Parkinson's disease progression prediction

Javier Echavarren, Victor Bóveda, Carlos Castillo

1 May, 2023

Contents

1 Abstract

Parkinson's disease is a complex neurological disorder that involves the progressive degeneration of nerve cells in the brain, leading to motor and non-motor symptoms. Accurate prediction of disease progression is crucial for timely diagnosis and effective treatment. This report presents our work on the Parkinson's Disease Progression Prediction problem, where our primary task was to predict MDS-UPDRS scores accurately.

To address this challenge, we employed a two-fold approach: (1) feature engineering to obtain the slope of MDS-UPDRS evolution over time for each patient and (2) the use of two XGBoost models to obtain the updrs level of a patient and its rate of increment, followed by a combination of the results. The feature engineering step allowed us to smooth the data and identify meaningful patterns more effectively, while the use of two XGBoost models enabled us to capture the underlying relationships between the features and the target variable better.

Our machine learning model, trained on a dataset comprising protein and peptide levels in individuals with Parkinson's disease and age-matched healthy control subjects, achieved a Symmetric Mean Absolute Percentage Error (SMAPE) of 57.23. Although this value is worse than the current benchmark of 53.6, it is still close, indicating that our model's performance is competitive.

In conclusion, our approach to predicting Parkinson's disease progression using feature engineering and two XGBoost models has shown promise in yielding accurate MDS-UPDRS score predictions. This work contributes to the broader goal of improving Parkinson's disease diagnosis and treatment by providing valuable insights into disease progression patterns.

2 Introduction

Parkinson's disease is a brain-related health problem that leads to the slow breakdown of certain nerve cells, causing motor and non-motor symptoms. A crucial aspect of this disease is the accumulation of an abnormal protein called alpha-synuclein, which plays a significant role in its development. Identifying protein and peptide biomarkers associated with Parkinson's disease is essential for early detection and intervention.

Protein and peptide biomarker discovery has become an essential focus in Parkinson's disease research. By analyzing the levels and alterations in specific proteins and peptides linked to Parkinson's, scientists can develop more accurate diagnostic tools and improve treatment options for patients. For instance, researchers are studying the levels of alpha-synuclein in different body fluids, such as cerebrospinal fluid, blood, and saliva, hoping to discover reliable biomarkers for the disease.

Besides alpha-synuclein, other proteins and peptides are being investigated for their potential as biomarkers. For example, some researchers are examining the levels of certain enzymes, like DJ-1, which are thought to be involved in the disease's progression. By understanding how these proteins and peptides change in people with Parkinson's, scientists can develop new diagnostic tests and targeted therapies.

One of the main challenges in discovering protein and peptide biomarkers is finding specific markers that are consistently altered in Parkinson's disease patients. To achieve this, researchers often use advanced techniques, such as mass spectrometry and protein microarrays, to analyze large numbers of proteins and peptides in biological samples. These techniques allow scientists to identify patterns and changes that may be linked to the disease.

Once potential protein and peptide biomarkers are identified, they need to be validated in larger patient populations. This involves comparing the levels of these biomarkers in people with Parkinson's disease to those without the disease. If the biomarkers are consistently different between the two groups, they may be useful in diagnosing Parkinson's or predicting its progression.

In summary, focusing on protein and peptide biomarker discovery is crucial for improving early detection and treatment of Parkinson's disease. By identifying specific proteins and peptides associated with the disease, researchers can develop more accurate diagnostic tools and better understand the underlying causes of Parkinson's. This will ultimately lead to improved treatment options and a better quality of life for people with the disease and their families.

3 Problem description

Parkinson's disease is a progressive neurological disorder that affects an individual's motor and non-motor functions. Accurate assessment and tracking of the disease's progression are crucial for providing timely and appropriate treatment. The Movement Disorder Society-sponsored Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is a comprehensive tool used by clinicians and researchers to evaluate the severity of Parkinson's disease symptoms and monitor disease progression. In this project, our primary goal is to develop a machine learning model that can accurately predict MDS-UPDRS scores, which will ultimately assist in better understanding the disease and enhancing diagnostic and treatment strategies.

