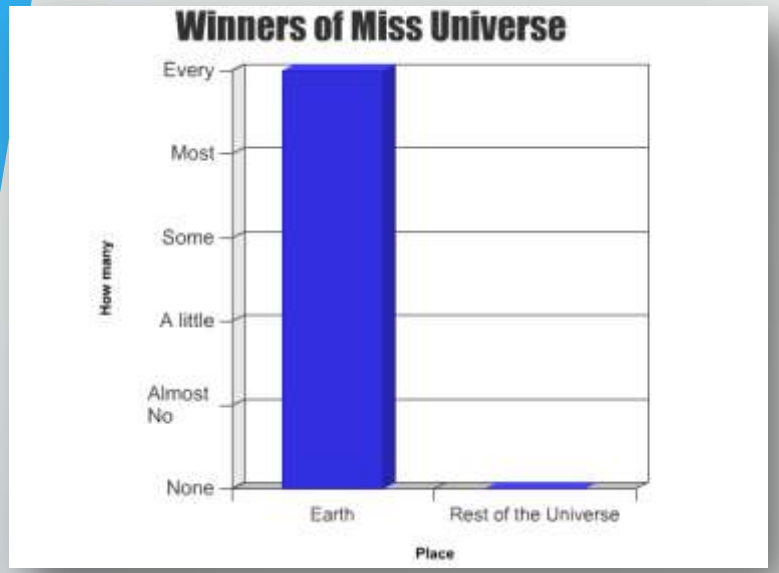
Introduction to Predictive Modeling

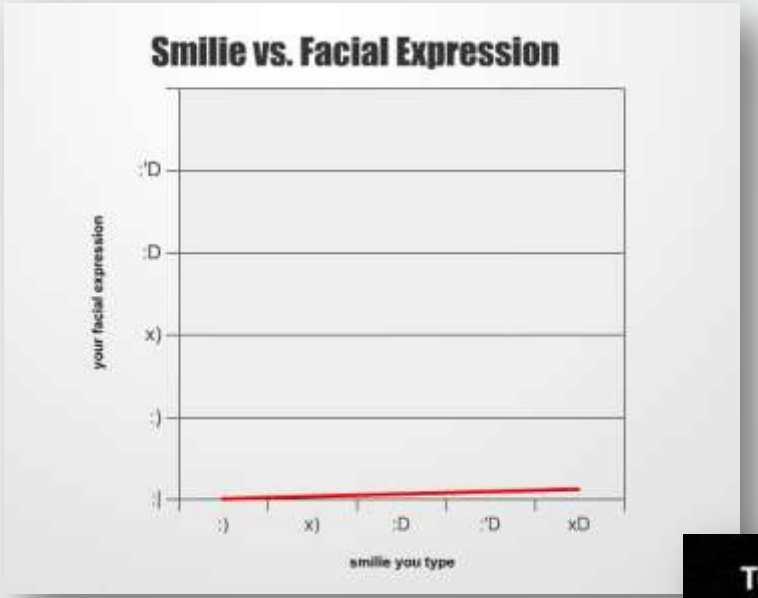
Venkat Reddy

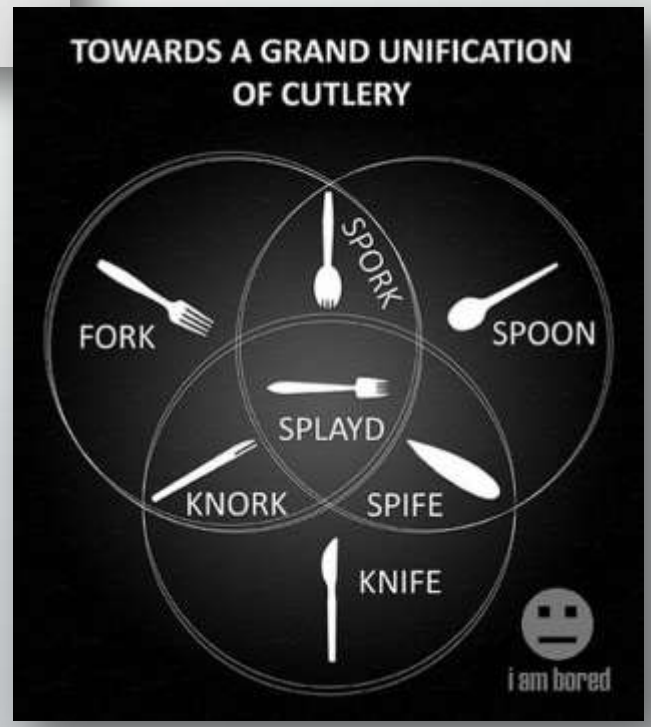
Contents

- Fancy visualizations to Predictive Modeling
- The Business Problem
- What is Predictive Modeling
- The Horse Race Analogy
- Credit Risk Model Building
- Other Applications of Predictive Modeling

