Data Science vs. Analytics

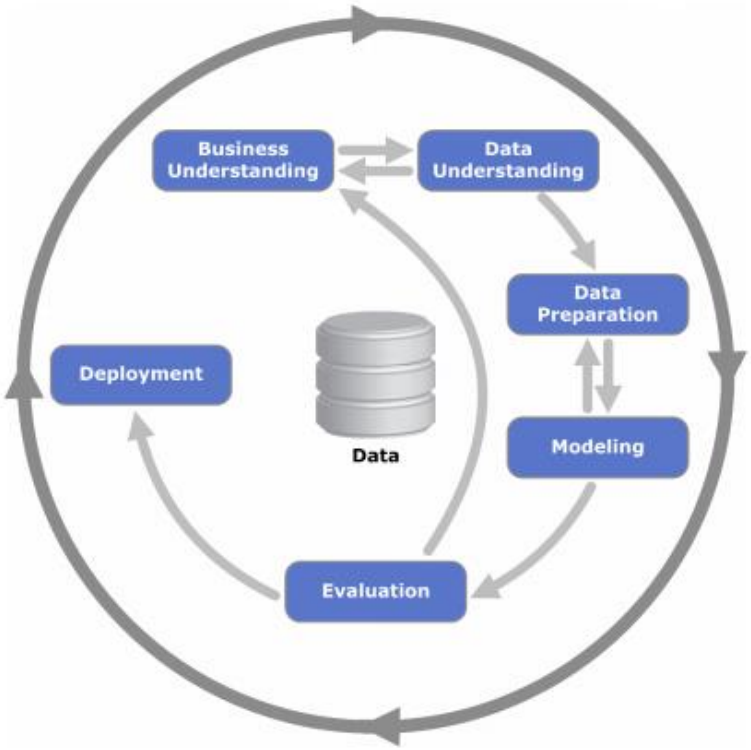
- Analytics takes one step further than data science into the implementation, the deployment and the strategies
    - Data scientists have computer science, statistics, and substantial expertise, and they can apply data science in a specific point in a specific area.
        - Coding + math & stats → machine learning
            - Taking the models in a vacuum
            - Understanding the models independent of your specific application area
        - Math + expertise → traditional research
        - Coding + expertise → danger
            - Not understanding the underlining of the model
        - Combining all of these together → data scientist
    - One becomes data scientist when they understand their specific phenomena (e.g. the output of the model and what the model is doing) so that they can apply the models in a transformative way
    - These techniques can all be learnt online, what really differentiate this from studying a master is how to apply data science effectively
        - Anyone can learn a deep learning algorithm
        - Very few people can take that model and turn it into a successful strategy
    - What makes an effective data scientist is not knowing more models, but turning the models into actionable strategies
        - Analytics worries about question, the data science models are tools
            - Use the tool to create something, answer something
            - In the real world, that something is creating knowledge

- Four chunks of business analytics
    - Descriptive analytics
        - Basic things like simple plots in excels,…
        - Answer the question: what happened?
    - Diagnostic analytics
        - Go further to investigate what the data doesn't tell you
        - Answer the question: why did it happen?
    - Predictive analytics
        - Create models that allow you to predict what will happen in the future
        - Answer the question: what will happen?
        - Traditional data science will stop here
    - Prescriptive analytics
        - Answer the question: how can we make the thing that we want to happen, happen?
        - How to improve collection rate, how do we stop people from defaulting, how to improve portfolio of loans, how to make things better
        - This objective goes against the previous objectives

    - If create a model that could predict
    - But when you create something that improves the situation will damage the model
    - The assumption that the future will like the past will be broken

Cross-Industry Standard Process for Data Mining

- CRISP-DM



- It helps you think the problem in the correct way
  - Start by business understanding
    - What are you necessities?
    - What do you need to do?
    - If a model is used for five years, the model will be biased
  - Then data understanding
    - Before touch the data, need to know what data are available
    - Should I buy the external data
    - What data should be used?
  - Data preparation
    - Now we start to work on the data
  - Modeling
  - Evaluation
  - Deployment
    - Based on the models, create strategy and create a system
    - Using algorithms is expensive, so we will always think about the simplest solution that gets the job done

- It can go both ways
  - AB testing: taking two randomized population, applying treatment on one, and nothing on the other; or several treatments to several pieces and have a control group with nothing
  - After evaluation, we may have a better understanding of the business

Intro to Credit Scoring

- Credit scoring is an estimate of the probability or the likelihood that an applicant will successfully repay their loan based on various information
- Several families of information
  - Applicant characteristics (e.g. age, income, employment status, time at address, …)
  - Credit bureau information
  - Application information of other applicants
  - Repayment behaviour of other applicants
    - Have previous loans? Late in credit cards repayment?
    - Alternative data: include utilities, up to date with their cell phone bill
    - Show one of the two things
      - Capacity of payment
      - Willingness to repay: how good they are at managing money?

