Neural Network

- Overview: Applied Feed Forward Neural Network with Batch Normalization to predict continuous UPDRS scores (1-4)
- Architecture:
  - Layers: 128 -> 64 -> 32 neurons, with ReLU activation, and 20% dropout for regularization
  - Batch Normalization: Added after each dense layer to stabilize training
  - Output: 4 neurons for UPDRS 1-4
- Training -> Optimizer: Adam, Loss, MSE, Epoch:50, Batch Size: 32, Validation Split: 20%
- Performance: R^2:0.166,MAE:4.64, depicting significant explanatory power over baseline.

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 256)	1,280
batch_normalization (BatchNormalization)	(None, 256)	1,024
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32,896
batch_normalization_1 (BatchNormalization)	(None, 128)	512
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8,256
batch_normalization_2 (BatchNormalization)	(None, 64)	256
dropout_2 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 32)	2,080
dense_4 (Dense)	(None, 4)	132

Total params: 46,436 (181.39 KB)
Trainable params: 45,540 (177.89 KB)
Non-trainable params: 896 (3.50 KB)



Model Training and Results

- Classification Training
  - Evaluated the model using precision, recall, and F1-score, essential for understanding model performance in imbalanced datasets
  - Confusion Matrices: help us visualize the accuracy of our predictions across different severity levels
  - Bar charts: these charts compare the actual vs predicted class distributions
- Regression Training
  - Trained regression models to predict continuous UPDRS (1-4)
  - Models Evaluated
    - Linear Regression (Baseline and Feature Engineered)
    - Random Forest Regressor
    - KNN Regressor
    - Feed Forward Neural Network (as of right now)
  - Evaluated model using MAE,MSE,RMSE,R^2,SMAPE

Result Tables

Training classifier for: updrs_1_bin
Classification Report for updrs_1_bin:

	precision	recall	f1-score	support
0	0.83	0.97	0.89	793
1	0.34	0.10	0.16	159
2	0.00	0.00	0.00	16
accuracy			0.81	968
macro avg	0.39	0.36	0.35	968
weighted avg	0.74	0.81	0.76	968

Classification Report for updrs_2_bin:

	precision	recall	f1-score	support
0	0.77	0.97	0.86	740
1	0.24	0.05	0.08	201
2	0.00	0.00	0.00	26
3	0.00	0.00	0.00	1
accuracy			0.75	968
macro avg	0.25	0.25	0.24	968
weighted avg	0.64	0.75	0.67	968

Classification Report for updrs_3_bin:

	precision	recall	f1-score	support
0	0.58	0.55	0.56	231
1	0.35	0.68	0.46	262
2	0.26	0.15	0.19	239
3	0.17	0.06	0.08	158
4	0.27	0.17	0.20	78
accuracy			0.37	968
macro avg	0.32	0.32	0.30	968
weighted avg	0.35	0.37	0.34	968

Training classifier for: updrs_4_bin
Classification Report for updrs_4_bin:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	963
1	0.00	0.00	0.00	5
accuracy			0.99	968
macro avg	0.50	0.50	0.50	968
weighted avg	0.99	0.99	0.99	968

Accuracy Score Table for Regressor Models:

Model	MAE	MSE	RMSE	R^2	SMAPE (%)
Linear Regression	5.16	61.46	7.84	0.056	99.61
Feat-Eng. Linear (5-Fold CV)	5.0	58.1	6.8	0.1	97.89
Random Forest	4.7	53.54	6.2	0.13	97.66
KNN Regressor	4.73	52.78	7.27	0.13	89.06
Neural Network	4.64	52.05	7.21	0.166	98.93

Proposed Solution (Best Model)

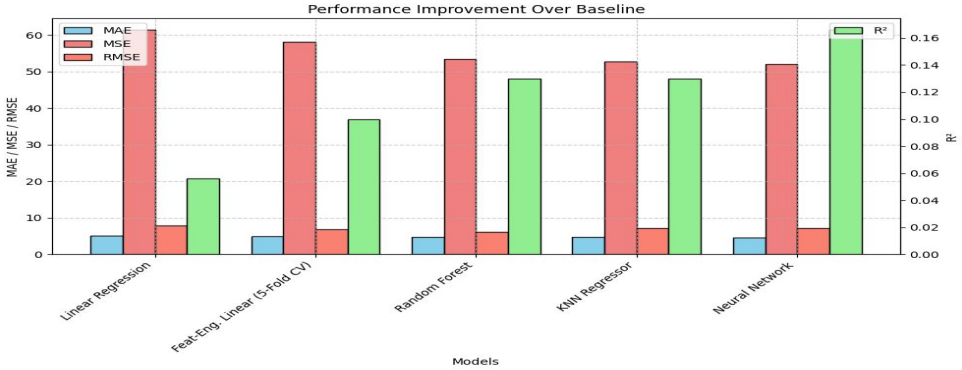
- Best Model: Neural Network with Batch Normalization
  - Reason: Achieves highest R^2(0.166) beating baseline model and all other models
  - NN Metrics: MAE (4.64), RMSE: 7.21, showcasing lowest error among regressors
- Complementary Classifier: Random Forest Classifier
  - Role: Categorizes UPDRS scores into severity levels (Mild, Moderate, Severe) for quick clinical insights
- Why preferring these 2?
  - NN captures complex non linear relationship in protein/peptide data that is ideal for UPDRS prediction
  - Combined regression and classification provide both precise scores and actionable severity categories for clinical use

