# Spin #5 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.**

The project scope has remained mostly unchanged. However, the analysis has been refined to consider the audience. A black box ensemble method will be used to identify important features. 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. For instance, bone mineral density values have been reduced for dimensionality using PCA to allow for other lifestyle variables to be the focus of the story. The target feature has been coded into binary classification (not predicting number of fractures, but if).

**Purpose of the story, domain question(s) to answer/problem(s) being solved**

The purpose of the data story is to analyze health and lifestyle factors that are related to low bone mineral density or increased fracture risk. The domain question is to identify specific current and past health-related issues or lifestyle factors affecting patients in the dataset. The actual question has changed from “What specific current and past health-related issues or lifestyle factors are related to low bone mineral density or increased fracture risk?” to “When combined with BMD data, what specific health-related issues or lifestyle factors are associated with increased fracture risk?”

**Identification of audience and their expectations and interests in the data story**

The data story is targeted primarily towards healthcare professionals and researchers interested in predicting fracture risk using lifestyle factors. However, there is hope that using feature importance there may be a path for a descriptive story for patients. For example, early findings suggest aspirin and low cholesterol diets are correlated with fracture risk. With our models so far we are generally able to achieve a very high recall rate for our target class at the cost of high false positives so there may be some potential for either doctors or patients to have personalized risk assessments.

**Discussion of data, methods, approaches used in creating the data story and is suitable for the audience**

Machine learning algorithms, in particular ensemble methods, are being used for both feature importance analysis (for potential descriptive story) and predictive modeling. The team is actively using XGBoost, Random Forest, and LightGBM to test different models for sensitivity and specificity of our target class. In conversations with our mentor, we are utilizing decision thresholds to try and maximize recall or F1 score for class 1 (non-trauma fracture prediction). Optuna and/or Gridsearch are being implemented for hyperparameter tuning.

**Potential or emerging results that answer the import domain questions/problems in the data story**

Work from this week led us to discover that the performance of the models drop after being tested by the validation set. Work will have to be done on further optimization, as well as analysis of what potential objectives the model may still be useful. If time allowed, we would explore additional dimensionality reduction techniques to address potentially overfitting and/or incorporate other datasets that could provide additional information.

**Brief description and inclusion of 3 -6 visualizations that you will be using in your data story**

**The following visuals are early and merely examples at this point. No final decisions have been made on best model so expect plots to change.**

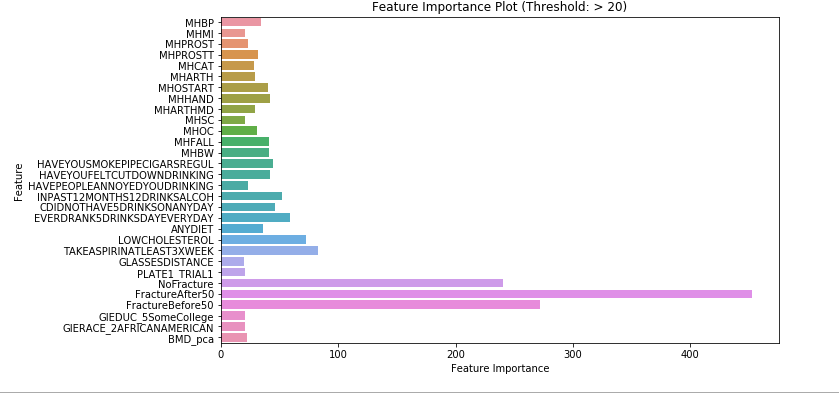
Early EDA on Bone Mineral Densities (BMD). The image shows the relationship between incidence of fractures and different BMD values for different body parts.

