# Pre-Final Check-In

**Reflection on project overall, what has worked well and what challenges need improvement in terms of content and process.**

Once we finalized our dataset, the arc of our project has been relatively straight-forward. We work well together as a team and each member has been able to utilize their strengths to make this project a success. Three different models were explored to enhance our chances at finding the best one. Unfortunately, what seemed like a straightforward exercise proved to be more challenging than what we had anticipated. None of the models we developed gave us the clear correlations that we anticipated. With more time, we might be able to overcome these limitations by exploring more hyperparameters and looking deeper into ancillary aspects of the dataset. Next steps for further improvement of the model will be incorporated into the final data story.

Process challenges include data availability, keeping everyone on the same train, val, test sets. Incorporating similar modeling techniques like oversampling when upstream work was different. Naturally, we are in much better shape here now than we were in the interim but this week did include work trying to tie everything together that do not always play nice with each other.

**Status update on final data story.**

The group will be using a power point presentation to review fracture risk in men over the age of 50. We will address the health and lifestyle factors and their relevant contribution to risk of fractures in this population. The primary deliverable is still planned as a predictive model, and due to the target outcome of fracture, we lowered the decision threshold to increase sensitivity at the cost of specificity. This led to the decision to pitch the model as an early screening tool for now. Next steps for further model improvement will be highlighted in the final presentation, as the team has many ideas for further success here but unfortunately are out of time.

**Status update on notebook workflow/pipeline leading to project reproducibility**

All Notebooks work from similar pipelined data and pickled objects. This will ensure that our work is reproducible and easily followed.

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

Tyler: Finished up an XGBoost model and tested it and incorporated into final artifact. Work on final deliverables. Work on a few things that didn’t pan out/ran out of time (two-step modeling).

Josh: Light GBM work to finish model components, and incorporated results into final artifact. Worked on 552 presentation, Pre-Final Check-In, and other final work.

David: Worked on Final Data Story and Final Data Journey. Incorporated his work into the final artifact.

Karen: Testing XGB model and making more visualizations. Worked on Pre-Final Check-in.

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

Our group has gotten to a point where we have completed modeling, EDA, and coding efforts. We are working to complete the final deliverables for the course and developing a strong, cohesive data story to represent our results. We have finished our final model testing notebook outside of any additional visuals (for instance, still having trouble getting decision tress visualized in our current environment). The group is taking an all hands on deck approach to collaborate on the presentation and data story.

Furthermore, our group has discussed future state of this project and the potential it could be translated to capstone work.

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

Yes.

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

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

**Discussion of any revisions to project SpIn artifacts (notebooks) are located in the TeamArtifacts\SpIn\_#\_Artifacts folder**

Yes

**Provided link to the "Final\_Team\_DataJourney\_TOC" notebook initiating the notebook workflow/pipeline documenting the team data journey**

[Final Team Data Journey - TOC](https://europa.dsa.missouri.edu/user/jwj8c8/notebooks/su23CaseStudy_Team03/TeamArtifacts/FinalArtifacts/Final_Team_DataJourney_TOC.ipynb)

**Provided link to the "Final\_Data Story\_TOC" notebook initiating the notebook workflow/pipeline documenting the final artifacts documenting the data story**

[Final Data Story - TOC](https://europa.dsa.missouri.edu/user/jwj8c8/notebooks/su23CaseStudy_Team03/TeamArtifacts/FinalArtifacts/Final_DataStory_TOC.ipynb)

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

Always.

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

Recording: [7-19-23\_Mentor\_Meeting\_Team03.mp4](https://mailmissouri-my.sharepoint.com/:v:/r/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/Zoom%20Meetings/7-19-23_Mentor_Meeting_Team03.mp4?csf=1&web=1&e=EPOQkT)

Notes: [07/19/2023 - Dr. Green](onenote:https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/SU23_DSA8080%20Casestudy/Mentor%20Meetings.one#07/19/2023%20-%20Dr.%20Green&section-id={3D12AE69-253A-46DF-A3B5-D58BA7327F52}&page-id={3CCFF679-1D4A-41E1-8470-A854846CC58B}&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%2F19%5C%2F2023%20-%20Dr.%20Green%7C3CCFF679-1D4A-41E1-8470-A854846CC58B%2F%29))

**Action Items**

Schedule 1 hour review meeting with Dr. Green next week

Everyone ensures that their models are pickled

Tyler creates one more notebook which calculates the model performance with Test data set

Continue work on all Final Artifacts

Record the presentation on Thursday 7/27

Define these low risk interventions and include them in the data story

**Data Story**

* Can pitch this as an early warning system because we do currently have lots of false positives
* The Data Story itself needs to be strong because that's how you pitch the benefits
* This is what our model can do
* How it benefits Patients
* How it benefits Users
* How we gather feedback
* What's important (recall and sensitivity)?

**Further analysis**

* Bone Mineral density delta over the visits
* Allows us to trend and get a slope (up or down is a feature)
* Median Bone Density of other age, gender is XX so can create new features
* Life style factors over time
* More Social Determinants of Health (SDOH)
* Can pull in "site" patients were seen at
* Augment with Census Data, Geographical data, access to affordable healthcare, etc.
* Calibration of the model for stratified risk probabilities
* Change the datastory that we'd be trying to predict

**Things to think of when refining a model**

* Look at new features
* More features
* Hyper parameter tunings
* Feature interactions

If specificity is not great, that can still be okay - think of risk vs. reward

* Intervention of the model is not expensive then having a number of false positives is not as big of an issue
* Maximized recall allows for early screening so they can recommend X, Y, Z based on the features we see
* Define these low risk interventions and include them in the data story
* If false positives of COVID, much more impactful because have to be isolated

**Other considerations**

* Course is not designed for us to have an AUC > 0.8 or recall > 0.8
* All of us need to be able to answer questions about the entire process
* Each of us should be able to speak intelligently about every step
* Probably smarter to leave calibration to the capstone since will take more effort to stratify
* Classification does not require the same level of calibration