- And then we will create a statistical model (scorecard) that estimates the probability of default of a customer
- And then compare them to everyone else in our data set
- We set up a cut-off
  - A point where you stop and say from here upwards not accepting

Judgmental vs. statistical approach

- In the 60s, loan applications were done and evaluated face to face with a credit risk analyst
  - Horrible way of lending to people
    - Racist, women were discriminated

- So generated a system that use statistics to create scorecards that do the analysis in a more transparent way to determine whether to grant a loan or not
  - The idea is the eliminate the judgment factor
  - Note that credit scoring is still discriminatory
    - Discriminate between people is the point of it
  - The point is we control how to discriminate them and we can test them
    - There are still racist, but now we can analyze it and control it
    - Some things are inevitable, because it reflects the societies that were living

- E.g. zip code

Problem

- Both methods assume that the future will resemble the past
  - But when a company grow, the behavior of their customers will change
  - The model downgrades faster and faster now because we are growing like crazy

- Statistical models that are well developed, well controlled, and well maintained will have faster and more accurate evaluations
  - The accuracy of the model will be a statistical reflection of your data set,
    - If your data set correctly reflects what you're seeing today, your model will work great
    - If you are changing those behavior, the model will stop working

- Judging people is difficult, because people are complex. Statistics, although can be lied, is consistent.

- Basel made credit scoring necessary in the 80s
  - Debt went down, operating costs went down, portfolio management improved, and credit availability went up

Types of Credit Scoring

- Application scoring
- Behavioral scoring
- Credit bureau scoring
- Profit scoring

- Generally, we will use Logistic regression to create our models
  - Because our response is binary
    - But not trying to differentiate two populations (probability model would be better)
    - We are trying to differentiate between two decisions

$$P(Default|x) = \frac{1}{1 + \exp(-\beta \cdot x)}$$

- So we have a vector of characteristics that are collected, assuming all the data cleaning are done, we train our model, get parameters estimates
- Then implement it in the future
  - So take the past, and calculate the probability of default using the logistic formula

- Default
  - Every bank, by law, will require to have its own credit score if they are regulated under Basel or

under the Dodd Frank Act in the US

- ➢ Basel says that default is one of
  - ✧ Obligor is 90 days late in paying the debt (some jurisdictions have 180 days, which require the model to be trained differently)
  - ✧ Credit institution considers that the obligor is unlikely to pay its debt obligations to the credit institution
- ➢ Note that if you default in one, you default in all of your debts
  - ✧ Banks hate this, because someone might be paying the mortgage but can't pay back the credit card, and all his products will be moved to the deteriorated portfolio. Then the person goes bankrupt (if applied) unless quickly covered

- Note that not every country has personal bankruptcy laws
  - ➢ In Canada, if a company defaults,
    - ✧ Someone will orderly liquidate, if necessary, everything to pay back debtors
    - ✧ And then either close up the company or resurrect the company where the unpaid debtors become the owner of the company
  - ➢ Person can declare bankrupt
    - ✧ This will determine the behavior of people
    - ✧ In ON, if bankrupt,
      - ▫ Get an assignment and people liquidate all the assets that are not first necessity
      - ▫ The record stays 7 years
  - ➢ Other jurisdiction will have a different personal bankruptcy law. This will affect what you see in your data

Application Scoring

- Application scoring
  - ➢ It is the most common type of credit scoring and is the one that determines whether you will be granted a loan or not
  - ➢ This only occurs when you have no past credit history with the company
    - ✧ You will be gone through a non-application scorecard
  - ➢ The application scorecard is bad for predicting because the bank will have limited behavioral history
  - ➢ It usually uses two sets of variables
    - ✧ Application variables: age, income, marital status, years at address, years with employer, …
      - ▫ In creating statistical modelling, we want to discriminate using behavior but never identity
      - ▫ Do people defining themselves as …
      - ▫ More strictly, so the variables should be that if a reasonable person could change if they want it
    - ✧ Credit bureau variables
      - ▫ A credit bureau is a company that collects data for the sole purpose of creating a credit score and sell it, they buy databases.

