INTRODUCTION TO CREDIT SCORING & DATA PREPARATION

Dr. Aric LaBarr

Institute for Advanced Analytics

INTRODUCTION TO SCORECARDS

What is a Scorecard?

- Common way of displaying the patterns found in a binary response model.
- Typically, people use logistic regression models.
- The main benefit is that a scorecard provides a clear and intuitive way of presenting the regression coefficients.

Scorecard Usage

these are just scorecards on logistic regression models.

- Credit Scoring
  - Equifax (http://www.equifax.com/home/en_us)
  - Experian (http://www.experian.com)
  - Transunion (http://www.transunion.com)
- Medicine / Healthcare
  - Trauma and Injury Severity Score (http://www.trauma.org/archive/scores/iss.html)
  - Coronary Heart Disease Risk Calculator (http://www.medcalc.com/heartrisk.html)
- Retail, IT and most cases where binary models can be applied.

CREDIT SCORING

Credit Scoring and Scorecards

- “One of the oldest applications of data mining, because it is one of the earliest uses of data to predict consumer behavior.”
- David Edelman – Credit Director of Royal Bank of Scotland

Credit Scoring and Scorecards

actual model itself

- **Credit scoring** is a statistical model that assigns a risk value to prospective or existing credit accounts.
- A **credit scorecard** is a statistical risk model that was put into a special format designed for ease of interpretation.
- Scorecards are used to make strategic decisions such as accepting/rejecting applicants and deciding when to raise a credit line, as well as other decisions.

layer put on top of model. interpretation of model. this is what we hand to regulators.

Credit Scoring and Scorecards

- The credit scorecard format is very popular and successful in the consumer credit world for a number of reasons:
  1. People at all levels within an organization generally find it easy to understand and use.
  2. Regulatory agencies are accustomed to credit risk models presented in this fashion.
  3. Credit scorecards are straightforward to implement and monitor over time.

regulators not used to looking at ML models, regulators not trained at statistical model, they trained to make sure no one is underserved.

Example 1 Scorecard

- Cut-off = 500
- New customer:
  - Months Since Last Miss Payment: 32
  - Home: OWN
  - Salary: $30,000

Variable	Level	Scorecard Points
MISS	$x < 24$	100
MISS	$24 \leq x < 36$	120
MISS	$36 \leq x < 48$	185
MISS	$x \geq 48$	200
HOME	OWN	225
HOME	RENT	110
INCOME	$x < 10{,}000$	120
INCOME	$10{,}000 \leq x < 25{,}000$	140
INCOME	$25{,}000 \leq x < 35{,}000$	180
INCOME	$35{,}000 \leq x < 50{,}000$	200
INCOME	$x \geq 50{,}000$	225

Example 1 Scorecard

- Cut-off = 500
- New customer:
  - Months Since Last Miss Payment: 32
  - Home: OWN
  - Salary: $30,000
- Total Points:

$$120 + 225 + 180 = 525$$

- **ACCEPT FOR CREDIT**

Variable	Level	Scorecard Points
MISS	$x < 24$	100
MISS	$24 \leq x < 36$	120
MISS	$36 \leq x < 48$	185
MISS	$x \geq 48$	200
HOME	OWN	225
HOME	RENT	110
INCOME	$x < 10{,}000$	120
INCOME	$10{,}000 \leq x < 25{,}000$	140
INCOME	$25{,}000 \leq x < 35{,}000$	180
INCOME	$35{,}000 \leq x < 50{,}000$	200
INCOME	$x \geq 50{,}000$	225

Example 2 Scorecard

- Cut-off = 500
- New customer:
  - Months Since Last Miss Payment: 22
  - Home: OWN
  - Salary: $8,000

Variable	Level	Scorecard Points
MISS	$x < 24$	100
MISS	$24 \leq x < 36$	120
MISS	$36 \leq x < 48$	185
MISS	$x \geq 48$	200
HOME	OWN	225
HOME	RENT	110
INCOME	$x < 10{,}000$	120
INCOME	$10{,}000 \leq x < 25{,}000$	140
INCOME	$25{,}000 \leq x < 35{,}000$	180
INCOME	$35{,}000 \leq x < 50{,}000$	200
INCOME	$x \geq 50{,}000$	225

