**Parkinson’s Disease Progression Measurement**

**Research Proposal**

**Background Research**

Parkinson's disease (PD) is a neurodegenerative disorder that primarily affects the dopamine-producing neurons in the brain. It is characterized by a variety of motor and non-motor symptoms, with movement impairments being the most prominent.

**Key Symptoms of Parkinson's Disease:**

Tremor: The most well-known symptom of PD is a resting tremor, typically observed in the hands, fingers, or limbs.

Bradykinesia: Slowness of movement and difficulty initiating and executing voluntary movements.

Rigidity: Stiffness and resistance to movement in the limbs, trunk, and muscles.

Postural Instability: Impaired balance and coordination, leading to difficulties in maintaining an upright posture and an increased risk of falls.

Non-motor Symptoms: PD also presents non-motor symptoms such as cognitive impairment, depression, anxiety, sleep disturbances, and autonomic dysfunction.

**Factors Influencing Parkinson's Disease Progression:**

Age: PD generally occurs in older individuals, and the risk of disease progression tends to increase with age.

Disease Duration: Longer disease duration is often associated with a more progressive course.

Genetics: Certain genetic factors can influence the severity and progression of PD. Mutations in genes like LRRK2, PARKIN, and PINK1 have been linked to both familial and sporadic forms of the disease.

Environmental Factors: Exposure to certain toxins, such as pesticides and industrial chemicals, may increase the risk of developing PD and potentially accelerate its progression.

Comorbidities: The presence of other health conditions, such as cardiovascular disease or diabetes, can impact the progression of PD.

**Methods of Measuring Parkinson's Disease Progression**

Clinical Assessments: Parkinson's disease progression is often assessed through various clinical scales and rating systems, such as the Unified Parkinson's Disease Rating Scale (UPDRS) or the Hoehn and Yahr scale. These assessments evaluate motor symptoms, functional abilities, and overall disease severity.

Imaging Techniques: Neuroimaging methods like magnetic resonance imaging (MRI) and positron emission tomography (PET) can provide insights into structural and functional changes in the brain associated with PD progression.

Biomarkers: Researchers are exploring various biomarkers that could indicate disease progression, such as specific proteins in cerebrospinal fluid (CSF) or blood samples, or genetic markers associated with PD progression.

Wearable Devices: The use of wearable devices, such as accelerometers and gyroscopes, can track movement patterns and provide objective measurements of motor symptoms and fluctuations in PD patients.

Machine Learning and Data Analysis: Advanced statistical and machine learning techniques, including PCA, SVR, or other deep learning models, can be applied to analyze large datasets and identify patterns associated with PD progression.

It is important to note that measuring PD progression is a complex task, and a combination of multiple approaches and assessments is often employed to capture the multidimensional nature of the disease. Researchers continue to explore new methodologies and refine existing techniques to improve the accuracy and reliability of measuring Parkinson's disease progression.

**Data Acquisition**

Taken from Github from a 4-year-old project. The target variable for model training is total\_UPDRS. Sex is the only categorical feature whereas the remaining features come under the numeric category.

**Data Cleaning**

There are no null values in the dataset, therefore, there is no need for data cleaning here.

**Exploratory Data Analysis (EDA)**

Perform EDA to gain insights into the dataset and understand the distribution and relationships between variables. Explore both the independent variables (features) and the target variable (Parkinson's disease progression) using various visualization techniques. Useful plots for EDA may include:

Histograms or density plots to examine the distribution of continuous variables.

Bar plots or count plots to visualize categorical variables.

Box plots to identify potential outliers or extreme values.

Scatter plots or correlation matrices to assess the relationships between variables.

**Feature Engineering**

Consider deriving new features from existing ones that may provide additional insights into Parkinson's disease progression. This could involve calculating ratios, creating interaction terms, or transforming variables (e.g., logarithmic, or polynomial transformations).

Principal Component Analysis can play a significant role in feature engineering for measuring Parkinson's disease progression. PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving the most important information in the data. It accomplishes this by finding the principal components, which are orthogonal (uncorrelated) linear combinations of the original features.

Here's how PCA can be beneficial in feature engineering for Parkinson's disease progression measurement:

**Reducing Dimensionality**: Parkinson's disease progression datasets can often contain many features, which might include various clinical assessments, imaging data, biomarkers, and wearable device measurements. High dimensionality can lead to computational challenges and may result in overfitting. By applying PCA, we can reduce the number of features while retaining the most relevant information, thus simplifying the modeling process and potentially improving model generalization.

**Identifying Relevant Features**: PCA ranks the principal components based on their importance in explaining the variance in the data. By analyzing the explained variance ratio, we can identify which components capture the most significant variability in the data. This can help us focus on the most informative features and discard less important ones.

