## Team #: 12

Team Members:
1. Trang Doan. Edx username: ngtra14@gmail.com  Data analyst at a logistic company.
2. Bharadwaj Kopparthi. Edx username: lakshminarsimha.kopparthi@gmail.com. Analytics developer at Salesforce.
3. Aaron Beach. Edx username: abeach0327@gmail.com. Mathematics teacher at public school.
4. Sidra Husain; sidrahusain. Bachelor's in electrical engineering working as a software consultant.
5. Xiaojie Yu. EdX username: yuxiaojie777@gmail.com. R&D scientist at a pharmaceutical company

# OBJECTIVE/PROBLEM

Project Title:  Credit Approval Analysis

Background Information on chosen project topic:

Credit is an important factor in our economy. Business needs credit to facilitate business activity, spend on capital expenditures, create jobs, and thus indirectly stimulate the economy. Individuals need credit to fund major expenses, such as: education, real estates, autos, etc. Therefore, it is very important to facilitate lending in the economy without incurring excessive risks.

During the last decade, with the widespread use of the internet and growth of technology, lenders are constantly developing and optimizing machine learning models to predict credit risk. This powerful technology can leverage banks and financial institutions to increase their market share in consumer lending while minimizing the risk of default and increasing their ROI.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

During the evaluation process, lenders look at socio-economic and demographic statistics of the applicant to make decision rules and minimize risk. The primary objective of this analysis is to explore which factors are the strongest indicators of lender's credit decisions and to assess whether these attributes are also statistically significant in predicting default.

State your Primary Research Question (RQ):

Which factors are the strongest ones influencing lender's credit decision? Given the weight that lenders put toward those factors in the decision process, are they modeling a reasonable risk management?
*( The second part of the question can be answered by examining a second dataset in which the significant factors are independent variables to predict "default". If these factors are statistically significant in managing credit risk, it should have very high prediction accuracy for " default" status.)*

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Ethnicity is a protected status and the decision to approve or deny an application cannot be based on the ethnicity of the applicant. Is there a statistically significant difference in how credit is granted between ethnicities that could indicate bias or discrimination? Are there any other attributes of a person in each ethnicity group influencing the decision? ( For example: does factor "educational level" influences the credit decision for ethnic group "A" the same way it affects ethnic group "B")
2. Although age is sensitive information and should not be considered in approving credit applications, lenders could use a proxy attribute such as "years of employment" to accept or

decline an application. There would be some level of correlation between a person's age and years of employment, as years of employment increases as the age of a person increases. If the correlation is high, a lender could simply remove the “age” attribute and include years of employment to model bias into the system.

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

The fall of the banking system in 2008 caused a lack of investor’s confidence and the rapid decline in credit availability for a long period of time. The crisis quickly spread into global economic shock, resulting in several banking failures and a global recession. Since then, banking regulation has changed, and risk management has been watched very closely by the Federal Reserve. With the growth of AI, machine learning and technology, it is very interesting to see how using these advanced techniques will improve the prediction accuracy, reduce the loss reserve that the bank needs in their balance sheet and increase their lending profit.

## DATASET/PLAN FOR DATA

Data Sources (links, attachments, etc.):

1. UCI Credit Approval dataset: The main dataset is the Credit Approval dataset taken from the archives of the machine learning repository of the University of California. It contains data from credit card applications in the US. The data is collected from the credit card application in the US. Data set has 690 records.
   There are 15 independent variables in the 1st dataset (below). There is one dependent variable: Approved. We think the most important factors in credit decisions are: credit score, prior default and employment status. The dataset is very clean so we dont expect to do a lot of cleaning. However, there are several categorical variables, such as ethnicity, we might have to create dummy variables during our modeling process.