Data Visualizations







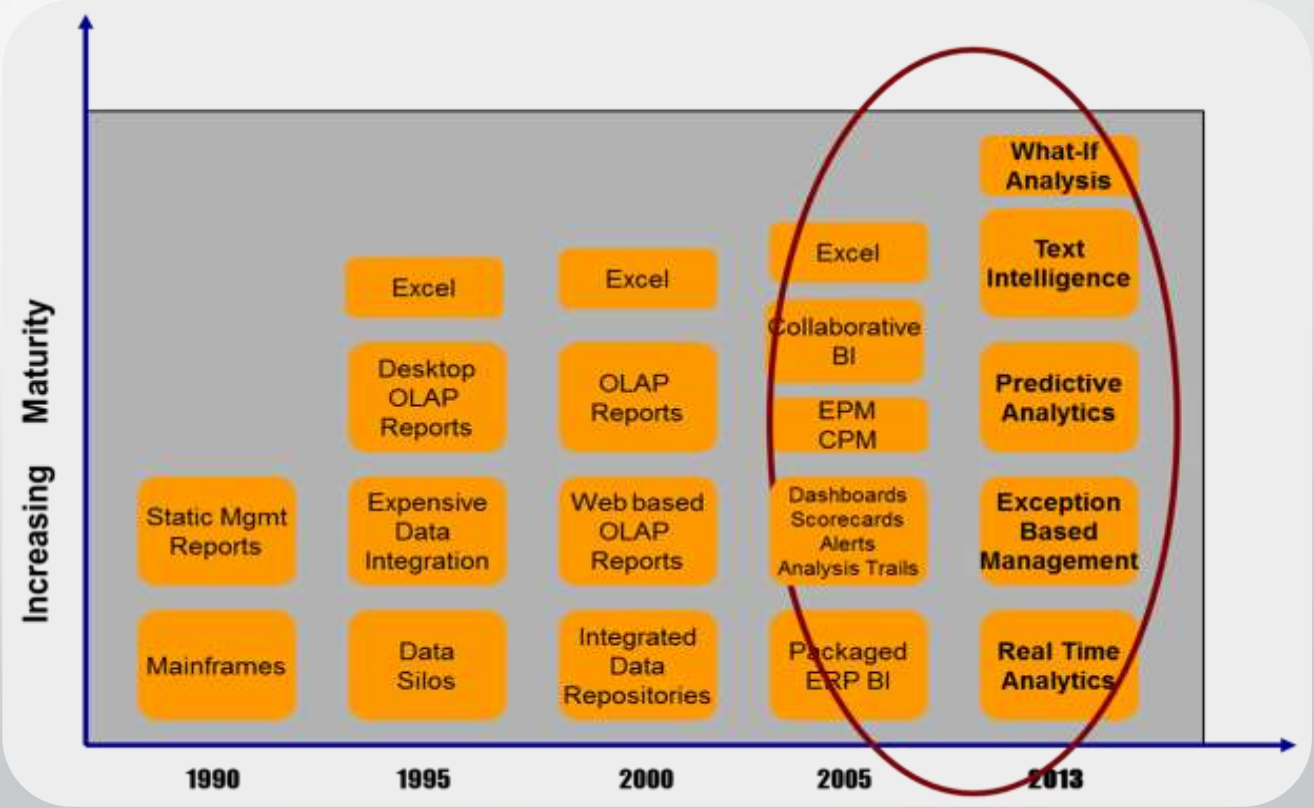
Data Visualizations

Data visualizations can help you make the right decisions quickly… that will ultimately lead to success.

Data Analysis & Predictive Modeling

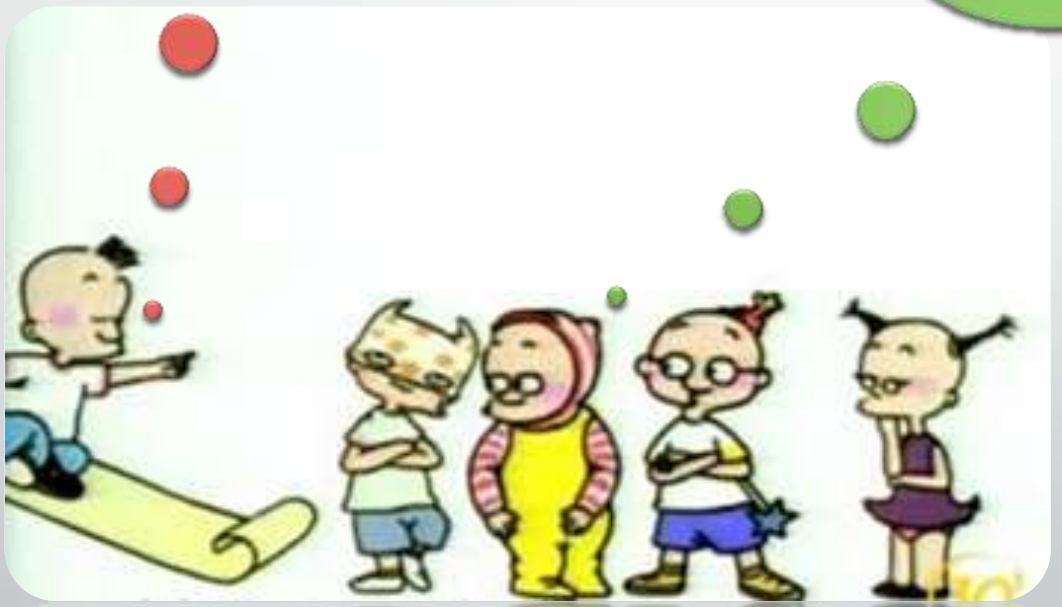
Increasingly, business rely on intelligent tools and techniques to analyze data systematically to improve decision-making.