**Handling Multicollinearity**: Parkinson's disease progression features might exhibit multicollinearity, which can make it challenging to interpret the effects of individual features on the target variable. PCA resolves multicollinearity by creating uncorrelated principal components, making it easier to understand the contributions of different features to the overall variance.

**Feature Combination**: PCA combines information from multiple correlated features into fewer uncorrelated components. This can be useful when dealing with highly correlated features that may be redundant or provide similar information. By combining such features into principal components, we can reduce noise and improve the robustness of the feature set.

**Visualization**: While not directly related to feature engineering, PCA can be employed for data visualization purposes. By reducing the data to two or three dimensions, PCA can help visualize the distribution of samples and explore potential clusters or patterns related to Parkinson's disease progression.

However, it's important to note that PCA may not be suitable for all cases. The choice to use PCA in feature engineering should be carefully considered and may depend on factors like the dataset size, the nature of the features, and the specific goals of the analysis. Additionally, PCA might not always improve interpretability, as the principal components themselves might be challenging to interpret in the context of the original features.

Overall, PCA can be a valuable tool in feature engineering for measuring Parkinson's disease progression, particularly when faced with high-dimensional and correlated datasets. However, it should be employed judiciously, and its impact on the overall modeling process should be carefully evaluated to ensure it aligns with the specific requirements and goals of the research project.

**Model Training**

In the context of measuring Parkinson's disease progression, the following models can be considered for model training.

**Multilayer Perceptron (MLP)**

MLP is a versatile deep learning model that can capture complex patterns and relationships in the data. It is widely used for regression tasks and can handle both numerical and categorical features effectively. MLP allows for flexibility in terms of architecture and can be trained using various optimization algorithms. It is a suitable choice when dealing with large datasets and nonlinear relationships.

**Random Forest Regression**

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It is particularly useful when dealing with complex interactions between features and can handle both numerical and categorical data. Random Forest provides feature importance scores, allowing for insights into the relative importance of features in predicting Parkinson's disease progression. It tends to be less prone to overfitting and can handle noisy data.

These two models, MLP and Random Forest Regression, offer a good balance between performance and interpretability, making them suitable for the task of measuring Parkinson's disease progression.

**Support Vector Regression (SVR)**

SVR can also be a viable algorithm for measuring Parkinson's disease progression, but its suitability depends on various factors and considerations. Here are some points to consider when evaluating whether SVR is a good fit for this specific task:

**Non-linearity:** Parkinson's disease progression is a complex process, and the relationship between disease progression and various factors may not be linear. SVR's ability to handle non-linear relationships using kernel functions can be beneficial in capturing these complex patterns.

**Feature Space**: SVR can be effective when dealing with a relatively high-dimensional feature space, especially if there are interactions and complex dependencies between features. It can handle both numerical and categorical features, making it versatile for different types of input data.

**Data Size**: SVR generally performs well with moderate to large datasets. If your dataset is relatively small, other algorithms, such as simpler regression models or Gaussian Processes, may be more appropriate, as they can handle smaller datasets effectively.

**Model Interpretability:** SVR's focus is on predictive performance, which might result in reduced model interpretability. If interpretability is a crucial requirement for your project, simpler linear regression models or decision tree-based models could be preferred.

**Hyperparameter Tuning:** SVR requires tuning hyperparameters like C and ε to achieve optimal results. This process may require careful experimentation and cross-validation to find the best parameter values, but it can significantly impact model performance.

**Feature Selection**: As SVR is sensitive to the number of features, feature selection becomes essential to avoid overfitting. Proper feature selection techniques, such as Recursive Feature Elimination (RFE) or L1 regularization (Lasso), should be employed to identify the most relevant features.

**Availability of Data:** For SVR to perform well, having a diverse and representative dataset that adequately covers the spectrum of Parkinson's disease progression stages and relevant factors is crucial.

**Comparison with Other Algorithms:** It's important to compare the performance of SVR with other regression models, including deep learning models like Multilayer Perceptron (MLP) and ensemble methods like Random Forest Regression. This will help determine which algorithm best meets the specific requirements of your project.

In conclusion, SVR can be a useful algorithm for measuring Parkinson's disease progression, especially when dealing with non-linear relationships and high-dimensional data. However, it is essential to consider the factors mentioned above and perform thorough evaluations and comparisons to select the best-suited algorithm for this specific research project.

**Feature Selection Techniques**

For feature selection, we need to choose suitable techniques which go in conjunction with Multilayer Perceptron (MLP), Support Vector Regression (SVR) and Random Forest Regression models. Few such techniques include Lasso (L1 regularization), tree-based feature selection and recursive feature elimination. A short description for each of them is given ahead.

