Our project comes from a kaggle competition and uses data from a company called Home Credit. Home Credit is a company that operates in several different countries including several countries in Europe and Asia in addition to the United States. The company's business model focuses on making loans to people with little or no credit. Their loan products primarily include point of sale loans (POS), cash loans, credit cards, debit cards, revolving loans, or car loans.

Our project uses supervised learning to predict a client's ability to repay a loan. Given the company's loan products, business model, and regions of operation, there are several unique challenges with this task. First, given that the company operates primarily by making loans to people with little or no credit, the ability to which we can use default on previous loans or credit history is limited. Moreover, since there is variation in which the company operates in, there is variation in measurement in various credit bureaus in each country. For example, lenders in the United States often use FICO scores as a large predictor in a borrower's ability to repay a loan. Given these challenges, we make use of several datasets provided by Home Credit.

To conduct this analysis, we will use data provided by Home Credit. The data contain information about each client such as age, employment status, living situation, and many other factors. The data also provide a binary variable for whether or not the client defaulted on the loan or not, which serves as our target. Home Credit also provides several other data about the client, including information on loans from other institutions and information on loans that the client has taken out from Home Credit before. Since Home Credit's business model relies on making loans to people with little or no credit, the number of observations in this samples is quite low.

In order to implement our analysis, we will make use of several different machine learning models: Support Vector Machine, Random Forest, and Naïve Bayes. We get sufficient background material from class notes, the internet, and *The Elements of Statistical Learning.* We judge the performance of each model based on several different metrics such as classification reports, confusion matrix, and ROC_AUC scores.