

# Project Steps

- Data Preparation:
  - Variable Encoding on Train/Validation/Test sets (See slide 2 for example)
  - Feature “State” -> OneHotEncoding
- Feature Selection
  - Run exhaustive search for feature selection using Logistic Regression model
  - Check for multicollinearity using VIF method
- Modeling (LR)
  - Run Logistic Regression on selected features
  - Check for fairness and do debiasing if needed
  - Report weights and confusion matrix
- Modeling (RF)
  - Run Random Forest on the selected features
  - Check for fairness and do debiasing if needed
  - Report feature importance and confusion matrix
  - (If time permits) Run RF on all features, and find overlapping features with LR model
- Model Selection
  - Choose between RF and LR based on Accuracy/Fairness trade-off
- Investigate “bank\_xyz” treatment
  - Answer the related question accordingly.
- Describe the rejection scenario
  - We use contrastive explanation for that.
- (If Time Permits) create a simple API for reporting the credit
- Writing Report and creating slides

- All predictors' values should be encoded into numbers 1,2,3,4 and 5. This can be done via percentiles.
  - If any predictors have NaN values, number "0" should be assigned.
  - Variable "ind\_acc\_XYZ" should be remained untouched (0,1).
  - Variable "States" should be one hot encoded.
  - Variable "Income" should be encoded within corresponding State.

Dataset 1

P1	P2	P3	Ind_acc_XYZ	isAZ	isNC	...	Default_ind
1	2	3	0	0	1	...	1
2	4	2	1	0	0		0
5	1	1	0	1	0		0

Dataset 2

P1	P2	P3	Ind_acc_XYZ	isAZ	isNC	...	Num_Defaulted	Num_Acc
1	2	3	0	0	1	...	38	90
2	4	2	1	0	0		58	120
5	1	1	0	1	0		90	200

# Project Steps

- Data Preparation: [\(Kusha or Nikhil\)](#)
  - Variable Encoding on Train/Validation/Test sets (See slide 2 for example)
  - Feature “State” -> OneHotEncoding
- Feature Selection [\(Mohammad\)](#)
  - Run exhaustive search for feature selection using Logistic Regression model
  - Check for multicollinearity using VIF method
- Modeling (LR) [\(Mohammad\)](#)
  - Run Logistic Regression on selected features
  - Check for fairness and do debiasing if needed
  - Report weights and confusion matrix
- Modeling (RF) [\(Mohammad\)](#)
  - Run Random Forest on the selected features
  - Check for fairness and do debiasing if needed
  - Report feature importance and confusion matrix
  - (If time permits) Run RF on all features, and find overlapping features with LR model [\(Kusha or Nikhil\)](#)
- Model Selection [\(Mohammad\)](#)
  - Choose between RF and LR based on Accuracy/Fairness trade-off
- Investigate “bank\_xyz” treatment [\(Mohammad\)](#)
  - Answer the related question accordingly. [\(Kusha or Nikhil with the help of Mohammad\)](#)
- Describe the rejection scenario [\(Mohammad\)](#)
  - We use contrastive explanation for that.
- (If Time Permits) create a simple API for reporting the credit [\(Kusha or Nikhil\)](#)
- Writing Report and creating slides [\(Mohammad, Kusha and Nikhil\)](#)

# Variables Encoding (Complete List)

5 Bins	5 Bins with missing (Bin 0)	Untouched Vars	One Hot Encoding
tot_credit_debt	rep_income (binnig per States)	Default_ind	States
avg_card_debt	uti_card_50plus_pct	ind_acc_XYZ	
credit_age		auto_open_36_month_num	
credit_good_age		card_open_36_month_num	
card_age		mortgages_past_due_6_months_num	
uti_max_credit_line		non_mtg_acc_past_due_6_months_num	
credit_past_due_amount		non_mtg_acc_past_due_12_months_num	
inq_12_month_num			
card_inq_24_month_num			
uti_card			
uti_50plus_pct			