**Lasso (L1 Regularization)**

Lasso regression, which applies L1 regularization, is compatible with both MLP and Random Forest Regression. Lasso regression shrinks coefficients to zero, effectively selecting the most relevant features while penalizing less important ones. By setting less important feature coefficients to zero, Lasso helps to identify a sparse set of features, enhancing model interpretability and reducing complexity.

**Tree-based Feature Importance**

Tree-based models like Random Forest Regression inherently provide feature importance scores. This estimation of feature importance can aid in identifying the most relevant features for prediction. The scores can be used to select a subset of important features, ensuring that the MLP and Random Forest Regression models focus on the most informative variables for Parkinson's disease progression.

It's important to note that feature selection methods are not exclusive to a single model type, and the mentioned methods can be applied to both MLP and Random Forest Regression. By using feature selection techniques alongside these models, you can improve their performance, reduce overfitting, and enhance the interpretability of the results.

**Recursive Feature Elimination (RFE)**

RFE is a general feature selection technique that can be applied to various models, including MLP and Random Forest Regression. It iteratively removes less important features and evaluates their impact on model performance, allowing for the identification of the optimal set of features. RFE can help improve model efficiency and reduce overfitting.

I found the computation time is excessively long for performing this technique, so I considered using feature selection techniques that are less computationally intensive or subsampling the dataset for faster processing which are already explained above. While the dataset is not exceptionally large, RFE can be applied effectively if you have sufficient computing resources. It must always be kept in mind that the appropriateness of the technique also depends on the characteristics of the data and the specific goals of our analysis.

On the other hand, we must beware of the limitations of this technique before using it in a project. Firstly, it can be computationally complex, particularly for large datasets or high-dimensional feature spaces, due to its iterative nature. Secondly, RFE's effectiveness is sensitive to the choice of the underlying model used for feature ranking, potentially leading to different feature selections with different models. Thirdly, RFE ignores feature interactions, which may result in missing relevant combinations of features. Additionally, it assumes linearity and may not capture complex non-linear relationships. Moreover, RFE's iterative process can introduce selection bias, favoring features with early strong associations to the target variable and overlooking other relevant features. Furthermore, if the dataset contains noisy or irrelevant features, RFE may not effectively eliminate them. Lastly, RFE's time-consuming nature and risk of overfitting, especially with limited samples, should be considered for time-sensitive or small-sample applications.

**Model training Findings**

**Tree based feature importance:**

**SVR**

Mean Squared Error: 10.763902528013379

R-squared: 0.9028643000369536

**Random Forest**

Mean Squared Error: 0.035944881985937135

R-squared: 0.9996756258928668

**MLP**

Mean Squared Error: 1.79869110349484

R-squared: 0.9837682365758544

**Lasso feature selection**

**SVR**

Mean Squared Error (MSE): 12.631139024405066

R-squared (R^2): 0.8860139687001992

**Random Forest**

Mean Squared Error (MSE): 0.12209548200126107

R-squared (R^2): 0.9988981849217184

**MLP**

Mean Squared Error (MSE): 10.315380203796705

R-squared (R^2): 0.906911858977445

**Descending Order of Model Performance: Random Forest > MLP > SVR**

1. Random Forest
2. MLP
3. SVR

**Interpretation and nature of Performance metrics**

**Mean Squared Error (MSE):**

**Definition:** MSE measures the average squared difference between predicted values and actual target values. A lower MSE indicates that the model's predictions are closer to the actual values, which is desirable.

**Reasoning:** When MSE has a greater value, it means that the model's predictions have more significant errors, and the average squared difference between the predicted and actual values is higher. This indicates that the model is less accurate and has larger discrepancies between predictions and actual outcomes, making it a less desirable choice.

**R-squared (R^2):**

**Definition**: R-squared is a statistical measure that represents the proportion of the variance in the target variable that the model can explain. It ranges from 0 to 1, with 1 indicating a perfect fit where the model explains all the variance, and 0 indicating that the model does not explain any variance better than the mean.

**Reasoning**: When R-squared has a greater value (closer to 1), it means that the model explains a higher proportion of the variance in the target variable, indicating a better fit to the data. A higher R-squared suggests that the model captures more of the underlying patterns and trends in the data, making it a more reliable and accurate model for predictions.

In summary, MSE is considered bad when it has a greater value (higher error) because it implies larger prediction errors and greater discrepancies between predicted and actual values. On the other hand, R-squared is considered good when it has a greater value (closer to 1) because it indicates a higher proportion of explained variance, implying a better fit of the model to the data and more reliable predictions.