- Retail sales analytics
- Financial services analytics
- Telecommunications
- Supply Chain analytics
- Transportation analytics
- Risk & Credit analytics
- Talent analytics
- Marketing analytics
- Behavioral analytics
- Collections analytics
- Fraud analytics
- Pricing analytics



Data analytics rocks & Predictive modeling is going to rule!!!

Dude we all know that

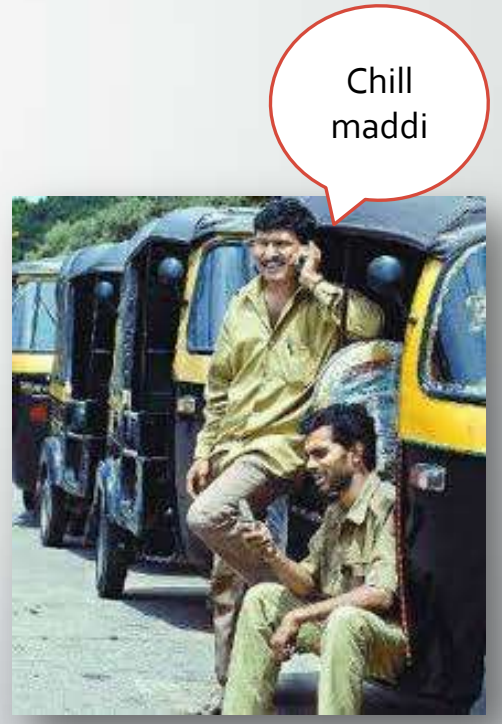


Business Problem

- Credit Cards & Loans- How does bank approve or decline CC and Loan applications
- **Citi Bank :** Present in more than 90 countries.
- More than **100,000** customers apply for credit cards/loans every month
- How do they approve or reject credit card applications?

How does bank approve or decline CC and Loan applications?

- Approve if the applicant is an auto driver in Bangalore
- Approve if the applicant's boy friend / girl friend is rich
- Approve if the applicant is a u-19 cricketer
- Decline if the applicant is a software professional from Bangalore



Any suggestions…

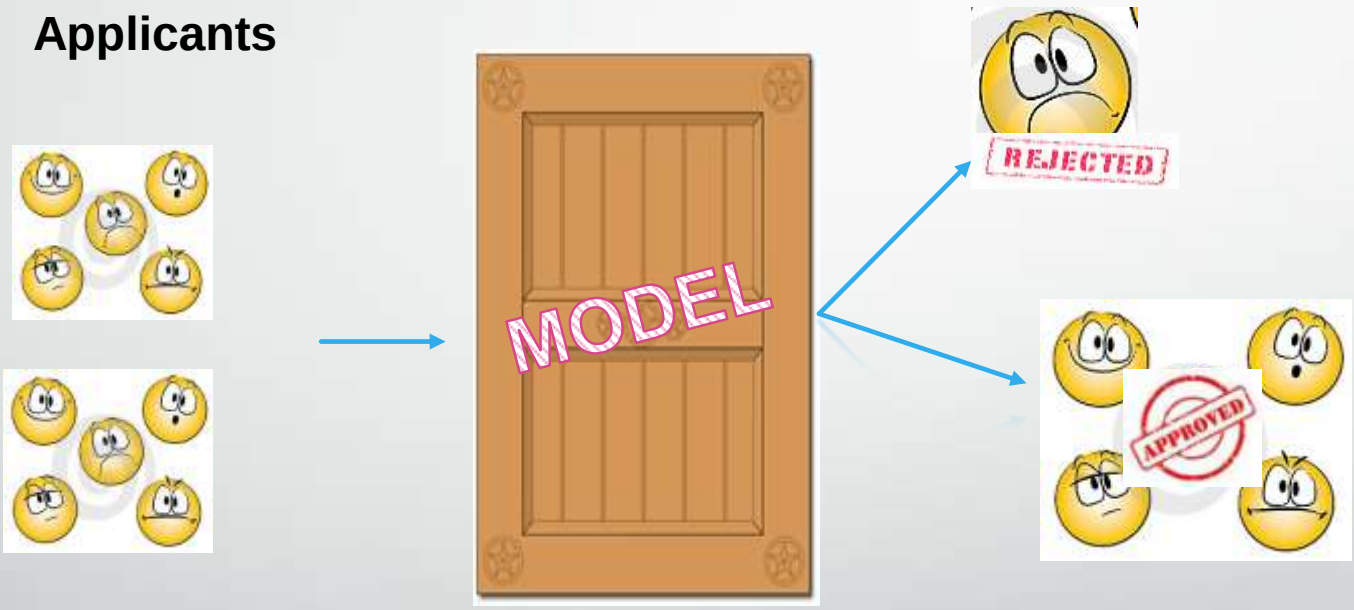
The Problem
Who will run away with my money?

