Credit Risk Modeling Proposal

Incorporating a machine learning system for credit risk modeling into Citi's existing loan infrastructure could be beneficial for the company's loan approval process. Since risk is currently assessed largely by manual review, the addition of a machine learning algorithm could provide a few core benefits:

- **Efficiency**: Automating the credit risk assessment process with a machine learning system can save time and reduce the workload for loan officers. This can lead to faster loan processing times and improved customer satisfaction.
- **Accurate risk management**: Machine learning models can analyze a large amount of data in a short amount of time, often leading to more accurate predictions of credit risk. This can help lenders to make better loan approval decisions and reduce the risk of defaults, resulting in improved profitability and reduced losses.
- **Adaptability to new data**: Machine learning models can be easily updated with new data, which means they can adapt rapidly to changes in the market.

Data Requirements

To create a credit risk modeling system, we'll need to decide on the variables to include in our system. Although we already have a lot of the data we need internally, we may also want to obtain data from other sources, such as credit bureaus.
Some variables we'll want to include in our credit risk model are loan amount, loan purpose, employment category, education level, income, credit score, debt-to-income ratio, loan repayment history, and outstanding loans. We may also want to explore other variables we can create using feature engineering.

Data Outputs

There are a few different outputs that our credit risk model can produce to help lenders assess whether to approve a loan. Two of the best options include:

- **Credit risk score**: A score ranging between 0 and 100, where 0 represents a low credit risk and 100 represents a high credit risk.
- **Probability of default**: A probability between 0 and 1, representing the likelihood that a loan applicant will default on their loan.

Rather than relying on a threshold score to automatically approve or deny a loan, loan officers will use the model's output to make informed loan approval decisions.

Architecture

There are a variety of model types and architectures that we can explore for use in this system. A few common choices are neural networks, random forests, gradient boosting, and support vector machines. We can also combine various models to increase the accuracy of our predictions. We will want to assess which option performs best within the constraints of our credit risk modeling system. The data output we ultimately choose will influence the best choice for the model architecture, as well as the metrics we should use to assess model performance.

Risks and Challenges

When creating our credit risk modeling system, there are a few risks and challenges that we need to keep in mind throughout our development process. Here are are a few examples:

- **Ethical considerations:** We must ensure that our credit risk modeling system is as unbiased as possible. We should always be searching for sources of bias, with the goal of creating an ethical risk modeling system.
- **Data quality**: Our model's accuracy depends on having access to high-quality data inputs. We need to keep an eye out for incorrect data, data shift with time, and data leakage.
- **Data drift**: Data drift over time can negatively impact model performance. We need the infrastructure to regularly retrain our model on new data.