Summary

Two approaches have been proposed: Sequential Modeling (Approach 1) and Non-Sequential Modeling (Approach 2). Each method offers unique advantages and challenges from a business perspective, which we will examine in detail to determine their alignment with Bree's goals and requirements.

Reducing Default Risk and Maximizing Lending Revenue

- Approach 1: By analyzing the detailed history of individual loan transactions and their outcomes, this approach directly models a borrower's behavior across multiple past loans. It reduces default risk by considering more precise loan-level details, such as whether previous loans were voluntarily paid back, how late they were paid, and patterns of spending/income during the loan period. This is an optimal way of analyzing a customer's behaviour and prevents Bree from making mistakes that it may have done in past loan applications with a customer.
- Approach 2: This approach provides a broader, aggregated view of the customer's financial behavior, offering insights into long-term patterns. It equally focuses on transactions as well as previous loan history (although not as detailed as Approach 1), which can be a better option than Approach 1 for a first loan approval.

Trade-offs and Risks

Approach 1 accepts input that is more detailed and gives a better overview of the customer's financial record, which means more nuanced and tailored predictions based on individual history. On the other hand, this approach comes with increased computational complexity due to the variable input size, which can make training and inference slower. Additionally, handling variable-length data requires more sophisticated preprocessing, which could introduce risks of errors or inefficiencies. From a scalability perspective, this approach might struggle as the volume of data grows, especially when dealing with customers who have extensive loan histories (although we could simply counter this by only taking the N most recent loan applications). With sequential models, which can be very complex, overfitting is a big risk, and further testing should be required to make sure that the model is generalizing well enough. Techniques such as gradient clipping can help keep training stable and effective.

Conversely, while Approach 2 provides a fixed input size that simplifies both preprocessing and model architecture, it sacrifices some of the insights by aggregating customer history into summary statistics. This trade-off could lead to less precise predictions, especially for customers with complex financial patterns. However, its simplicity makes it more scalable and efficient to deploy, especially in high-throughput or resource-constrained environments. The risk lies in oversimplification, where key nuances in individual customer behavior might be overlooked, potentially increasing the likelihood of misjudging default risk or lending capacity.

Strategies for Scalability

- **Distributed Training**: Leveraging distributed training across GPUs becomes important for large scale models. Additionally, techniques such as gradient checkpointing can further optimize memory usage during training, ensuring scalability even for deep sequential architectures.
- **Optimized Data Pipeline Design**: Building an efficient data pipeline is essential for preprocessing transaction data. For real-time predictions, the pipeline should minimize latency by incorporating parallelized processing steps, such as batch data loading into memory and feature extraction. Using frameworks like Apache Spark can help process large streams of incoming transaction data at scale.
- **Model Adaptability with Continuous Learning**: Regularly updating models with new data ensures they remain adaptable across changing customer behaviors, demographics, and markets. This can be achieved by implementing a continuous learning pipeline, where the model retrains periodically or incrementally based on the latest data. Fine-tuning models for specific markets or customer segments ensure predictions remain accurate and culturally relevant, especially in international expansion efforts.
- **Hybrid Approach for Prediction**: A hybrid approach can be explored where Approach 1 is used for high-value or high-risk loans requiring more detailed analysis, while Approach 2 is applied to lower-risk cases or batch processing.
- **Model Compression Techniques**: To handle growing datasets, techniques like model quantization, or knowledge distillation can be used to reduce the size and computational complexity of trained models. This can be useful for deploying models on computers with limited power, such as mobile devices.
- **Explainability and Monitoring Tools**: Scaling up predictions also necessitates robust tools for monitoring and explaining model outputs. Implementing Explainable AI methods such as SHAP values and Integrated Gradients can help detect potential biases or inaccuracies.

Success Criteria

Here is a list of important and relevant criteria:

- **Mean Squared Error**: Evaluate model accuracy in predicting the balance after repayment on a robust testing set.
- **Default Rate**: Measure the decrease in loan defaults post-deployment compared to historical baselines.
- **Revenue Impact**: Assess the improvement in lending revenue from optimized loan amounts and reduced defaults.
- **Inference Speed**: Ensure predictions can be made within milliseconds for real-time decision-making.
- **Scalability**: Demonstrate the model's ability to handle increasing data volume without significant latency, while maintaining high levels of accuracy.