INSEAD

The Business School for the World®

Prof. Anton Ovchinnikov

Prof. Spyros Zoumpoulis

DSB Classes 05-06, January 30, 2018

- Introduction to Classification

Plan for the day
Learning objectives

- Conceptual introduction to classification: metrics
- Data science methodologies for classification:
  - Stats: logistic regression (generalized linear model, `glm`) + variable selection
  - Machine Learning: CART (classification and regression tree)
  - in Session 7: additional methods: regularizations (LASSO), random forest, gradient boosting machines (xgboost), support vector machines (SVM)

What is classification, and why do we need it?

INSEAD

The Business School for the World®

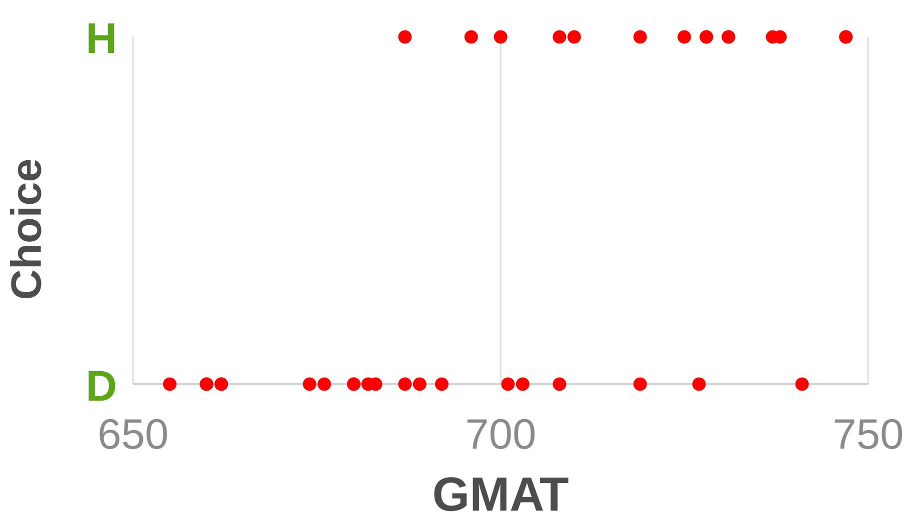
- In sessions 1-4 we considered a task of predicting a quantity (price of a diamond, electricity rate, number of website users)
- But an equally* common task is to predict an outcome of an event:
  - Binary outcomes:
    - Will a customer churn? Will a customer default on a loan?
    - Will an employee/student accept a job/school offer?
  - Multi-nomial outcomes:
    - Will a person walk/drive/bike/take a public transit?
    - Will a customer buy iPhone X/8/8+/nothing?

Task(s): predict the event(s) per se
+ understand the actionable drivers

Predicting events: what if “Y” is categorical?

INSEAD
The Business School for the World®

- Examples of categorical dependent variable?
- Customer choice:
  - b/school "D" versus "H" as a function of GMAT score



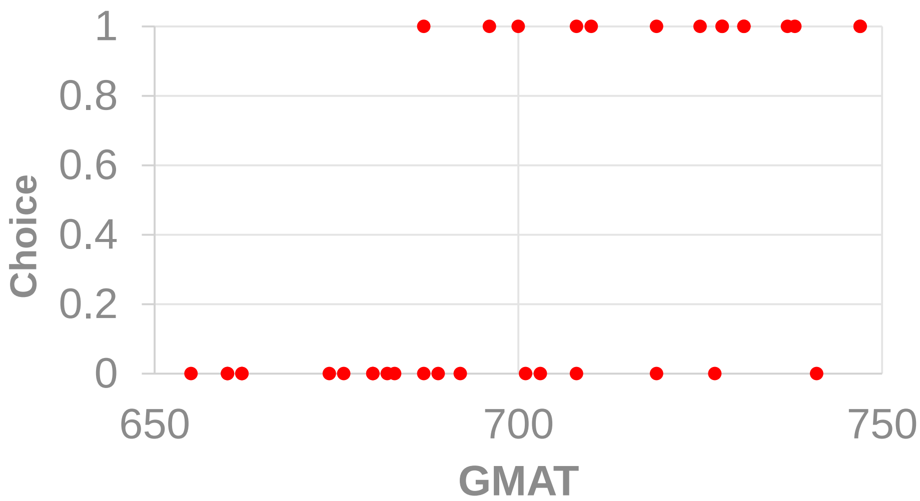
If we know GMAT, can we predict choice?

	A	B	C
1	ID	GMAT	Choice
2	1	655	D
3	2	660	D
4	3	660	D
5	4	662	D
6	5	662	D
7	6	674	D
8	7	676	D
9	8	680	D
10	9	680	D
11	10	682	D
12	11	683	D
13	12	687	H
14	13	687	D
15	14	689	D
16	15	692	D
17	16	696	H
18	17	700	H
19	18	701	D
20	19	703	D
21	20	708	H
22	21	708	D

Source: 0506 Data and model from the Modeling Discrete Choice reading.xls

Predicting choice: Regression?

- Step one: Transform D/H into a dummy variable (0,1)

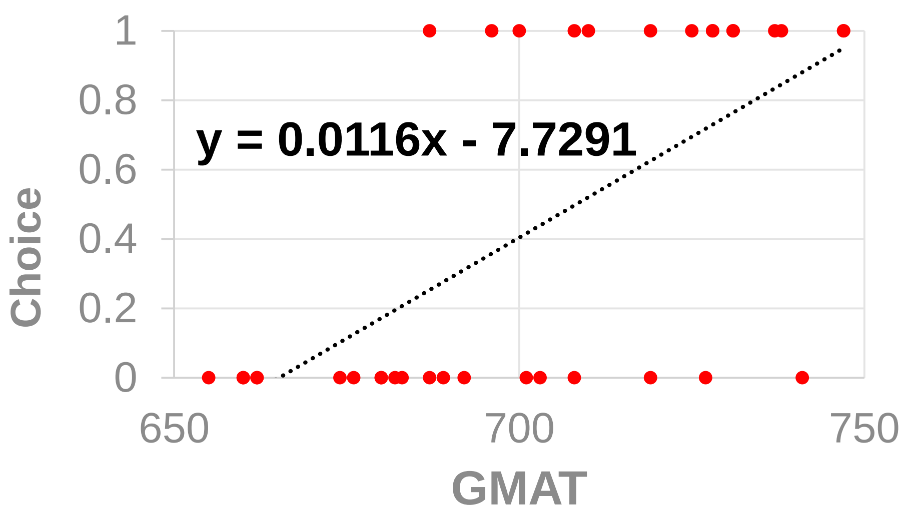


- Step two: Run a (linear) regression

Predicting choice: Regression?

INSEAD
The Business School for the World®

- Step one: Transform D/H into a dummy variable (0,1)



Multiple Regi Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate		
	0.6385	0.4076	0.3897	0.392249		
ANOVA Table	**Degrees of Freedom**	**Sum of Squares**	**Mean of Squares**	**F-Ratio**	**p-Value**	
Explained	1	3.494068	3.494068	22.7095	< 0.0001	
Unexplained	33	5.07736	0.153859			
Regression Ta	**Coefficient**	**Standard Error**	**t-Value**	**p-Value**	**Confidence Interval 95% Lower**	**Upper**
Constant	-7.72912	1.713126	-4.5117	< 0.0001	-11.2145	-4.24374
GMAT	0.011619	0.002438	4.7654	< 0.0001	0.006659	0.01658

- Step two: Run a (linear) regression
  - How should we interpret the Y variable?
    - E.g. for GMAT=700, Y = 0.4… [of what?]
    - What about GMAT = 650?

Predicting Probability of Choice: Logistic Regression

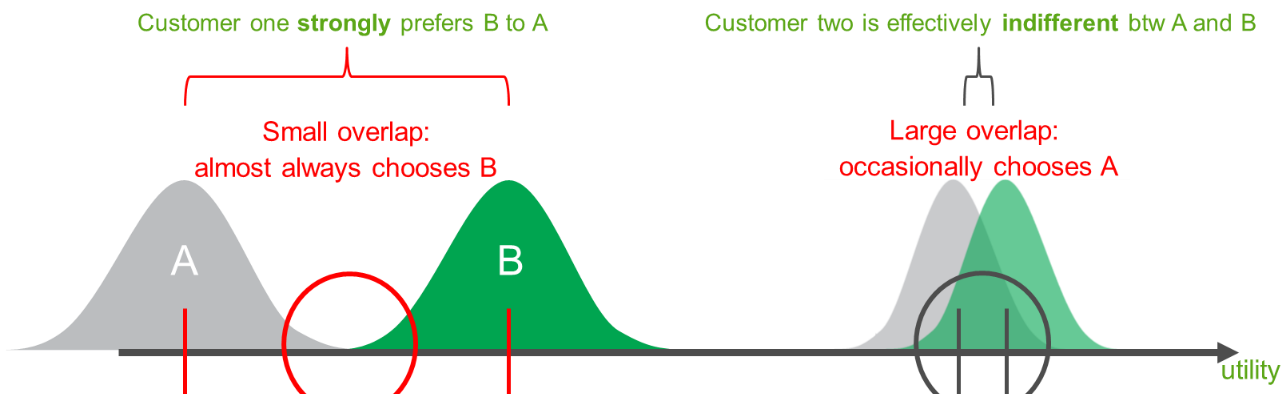
INSEAD
The Business School for the World®

- It is natural to interpret the "Y" variable in the preceding example as a probability of choice
- Hence we are predicting the probability of choice, not the choice itself
- But a linear model is not suitable to predict probabilities (e.g., because it cannot guarantee probability >0 or <1)
- We need a better model, one that explicitly accounts for the fact that a predicted quantity is a probability
- Logit model (hence "logistic" regression) is one such model [a popular one]
- The term "logit" refers to the Log of odds prob/(1-prob).
  - Logit is not the only model used to model choice
  - Probit is another commonly used model