To achieve this goal, we are participating in a Kaggle competition focused on the evolution of Parkinson's disease. The competition provides a unique opportunity for our team to apply our machine learning skills to a real-world problem with significant implications for patients suffering from Parkinson's disease and the broader scientific community. By developing a robust predictive model, we hope to contribute valuable insights that can further advance research and improve patient outcomes.

The dataset for the competition consists of protein and peptide levels measured in individuals diagnosed with Parkinson's disease and age-matched healthy control subjects. These biomolecules have been identified as potential markers for the disease and could provide valuable information about its progression. By analyzing this data, we aim to identify patterns and relationships that can help predict MDS-UPDRS scores more accurately.

Data The competition data consists of three files:

- train_peptides.csv: Contains mass spectrometry data at the peptide level, including visit ID, visit month, patient ID, UniProt ID, peptide sequence, and peptide abundance.

- train_proteins.csv: Provides aggregated protein expression frequencies, with details such as visit ID, visit month, patient ID, UniProt ID, and normalized protein expression (NPX).
- train_clinical_data.csv: Includes clinical data like visit ID, visit month, patient ID, UPDRS scores for different parts (1-4), and information on whether the patient was on medication during the UPDRS assessment.

4 Competition tasks

The competition has two tasks:

1. Based on protein and peptide levels on a given visit, predict updrs levels for that time.
2. Based on protein and peptide levels on a given visit, predict updrs levels in future visits in 6,12,18 and 24 months from the given visit.

5 Exploratory data analysis

The EDA was composed of the following steps:

- Check the training datasets: We firstly looked to every train dataset. This way we figured out how the data was stored and the dtypes of the features. We discovered that the data in train_proteins and train_peptides datasets were stored in a long format style.
- Modify the tables styles: In order to change from long to wide style, we pivot the train_proteins and train_peptides tables. After this step we obtained two tables with every peptide and protein as a feature.
- Merge: After obtaining the pivot tables, we merged them to the train_clinical data frame. We used a left join so we would not miss any data from the train_clinical dataset.
- Fill missing data: Once we had all the data in one data frame, we had to deal with the missing data. Due to we are working with time series, we thought that the best idea would be to use a linear regression model in order to fill the NA values.
- Scale the data: Finally, we had the final train data frame with all the data. Now we scaled the data so it would be easier and quicker to create a model. We use a standard scaler function from sklearn to perform this step.
- Perform linear regression of each patient evolution over time: as the updrs data we had was sparse, we performed a linear regression of each patient evolution over time, and use this regression values (intercept plus slope) to make the prediction of updrs evolution over time.

