Parkinson's Disease Analysis: A Multifaceted Data Science Approach

Team: Min Shi, Sayantan Roy Chowdhury, E.G.K.M.Gamlath

Github: https://github.com/blackhole-hope123/Parkinson-Disease-Analysis

Introduction: Parkinson's Disease (PD) is a progressive neurodegenerative disorder that impairs movement, cognition, sleep, and other essential functions. Over the past 25 years, the global prevalence of Parkinson's disease has more than doubled, and the associated mortality risk has also increased significantly. This growing burden underscores the critical need for improved methods to understand and manage disease progression.

Datasets: We used a longitudinal dataset from the Accelerating Medicines Partnership® Parkinson's Disease (AMP® PD), Version 4.0, obtained through by request from their knowledge platform. AMP PD is a large-scale, collaborative effort that brings together clinical, genetic, and biospecimen data from four major cohort studies: BioFIND, PPMI, PDBP, and HBS.

Our analysis is based on the Unified Parkinson's Disease Rating Scale (UPDRS) scores, which assess three key areas: cognition, self-care ability, and motor function.

Goals:

I. Predicting the Parkinson's disease progression.
II. Identify the distinct patterns of how the Parkinson's disease progresses.
III. What factors influence the time to freezing of gait.

Preliminary EDA: (These are the reasons why we choose the method in goal I) Our initial exploratory data analysis focused on understanding the relationship between demographic, clinical, imaging, and biospecimen variables with Parkinson's disease progression, as measured by UPDRS scores.

The main points we observed from EDA are:

1. The correlation factors of existing features with targets are very low
2. The target does not display a clear seasonality or linear trend
3. The progression of the target variables exhibited severe heterogeneity, and the distributions of the target variables are very skewed.
4. The target variables have internal correlations.

Goal I: To make a Prediction of Parkinson's disease progression.

Stakeholders: Parkinson patients and healthcare providers

Feature variables: static and time dependent clinical results, the demographic, family history, brain scan and biospeciman data, in total about 1400 columns.

Target Variables: UPDRS scores, more specifically part I, II, III of the UPDRS scores.

KPI: MAE: Mean Absolute Error in the predicted value and true test values

Methods:

1. **Data preparation:** We noticed that there are different participants ids in the table for UPDRS scores. We deleted such participant ids in fear that they may be split into train and test set separately, causing data leakage. The final data set consists of over 20000 observations from 4000 participants, with about 1400 features.
2. **Train Test Split & Data Preprocessing:** We used GroupKFold for the train test split, to ensure that a participant can only appear in the training set or test set to avoid possible data leakage. We preprocess by grouping some rare categories together and then apply one-hot encoding.
3. **Model Selection:** The non-linear trend, heterogeneity and highly-skewed distribution of the target variables implies that linear models are not reliable here, so we decide to start with tree models.
4. **Lagged Feature Creation:** From the weak predictive power of explicit features, it is most possible that past values of the target variables contain hidden signals, therefore we create lagged features.
5. **Imputation**: Lagged features will always have missing values, but to avoid data leakage, we cannot access future values so we cannot impute with "mean" or "median". We chose to impute with filling by 0 first and then use Multiple Imputation by Chained Equations (MICE) on the lagged updrs features to exploit their deep correlations.
6. **Biospecimen Data Exclusion:** To prevent curse of dimensionality, all the above methods are implemented without including the biospecimen data, which accounts for over 1300 features. We did a permutation test to check the signal strength in the biospecimen data, with the null Hypothesis being that the biospecimen data contain little signal. The result is a p-value of 0.46, so we fail to reject the null hypothesis and exclude the biospecimen data.
7. **Feature Engineering & Hyperparameter tuning:** The feature importance dataframe and plot shows that almost all the important features are time aware features. So we engineered more lagged features, created slope and rolling window features to better capture the change of target variables over time. Finally, we selected top 75 feature as the final feature set, followed by a hyperparameter tuning.
8. **Final Evaluation:** One thing we need to treat carefully here is how we should do one hot encoding for the test set as we did in the preprocessing for the training data. To avoid data leakage, we define a function which can return both the encoded dataframe and a list of rare categories. Then we grouped categories in test data by the learned list of rare categories from the training data. This ensures that the model does not access any information from the test set when learning, maintaining a strict separation between training and evaluation. We also created lagged variables and did the same feature engineering to the test set.

Results: The tuned LightGBM model achieved a MAE of 3.53 on average, representing a 49.4% improvement from the strongest baseline model. The feature importance dataframe again confirmed the importance of time-aware features. Moreover, the final test metrics are within one standard deviation of the Cross-Validation scores, showing that our model avoids overfitting and learned some meaningful patterns from the dataset.

Goal II :Clustering and Medication Comparisons

We investigate distinct patterns of disease progression by clustering trajectories of UPDRS summary scores I, II, and III from baseline to month 24 for patients with at least three visits, resulting in a dataset of 1,055 patients. From each trajectory, we extract seven statistical and temporal features per UPDRS score, creating a 1055×21 feature matrix. Four clustering algorithms—K-Means, Agglomerative Clustering, Gaussian Mixture Models (GMM), and DBSCAN—were evaluated using a combination of Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Score. K-Means with two clusters performed best.

Cluster 0 (429 patients) and Cluster 1 (629 patients) showed distinct progression patterns: UPDRS I and II scores increased faster in Cluster 0, while UPDRS III scores rose more steeply in Cluster 1. Biologically, patients in Cluster 0 were more likely to be on Levodopa (85.4% vs. 79.7%), whereas those in Cluster 1 had higher usage of dopamine agonists (58.9% vs. 48.2%) and other Parkinson's medications (74.9% vs. 64.5%). These differences were statistically significant ($p < 0.001$). Future work will explore additional biological markers to better understand this two-cluster structure.

Goal III : Freezing Time Analysis

Freezing of gait, a common Parkinson's symptom causing difficulty in starting or continuing to walk, is defined here as a non-zero response to UPDRS Part II Question 13 or Part III Question 11. To investigate factors influencing its onset, we apply a Cox proportional hazards model on 3255 patients using baseline features. The data were split evenly by patient into training and test sets to avoid data leakage. Fitting the Cox model on the training set identified 17 covariates as statistically significant predictors of hazard ($p < 0.05$) that also satisfied the proportional hazards assumption ($p > 0.05$ in the PH test). Applying the model to the test set with these 17 covariates reduced the set to 11 that met both criteria.

The most significant predictor in both datasets was whether the patient was on Parkinson's medication. With an extremely small p-value (1.91×10^{-11}), patients on medication had approximately a 50% higher risk of experiencing freezing of gait compared to those not on medication, controlling for other factors.