Understanding a Logit Model: Concept of Utility

INSEAD
The Business School for the World®

- **Q**: Why is a consumer buying a product?
  - A: Because the utility (pleasure, enjoyment, "smiles") from buying/consuming the product is larger than the product's price
- **Q**: Why is a consumer buying product A and not product B?
  - A: Because the utility from buying A is larger than from buying B
- **Q**: suppose your utility for A > utility for B [e.g., you like orange juice more than apple juice]? Will you <u>always</u> buy A?
  - A: No, but how often you "deviate" depends on the strength of preferences for A vs B and the noise/uncertainty (ε) in utility



Logistic/Logit Model (Gumbel distribution of ε)

- General form:

$$Prob(A\ is\ chosen\ from\ set\ of\ S\ alternatives) = \frac{\exp(utility\ of\ A)}{\sum \exp(utilties\ of\ all\ alternatives)}$$

- Note that "all" must also include an alternative to buy nothing
- With only two alternatives:

$$Prob(A\ is\ chosen\ over\ B) = \frac{\exp(utility\ of\ A)}{\exp(utility\ of\ A) + \exp(utility\ of\ B)}$$

- Further, since only relative utility matters, we can normalize $utility\ of$ B=0, and then noting that $\exp(0) = 1$

$$Prob(A\ is\ chosen\ over\ B) = \frac{\exp(utility\ of\ A)}{1 + \exp(utility\ of\ A)}$$

Back to Our Example: School H Versus D

INSEAD
The Business School for the World®

Let $utility\ of\ D = 0$ [arbitrarily]

Let $utility\ of\ H\ =\ a * GMAT + b$

We can then express the probabilities of choices

And estimate utility coefficients a and b

uH=a*GMAT+b

a= 0.07

b= -48

=$G$2+$G$1*B

=F4/(1+F4)

ID	GMAT	Choice	Dummy	uH	EXP(uH)	Prob(H is chosen)
1	655	D	0	-2.1500	0.1165	0.1043
2	660	D	0	-1.8000	0.1653	0.1419
3	660	D	0	-1.8000	0.1653	0.1419
4	662	D	0	-1.6600	0.1901	0.1598
5	662	D	0	-1.6600	0.1901	0.1598
6	674	D	0	-0.8200	0.4404	0.3058
7	676	D	0	-0.6800	0.5066	0.3363
8	680	D	0	-0.4000	0.6703	0.4013
9	680	D	0	-0.4000	0.6703	0.4013
10	682	D	0	-0.2600	0.7711	0.4354
11	683	D	0	-0.1900	0.8270	0.4526
12	687	H	1	0.0900	1.0942	0.5225
13	687	D	0	0.0900	1.0942	0.5225

Estimating Utility Coefficients: (Log)Likelihood

INSEAD
The Business School for the World®

- For customer ID1, the choice is D and the predicted probability of choosing H is 0.1043
    - Hence the likelihood that ID1 indeed chooses D in our model is 1-0.1043=0.8957
- For ID2: Choice is D, predicted prob=0.1419, hence the likelihood is 0.8581
- The likelihood of ID1 choosing D and ID2 choosing D is 0.8957*0.8581, etc…
- We would like to select a and b such that the likelihood is maximized (Maximum Likelihood Estimation, MLE)
- Note:
    - With many datapoints such product will be very small - inconvenient for optimization
    - However, Log (X*Y*Z)=Log(X)+Log(Y)+Log(Z)
- Hence instead of maximizing likelihood, we maximize log-likelihood (LL)

=$G$2+$G$1*B

=F4/(1+F4)

=IF(D4=1,G4,1-G4)

=LN

	A	B	C	D	E	F	G	H	I
1		uH=a*GMAT+b				a=	0.07		
2						b=	-48		
3	ID	GMAT	Choice	Dummy	uH	EXP(uH)	Prob(H is chosen)	Likelihood	Log(Likelihood)
4	1	655	D	0	-2.1500	0.1165	0.1043	0.8957	-0.1102
5	2	660	D	0	-1.8000	0.1653	0.1419	0.8581	-0.1530
6	3	660	D	0	-1.8000	0.1653	0.1419	0.8581	-0.1530
7	4	662	D	0	-1.6600	0.1901	0.1598	0.8402	-0.1741
8	5	662	D	0	-1.6600	0.1901	0.1598	0.8402	-0.1741
9	6	674	D	0	-0.8200	0.4404	0.3058	0.6942	-0.3649
10	7	676	D	0	-0.6800	0.5066	0.3363	0.6637	-0.4099
11	8	680	D	0	-0.4000	0.6703	0.4013	0.5987	-0.5130
12	9	680	D	0	-0.4000	0.6703	0.4013	0.5987	-0.5130
13	10	682	D	0	-0.2600	0.7711	0.4354	0.5646	-0.5716
14	11	683	D	0	-0.1900	0.8270	0.4526	0.5474	-0.6027
15	12	687	H	1	0.0900	1.0942	0.5225	0.5225	-0.6492
16	13	687	D	0	0.0900	1.0942	0.5225	0.4775	-0.7392

Results:
School H Versus D example

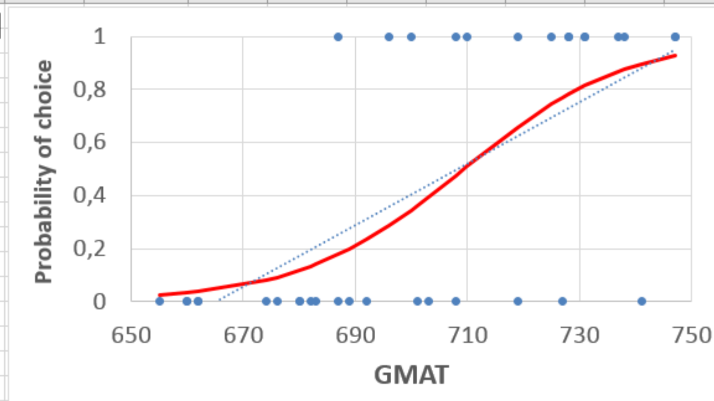
=$G$2+$G$1*B

=F4/(1+F4)

=IF(D4=1,G4,1-

=LN(H4)

	D	E	F	G	H	I
1			a=	0,068325198		
2			b=	-48,47037424		
3	Dummy	uH	EXP(uH)	Prob(H is chosen)	Likelihood	Log(Likelihood)
4	0	-3,7174	0,0243	0,0237	0,9763	-0,0240
5	0	-3,3757	0,0342	0,0331	0,9669	-0,0336
6	0	-3,3757	0,0342	0,0331	0,9669	-0,0336
7	0	-3,2391	0,0392	0,0377	0,9623	-0,0385
8	0	-3,2391	0,0392	0,0377	0,9623	-0,0385
9	0	-2,4192	0,0890	0,0817	0,9183	-0,0853
10	0	-2,2825	0,1020	0,0926	0,9074	-0,0971
11	0	-2,0092	0,1341	0,1182	0,8818	-0,1258
12	0	-2,0092	0,1341	0,1182	0,8818	-0,1258
13	0	-1,8726	0,1537	0,1332	0,8668	-0,1430
14	0	-1,8043	0,1646	0,1413	0,8587	-0,1524
15	1	-1,5310	0,2163	0,1779	0,1779	-1,7268
16	0	-1,5310	0,2163	0,1779	0,8221	-0,1958
17	0	-1,3943	0,2480	0,1987	0,8013	-0,2215



T	U	V	W	X
Logistic Regression for Dummy				
Summary Measures				
Null Deviance	47,80356733			
Model Deviance	30,96154543			
Improvement	16,8420219			
p-Value	< 0.0001			
Regression Coefficients	Coefficient	Standard Error	Wald Value	p-Value
Constant	-48,47037424	15,38526195	-3,150441923	0,0016
GMAT	0,068325198	0,021740921	3,142700311	0,0017
Classification Matrix	1	0	Percent Correct	
1	11	4	73,33%	
0	3	17	85,00%	

$$Prob(H\ is\ chosen|GMAT) = \frac{\exp(utility\ of\ H)}{1+\exp(utility\ of\ H)} = \frac{\exp(-48{,}47 + 0{,}0683 * GMAT)}{1+\exp(-48{,}47 + 0{,}0683 * GMAT)}$$

- With GMAT=700:
  - Utility of H = -48,47+0,0683*700 = -0,66 **[why is it negative?]**
  - Prob of H = exp(-0.66)/(1+exp(-0.66)) = 0,5168/1,5168 =0,34

Logistic Regression in R: School H vs D example

INSEAD
The Business School for the World®

ChoiceData<-read.csv(file.choose()) #load data

str(ChoiceData) #make sure that the field types are interpreted correctly (as numbers/integers, factors, etc.)

Logistic_Model<-**glm**(Choice ~ GMAT, data = ChoiceData, family="binomial"(link="logit")) #logistic regression is part of the "generalized linear models" family, hence glm

summary(Logistic_Model) #summary of the model

par(mfrow=c(1,4)) # This command sets the plot window to show 1 row of 4 plots

plot(Logistic_Model) # check the model using diagnostic plots

predict(Logistic_Model, newdata=data.frame("GMAT"=700),type="response") #predict the probability of choice as a function of GMAT

```
Call:
glm(formula = Choice ~ GMAT, family = binomial(link = "logit"),
    data = ChoiceData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1298  -0.5889  -0.2593   0.6726   1.8584

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -48.47108   15.38544  -3.150  0.00163 **
GMAT          0.06833    0.02174   3.143  0.00167 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> predict(Logistic_Model, newdata=data.frame("GMAT"=700),type="response")
#predict the probability of choice as a function of GMAT
        1
0.3446266
```

