**DSC450 NO SECOND STAY\_PROJECT 3: STAKEHOLDER Q&A**

1. **What features were the most important in predicting 30-day hospital readmissions?**

The features that stood out the most were discharge disposition, number of medications, and whether the patient had any previous readmissions. Other important factors included age, hypertension, and where the patient was sent after discharge. These variables gave the models strong signals about who was likely to come back within 30 days.

1. **How does the place a patient is discharged to affect their chance of coming back?**

Where a patient goes after discharge plays a big role. Patients sent to rehab, nursing homes, or long-term care facilities had much higher chances of readmission. That suggests those patients may already be at higher risk or aren’t getting the follow-up support they need. On the other hand, those discharged to home with self-care had better outcomes overall.

1. **Why might AUC not be the best way to measure model performance here?**

Because the data is imbalanced, AUC can be misleading. A model could predict “not readmitted” most of the time and still score well on AUC. But that doesn’t help us catch the real readmissions. That’s why precision, recall, and F1-scorematter more here—they focus on how well the model finds actual high-risk patients.

1. **What data preparation steps made the biggest difference for modeling?**

Some of the most effective steps included encoding categorical variables, handling missing values, and using balanced class weights during model training. These steps helped the models perform better by giving them cleaner inputs and ensuring the minority class—readmitted patients—wasn’t ignored.

1. **Why did we choose Random Forest and XGBoost instead of other models?**

We picked Random Forest and XGBoost because they work well with complex, messy datasets like this one. They handle non-linear patterns, categorical data, and help us see which features matter most. XGBoost in particular is great at squeezing out better performance through boosting, and both models are less likely to overfit than simpler ones.

1. **What could we try to improve the model since AUC scores were low?**

To improve performance, we could try SMOTE or under sampling to better balance the classes, do more hyperparameter tuning, or test more advanced models like neural networks. We could also add new features or create better ones based on domain knowledge, like length of stay or follow-up adherence.

1. **What other types of data might help the model do better?**

Bringing in more detailed or real-world data could really help. Things like clinical notes, time-series lab values, or even social factors (like caregiver support or housing stability) could provide deeper context. These would help the model better understand who’s at risk before they return.

1. **How could hospitals actually use these model results to help patients?**

Hospitals could use this kind of model to flag high-risk patients before discharge, prioritize them for follow-up calls, or enroll them in care transition programs. It also helps with planning resources—knowing who’s most likely to return means teams can step in earlier and potentially prevent the readmission altogether.

1. **What are the risks of using machine learning to predict patient outcomes?**

There are a few key risks. First, bias—if the data isn’t fair, the model won’t be either. Second, there’s a risk of over-relying on the model and ignoring clinical judgment. And finally, false predictions (especially false negatives) could cause patients to miss out on help they actually need.

1. **How does using synthetic data limit how we apply these results in the real world?**

Since we used synthetic data, the model hasn’t been tested on real patients. That means we can’t assume it’ll perform the same in a real hospital setting. The trends are useful for exploration, but we’d need real-world validation before turning it into an operational tool.