## Credit Card Scenario

Customer Problem

1. Describe the problem, related to the topic you selected.
   1. Some credit card customers regularly don't make the minimum payment each month (for arguments sake we'll say it's a problem if they miss more than 3 payments in a 6 month period).
   2. Supporting customers in financial difficulty is costly and stressful for the customer - if there was a way of knowing they'd struggle then we should turn them down when they apply. (Alternatively we could offer them a card still, but at a higher interest rate to mitigate the potential cost associated with their risk of defaulting).
2. Phrase the problem as a question to be answered using data.
   1. Can we reduce the proportion of customers who miss more than 3 payments in a 6 month window by predicting whether they will default or not at their application stage?

Possible Data Science project approach

1. Analytic Approach
   1. I will need to compare the proportion of defaults before and after a new model is used in order to judge whether the project has been successful.
   2. Investigate the causes of customers defaulting - make sure it's not caused by a difficulty with our payment system for example (a non-data solution to the problem!), or the relationship between default rates and measures of the wider country's economy.
   3. I will need to build a classification model that inputs key features that can be gathered at the stage a customer applies for a new cards and predicts default/no-default.
   4. If the model predicts default then they can be turned down for the loan.
   5. We will want to minimise the chance of falsely turning an applicant down though so we don't annoy valid potential customers.
   6. It would be helpful (but not essential to the question currently defined) to give a probability of default (e.g. from a decision tree or logistic regression) so that this could be used to determine a risk level. Some slightly riskier customers could still be accepted if the interest is higher - sufficient to cover the cost of support or overall probable losses for a group of risky customers.
2. Data Requirements
   1. Data could come from the application form, or external sources like paying to access credit scores, national statistics about the economy.
   2. Useful features are likely to include current savings, income, job type, length of time in work, credit score, family ties/responsibilities, and more should be identified in discussion with key human decision makers who understand the domain in depth e.g. experienced bank managers.
   3. Would need to check the financial sector regulations as to what information can be used for decision making. Also company policies. For example, if regulation or company policy prevents sex or gender being used to determine customer application success then a) that data must not be used to create the decision making model BUT b) should check if we are able to use that data in order to create subsets of the test data - to determine model fairness.
3. Data Collection
   1. Once in deployment the application form could be online and data could then be automatically processed. External data would need to be processed from an API.
   2. But at this stage we need to see how the information is currently gathered. When creating the data set we should think about when data would really be available (e.g. how long does it take to get a credit score for an applicant - once deployed would it actually need to use a credit score that was updated the previous month).
   3. Data sources would need to be merged into a dataset with at least one column being a 1/0 (boolean) column for 'default' which is calculated by looking at the number of missed payments in the previous 6 months.
4. Data Understanding and Preparation
   1. Check for correlations/associations with the credit score column - this might save us asking more questions on the application form if we don't use features that don't give us any additional insight than credit score (or perhaps we could remove credit score and save the cost of paying for that external data).
   2. Check for missing values or any outliers/anomalies - investigate what is causing them (e.g. why might someone not have a credit score e.g. if they've lived abroad or are too young to have a history? Do we need a separate decision making process or separate model for deciding in those cases?
   3. Normalise/standardise the data in preparation for modelling.
5. Modeling and Evaluation
   1. There's a number of different classifier models that could be tried with various settings. For example decision trees.
   2. Because the decision process is likely to be automated (so the customer gets a quick decision) don't need the advantage of decision trees in terms of company staff having a 'flow chart'.
   3. Need to use the test data to check the accuracy, false positives, false negatives - if the test data can be divided into customers from different protected characteristic groups then we could check if the model has the same accuracy regardless - to check it's making fair decisions.
   4. Remember that the data used to develop and test the model was only from customers who were previously accepted under the old decision making process - so there needs to be a sense check with stakeholders to make sure the new model wouldn't approve in cases that would previously have been rejected - potentially letting through more 'false positives' once deployed.
   5. Get feedback from stakeholders, deploy in a small way to test for any unforseen problems. After 6 months from deployment compare the default rates to the 6 months before deployment - has the model succeeded in reducing the number of customers who miss 3 or more payments in those 6 months?