Summary of Logistic Regression

INSEAD

The Business School for the World®

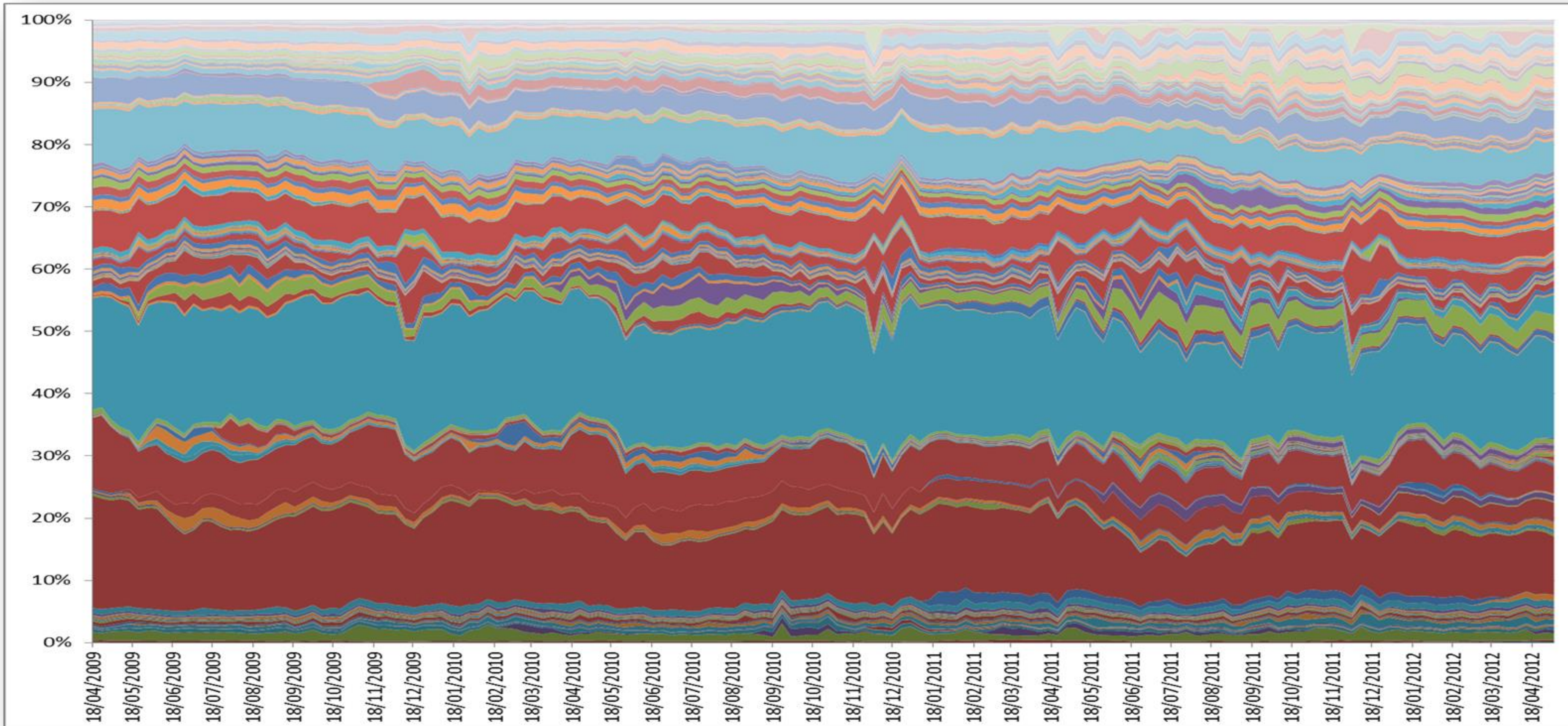
- A common analytics task involves building a regression model to predict a categorical variable (e.g. will customer buy or not)
  - Rather than predicting the choice itself, it is natural to predict the probability of choice
  - Linear regression is not quite suitable for that; we need a special model for predicting probabilities
- Logistic regression is such a model:
  - It builds a linear model for the utility of choice
  - And then combines those utilities with a exp formula to obtain a probability estimate
- We saw R implementation, Excel add-on statistical packages (StatTools) can also run logistic regression
- The D vs H example was of a binary logit model, a more general case with multiple options is called Multinomial Logit Model (MNL, a "workhorse" of customer analytics) – the beer sales example next

Logit models are rather accurate

Nested MNL, beer sales, 35m observations, <500 variables [prices, promos…]

INSEAD

The Business School for the World®

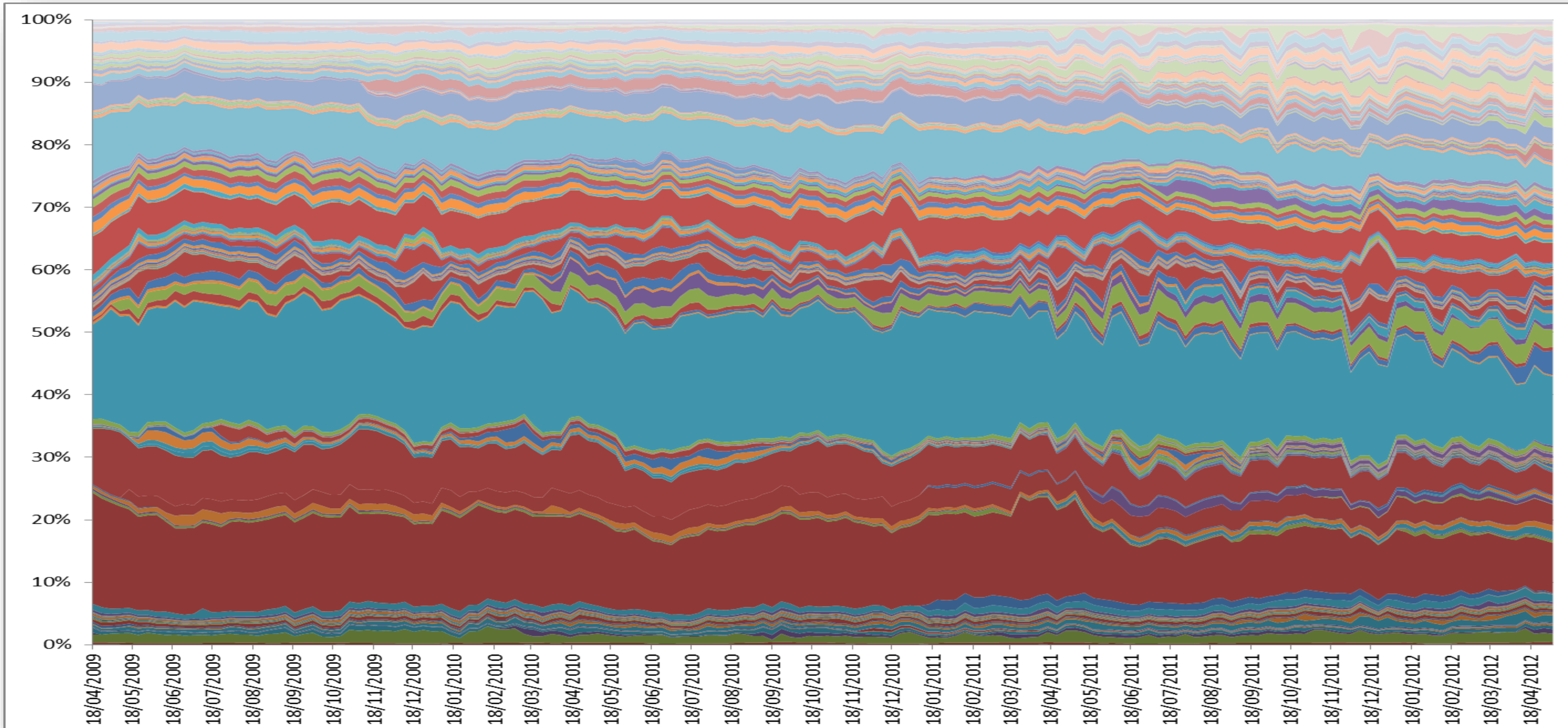


actual

Logit models are rather accurate

Nested MNL, beer sales, 35m observations, <500 variables [prices, promos…]

INSEAD
The Business School for the World®



predicted

Back to classification: metrics

INSEAD
The Business School for the World®

- For models with continuous quantities we discussed multiple metrics:
  - r^2, MAPE, (R)MSE
- For classification models we need other metrics, that specifically account for the fact that the predicted object is an event:
  - Confusion matrix and its measures
  - ROC ("receiver operating characteristic") curve
  - AUC ("area under curve) and Gini coefficient
  - Lift chart / Gains chart

Confusion Matrix [customer retention example]

INSEAD
The Business School for the World®

	Predicted Retained	Predicted Not Retained
Actual Retained	a (TP)	b (FN)
Actual Not Retained	c (FP)	d (TN)

TP stands for True Positive

FN stands for False Negative, etc.

Confusion Matrix [customer retention example]

INSEAD
The Business School for the World®

	Predicted Retained	Predicted Not Retained		
Actual Retained	a (TP)	b (FN)	Positive Predictive Value	a/(a+b)
Actual Not Retained	c (FP)	d (TN)	Negative Predictive Value	d/(c+d)
	Sensitivity [TPR] a/(a+c)	Specificity [FNR] d/(b+d)		

Confusion Matrix [customer retention example]