Example 2 Scorecard

- Cut-off = 500
- New customer:
  - Months Since Last Miss Payment: 22
  - Home: OWN
  - Salary: $8,000
- Total Points:

$$100 + 225 + 120 = 445$$

- **DO NOT ACCEPT FOR CREDIT**

Variable	Level	Scorecard Points
MISS	$x < 24$	100
MISS	$24 \leq x < 36$	120
MISS	$36 \leq x < 48$	185
MISS	$x \geq 48$	200
HOME	OWN	225
HOME	RENT	110
INCOME	$x < 10{,}000$	120
INCOME	$10{,}000 \leq x < 25{,}000$	140
INCOME	$25{,}000 \leq x < 35{,}000$	180
INCOME	$35{,}000 \leq x < 50{,}000$	200
INCOME	$x \geq 50{,}000$	225

Example 2 Scorecard

- Cut-off = 500
- New customer:
  - Months Since Last Miss Payment: 22
  - Home: OWN
  - Salary: $8,000
- Total Points:

$$100 + 225 + 120 = 445$$

- **DO NOT ACCEPT FOR CREDIT**

Variable	Level	Scorecard Points
MISS	$x < 24$	100
MISS	$24 \leq x < 36$	120
MISS	$36 \leq x < 48$	185
MISS	$x \geq 48$	200
HOME	OWN	225
HOME	RENT	110
INCOME	$x < 10{,}000$	120
INCOME	$10{,}000 \leq x < 25{,}000$	140
INCOME	$25{,}000 \leq x < 35{,}000$	180
INCOME	$35{,}000 \leq x < 50{,}000$	200
INCOME	$x \geq 50{,}000$	225

Example 2 Scorecard

- Cut-off = 500
- New customer:
  - Months Since Last Miss Payment: 22
  - Home: OWN
  - Salary: $8,000
- Total Points:

$$100 + 225 + 120 = 445$$

- **DO NOT ACCEPT FOR CREDIT**

Variable	Level	Scorecard Points
MISS	$x < 24$	100
MISS	$24 \le x < 36$	120
MISS	$36 \le x < 48$	185
MISS	$x \ge 48$	200
HOME	OWN	225
HOME	RENT	110
INCOME	$x < 10{,}000$	120
INCOME	$10{,}000 \le x < 25{,}000$	140
INCOME	$25{,}000 \le x < 35{,}000$	180
INCOME	$35{,}000 \le x < 50{,}000$	200
INCOME	$x \ge 50{,}000$	225

Example 2 Scorecard

- Cut-off = 500
- New customer:
  - Months Since Last Miss Payment: 22
  - Home: OWN
  - Salary: \$8,000
- Total Points:

$$100 + 225 + 120 = 445$$

- **DO NOT ACCEPT FOR CREDIT**

Variable	Level	Scorecard Points
MISS	$x < 24$	100
MISS	$24 \leq x < 36$	120
MISS	$36 \leq x < 48$	185
MISS	$x \geq 48$	200
HOME	OWN	225
HOME	RENT	110
INCOME	$x < 10{,}000$	120
INCOME	$10{,}000 \leq x < 25{,}000$	140
INCOME	$25{,}000 \leq x < 35{,}000$	180
INCOME	$35{,}000 \leq x < 50{,}000$	200
INCOME	$x \geq 50{,}000$	225

Discrete because you paid it doesnt matter Jan 6 or Jan 23..Jan payment is done and made.

Question when ask are you gonna default is are you gonna pay me next month? I dont care if first or last day not gonna pay you. Point is are you gonna miss pmt?

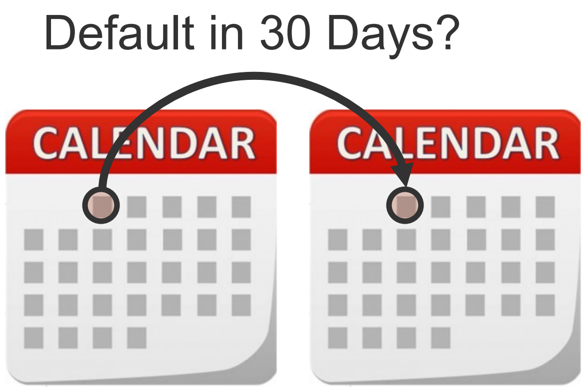
Discrete vs. Continuous Time

- Credit scoring typically tries to understand the probability of default on a customer (or business).
- However, default is also dependent on **time**.
- When will someone default? → **JUST AS IMPORTANT!**

- **Discrete** – Evaluating binary decisions on predetermined intervals of time.
- **Continuous** – Evaluating probability of default as it changes over continuous points in time (survival analysis).

