Machine Learning project

Micro credit project:

-Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan.

Enter

kaggle

- Customer's geographical, personal, Financial Information & HomeOwnership details

- Quote for every customer

Customer's **likelihood** of buying that Insurance contract

Data shared: Training

	A	B	C	D
1	Variable	Definition	Comment	
2	label	Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}		
3	msisdn	mobile number of user		
4	aon	age on cellular network in days		
5	daily_decr30	Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)		
6	daily_decr90	Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)		
7	rental30	Average main account balance over last 30 days	Unsure of given definition	
8	rental90	Average main account balance over last 90 days	Unsure of given definition	
9	last_rech_date_ma	Number of days till last recharge of main account		
10	last_rech_date_da	Number of days till last recharge of data account		
11	last_rech_amt_ma	Amount of last recharge of main account (in Indonesian Rupiah)		
12	cnt_ma_rech30	Number of times main account got recharged in last 30 days		
13	fr_ma_rech30	Frequency of main account recharged in last 30 days	Unsure of given definition	
14	sumamnt_ma_rech30	Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)		
15	medianamnt_ma_rech30	Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)		
16	medianmarechprebal30	Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)		
17	cnt_ma_rech90	Number of times main account got recharged in last 90 days		
18	fr_ma_rech90	Frequency of main account recharged in last 90 days	Unsure of given definition	
19	sumamnt_ma_rech90	Total amount of recharge in main account over last 90 days (in Indian Rupee)		
20	medianamnt_ma_rech90	Median of amount of recharges done in main account over last 90 days at user level (in Indian Rupee)		
21	medianmarechprebal90	Median of main account balance just before recharge in last 90 days at user level (in Indian Rupee)		
22	cnt_da_rech30	Number of times data account got recharged in last 30 days		
23	fr_da_rech30	Frequency of data account recharged in last 30 days		
24	cnt_da_rech90	Number of times data account got recharged in last 90 days		

Binary classification

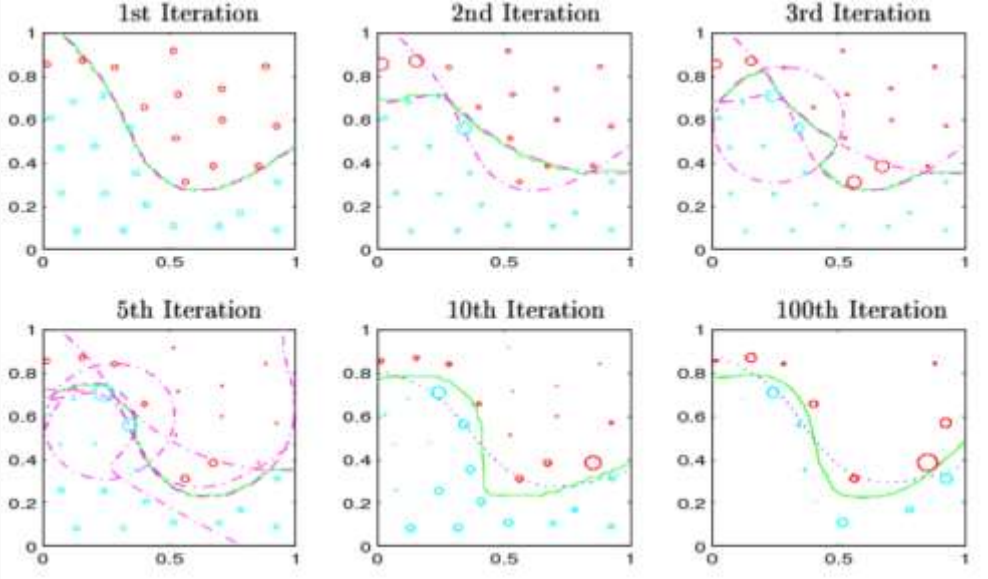
nts

Data cleaning steps

1. Removing Constants
2. Removing Identifier rows
3. Synthesizing Date column
4. Treating NA variables
5. Treating bad levels (-1)
6. Treating false categorical
7.