INSEAD
The Business School for the World®

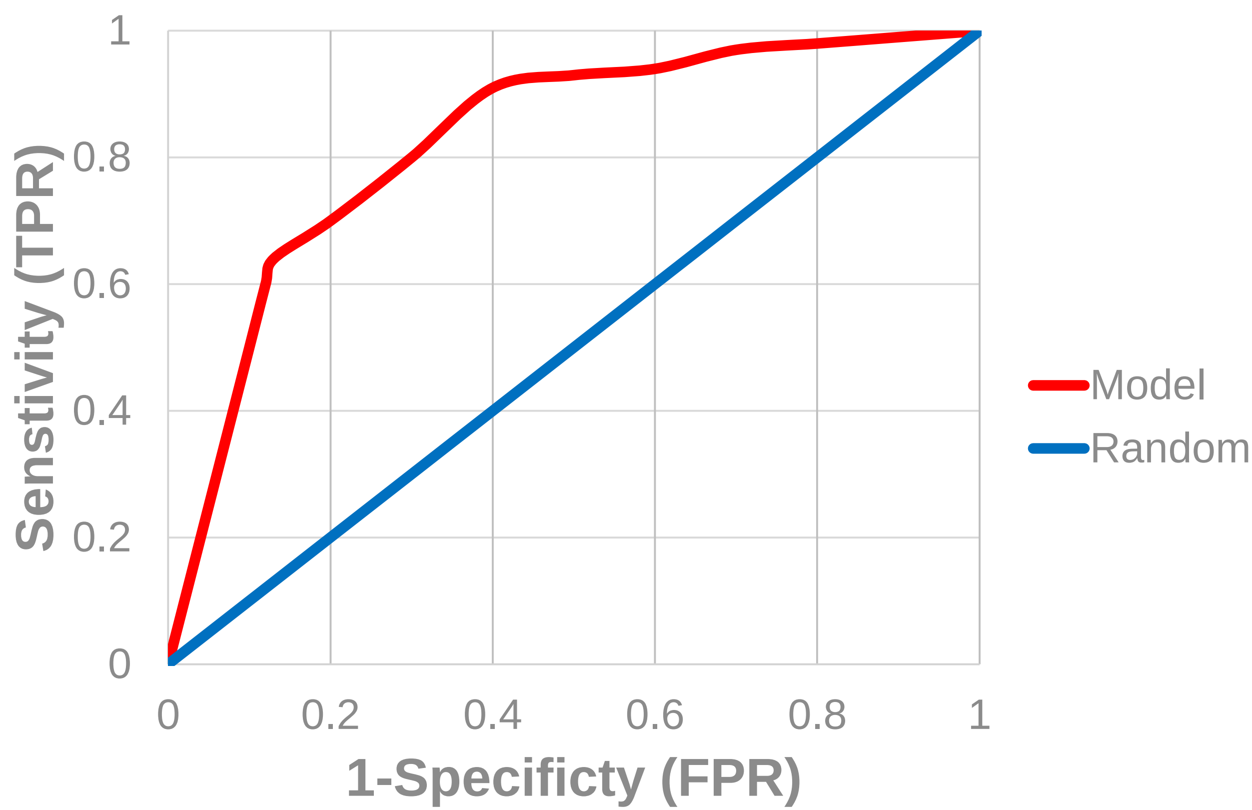
	Predicted Retained	Predicted Not Retained		
Actual Retained	a (TP)	b (FN)	Positive Predictive Value	a/(a+b)
Actual Not Retained	c (FP)	d (TN)	Negative Predictive Value	d/(c+d)
	Sensitivity [TPR] a/(a+c)	Specificity [FNR] d/(b+d)		

Overall measure:
Accuracy=(a+d)/(a+b+c+d)
Misclassification error = 1- accuracy

ROC Curve

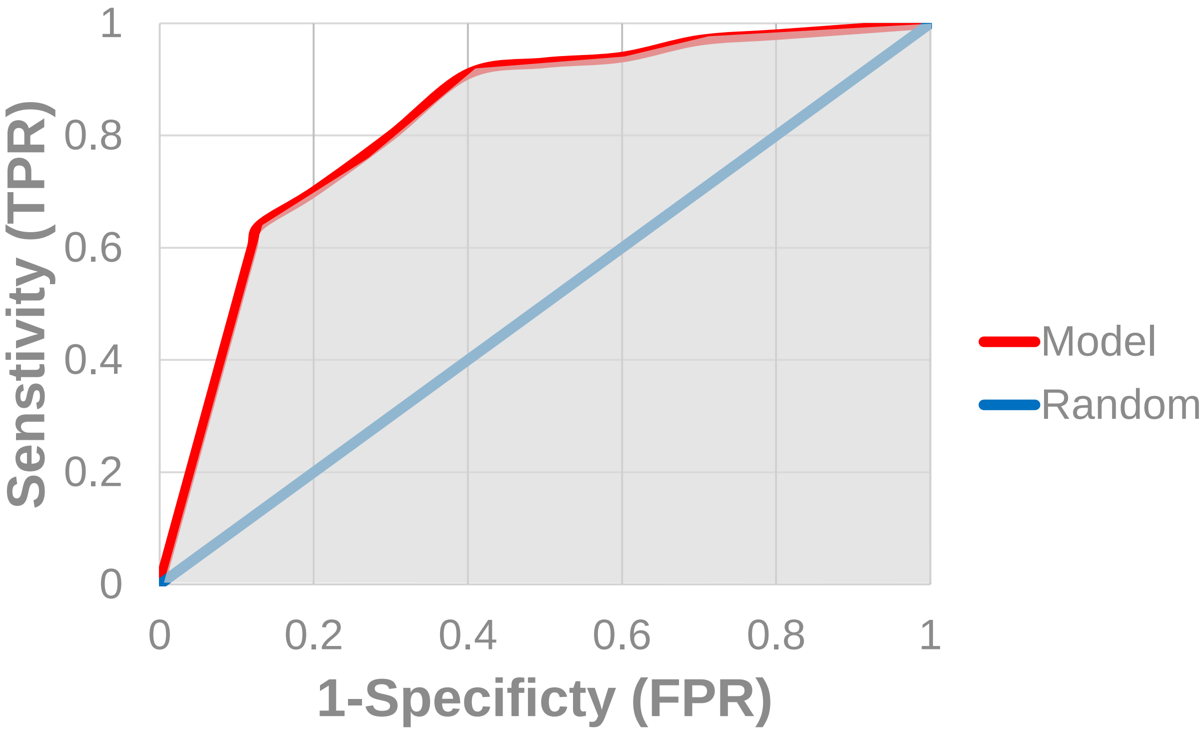
INSEAD
The Business School for the World®

- ROC stands for "receiver operating characteristic" and roots to analyzing radar signals during WWII



AUC (Area Under Curve)

INSEAD
The Business School for the World®

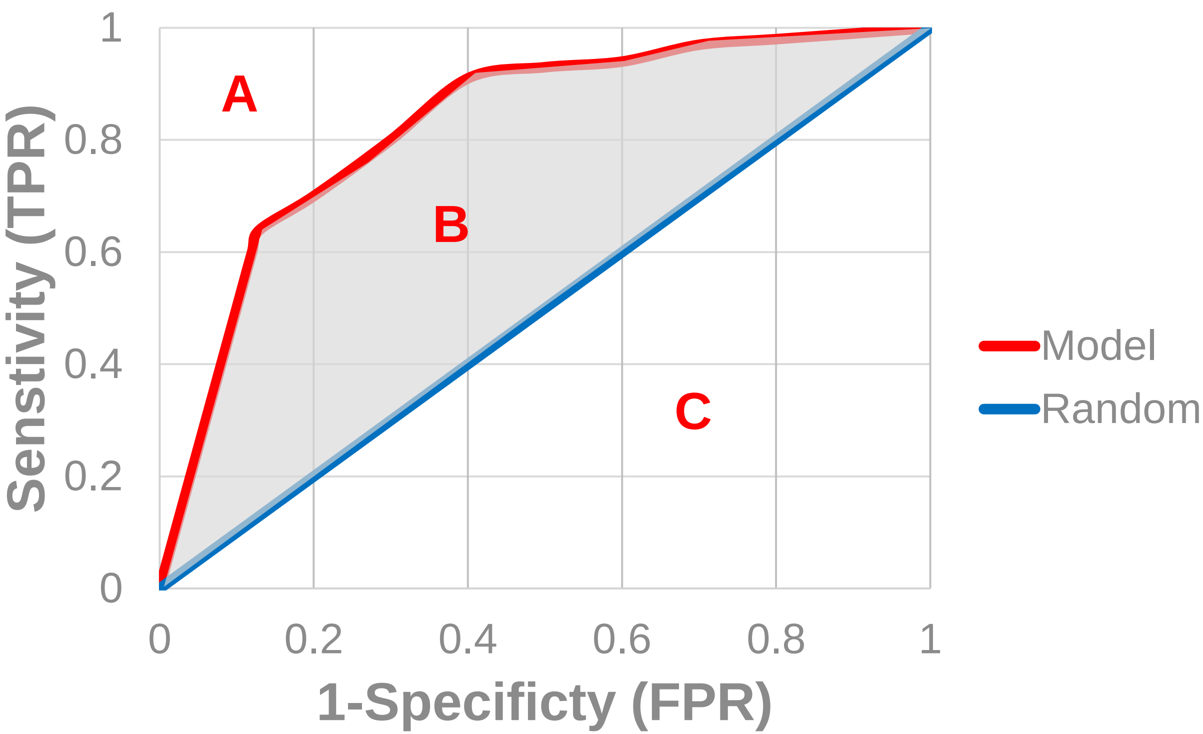


AUC	Quality of Prediction
0.50	Random
0.50-0.60	Fail
0.60-0.70	Poor
0.70-0.80	Fair
0.80-0.90	Good
0.90-1	Excellent

*context specific (driverless car vs fin. instrument)

Gini Coefficient

INSEAD
The Business School for the World®



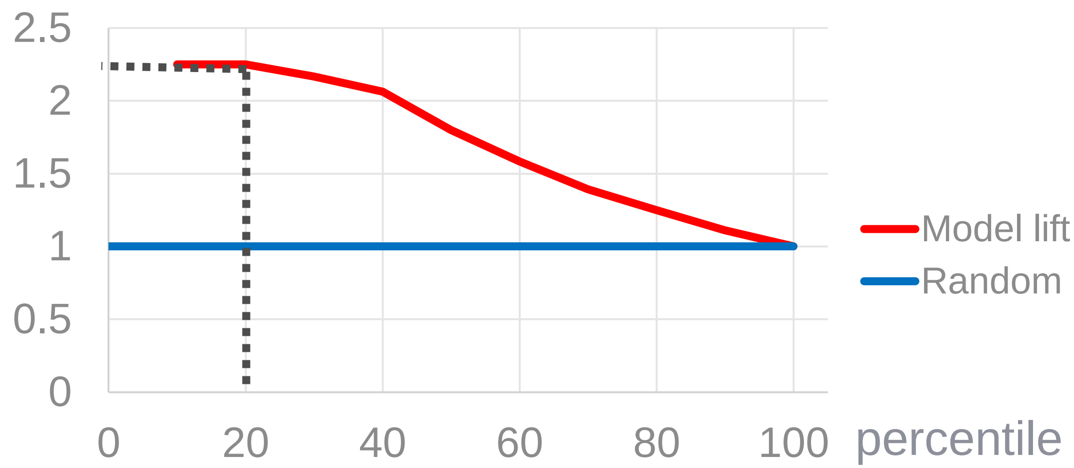
Gini coefficient (index, ratio): Common measure of income distribution, named after the Italian statistician's 1912 paper