        - Problem is, they cannot get everyone's income, so they average it for the age range and location
        - Bank can get your income
        - Controversial variable: number of credit inquiries with the last 12 months; this impedes the competition among banks
        - Definition of default

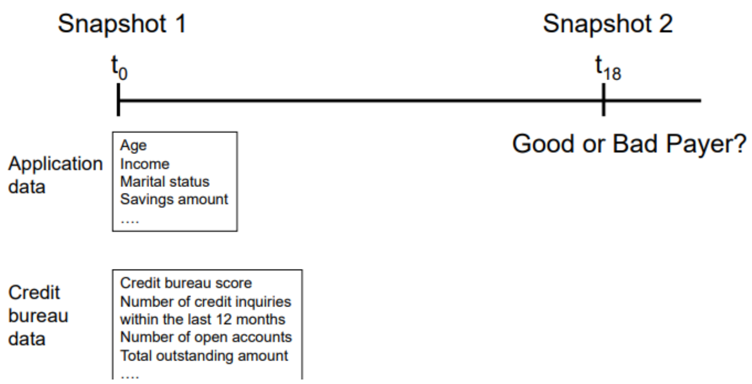
You are **Bad** if during the next 12 months:
- There is more than 1 instalment with more than 90 days in arrears.
- There are more than 2 instalments with more than 60 days in arrears.
- There are more than 3 instalment with more than 30 days in arrears.

You are **Good** if during the next 12 months:
- There is no more than 1 instalment with more than 30 days in arrears.

You are **Undetermined** if you are neither good or bad.

    - Banks use these credit bureau scores, but have different weight applied to it based on their customers

- It is a snapshot

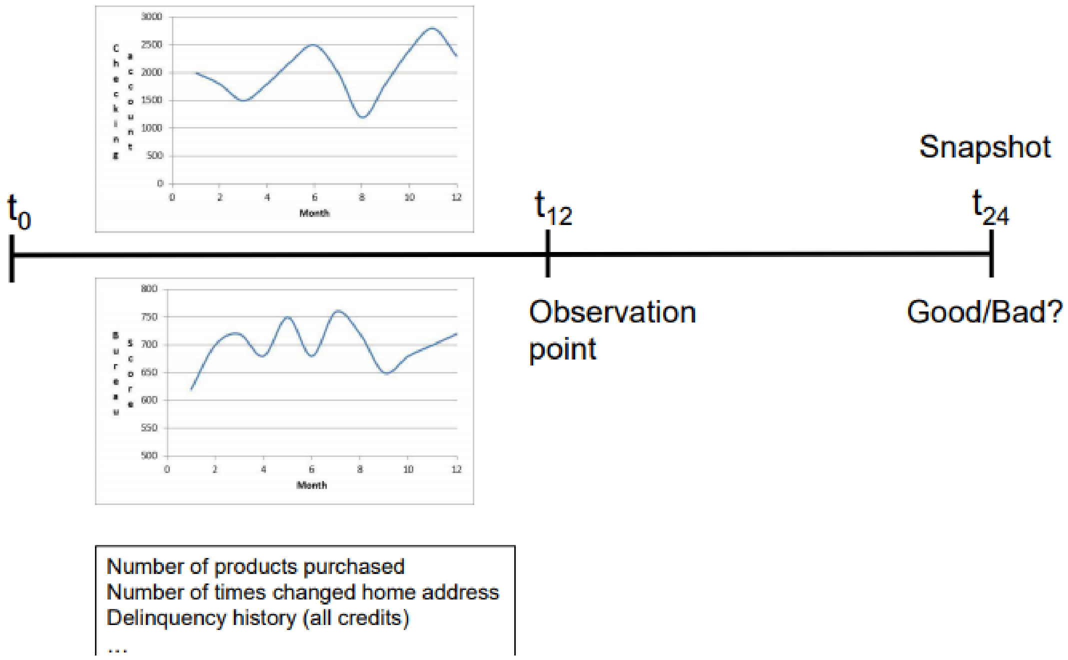


    - I take a snapshot of your information, and 12 months into the future, have you defaulted then?
        - If answer is no, you are good
        - If the answer is bad, you are bad
    - Question: If I am creating the model today and I have the bank's data of 3 months old, should I use the new data (e.g. new age)?
        - No, I must use the data at the time of application
        - It's target information leak which improve the model due to the wrong reason
    - During time between 0 and 12 months, I do not collect variables, I just look for the occurrence of default
        - If the person is in default 12 months into the future, we classify them as a default
            - This is an application scorecard
            - It is the most basic scorecard that we create
- A big problem: when building the statistical models, they are old the first day they are used
    - Since we are building with past credit history, and we need to wait for a year in order for people to build history so we can have the objective variable
    - Question: how old is the youngest loan in an application scoring database
        - One year back
        - When the world is in a different place than today

