### Making Real-Time Predictions on Home Credit Default Risk using AWS S3, AWS Sagemaker, and AWS Quicksight

**Team Members:**

* Anjali Chikkudukaila
* Ariana Irani
* Pranav Chakradhar
* Rohan Rajeev
* William Missling
* Zhewen Zhang

**Abstract**

One of the main problems for loan providers is to find out if a potential applicant is likely to repay the loan or not. Knowing this in advance safeguards them against huge losses. Credit risk arises when a corporate or individual borrower fails to meet their debt obligations. The risk specifically is the probability that the lender will not receive the principal and interest payments of a debt required to service the debt extended to a borrower. It is this risk that inhibits the lender from providing a borrowing experience, for fear of incurring losses. Many new age customers are also unable to borrow from traditional banks due to a lack of credit history. Hence, some of the leading fintech unicorns often use predictive models from historical data to estimate their risk and maximize their returns.

**Use Case:**

Financial companies want to maximize their revenue earned through loans. Their goal is to identify potential consumers in need of loans and maximize their returns on them, while also identifying customers who could potentially default and lose them money. They would not want to administer loans to these customers likely to default.

With millions of potential and existing customers currently utilizing loans, we will build a model leveraging big data capabilities that can determine the likelihood a customer will default on a loan. This will help banks provide real-time responses to customers applying for a loan.

We intend to utilize Amazon S3 to store historical data, AWS Sagemaker to build and implement the Machine Learning model, and AWS Quicksight to provide a small overview of the decisions made on historical applications.

To simulate the real-world nature of loan processing, we also intend to batch the application data to display the streaming/live nature of loan inquiries.

**Dataset :**

[**Home Credit Default Risk | Kaggle**](https://www.kaggle.com/competitions/home-credit-default-risk/overview)

Home Credit is a company that offers simple and fast loans for a range of Home Appliances, Mobile Phones, Laptops, and a variety of personal needs. Here we are using Home Credit’s dataset found on Kaggle.

There are 7 different sources of data:

* application\_train/application\_test: The main training data with information about each loan application at Home Credit. The training application data comes with the TARGET variable indicating 0: the loan was repaid or 1: the loan was not repaid
* bureau: Consists of data concerning client’s previous credits from other financial institutions
* bureau\_balance: Consists of monthly data about the previous credits in the bureau
* previous\_application: The data of previous applications for clients with loans at Home Credit
* pos\_cash\_balance: Consists of monthly data about previous point of sale or cash loans clients have had with Home Credit
* credit\_card\_balance: The monthly data about previous credit cards clients have had with Home Credit
* installments\_payment: The data of payment history for previous loans at Home Credit

**Technologies to be used:**

* AWS S3 (for data storage)
* AWS Sagemaker (for model building)
* AWS Quicksight (for visualization)

**Set-up Instructions:**

1. Create a free Amazon S3 account which has a free usage period
   1. [Link](https://aws.amazon.com/pm/serv-s3/?trk=fecf68c9-3874-4ae2-a7ed-72b6d19c8034&sc_channel=ps&sc_campaign=acquisition&sc_medium=ACQ-P|PS-GO|Brand|Desktop|SU|Storage|S3|US|EN|Text&s_kwcid=AL!4422!3!488982706722!e!!g!!aws%20s3&ef_id=CjwKCAjw682TBhATEiwA9crl36lYJQUNUu3wfQhSMUDQfPIVOjMN8RHvlCBEDLhg5jaSDdONgpn8bRoCYUkQAvD_BwE:G:s&s_kwcid=AL!4422!3!488982706722!e!!g!!aws%20s3)
2. Upload the data into the S3 bucket
3. Navigate to AWS Home and create a new instance of Sage maker
   1. Create a new jupyter notebook
   2. Use the code in the notebook “Training\_the\_model.ipynb” to train the model
   3. The data to be used is based on the name of the S3 bucket assigned by you
   4. AWS helps us understand how to link S3 to Sagemaker. [Link](https://aws.amazon.com/blogs/machine-learning/secure-amazon-s3-access-for-isolated-amazon-sagemaker-notebook-instances/)
   5. Post training the model, utilize “Testing\_Data.ipynb” notebook to test the model and make predictions.
   6. The resulting data can be further stored in AWS S3 for future usage
4. We have utilized a Python function to replicate the real time data of a few applicants at a time, this can also be done using Flask, AWS Kinesis etc.
5. The final data has been downloaded as a CSV file and visualized using AWS Quicksight.
   1. Quicksight tutorials can be found [here](https://aws.amazon.com/quicksight/)