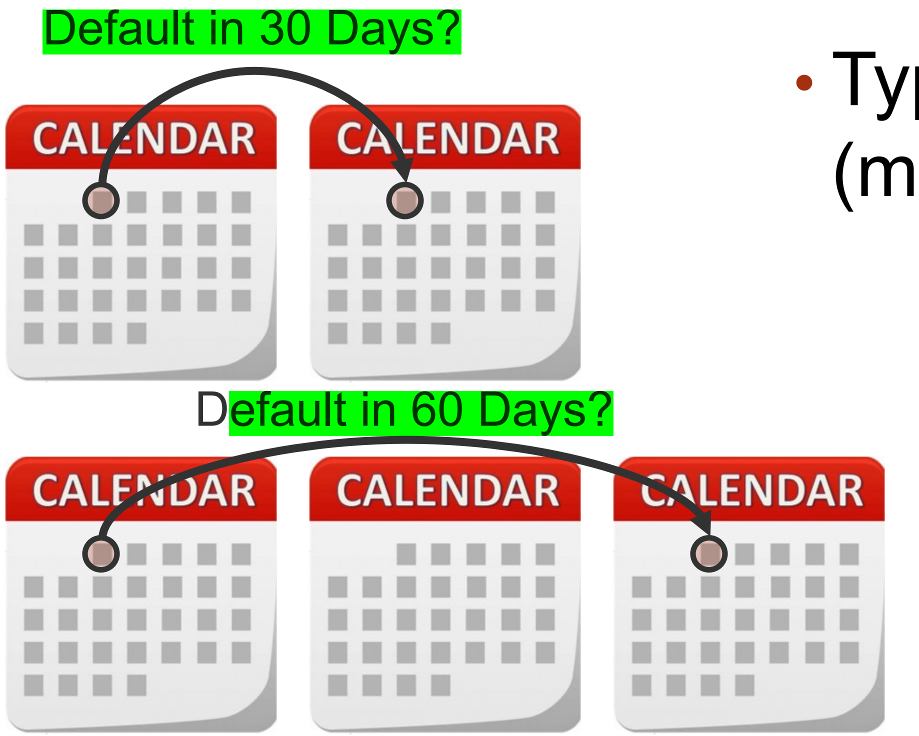
Discrete Time

Discrete Time



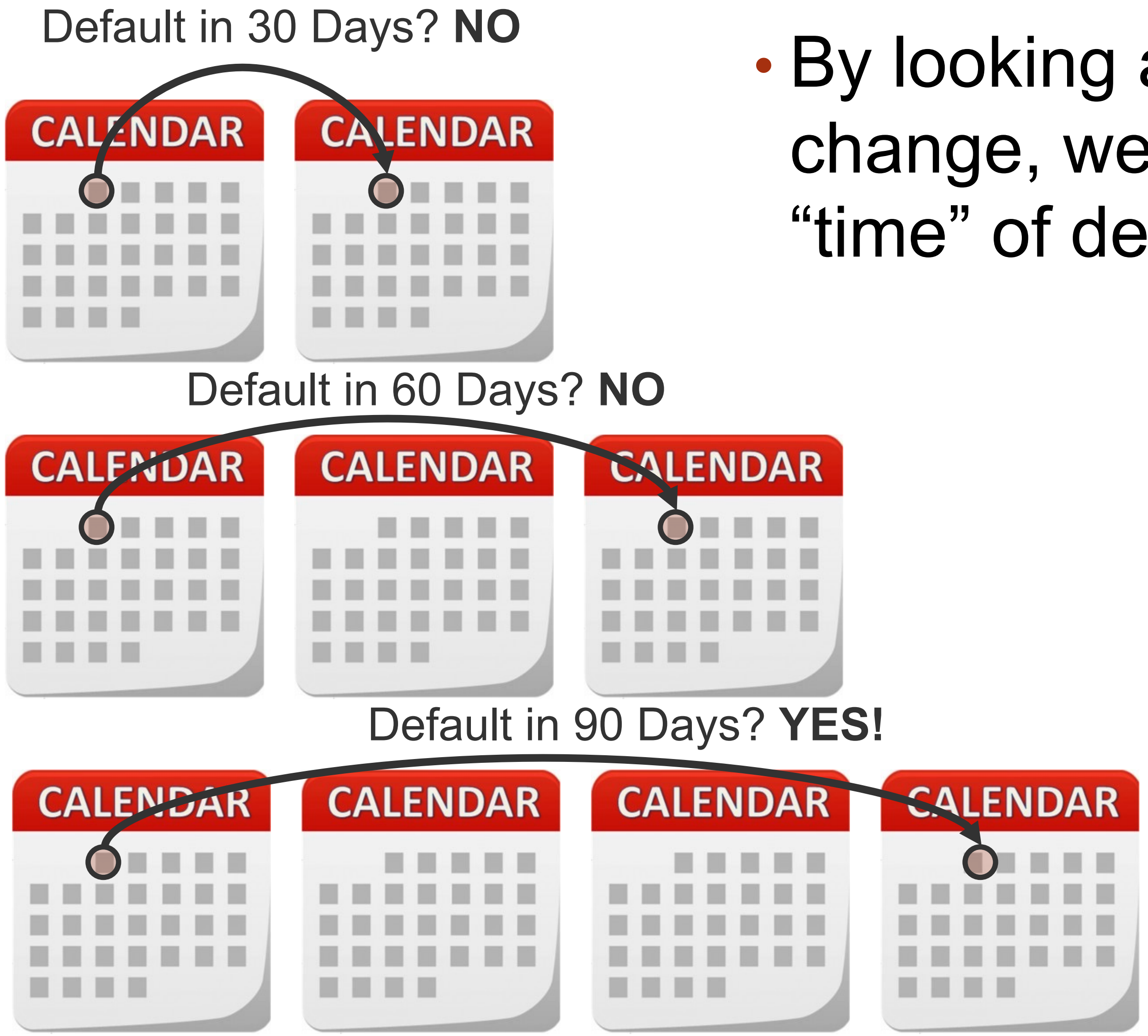
- Discrete time models evaluate the probability of default within a window of time.