A screenshot of a graph

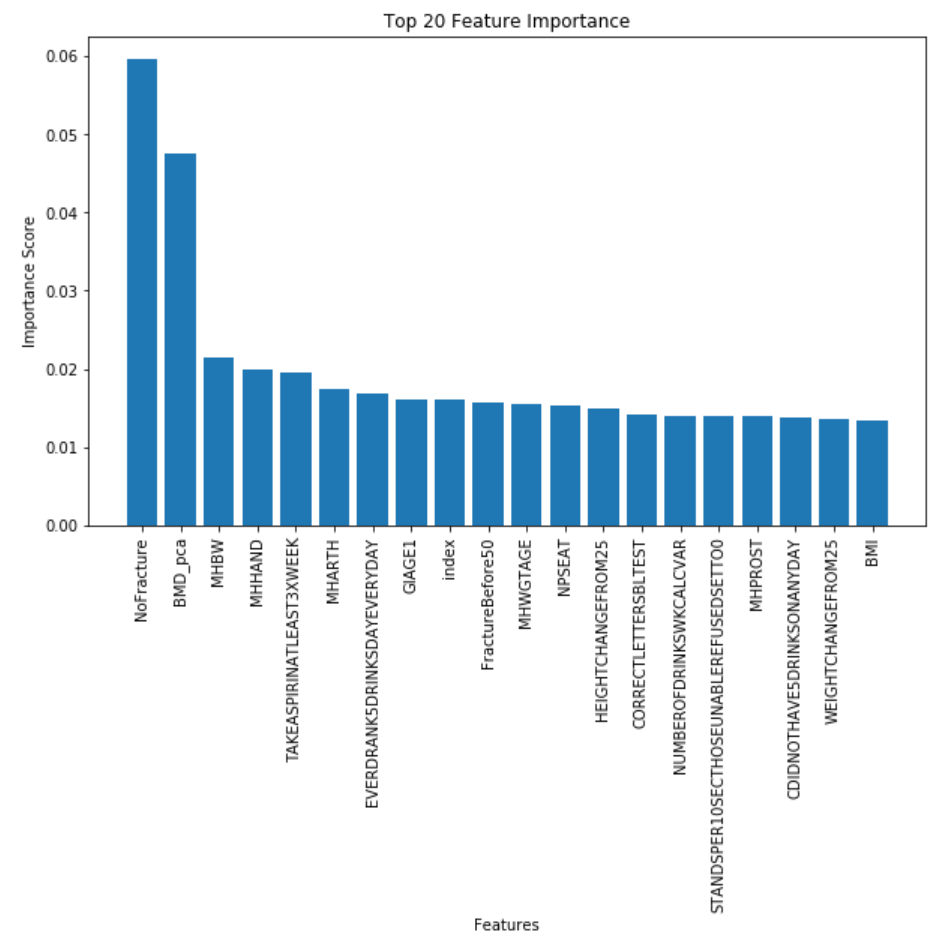
Description automatically generated

Feature importance plots

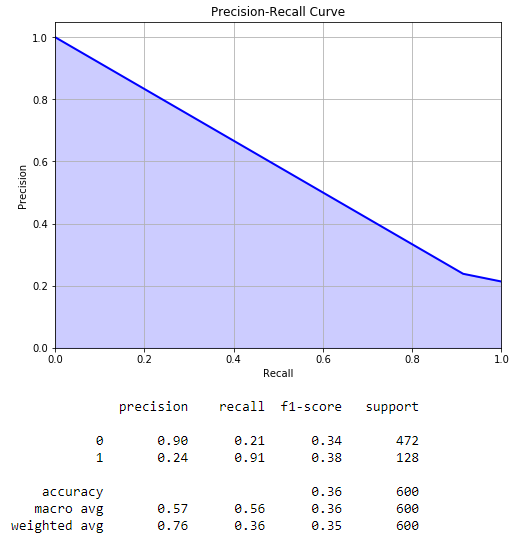
XGBoost



Random Forest



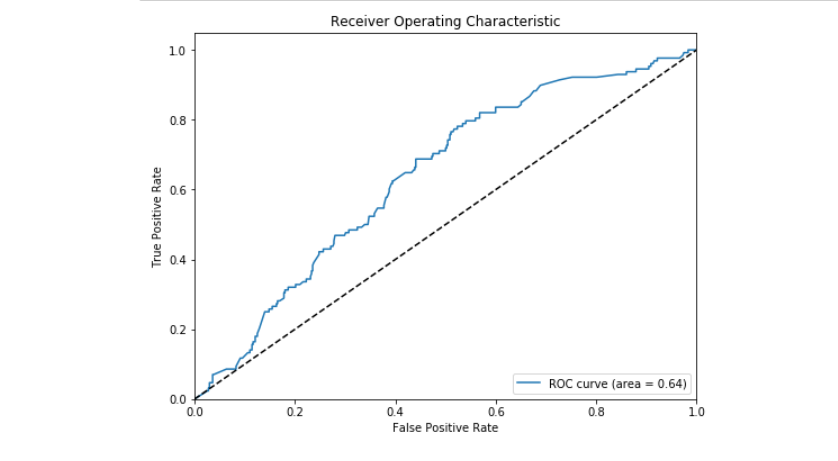
Precision-recall curves



Decision tree visualization

Classification table that includes all models performance on the final test set.

ROC curve after GridSearch



**Data story conclusions or takeaways**

So far, analysis of the entire dataset has not provided features that are as clearly predictive as we had hoped. It’s still too early to develop conclusions but further refining of our data should reveal predictive features that we can demonstrate in our final story. Preliminary results indicate some potential of stratifying risk – ie. Someone in the low risk category experiences fractures less than the general public, someone in the high risk experiences fractures more than general public. So there is some potential for something like a personalized risk assessment considering we have high true negatives but at the cost of false positives.