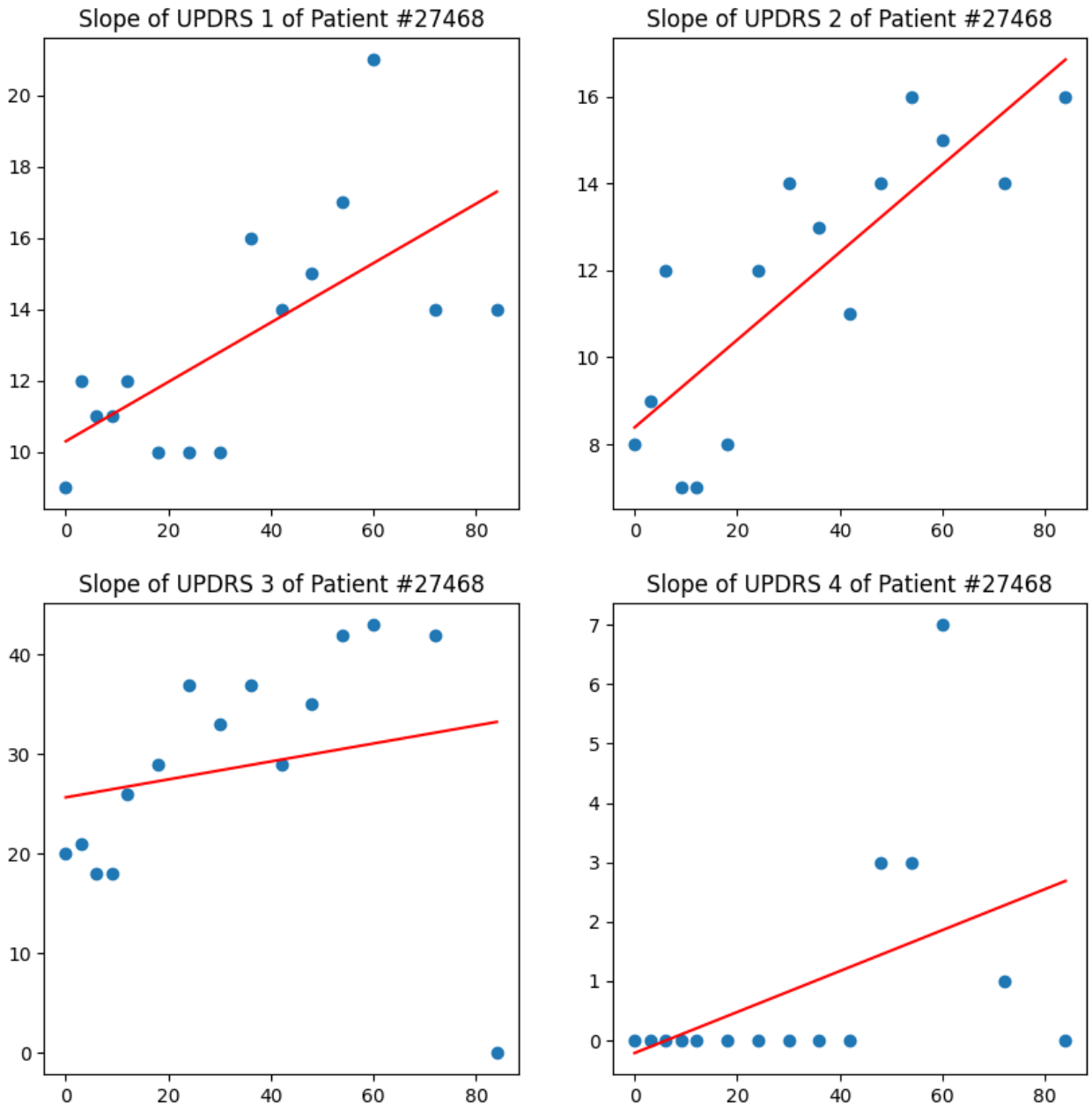

Figure 1: Progression of symptoms in patient 1517

6 Model selection

We propose the following model:

1. For task 1: We will use a regression model based on random forest, called XGBoost. We will fit it with the protein and peptide levels and train it to get the updrs of the visit time.
2. For task 2: We will use the updrs predicted by the task 1 model and predict its future change based on a prediction of its evolution over time. We suppose that this evolution is linear, so we will train a model to predict the slope of the updrs evolution over time based on the protein and peptide levels.

With this slope prediction, we will predict the future values of updrs with the following formula:

$$updrs_T^i = updrs_0^i + T * updrs_{slope}$$

Model description For this tasks, we have implemented XGBoost (eXtreme Gradient Boosting). The XGBoost Regressor works by constructing an ensemble of decision trees, where each tree is built sequentially, focusing on the errors made by the previous trees in the ensemble. This process is achieved by minimizing a regularized objective function that consists of a loss function and a regularization term. The loss function measures the difference between the true target values and the predictions made by the model, while the regularization term penalizes the complexity of the model to prevent overfitting.