- **Citi Bank :** Present in more than 90 countries.
- More than **100,000** customers apply for credit cards/loans every month
- All of them have different **characteristics**
- Out of 100,000 customers, who all have the higher probability of default/ Charge off?
- Basically, who will run away with my money?
- We need to predict the probability of "Running away"

How to predict?

We have historical data and we have statistics

Bank builds a model that gives a score to each customer



"Developing set of equations or mathematical formulation to forecast future behaviors based on current or historical data."

Predictive Modeling
Lets try to understand predictive modeling

Predictive Modeling – Fitting a model to the data to predict the future.

- Predicting the future –and it is so easy some times
- Who is going to score more runs in IPL-2014?
- That's it you predicted the future..
- BTW how did you predict?
- Predicting the future based on historical data is nothing but Predictive modeling

Predictive Modeling
Lets try to understand predictive modeling

Predictive Modeling – Fitting a model to the data to predict the future.

- Who is going to score more runs in IPL 2014?
- Predicting the future …well it is not that easy …

Predictive Modeling
Horse Race Analogy

How to bet on best horse in a horse race

The Historical Data
Win vs. Loss record in past 2 years

- **Long legs:** 75% (Horses with long legs won 75% of the times)
- **Breed A**: 55%, **Breed B**: 15 % **Others** : 30%
- **T/L (Tummy to length) ratio** <1/2 :75 %
- **Gender**: Male -68%
- **Head size:** Small 10%, Medium 15% Large 75%
- **Country**: Africa -65%

Given the historical data
Which one of these two horses would you bet on?

	Kalyan	Chethak
Length of legs	**150 cm**	110 cm
Breed	**A**	F
T/L ratio	**0.3**	0.6
Gender	**Male**	Female
Head size	**Large**	Small
Country	India	India

Given the historical data
Which one would you bet on….now?

	Kalyan	Chethak
Length of legs	110 cm	**150 cm**
Breed	C	**A**
T/L ratio	**0.45**	0.60
Gender	**Male**	Female
Head size	Small	**Large**
Country	**Africa**	India

Given the historical data
What about best one in this lot?

	Horse-1	Horse-2	Horse-3	Horse-4	Horse-5	Horse-6	Horse-7	Horse-8	Horse-9	Horse-10
Length of legs	109	114	134	130	149	120	104	117	115	135
Breed	C	A	B	A	F	K	L	B	C	A
T/L ratio	0.1	0.8	0.5	1.0	0.3	0.3	0.3	0.6	0.7	0.9
Gender	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Head size	L	S	M	M	L	L	S	M	L	M
Country	Africa	India	Aus	NZ	Africa	Africa	India	India	Aus	Africa

Citi has a similar problem?
Who is going to run away with my money?

- Given Historical of the customers we want to predict the probability of bad
- We have the data of each customer on
  - Customer previous loans, customer previous payments, length of account credit history, other credit cards and loans, job type, income band etc.,
- We want to predict the probability of default

Credit Risk Model Building-Four main steps

1. Study historical data
   - What are the causes(Customer Characteristics)
   - What are the effects(Charge off)

2. Identify the most impacting factors

3. Find the exact impact of each factor(Quantify it)

4. Use these coefficients for future

The Historical Data of Customers

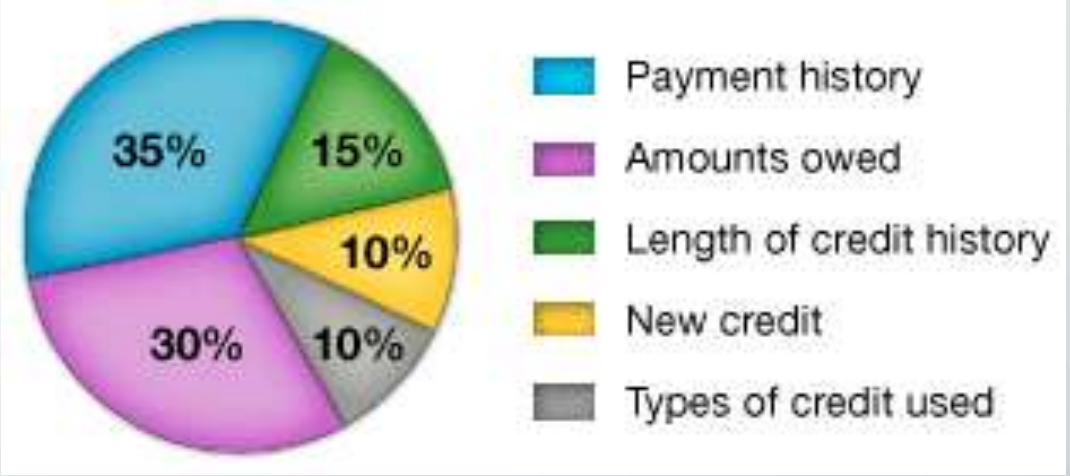
- Contains all the information about customers
- Contains information across more than 500 variables
- Portion of data is present in the application **form**
- Portion of the data is available with **bank**
- Lot of data is maintained in **bureau**
  - Social Security number –in US
  - PAN Number in India