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

Tyler: XGBoost. Feature importance. Test on validation. Optimize decision threshold. Optimize hyperparameter using optuna. SPIN5 work.

Josh: Light GBM. Hyper parameter tuning with random search and grid search, model performance and evaluation. 552 presentation foundation. SpIn 5 work.

David: Random Forest model with additional feature reduction and hyper parameter tuning with Grid Search. SpIn5 Artifact consolidation and initial work on Final\_Team\_DataJourney\_TOC.

Karen:Created models using XGBoost, adjusted hyperparameters and checked with validation set using GridSearchCV, RandomizedSearchCV, and Optuna.

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

While some team members have moved beyond testing on the validation set, we plan to have a deliverable next week that includes a notebook with evaluation metrics of each model built on unseen test data. This way we can compare each model’s performance and visualize comparisons.

Upcoming work for the week includes final optimization, final model testing, visualization building, model interpretability, and work on our presentation.

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

Confirmed

**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\_5\_Artifacts folder (provide link (within Europa) to the "Start\_Here" notebook containing links to the notebook workflow/pipeline). Final Artifacts indexes weekly SpIns (data journey) and data story "TOC".**

SpIn5 Start Here <https://europa.dsa.missouri.edu/user/ejm301/notebooks/su23CaseStudy_Team03/TeamArtifacts/SpIn_5_Artifacts/SpIn5-StartHere.ipynb>

Data Journey:

<https://europa.dsa.missouri.edu/user/ejm301/notebooks/su23CaseStudy_Team03/TeamArtifacts/FinalArtifacts/Final_Team_DataJourney_TOC.ipynb>

Data Story (progress will start this week)

<https://europa.dsa.missouri.edu/user/ejm301/notebooks/su23CaseStudy_Team03/TeamArtifacts/FinalArtifacts/Final_DataStory_TOC.ipynb>

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

Confirmed

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

[7-12-23\_Mentor\_Meeting\_Team03.mp4](https://mailmissouri-my.sharepoint.com/:v:/r/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/Zoom%20Meetings/7-12-23_Mentor_Meeting_Team03.mp4?csf=1&web=1&e=UZzdhb)

Mentor meeting notes: [07/12/2023 - Dr. Green](onenote:https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/SU23_DSA8080%20Casestudy/Mentor%20Meetings.one#07/12/2023%20-%20Dr.%20Green&section-id={3D12AE69-253A-46DF-A3B5-D58BA7327F52}&page-id={4642BBA4-6189-4D02-B464-6998954EEB6A}&end) ([Web view](https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/_layouts/OneNote.aspx?id=%2Fpersonal%2Fjwj8c8_umsystem_edu1%2FDocuments%2FSU23_DSA8080%20Casestudy%2FSU23_DSA8080%20Casestudy&wd=target%28Mentor%20Meetings.one%7C3D12AE69-253A-46DF-A3B5-D58BA7327F52%2F07%5C%2F12%5C%2F2023%20-%20Dr.%20Green%7C4642BBA4-6189-4D02-B464-6998954EEB6A%2F%29))

Goup meeting notes: [2023-07-12 - Team Meeting](onenote:https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/SU23_DSA8080%20Casestudy/Group%20Meetings.one#2023-07-12%20-%20Team%20Meeting&section-id={22F8CF55-8C19-4BDF-B116-5B29E397A512}&page-id={F42E718E-C037-451E-9979-D6D51DF85EDB}&end) ([Web view](https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/_layouts/OneNote.aspx?id=%2Fpersonal%2Fjwj8c8_umsystem_edu1%2FDocuments%2FSU23_DSA8080%20Casestudy%2FSU23_DSA8080%20Casestudy&wd=target%28Group%20Meetings.one%7C22F8CF55-8C19-4BDF-B116-5B29E397A512%2F2023-07-12%20-%20Team%20Meeting%7CF42E718E-C037-451E-9979-D6D51DF85EDB%2F%29))

* Do we want to move forward with just classification or should we try to calibrate so we can provide probabilities?
* Future state we'd look at calibration to determine the actual level of risks but do not think we'll have time
* Review our thresholds of sensitivity and specificity
* Set up our pipelines to try class balancing before and as part of the model
* Can scale to 2:1 by reducing majority
* We've already determined when modeling that there was limited benefit of running as part of the model so we'll be including SMOTE in the pipeline
* Agree on which dataset we'll all be inputting into the model in the mode
* Prepare visualizations by Saturday evening
* Begin work on Final Artifacts