- Gini = B/(A+B)
- Note that AUC = B+C
- Because A+B+C=1, A+B=C=1/2:
- Gini = 2*AUC-1

Lift Chart / Gains Chart

INSEAD
The Business School for the World®

- Lift is a common metric of cumulative model performance (especially relevant in marketing analytics)
- It evaluates the model on a proportion of the population and depicts cumulative responses by percentile. Example:
  - 20% of random customers correspond to 20% of those who are retained: random lift at 20th percentile = 20/20=1
  - if 20% of the "best" customers per the model correspond to 45% of all retained customers, then model lift at 20th percentile = 45/20 = 2.25



Now lets practice STC(A) case

INSEAD
The Business School for the World®

- Files on portal:
  - R-code 0506 STC (A) Logistic.R
  - CSV data 0506 STC(A) data_numerical dates.csv
    - BTW, how to generate the CSV datafile from an Excel case exhibit?
- The general structure of the code has the following steps:

1. Packages & libraries: package for managing packages, `pacman`
2. Load data
3. "Clean" data: formats, missing values (custom function `fixNAs`)
4. Split the dataset into testing vs training
5. Run ("train") a model on the <u>training</u> data: `stepAIC` variable selection
6. Obtain model prediction for the <u>testing</u> data
7. Obtain metrics (classification matrix, ROC curve, AUC, lift chart) for the testing data

Missing values

INSEAD

The Business School for the World®

- VERY often, some of the data entries will be missing
- What should we do about it?
  - Ignore? Bad idea: missing is often not random
  - Categorical variables [easy] <u>add</u> a missing category
  - Continuous variables [harder]
    - <u>replace</u> (with 0, mean, median, etc.), or <u>impute</u> (create a separate model to predict the missing values based on what's not missing)
    - and add a "surrogate" dummy for each missing value

AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
Poverty C	Region	CRM Segm	School Typ	Parent Me	Parent Me	Parent Me	MDR Low	MDR High	Total Scho	Income Le
B	Southern	4	PUBLIC	1	#########		K	5	927	Q
C	Other	10	PUBLIC	1	#########	#########	7	8	850	A
C	Other	10	PUBLIC	1	#########		6	8	955	O
	Other	7	CHD	0					0	
D	Other	10	PUBLIC	1	#########		6	8	720	C
C	Other	8	PUBLIC	1	#########		10	12	939	I
	Other	8	Catholic	1	#########		9	12	225	G
	Other	7	CHD	1	9/8/2010				0	
	Other	5	CHD	1	9/8/2010		6	12	500	K
	Houston	5	Private nc	1	#########		PK	8	635	K
	Other	10	CHD	1	9/9/2010		K	12	746	O
	Other	10	CHD	1	#########		PK	8	650	L
A	Northern	5	PUBLIC	1	#########		6	8	670	Q
B	Northern	5	PUBLIC	1		9/1/2010	6	8	750	L
	Northern	7	PUBLIC	1		9/9/2010			0	P5
B	Other	6	PUBLIC	1	#########	#########	6	8	753	I

Handling missing values in R custom "fixNAs" function

INSEAD

The Business School for the World®

```
fixNAs<-function(data_frame){                 # Crete a function to fix NAs and preserve the NA info as surrogate variables
integer_reac<-0                               # Define reactions to Nas for different classes of variables as shown in your data structure (str command)
factor_reac<-"FIXED_NA"
character_reac<-"FIXED_NA"
date_reac<-as.Date("1900-01-01")
for (i in 1 : ncol(data_frame)){              # Loop through columns in data frame and depending on which class the variable is, apply the
                                              defined reaction and create a surrogate
  if (class(data_frame[,i]) %in% c("numeric","integer")) {
   if (any(is.na(data_frame[,i]))){
    data_frame[,paste0(colnames(data_frame)[i],"_surrogate")]<-
     as.factor(ifelse(is.na(data_frame[,i]),"1","0"))
    data_frame[is.na(data_frame[,i]),i]<-integer_reac     }
  } else
   if (class(data_frame[,i]) %in% c("factor")) {
    if (any(is.na(data_frame[,i]))){
     data_frame[,i]<-as.character(data_frame[,i])
     data_frame[,paste0(colnames(data_frame)[i],"_surrogate")]<-
      as.factor(ifelse(is.na(data_frame[,i]),"1","0"))
     data_frame[is.na(data_frame[,i]),i]<-factor_reac
     data_frame[,i]<-as.factor(data_frame[,i])       }
   } else {
    if (class(data_frame[,i]) %in% c("character")) {
     if (any(is.na(data_frame[,i]))){
      data_frame[,paste0(colnames(data_frame)[i],"_surrogate")]<-
       as.factor(ifelse(is.na(data_frame[,i]),"1","0"))
      data_frame[is.na(data_frame[,i]),i]<-character_reac       }
    } else {
     if (class(data_frame[,i]) %in% c("Date")) {
      if (any(is.na(data_frame[,i]))){
       data_frame[,paste0(colnames(data_frame)[i],"_surrogate")]<-
        as.factor(ifelse(is.na(data_frame[,i]),"1","0"))
       data_frame[is.na(data_frame[,i]),i]<-date_reac        }}}}   return(data_frame)      }
```

Do you need to know how to write such custom functions?

NO!

But you certainly can copy-paste this function and use it anytime you need to deal with missing values

Now lets practice STC(A) case

- Files on portal:
  - R-code 0506 STC (A) Logistic.R
  - CSV data 0506 STC(A) data_numerical dates.csv
    - BTW, how to generate the CSV datafile from an Excel case exhibit?
- The general structure of the code has the following steps:

1. Packages & libraries: package for managing packages, `pacman`
2. Load data
3. "Clean" data: formats, missing values (custom function `fixNAs`)
4. Split the dataset into testing vs training
5. Run ("train") a model on the <u>training</u> data: `stepAIC` variable selection
6. Obtain model prediction for the <u>testing</u> data
7. Obtain metrics (classification matrix, ROC curve, AUC, lift chart) for the testing data

STC(A) results confusion matrix

INSEAD

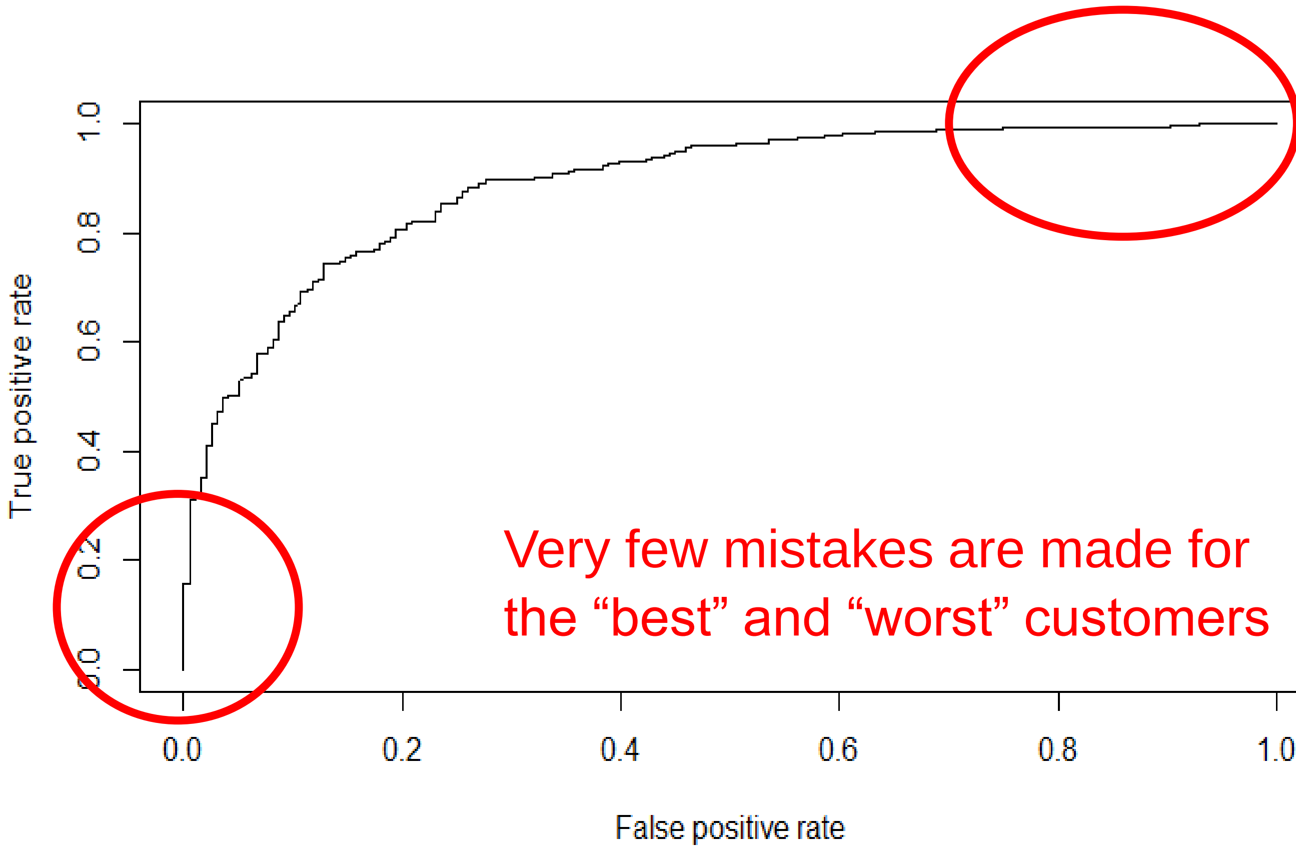
The Business School for the World®

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 157  59
         1  39 245

               Accuracy : 0.804
                 95% CI : (0.7664, 0.8379)
    No Information Rate : 0.608
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5961
 Mcnemar's Test P-Value : 0.05495

            Sensitivity : 0.8010
            Specificity : 0.8059
         Pos Pred Value : 0.7269
         Neg Pred Value : 0.8627
             Prevalence : 0.3920
         Detection Rate : 0.3140
   Detection Prevalence : 0.4320
      Balanced Accuracy : 0.8035