Attribute	Value
SSN	*111259005*
Age	*27*
Number of dependents	*2*
Number of current loans	*1*
Number of credit cards	*1*
Number Installments 30days late in last 2 years	*4*
Average utilization % in last 2 years	*30%*
Time since accounts opened	*60 months*
Number of previous applications for credit card	*2*
Bankrupt	***NO***

The Historical Data of Customers: Types of variables

Payment History

- **Account payment** information on specific types of accounts (credit cards, retail accounts, installment loans, finance company accounts, mortgage, etc.)
- Presence of adverse public records (bankruptcy, judgements, suits, liens, wage attachments, etc.), collection items, and/or delinquency (past due items)
- Severity of delinquency (how long past due)
- Amount past due on delinquent accounts or collection items
- Time since (recency of) past due items (delinquency), adverse public records (if any), or collection items (if any)
- Number of past due items on file
- Number of accounts paid as agreed



More variables

Amounts Owed

- Amount owing on accounts
- Amount owing on specific types of accounts
- Lack of a specific type of balance, in some cases
- Number of accounts with balances
- Proportion of credit lines used (proportion of balances to total credit limits on certain types of revolving accounts)
- Proportion of installment loan amounts still owing (proportion of balance to original loan amount on certain types of installment loans)

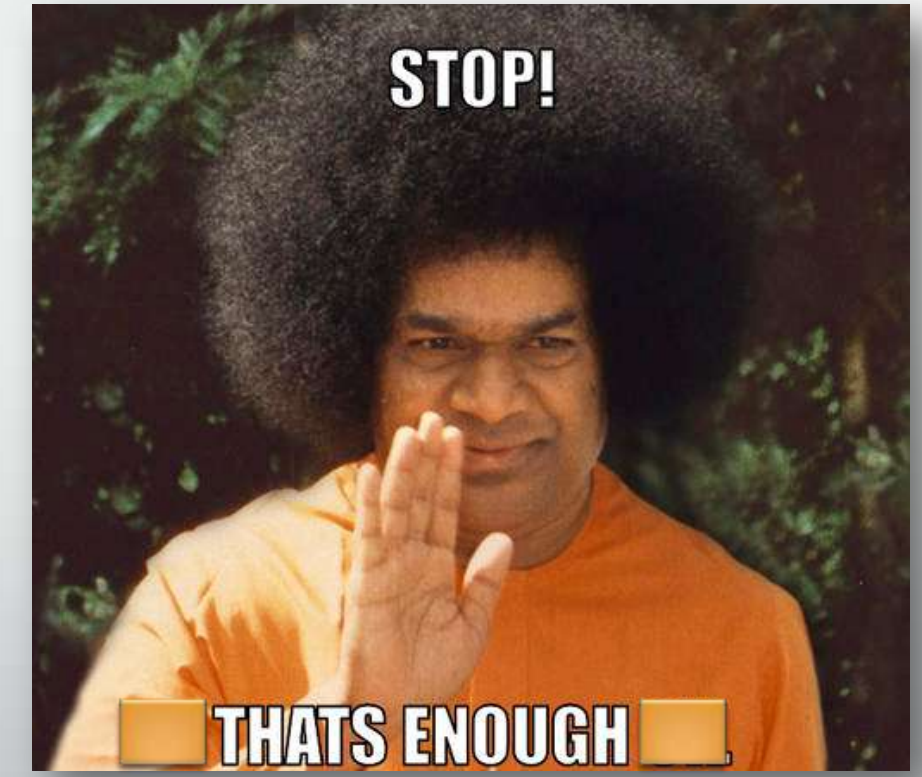
Length of Credit History

- Time since accounts opened
- Time since accounts opened, by specific type of account
- Time since account activity