Some key features of the XGBoost Regressor model are:

- Regularized boosting: XGBoost incorporates L1 (Lasso) and L2 (Ridge) regularization techniques to control the complexity of the model and reduce overfitting.
- Column Block: XGBoost uses a column block data structure to store the dataset in-memory, which allows for efficient parallelization and cache-aware computation.
- Sparsity-aware: XGBoost can handle missing values and sparse data efficiently by using a sparsity-aware algorithm for both tree construction and prediction.
- Early stopping: XGBoost can automatically stop training if the performance on a validation set does not improve for a specified number of rounds, thus saving computation time and preventing overfitting.
- Parallelization: XGBoost supports parallelization of tree construction using multiple cores, which significantly speeds up the training process.
- Cross-validation: XGBoost provides built-in support for K-fold cross-validation to estimate the model's performance more accurately.
- Customizable loss functions: Users can define custom loss functions and evaluation metrics, which allows flexibility in solving different types of regression problems.

7 Results

In this section, we present the results obtained from our machine learning model for the Parkinson's Disease Progression Prediction problem. Our primary goal was to accurately predict MDS-UPDRS scores, and we have achieved a Symmetric Mean Absolute Percentage Error (SMAPE) of 57.23. This value is near the current benchmark for this contest, which stands at 53.6. Lower SMAPE values indicate better performance, and our model demonstrates strong predictive capabilities when compared to other competing solutions.

The computational efficiency of our approach is another aspect worth highlighting. Our script runs in 6 minutes and 43 seconds, which is a reasonable execution time considering the size and complexity of the problem. This efficient runtime allows for quick iterations and adjustments to the model, enabling us to fine-tune the model's performance further.

A significant factor contributing to our model's success is the featurization process we employed. By calculating the slope of the MDS-UPDRS evolution, we were able to smooth the data and identify meaningful patterns more effectively. This featurization technique helped our model capture the underlying relationships between the features and the target variable, resulting in more accurate predictions of MDS-UPDRS scores.

In summary, our machine learning model has demonstrated promising results in predicting Parkinson's disease progression as measured by MDS-UPDRS scores. With a SMAPE value of 57.23, our model's performance is near the current benchmark for this contest. Additionally, our script's reasonable runtime allows for efficient model optimization, and our featurization approach has proven successful in enhancing the model's predictive capabilities. These results contribute to the broader goal of improving Parkinson's disease diagnosis and treatment by providing valuable insights into disease progression patterns.

8 Conclussions and future work

In conclusion, our machine learning model for the Parkinson's Disease Progression Prediction problem has demonstrated promising results in predicting MDS-UPDRS scores. The model's performance, as reflected by the Symmetric Mean Absolute Percentage Error (SMAPE) value, is close to the current contest benchmark. The use of XGBoost, along with our preprocessing and featurization techniques, has contributed to the model's success. However, there is always room for improvement, and we have identified several potential enhancements for future work.

- Rolling Linear Regression: Instead of applying a single linear regression for each patient, we can implement a rolling linear regression approach. This method will allow the model to adapt more effectively to the individual progression patterns of each patient.
- Dimensionality Reduction: To decrease the number of features, we can employ techniques such as Principal Component Analysis (PCA) or autoencoders. By reducing the dimensionality, we can simplify the model, potentially improving its performance and interpretability.
- Missing Value Imputation: We can explore alternative methods for handling missing val-

ues, such as using autoencoder-based imputation or higher-order interpolation techniques. These methods may result in more accurate predictions and a better understanding of the underlying data patterns.

- Model Selection: We can experiment with different machine learning models, such as recurrent neural networks (RNNs), to see if they can yield better results than the current XGBoost model.
- Handling Proteins with High Missing Values: We can consider eliminating proteins that have a large number of missing values, as they might not contribute significantly to the model's predictive power and could even introduce noise.
- Adjusting Slope Values: For patients with consistently low MDS-UPDRS scores throughout the study period, we can assign a slope value of 0. This adjustment may help the model capture the disease's progression more accurately for patients with minimal symptom changes.

In summary, our machine learning model has shown potential in predicting Parkinson's disease progression using MDS-UPDRS scores. Future work will focus on refining the model and exploring alternative techniques to enhance its performance further. By continuing to improve our model, we hope to contribute valuable insights that can support the scientific community in their ongoing efforts to understand Parkinson's disease better and develop more effective diagnostic and treatment options.

9 References

https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction

https://github.com/carloscastmar/Parkinson-s-Disease-Progression-Prediction.git