# Spin #4 Retrospective Document

**Describe any changes/updates to the project action plan including but not limited to project scope, domain questions, goals and measurable objectives, etc.**

Our scope has remained mostly unchanged. However, we are beginning to refine our analysis to consider our audience. This will entail using a black box ensemble method to identify important features. In response to the analysis carried out on the V1 dataset, we are further refining our approach to consider health-related issues or lifestyle factors related to low bone mineral density or increased fracture risk. Our goal is to create a machine learning algorithm that can accurately predict fracture risk based on health and lifestyle factors, and this objective has been refined through the analysis of various feature importance analysis seen in our work this week.

The domain questions have remained consistent with the previous week, focusing on identifying the specific current and past health-related issues or lifestyle factors that are related to low bone mineral density or increased fracture risk. Our measurable objectives have been refined as we've gained more insights from the data. We are narrowing down the lifestyle factors, diseases, and health history aspects most commonly affecting patients in our dataset.

**Discuss any additional data, data shaping and carpentry.**

We have potentially finished cleaning our dataset entirely this week. All nulls imputed either with 0 due to pattern logic of chained responses or otherwise in some cases imputed with K nearest neighbors. The cleaned dataset has been processed into independent and dependent variables, as well as split along train, test, and validation sets plus additional resampled train sets.

**Discuss any changes/updates regarding data storage location structure, use of database, and data curation and management strategy**

Our team has transitioned to working on .csv files stored on the team drive. This centralized storage system allows us to share and update our work more effectively since we no longer have challenges with table locks and schema permission challenges.

**Discuss any changes/updates related to data provenance and assessment**

Further work was done this week assessing the B1 (bone mineral density file) dataset for reasons of missing values. While the V1 (lifestyle) factors were well documented reasons for missing values, the B1 file only specified 1 missing value as ‘couldn’t perform’ and the other 30 as ‘imaging issues’. Because imaging issues could be a potential indicator of risk (“why were there imaging issues in these participants?”) further analysis was performed to examine the target variables for these participants. The missing values here were significantly associated with increased fractures so in order to keep these observations we imputed with KNN.

**Discuss any changes/updates to data acquisition and characteristics for data sets used in project**

Data acquisition and characteristics have remained unchanged and are sourced from the MrOS public study (not Dr. Leary).

**Description of any exploratory ML modeling and important methodological details (e.g. approach, target, features, training, testing and validation)**

We have utilized SMOTE to rebalance our dataset along target variables and in discussion with our mentor decided a train, test, validation split to be used going forward (80/10/10). We have started with ML modeling and our current goal is to use feature importance for reducing the dimensionality of our data. We have utilized various ensemble methods such as Random Forests, Light GBM, and XGBoost for this purpose. Target variables have been changed from ordinal to binary, aligning with our scope of predicting incidence of fracture, not the number of fractures.

**Discuss any changes/updates to the vision for final data story including intended audience and format/style medium of data story/product**

As we eliminate features the descriptive nature of the story is changing. Either due to the factors being insignificantly related to our target variable or without enough data to contribute. It is possible that we have to adjust our hopes for our descriptive story to focus more or entirely on the predictive analytics of our model. Further work on feature importance and elimination is being performed this week as we continue to explore methods like Boruta (as recommended by our mentor) and Recursive Feature Elimination.

**Weekly individual team member accountability/contribution assessment/evaluation included**

Karen-Cleaned up and performed EDA on target variable. Created merged dataset with target. Started performed linear regression on merged dataset.

David- Contributed code for cleaning of merged datasets, exploration of BMD null data, SpIn #4 artifact organization, Random Forest modeling. Feature reduction discovery with Boruta.

Josh- Updated our ETL processes, performed EDA on target variables, and did initial exploration of feature importance using Light GBM. Additionally, worked on SpIn #4 document.

Tyler- Finished cleaning up datasets. Utilized KNN imputation to fill in some missing values. Resampled our datasets using synthetic minority oversampling technique. Initial exploration of logistic regression on resampled datasets. Initial exploration of feature importance using XGBoost.

**Project work success status evaluated and future work/tasks discussed**

This upcoming week will be primarily focused on model building. Each team member is utilizing a different model to examine feature importance as well as predictive capabilities on the test set. We currently have one member working on random forest, one on lightGBM, and two using XGBoost.

**Jupyter Notebook(s) fully internally documented**

**All paths to data files map to DSA team shared folders**

All data is being stored in /dsa/groups/casestudy2023su/team03/

**Current project SpIn artifacts (notebooks) are located in the TeamArtifacts\SpIn\_4\_Artifacts folder (provide link (within Europa) to the first notebook in the pipeline)**

**Jupyter Notebook(s) execute without exceptions**

**Link to Mentor recorded mentor meeting and key meeting takeaways provided**

[7-5-23\_Mentor\_Meeting\_Team03.mp4](https://mailmissouri-my.sharepoint.com/:v:/r/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/Zoom%20Meetings/7-5-23_Mentor_Meeting_Team03.mp4?csf=1&web=1&e=1ZDDm9)

* Explore different approaches
  + We’re going to continue with various techniques and experiment with different strategies to find the best approach for analyzing the dataset and dealing with the 20% class imbalance. This will involve tweaking hyperparameters, considering XGBoost or LightGBM, and exploring methods like SMOAT or nlearn.
* Consider under sampling
  + If the feature set is limited, it may be worth exploring under sampling the minority class. This approach can be helpful in addressing the class imbalance issue.
* Dataset splitting
  + Given the total row count of 5500, it is recommended to allocate around 5% (or up to 10%) for the validation set
* Model evaluation and retraining
  + With no production data available, maintaining a train/test split is advisable to avoid overfitting the model. However, it is also important to consider retraining the model with all available data if the main objective is to gain a better understanding and explanation of the data. Boruta, a feature selection technique, can be used to compare its performance with the existing approaches.