New Credit

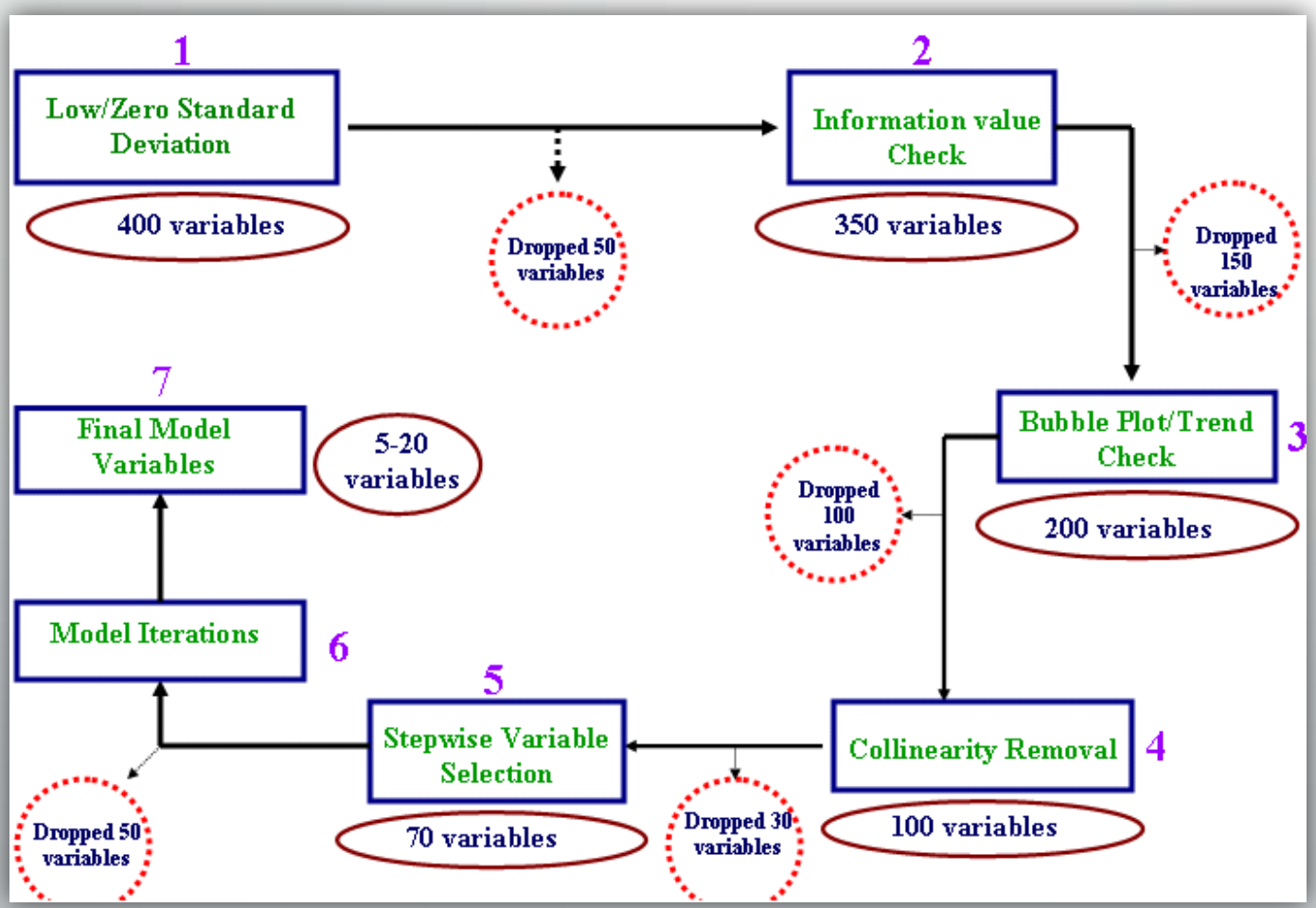
- Number of recently opened accounts, and proportion of accounts that are recently opened, by type of account
- Number of recent credit inquiries
- Time since recent account opening(s), by type of account
- Time since credit inquiry(s)
- Re-establishment of positive credit history following past payment problems

Even more variables



Identifying most impacting factors

Out of 500 what are those 20 important attributes



Selecting the variables- How do you do that?

Num_Enq	Good	Bad
0	6400	5
1	5800	69
2	5445	134
3	4500	250
4	4070	375
5	3726	470
6	2879	650
7	1893	876
8	1236	987
9	354	1298
	36303	**5114**

Num_companies	Good	Bad
0	4858	594
1	4291	602
2	4410	577
3	4738	529
4	4506	707
5	4976	796
6	4723	770
7	4407	583
8	4157	825
9	4970	964
	46036	**6947**