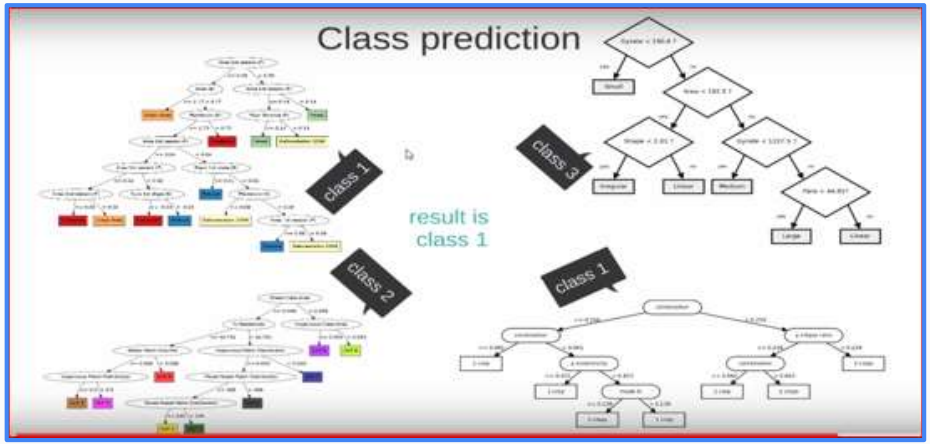
Gradient Boosting (Iterative corrections)

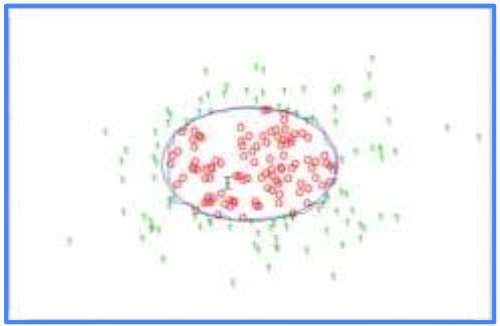
- Learning from past mistakes
- Could get nearly 0 training error
- Weighted scoring of multiple trees
- Hard to tune, as there are too many parameters to adjust
- Often overfit and hard to decide the stopping point



Random Forests (Majority wins)

- Handles missing data
- Handles redundancy easily
- Reduces variations in results
- Produces Out of Bag error rate
- Produces De-correlated trees
- Random subspace & split





Bias sometimes Increases as Trees are shallower

Gradient Boosting + Random Forest

- Handles missing data
- Handles redundancy easily
- Reduces variations in results
- Produces Out of Bag error rate
- Produces De-correlated trees
- Random subspace & split
- Does not overfit
- Little bias, due to correction
- Easy to tune

Our Score
AUC = .95

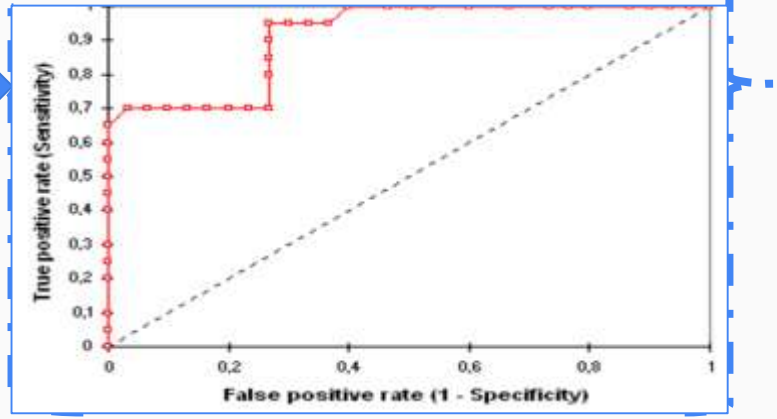
Quite slow & Computational expensive, optimizing these constraints could be an excellent area for research

Calculating AUC

ID	True class	Predicted probability
1	1	.8612
2	0	.2134
3	0	.1791
4	0	.1134
5	1	.7898
6	0	.0612

AUC

- Randomly decide a threshold
- Calculate True Positive Rate (y) & False Positive Rate (x)
- Based on (x,y) plot the point
- Repeat steps for each value of threshold [0,1]
- We now have a curve and we call it **ROC**
- **Area under this curve becomes AUC**



What we have already employed

- Categorical to Continuous conversion
- Continuous to Ordinal conversion
- Variable bucketing
- SVM / Logistic Regression
- Random Forest/ Trees
- Lasso / Ridge / Elastic Net
- Gradient Boosting
- Multicollinearity elimination
- Outlier treatment
- K-Fold Cross validation

What we look forward to use

- Imputation for NA's
- Model tuning
- Variable transformation
- Most importantly, **Your Suggestio**

THANK YOU