Discrete Time



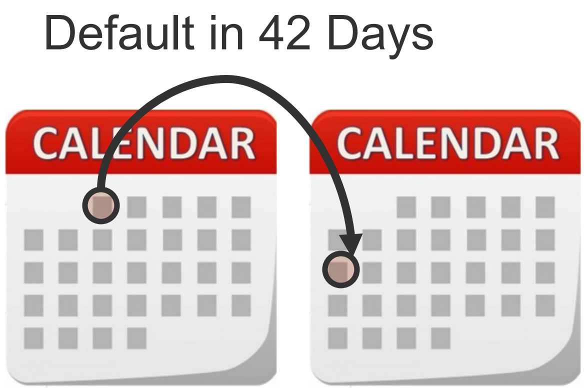
- Typically pick multiple windows (models) to evaluate across.

Discrete Time



- By looking at where decisions change, we can evaluate a “time” of default.

Continuous Time



- Continuous time models provide a probability of default for **every** day.
- From this more exact times of default are possible.

Process Flow

Data Collection
- Variable Selection
- Sample Size
- Sample / Performance Window

Data Cleaning
- Eliminate Duplicates
- Examine / Remove Outliers

Variable Grouping and Selection
- Weights of Evidence (WOE)
- Information Value (IV)
- Gini Criterion

Initial Scorecard Creation
- Logistic Regression
- Accuracy
- Threshold
- Assessment

Reject Inference
- Remove bias resulting from exclusion of rejects

Final Scorecard Creation
- Final Model Assessment

DATA DESCRIPTION

ACCEPTS Data Set

- Type of Product: Auto Loans
- Information available on customers with performing or non-performing loans.
- 5,837 cases of individuals who applied for and were granted an automobile loan.
- 22 variables in all.

Data Dictionary

Variable Name	Description
Age_oldest_tr	Age of oldest trade
App_id	Application ID
Bad	Good/Bad Loan
Bankruptcy	Bankruptcy or Not
Bureau_score	Bureau Score
Down_pyt	Amount of down payment on vehicle
Loan_amt	Amount of Loan
Loan_term	How many months vehicle was financed
Ltv	Loan to Value
MSRP	Manufacturer suggested retail price
Purch_price	Purchase price of vehicle

Variable Name	Description
Purpose	Lease or own
Rev_util	Revolving utilization (balance/credit limit)
Tot_derog	Total number of derogatory trades (go DPD)
Tot_income	Applicant's income
Tot_open_tr	Number of open trades
Tot_rev_debt	Total revolving debt
Tot_rev_line	Total revolving line
Tot_rev_tr	Total revolving trades
Tot_tr	Total number of trades
Used_ind	Used car indicator
Weight	Weight variable

REJECTS Data Set

- Type of Product: Auto Loans
- 4,233 cases of individuals who applied for and were **NOT** granted an automobile loan.
- 21 variables in all – BAD variable not part of data set and should be inferred.
- Used for **reject inference** later in the analysis.