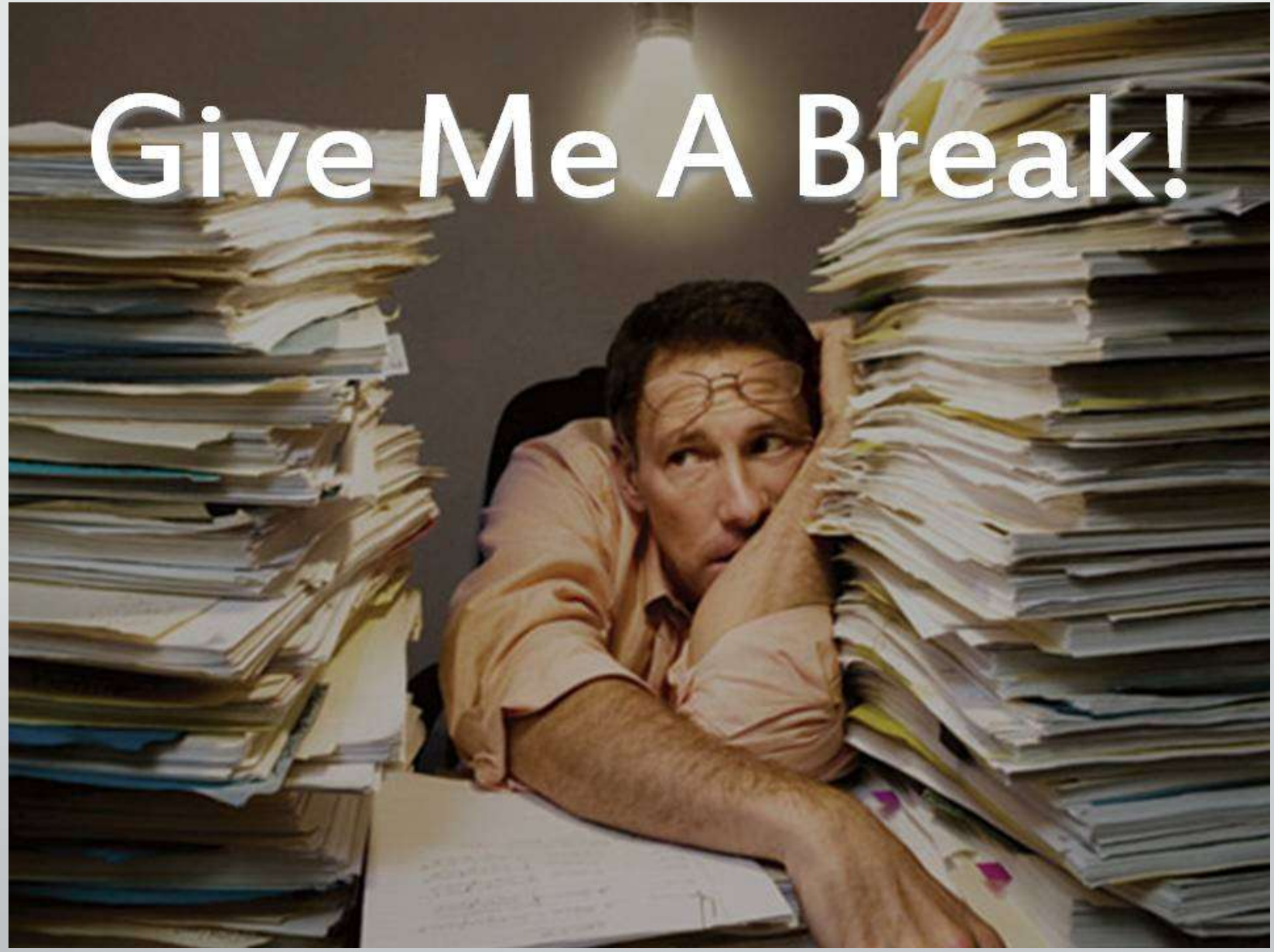
Bad Rate

Num_Enq	Good	Bad	Bad Rate
0	6400	5	0%
1	5800	69	1%
2	5445	134	3%
3	4500	250	5%
4	4070	375	7%
5	3726	470	9%
6	2879	650	13%
7	1893	876	17%
8	1236	987	19%
9	354	1298	25%
	36303	**5114**	

Num_companies	Good	Bad	Bad Rate
0	4858	594	9%
1	4291	602	9%
2	4410	577	8%
3	4738	529	8%
4	4506	707	10%
5	4976	796	11%
6	4723	770	11%
7	4407	583	8%
8	4157	825	12%
9	4970	964	14%
	46036	**6947**	

Information value

Utilization %	# of Good	# of Bad	%Good [x]	%Bad [Y]	%Good / %Bad	X -Y	WOE = Log (X/Y)	IV
< 5	1850	150	29%	5%	6.31	0.25	0.80	0.20
5-30	1600	400	25%	12%	2.05	0.13	0.31	0.04
31 - 60	1200	600	19%	18%	1.02	0.00	0.01	0.00
60 - 90	900	900	14%	28%	0.51	(0.14)	-0.29	0.04
>= 91	800	1200	13%	37%	0.34	(0.24)	-0.47	0.11
Total	**6350**	**3250**						**0.39**


Give Me A Break!