| Gender | Age | Debt | Married | BankCusto | Industry | Ethnicity | YearsEmpl | PriorDefau | Employed | CreditScor | DriversLicense | Citizen | ZipCode | Income | Approved |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 30.83 | 0 | 1 | 1 | Industrials | White | 1.25 | 1 | 1 | 1 | 0 | ByBirth | 202 | 0 | 1 |
| 0 | 58.67 | 4.46 | 1 | 1 | Materials | Black | 3.04 | 1 | 1 | 6 | 0 | ByBirth | 43 | 560 | 1 |
| 0 | 24.5 | 0.5 | 1 | 1 | Materials | Black | 1.5 | 1 | 0 | 0 | 0 | ByBirth | 280 | 824 | 1 |
| 1 | 27.83 | 1.54 | 1 | 1 | Industrials | White | 3.75 | 1 | 1 | 5 | 1 | ByBirth | 100 | 3 | 1 |

   Masked data: https://www.kaggle.com/datasets/echo9k/uci-credit-approval/code
   Unmasked Data: https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data?select=clean_dataset.csv
2. Credit risk dataset: The second data set is the Credit Risk dataset which contains information about the loan's applicant and its current status. There are 11 independent variables in this data set. 1 dependent variable: Loan Status. Dataset has 32852 records.

   https://www.kaggle.com/datasets/laotse/credit-risk-dataset

| person_age | person_income | person_home_ownership | person_emp_length | loan_intent | loan_grade | loan_amnt | loan_int_rate | loan_status | loan_percent_income | cb_person_default_on_file | cb_person_cred_hist_length |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 59000 | RENT | 123 | PERSONAL | D | 35000 | 16.02 | 1 | 0.59 | Y | 3 |
| 21 | 9600 | OWN | 5 | EDUCATION | B | 1000 | 11.14 | 0 | 0.1 | N | 2 |
| 25 | 9600 | MORTGAGE | 1 | MEDICAL | C | 5500 | 12.87 | 1 | 0.57 | N | 3 |
| 23 | 65500 | RENT | 4 | MEDICAL | C | 35000 | 15.23 | 1 | 0.53 | N | 2 |

# APPROACH/METHODOLOGY

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

Our approach can be divided into 2 major steps:

1. Identify significant factors influencing the lender's decision in approving credits.
   - Cleaning & prep data: examining the dataset to find outliers and missing values. Create imputation if needed. We have several categorical variables that we might need to create dummy variables.
   - Variable selection: checking for multicollinear. If there is none, we can proceed with at least 2 methods(stepwise regression, random forest) to select variables.
2. Based on those factors, we are going to do 2 things:
   - Predict the application's decision :
     - ✓ First approach: Logistic regression model to evaluate the probability of being approved. We then can set the threshold: if the probability above it, we will say the application is likely to get approved. We will split data into training and testing.
     - ✓ Second approach: k-nearest neighbors or CART

     The metric that will be used to measure the "goodness" of the models will be confusion matrix (since our dependent variable "decision" is binary)
   - Examine whether these factors are good indicators of credit risk management by predicting the default probability.
     - ✓ First approach: build a logistic regression model to test the hypothesis that these above factors are good indicators of default risk.
     - ✓ Second approach: analyze the credit risk dataset using different methods like random forest, CART to identify significant factors in predicting default and compare w/ significant factors found in the first dataset.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

Based on our analysis, we would likely see that prior default, employment status and credit score are the strongest predictors of whether an application will get approved. Age and income aren't the most significant attributes in predicting credit application decisions, but they are key indicators of " default" risk, thus is it very important for lenders to put more weight toward those factors in their decision model.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

By improving the credit decision process, financial institutions can improve their net operating revenue on lending segment, reduce bad-debt write-off and maintain a healthy balance sheet to stabilize the whole economy. Indeed, human's behavior changed so it is very important for lenders to fine tune their attributes and model to reflect the changes and improve the decision process.

## PROJECT TIMELINE/PLANNING

Project Timeline/Mention key dates you hope to achieve certain milestones by:

| Proposed Timeline | Actual Due Date | Milestone |
|---|---|---|
| June 30th | N/A | Basic cleaning data & imputation, Variable selection- dataset is ready for training |
| June 30th | July 2nd | Video project proposal |
| July 5th | July 7th | Progress report due: models should be run and ready to analysis result |
| July 12th | N/A | Reassess coding parts and draft of analysis |
| July 16th | July 20th | 1st draft submission to TA for feedback- review and editing if needed |
| July 20th | July 23rd | Video final project |