Reject Inference

- **Reject inference** is the process of inferring the status of the rejected applicants based on the accepted applicants model in an attempt to use their information to build a scorecard that is representative of the entire applicant population.
- Reject inference is about solving sample bias so that the development sample is similar to the population to which the scorecard will be applied.

Rejected Inference

- Can we develop a scorecard without rejected applications?

technically can still build scorecard wihtout rejected, legally permissible currently, but legislature in place to push through against it. Should be using entire applicant pool not just good candidate data set other get bias.

Weight assignment not work here cuz weight assigned on what data set seen by model (need model to see it ALL).

Rejected Inference

- Can we develop a scorecard without rejected applications? **YES!**
- Is it **legally permissible** to develop a scorecard without rejected applications?

Rejected Inference

- Can we develop a scorecard without rejected applications? **YES!**
- Is it **legally permissible** to develop a scorecard without rejected applications? **YES!**
- If yes, then how **biased** would the scorecard model be?

Rejected Inference

- Can we develop a scorecard without rejected applications? **YES!**
- Is it **legally permissible** to develop a scorecard without rejected applications? **YES!**
- If yes, then how **biased** would the scorecard model be? **DEPENDS!**

- *“My suggestion is to develop the scorecard using what data you have, but start saving rejected applications ASAP.”*
  Raymond Anderson, Head of Scoring at Standard Bank Africa, South Africa

DATA COLLECTION AND CLEANING

Process Flow

Data Collection

- Variable Selection
- Sample Size
- Sample / Performance Window

Data Cleaning

- Eliminate Duplicates
- Examine / Remove Outliers

Variable Grouping and Selection

- Weights of Evidence (WOE)
- Information Value (IV)
- Gini Criterion

Initial Scorecard Creation

- Logistic Regression
- Accuracy
- Threshold
- Assessment

Reject Inference

- Remove bias resulting from exclusion of rejects

Final Scorecard Creation

- Final Model Assessment

Defining Our Target

decide default on loan based on how many payments missed. Miss 1 month out of 2 month loan. But miss 1 month out of 5y loan, doesnt not mean defaulted.

Banking has 90 dpd - days passed origiinal payment date. equates 3 pmts. Nowadays they have varying degrees of past due pmt. Bank write off loan after 6 months past due.

- When does someone actually default?
  - Is it when the loan is charged-off?
  - Probably signs of stopped paying before then

- Need to define target variable
  - 90 days past due (DPD) for everything (old approach)
  - 90-180 DPD based on types of loans, business sector, country regulations, etc. (current approach)
    - For example: US mortgages – 180 DPD

Variable Selection

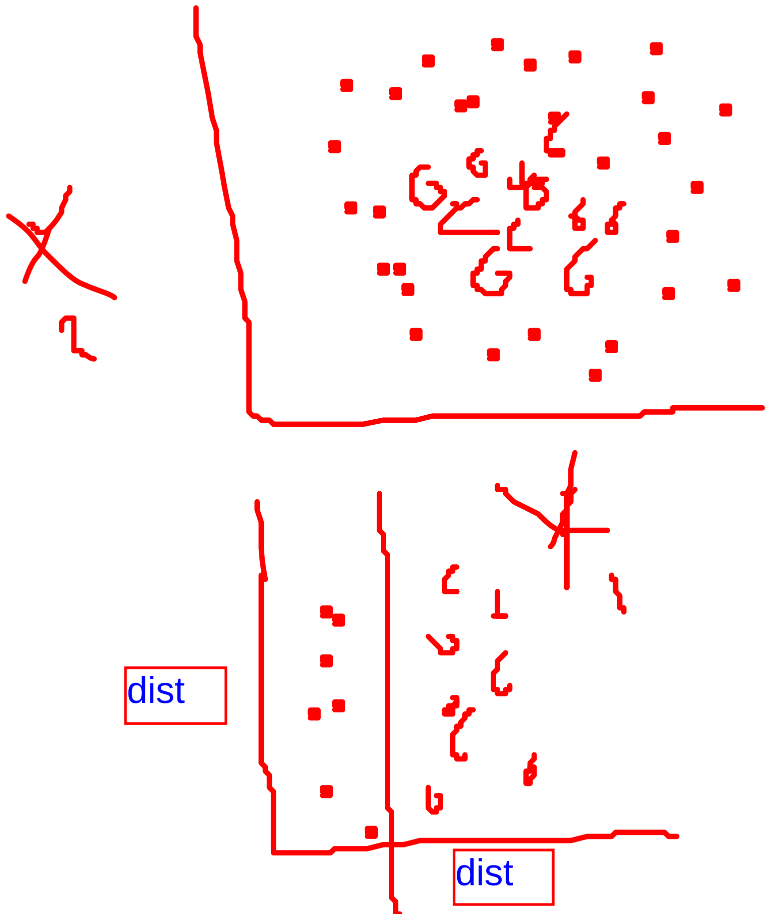
- Criteria for explanatory variables:
  - Expected predictability power
  - Business interpretation
  - Reliability
  - Legal issues
  - Ease in collection
  - Future availability