Model Building

Historical Data

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

The model equation

Model Building

logistic regression model to predict the probability of default

- Probability of bad = w1(Var1)+w2(Var2) +w3(Var1)……+w20(var20)
- Logistic regression gives us those weights
- Predicting the probability
  - **Odds of bad**=0.13(number of cards)+ 0.21 (utilization)+…….+0.06(number of loan applications)
- That's it ….we are done

Credit Risk Model Building-Example

Attributes used on the model

1. Age → Application form
2. MonthlyIncome1→ Application form
3. Dependents → Application form
4. Number of times 30DPD in last two years → Bureau
5. Debt Ratio →Bureau /Bank
6. Utilization → Bureau /Bank
7. Number of loans →Bureau /Bank

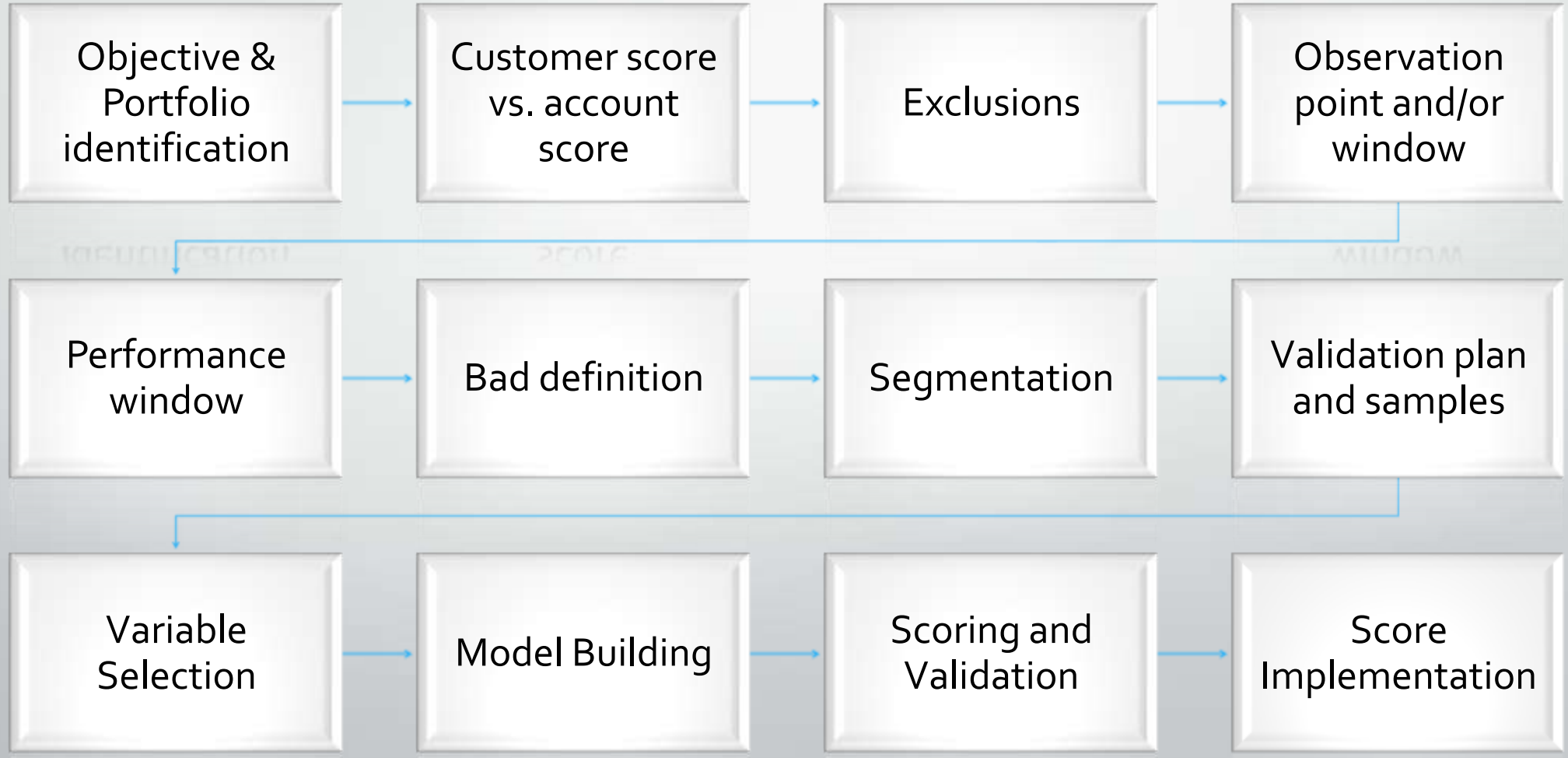
Model building and Implementation

Data: http://www.kaggle.com/c/GiveMeSomeCredit

Model building script:

The final Model

Actual model Building Steps



Marketing Example

Predicting the response probability to a marketing Campaign

- Selling Mobile phones – Marketing campaign
- Who should we target?
  - Consider historical data of mobile phone buyers
  - See their characteristics
  - Find top impacting characteristics
  - Find weight of each characteristic
  - Score new population
  - Decide on the cut off
  - Try to sell people who score more than cut off

Other Applications of Model Building

- **Fraud transactions scorecard** – Fraud identification based on attributes like transaction amount, place, time, frequency of transactions etc.,
- **Attrition modeling** – Predicting employee attrition based on their characteristics

Thank You

Venkat Reddy

http://www.trendwiseanalytics.com/training/venkat.php