Improvement over baseline

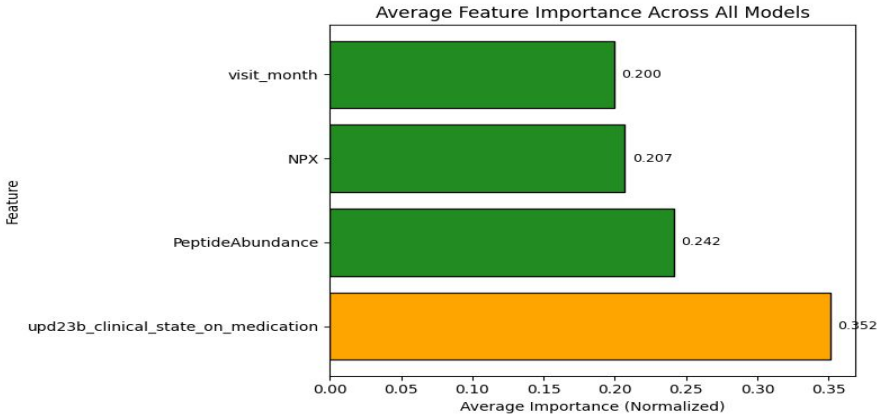
Reduced Errors: Enhanced models lower MAE from 5.16 (Linear Regression) to 4.64 (Neural Network) and RMSE from 7.84 to 6.20 (Random Forest).

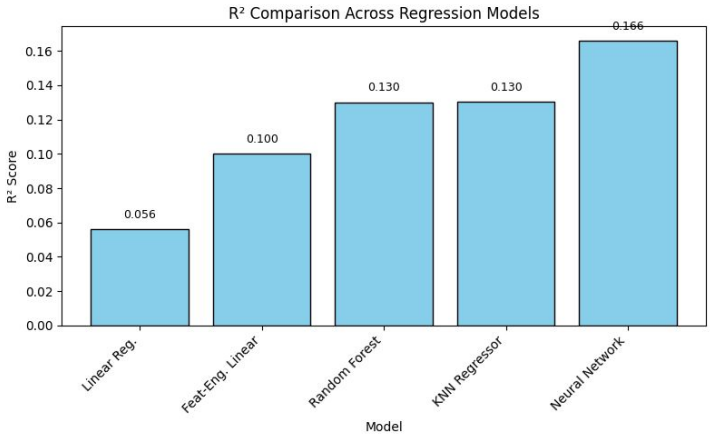
Better Variance Fit: R^2 improves from 0.056 (baseline) to 0.166 (Neural Network), a nearly threefold increase.

Robust Enhancements: Feature engineering and advanced models (e.g., Neural Network, Random Forest) consistently outperform the baseline.



Insights - Visualization





Neural Network Leads: Highest R^2 (0.166), showing the best fit among models.
Baseline Outperformed: Linear Regression's R^2 (0.056) is surpassed by all models, with Random Forest and KNN at 0.13.
Medication feature dominates: Has highest importance (0.352) showing its essential role in Parkinson's progression

Key Takeaways

- **Medical Status Importance:** Since, upd23b_clinical_state_on_medication (Medical Status) rose as the most important predictor across all models, this means it is critical in determining Parkinson's disease progression
- **Best Model:** Neural Network outperforms every model with highest R^2 (0.166) depicting best explanatory power among all models
- **Error reduction across models:** Enhanced models have reduced errors compared to the initial baseline with MAE dropping from 5.16 to 4.64 and SMAPE% from 99.61% to 89.06% showing better predictive accuracy.
- **Challenges and future take:** High SMAPE value (98.3% for NN) and class imbalance suggest areas of improvement, future consideration would be addressing scaling issues and data imbalance with technique like SMOTE

References

Works Cited

"AMP®-Parkinson's Disease Progression Prediction." *Kaggle*, www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction. Accessed 4 Feb. 2025.

Murphy, Kevin P. *Probabilistic Machine Learning: An Introduction*. The MIT Press, 2022.