Feature Engineering

- Variable creation based on business reasoning:
  - Loan to value ratio
  - Number of delinquent accounts
  - Expense to income ratio
  - Credit line utilization
- Omit variables that are highly dependent:
  - Variable clustering!
- Review / remove outlier and abnormal values

Variable features will always make a model better comapred to technique. They beat technique.

logistic regression is a linear separator will be worse for if data was like this. Below can feature engineer distance from centre, then data look like this 2nd graph below:



Rich ppl doesnt mean they will pay it back cuz they take higher amt loan default prob could be same. Look at loan to income ratio, these features engineered make it better.

Still have to do multi coll.

Sample Size

- *"There are no hard and fast rules, but the sample selected normally includes at least 1,000 good, 1,000 bad, and about 750 rejected applicants."* FDIC, Credit Card Activities Manual (https://www.fdic.gov/regulations/examinations/credit_card/index.html)
- No exact answer on the correct sample size.
- Sample size depends on the overall size of the portfolio, the number of explanatory variables you are planning to use, and the number of defaults or claims filled.

Sample and Performance Window

- The sample must be characteristic of the population to which the scorecard will be applied.
- Example:
  - If the scorecard is to be applied in the subprime lending program, then use a sample that captures the characteristics of the subprime population targeted.

Sample and Performance Window

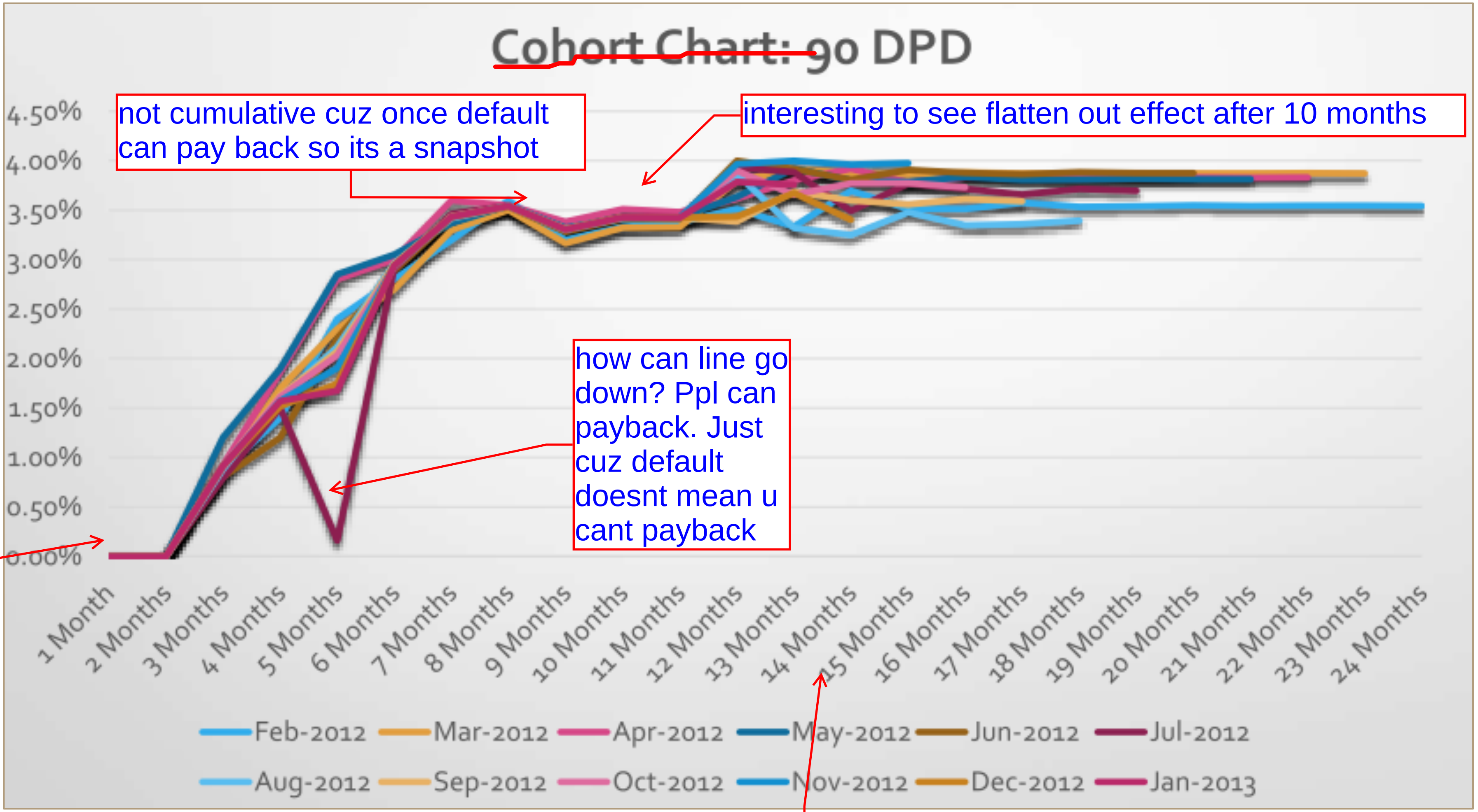
- Objective:
  - Gather data for accounts opened during a specific time frame.
  - Monitor their performance for another specific length of time to determine if they were good or bad.
- Problems:
  - Accounts opened recently are more similar to accounts that will be opened in the near future.
  - Want to minimize the chances of misclassifying performance – accounts must be monitored long enough to not underestimate expected bad rates.

how many ppl in default scneario over TIME

Sample and Performance Window

cohort chart each line represent cohort of people given loan a given month. We are using this to figure how long to give someone before they default ie become problem. "Cut off lenght of time"

Y axis is default rate. What % of ppl defaulted at that month out on x axis. This is defined by 180days



not cumulative cuz once default can pay back so its a snapshot

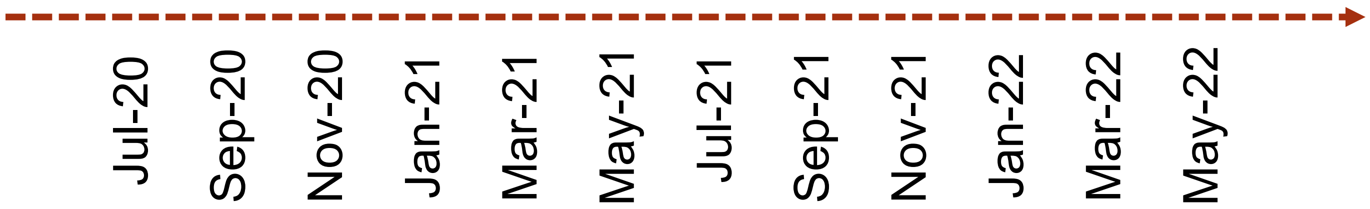
interesting to see flatten out effect after 10 months

how can line go down? Ppl can payback. Just cuz default doesnt mean u cant payback

havent had time to default After 3 months defaulted ie given loan but never pmt

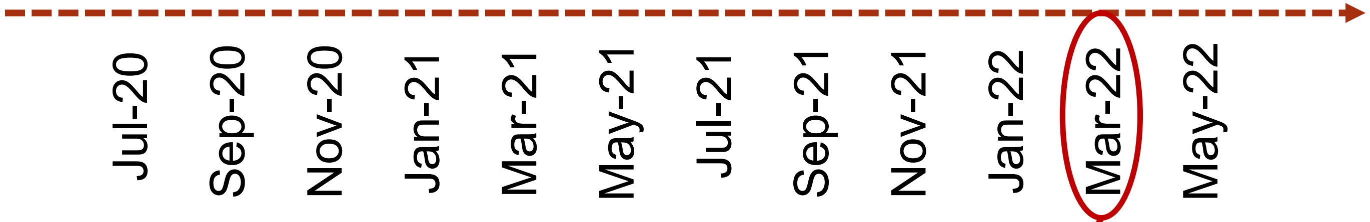
X axis is #of months since they given loan. Think survival analysis. tenure was same between 2 ppl but physical time wasnt same. Same idea here.