- Question: what is exactly a credit scoring model giving?
    - Although the output of a logistic regression looks like a probability, they are not probabilities
    - A scoring system in general gives you an ordered ranking of relative behavior
    - We need to do something else to get well calibrated probabilities
        - It's a challenge of using data science models

Behavioral Scoring

- Behavioral scoring vs. application scoring
    - While some variables in the application scoring could be identified as behavior, the behavioral scoring focuses exclusively on behavior of the operation that you already behave
        - Application scoring could include information from other loans that have been closed in the past
            - E.g. she was a customer of my company and they were a customer, they had these behaviors
        - Application scoring typically not include like average number of arrears in loans because you don't have the loans there
            - What you have is other past information
    - In behavioral scoring, the person has to have an active product with you
        - They have a loan that they are paying
        - They have a credit card that is generating installments every month
    - We define a customer with behavior once they have had at least 3/6 months of operations of ongoing payments with the bank
        - Before that, you are considered as an unknown customer
        - If you got nothing but a bank account, then you are not a customer that's showing credit behavior.
        - You got a credit card, you are building credit history
    - The difference between one and the other is the type of variables that we use and the quality of the scorecard
        - A behavioral scorecard system is a lot harder and longer to build, but it is a lot more accurate
- After 6 months of credit history
    - You not only have demographic variables have for the application scoring
    - Not only the variables come from the credit Bureau
    - You have a much stronger signal: actual repayment behavior
        - You know when they pay, how they pay, how late or early they pay
        - You know how the behavior has changed and iterated when compared to older behaviors
    - When you observe a deterioration in the repayment behavior
        - The provisions should start immediately climbing to cover for the increased risk
    - This implies that the variables in the model will be huge
        - Application scoring: 7-15 variables
        - Bureau scoring: 15 variables
        - Behavioral scoring: 10 – 30 variables

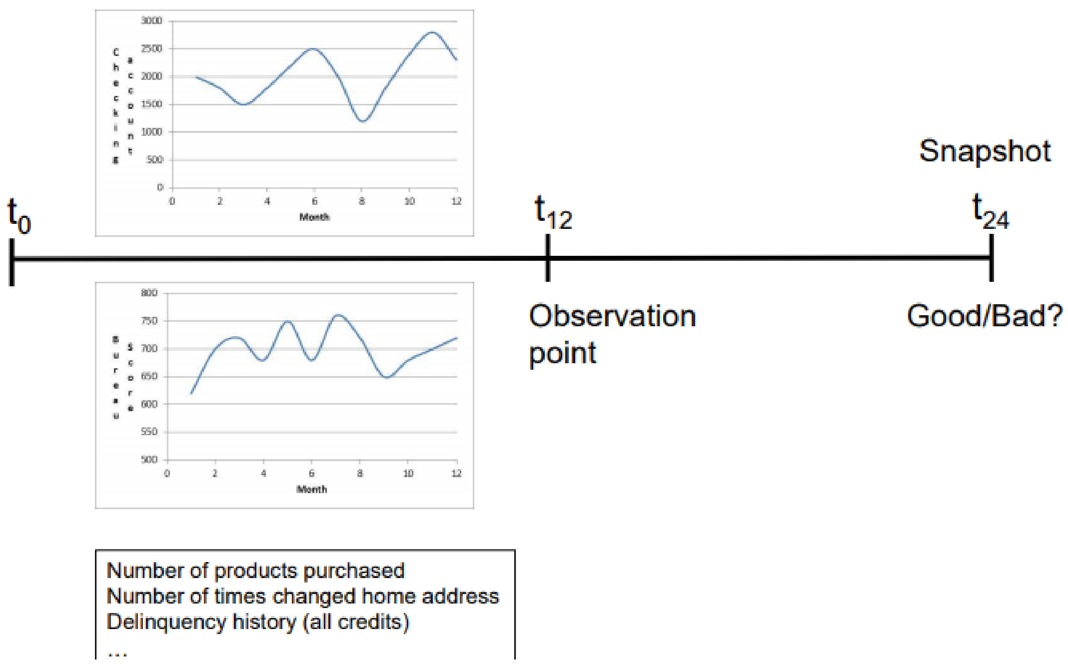


- Now you stop at a certain point (today is the observation point)
  - You observe back at performance window (from time 0 to today): 3-6 months
  - From today, I project forward 1 year to see if that person defaults or not
  - I do this by looking at every installment I have in my database