       'Positive' Class : 0
```

Our model is ~80% accurate

Accuracy is balanced

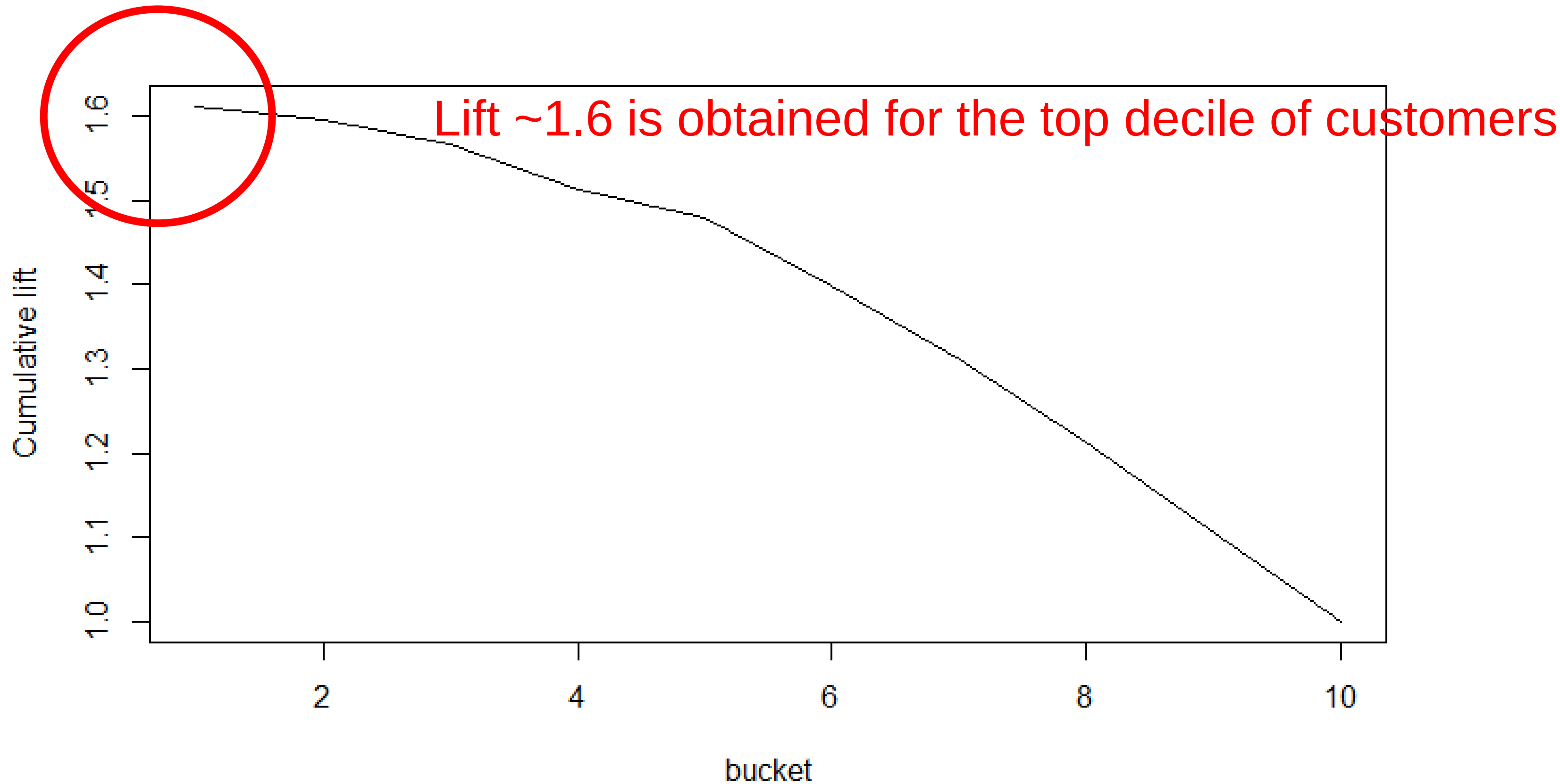
STC(A) results
ROC curve and AUC



AUC = 89% ~ “Excellent”

STC(A) results
ROC curve and AUC

INSEAD
The Business School for the World®



Where is 1.6 coming from?

On average 60.73% of customers are retained. So from the a decile of the testing data (50 customers), ~30 are expected to be retained. But in the top decile, all 50 were retained, we see this from the ROC curve [how?], which is 50/30~1.6 times more than average

Intermediate summary: classification metrics logistic regression and STC(A) case

INSEAD
The Business School for the World®

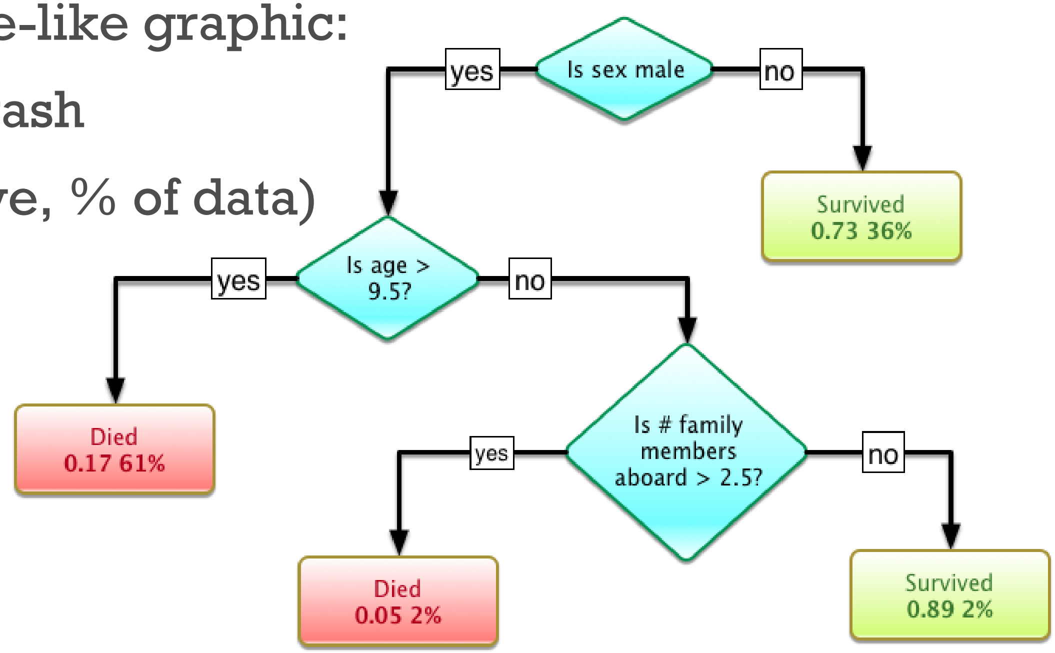
- Classification ~ predicting events
- STC(A) case: need to predict which customers will purchase next year
- Logistic regression: predicting the probability of a purchase
- R "tricks":
  - Data pre-processing: fixing types and missing values
  - Stepwise variable selection
  - Holdout: training the model on one subset of data, testing on another
- STC(A) case results so far with logistic regression:
  - Pretty good: overall accuracy ~80%, very few errors on top and bottom 30% of customers; clear guidance to marketing/operations
  - Structure of the model (significant variables) give insight into why some customers may not purchase

Next: CART
classification and regression tree

INSEAD
The Business School for the World®

- Main idea: set of questions (business/decision rules) which partition data into pockets ("clusters") with similar characteristics

- These rules/questions form a tree-like graphic:
- Example: surviving the Titanic crash
  - #s in parenthesis: (prob. survive, % of data)



- Several way to "build" trees
  - We will look at two:
    Conditional inference, `ctree`
    Recursive partitioning, `rpart`
- CART is a "mother" (father?;) of many machine learning methods, e.g., random forest, gradient boosting machines (xgboost) [Session 7]

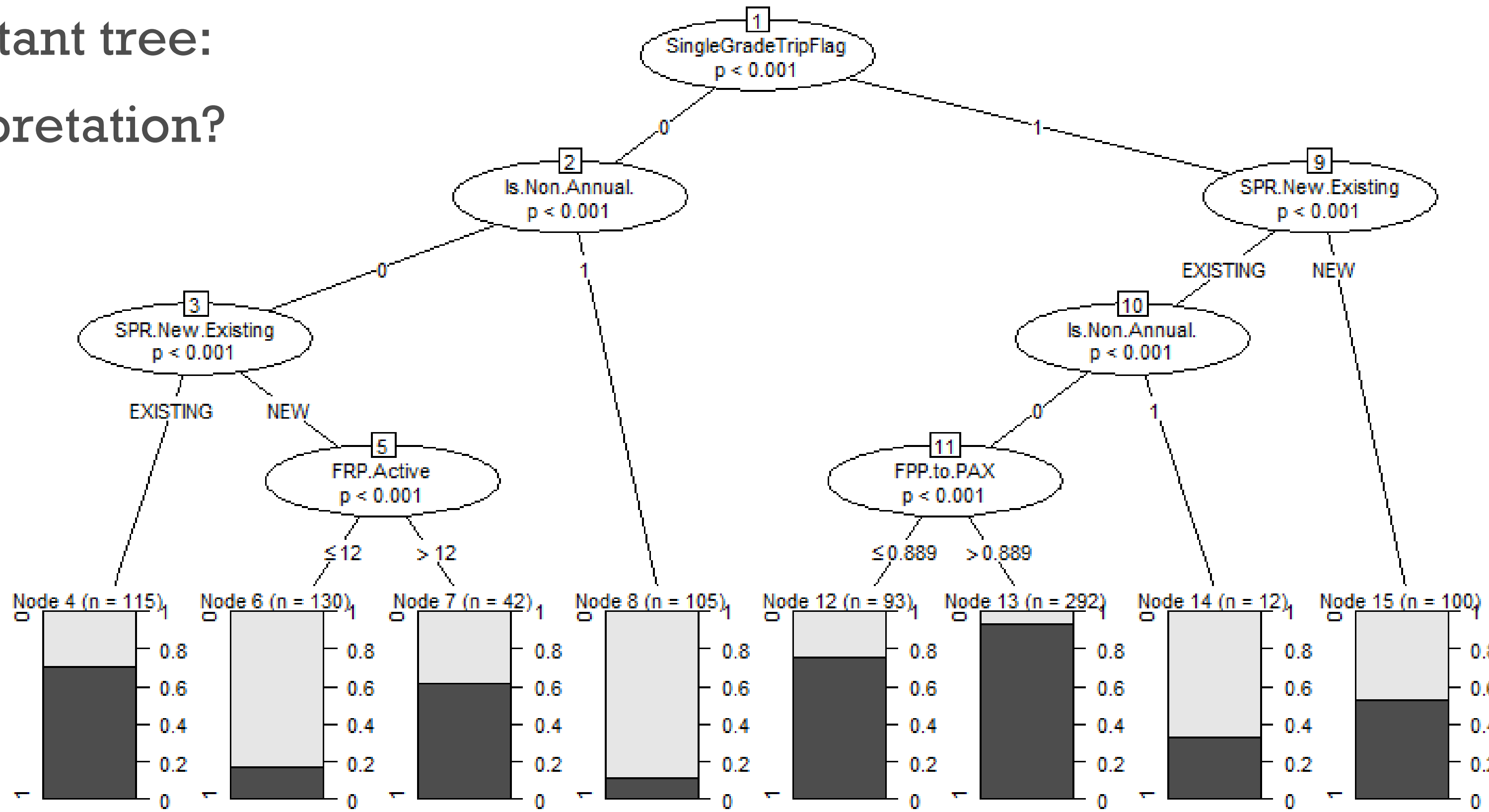
Source: https://en.wikipedia.org/wiki/Decision_tree_learning

STC(A), `ctree` CART

Remarks:

- `ctree` is slow and takes lots of memory when dealing with high-dimensional categorical data: combine categories or shrink training set
- For “apples-to-apples” comparison with logistic, keep same testing subset (500 datapoints), but sample 889 out of 1889 from training

- Resultant tree:
- Interpretation?



Some technical R remarks

INSEAD
The Business School for the World®

- Running a model with all variables included (use "dot" **.** for independent variables:

```
glm(Retained.in.2012.~ ., data=training, family="binomial"(link="logit")) # for logistic
ctree_tree<-ctree(Retained.in.2012~.,data=training) # for CART
```

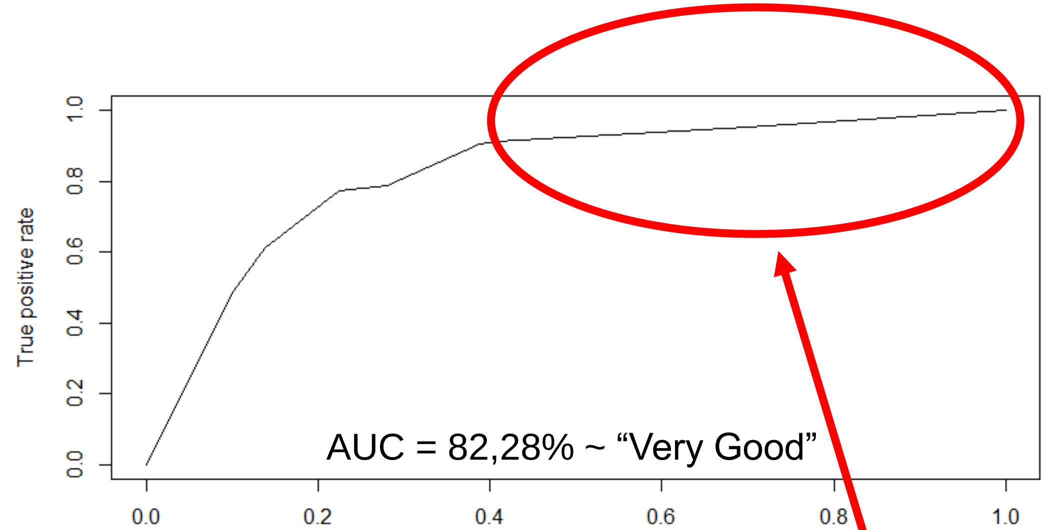
- Combining categories (this example, with less than 10 datapoints):

```
combinerarecategories<-function(data_frame,mincount){ #custom function to combine rare categories
 for (i in 1 : ncol(data_frame)){
  a<-data_frame[,i]
  replace <- names(which(table(a) < mincount))
  levels(a)[levels(a) %in% replace] <-paste("Other",colnames(data_frame)[i],sep=".")
  data_frame[,i]<-a  }
return(data_frame) }
STCdata<-combinerarecategories(STCdata,10) #combine categories with <10 values in STCdata into "Other"
```

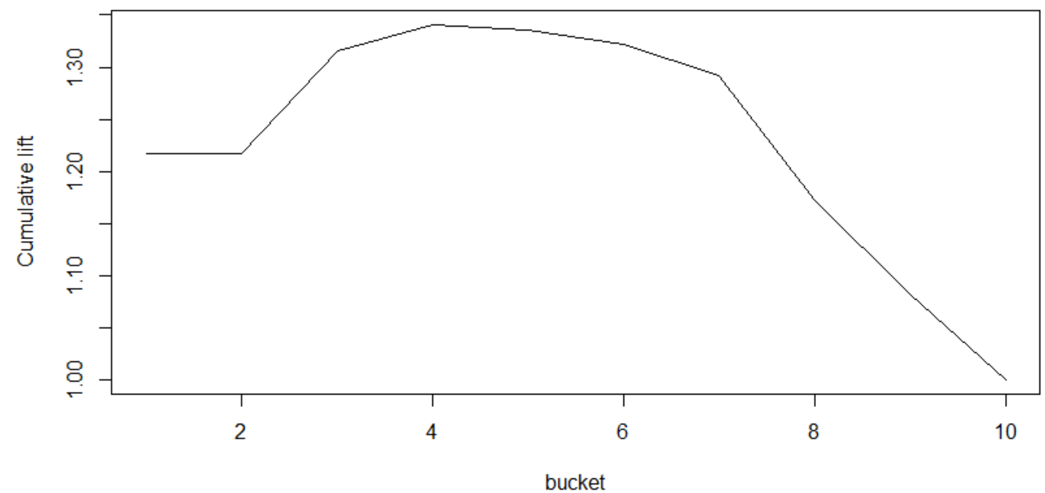
STC(A) results: ctree CART

INSEAD
The Business School for the World®

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 120  29
         1  76 275

               Accuracy : 0.79
                 95% CI : (0.7516, 0.8249)
    No Information Rate : 0.608
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5398
 Mcnemar's Test P-Value : 7.151e-06

            Sensitivity : 0.6122
            Specificity : 0.9046
         Pos Pred Value : 0.8054
         Neg Pred Value : 0.7835
             Prevalence : 0.3920
         Detection Rate : 0.2400
   Detection Prevalence : 0.2980
      Balanced Accuracy : 0.7584

       'Positive' Class : 0
```



Disbalanced: much better at predicting who will <u>not</u> purchase: under 10% mistakes in bottom 60%

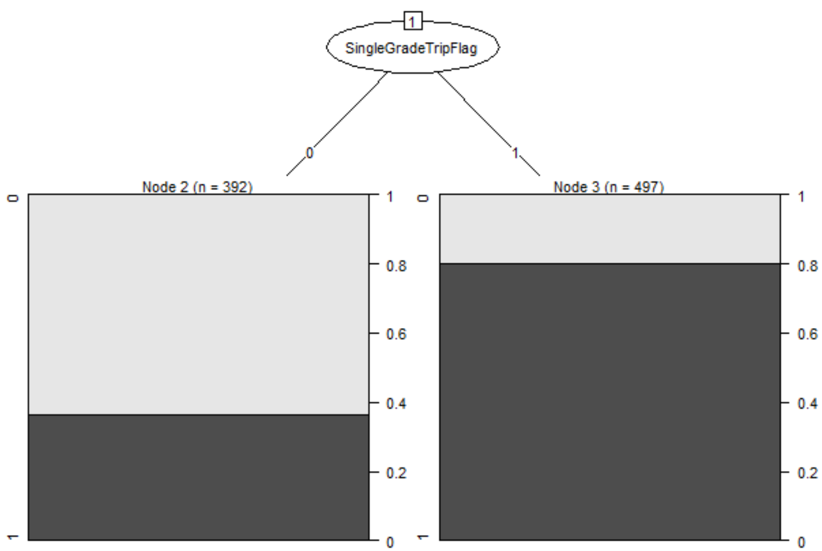


STC(A), `rpart` CART

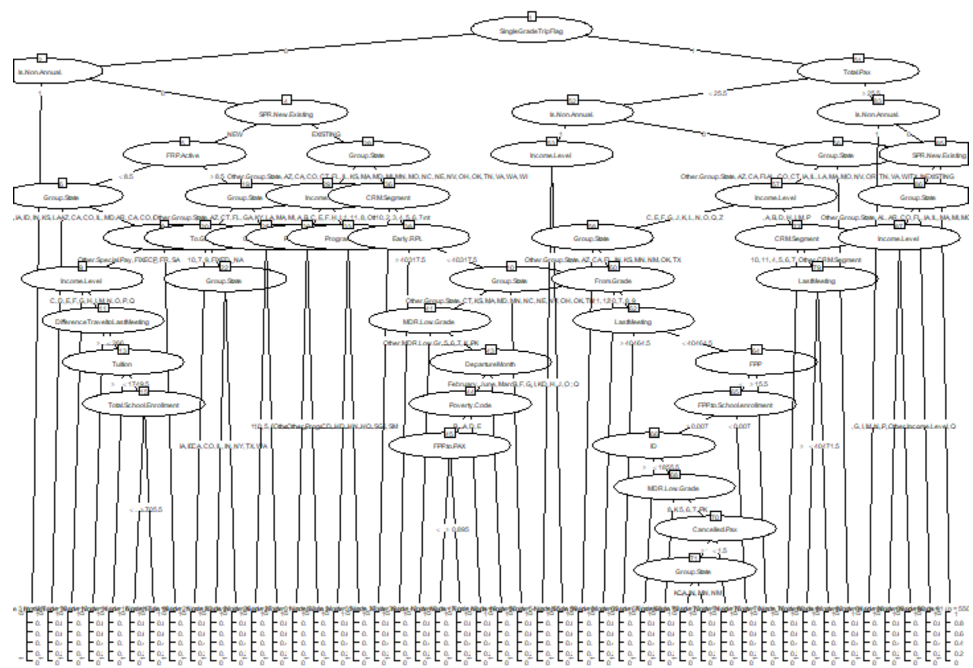
INSEAD
The Business School for the World®

Remarks:

- Unlike ctree, rpart methodology relies on a user-specified "cost paramter" (cp) to decide how to prune the tree
  - High cp: small tree, possible loss of precision on training and testing
  - Low cp: large tree, better fit on testing, but overfitting on training
- Interpretation?