Sample and Performance Window



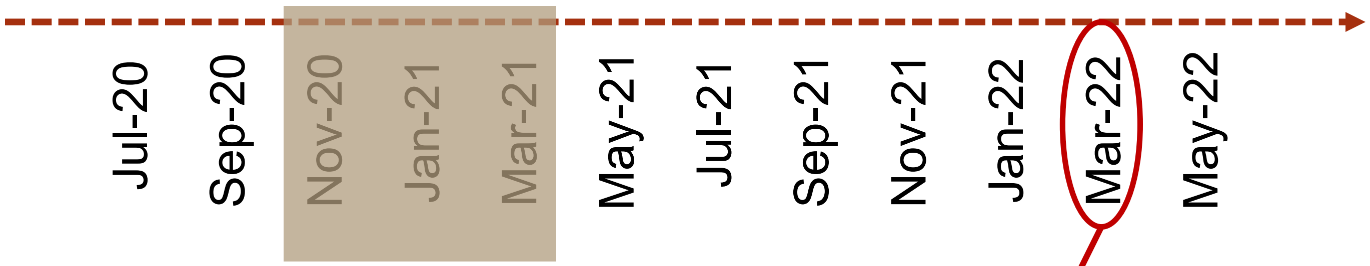
- Based on cohort graph: “Bads” level off about 14 months after loan origination.
- If the analysis is to be performed on March 2022, we select our **sample** from 12-16 months back; this will give an average of 14 months **performance window**.

Sample and Performance Window



- Based on cohort graph: “Bads” level off about 14 months after loan origination.
- If the analysis is to be performed on March 2022, we select our **sample** from 12-16 months back; this will give an average of 14 months **performance window**.

Sample and Performance Window

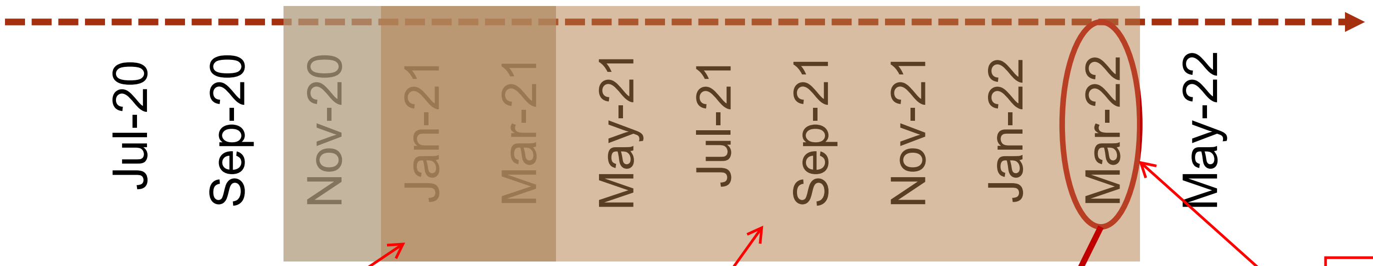


sample window where i get ppl from. Start 12-14 months back

- Based on cohort graph: “Bads” level off about 14 months after loan origination.
- If the analysis is to be performed on March 2022, we select our sample from 12-16 months back; this will give an average of 14 months **performance window**.

Sample and Performance Window

window changes based on recession time etc etc



sample window where get obs from

performance window. Waiting long enough to see default based on

cut off, 1 and 0s at this point

- Based on cohort graph: “Bads” level off about 14 months after loan origination.
- If the analysis is to be performed on March 2022, we select our **sample from 12-16 months back**; this will give an average of 14 months **performance window**.

if u buy sth too expensive for you, you will quickly if you default. Credit card tell 1-2 years if defaulting. 30y mortgage within 5y.

Sample and Performance Window

- The exact length of the performance window depends on the product.
  - Credit Cards: Typically 1 – 2 years
  - Mortgages: Typically 3 – 5 years
- Sample window length can vary based on data availability as well.