- Behavioral scoring will require a very strong sampling process to not bias my model
  - If my company is doing great, and keeps growing, the last five years of loans will contain more newer loans than older loans, this is usually fine
    - But when my company is staggering, then I need to sample effectively so I leave sufficient cases in my database for it to be balanced towards newer customers
  - If I have one record per installment, what will I have the most in my dataset, the longer term loans or shorter term loans?
    - Longer term loans, a loan that is 7 years will have 84 appearances in my dataset
    - So without proper sampling, it will overrepresent longer term loans
  - Given I am having more longer term loans, will I have more newer loans or more older loans?
    - Older loans, because they are there for more times
  - What banks usually do is to weight cases
    - Older loans weight less

- Behavioral scoring also has a huge problem that circumstances change and people do not tell their bank all the time when their circumstances change
  - So as behavioral scorecard has more variables, it has to deal with relatively old data
  - It got even worse for mortgages (except Canada, it forces your mortgage to expire every five years)
    - In Canada, if you got a mortgage with term 30 years, after 5 years, you eighter pay everything of renegotiate
      - It provides certain stability in the mortgage market because it reacts very quickly to the situation

        - If rates are low, people will probably refinance and extend term (looking to pay less interest)
        - If rates are high, people will probably put hand in their savings and prepay as much as they can to the mortgage (so looking to pay less interest)
    - That means in Canada, customers in your mortgage are not that old
    - If you have 40 years mortgage
        - You don't know the current condition of the old customers (e.g. where they are)
  - Since this is a regulatory question, regulators will say, they don't care, you need to have as much data as you can in your database
    - So you need to justify very well how to create your models so that your scoring system is correctly representing your customers
- Behavioral scoring the one of the most strategic models that a bank has because they use it for everything



- In the diagram, the observation point starts in 1 year, but usually, we don't wait that long to transform customers from application group into behavioral group
  - Because behavioral groups are a lot more accurate and the provisions required for behavioral groups are usually less than that of application score system
  - But not too early, because they won't have enough time to build history, which is the part that actually brings you profits
- Again, you need to wait for a year before start seeing results, the youngest installment paid was paid a year ago
- With Covid, credit scoring systems stopped working
  - Because none of the variables that traditionally apply, apply during this time period
    - The government support was many times above (part-time) minimum wage
  - How to fix it? Don't know, there is no history of this

- As mentioned, behavioral scoring is for everything
  - To set new limits
  - To decide new loans
  - To set limits to any renewals, to any collections
- There could be more than 1,000 candidate variables that need to do variable selection on
  - You need to define the variables that measure the performance
    - E.g. if someone is always one, two, or three days late, it means the person is relatively unwilling to pay this
- A behavioral scoring system is built at the company level, not at the produce level
  - It means that you will have one scorecard across all your products
- The best way to increase a model performance is getting better data, not using more sophisticated models

Aggregation Functions

- Trends are important
  - Is it going up?
  - Is it good (always pays on time) or is it bad (always late) or is it in the middle (goes up and down and up and down, …, this is what should be worried)
- The other important thing is that you need to build variables
  - One of the best ways to create new performance is creating variables
  - So start playing with ratios, start thinking what variables could be built, mixing other variables, should try them out if there is a good reason to do so
    - Remember that I need to justify those variables to the regulators
    - Need to explain why that variable chose a behavior
  - Don't be afraid to mix different services, different products

Building a behavioral Scorecard

- Step 1: taking monthly closing stock for the last five years at least
  - Closing stock: every loan that's active and not written off at the end of the month
  - This contains snapshots at the closing of the last five years, total of 50 portfolios
- Step 2: build the behavioral variables that represent the customers' behavior
- Step 3: sample from the snapshots so it represents correctly of all loans, and maybe apply weights
  - Calculate all the trends on variables
- Step 4: Train the model
  - Be extra careful with correlations and variable selection
  - Validate all variables that they make financial sense: explain them to the regulators
  - Validate scorecard
- Step 5: monitor performance to detect degradation of the model