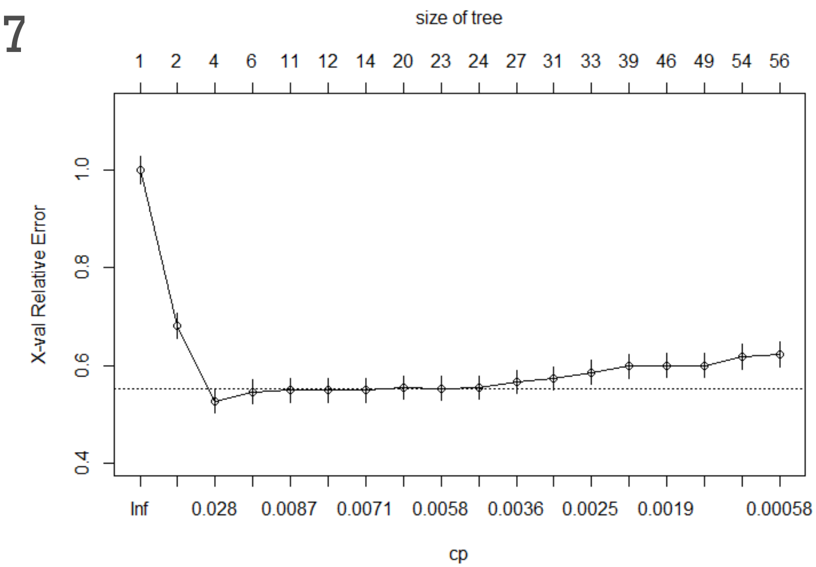
cp=0.2


cp=0.002

STC(A), `rpart` CART

Remarks:

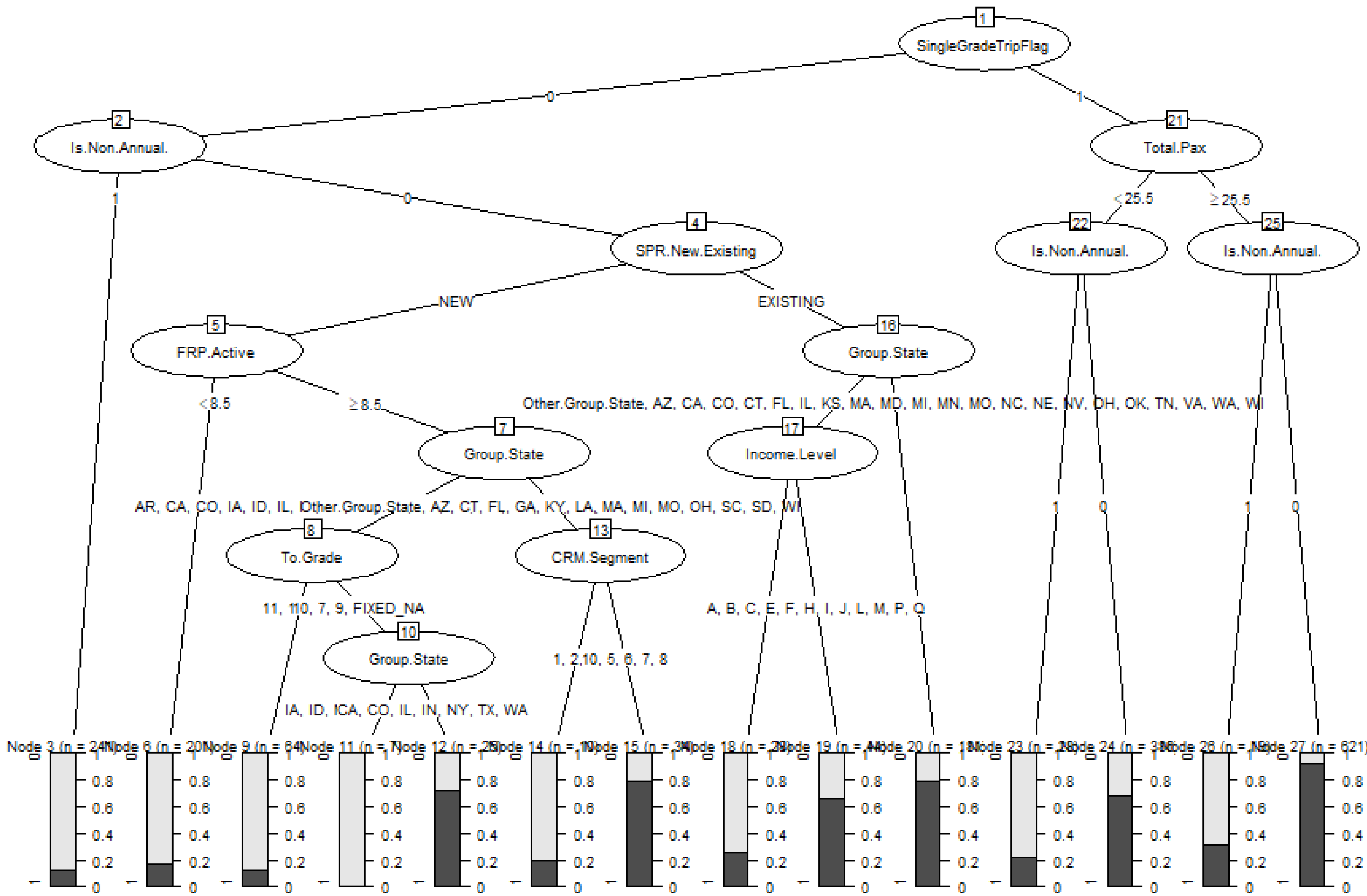
- Unlike ctree, rpart methodology relies on a user-specified "cost paramter" (cp) to decide how to prune the tree
  - High cp: small tree, possible loss of precision on training and testing
  - Low cp: large tree, better fit on testing, but overfitting on training
- Which cp to use?
- plotcp(rpart_tree) # rule of thumb: pick the largest cp at which error crosses dotted line
- In our case, ~0.007



STC(A) results:
`rpart` CART with cp=0.007

INSEAD
The Business School for the World®

- Interpretation? Does the tree "make sense"?

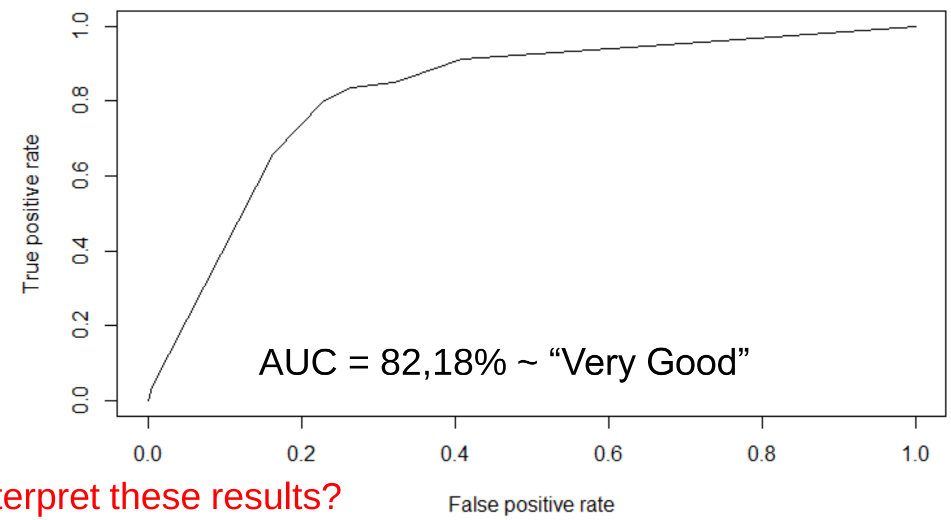


STC(A) results: `rpart` CART with cp=0.007

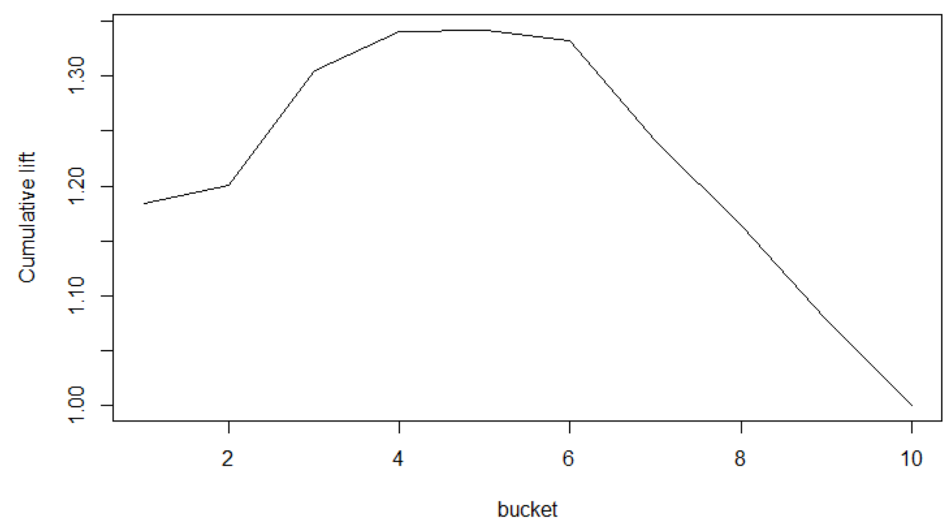
INSEAD

The Business School for the World®

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 133  46
         1  63 258

               Accuracy : 0.782
                 95% CI : (0.7432, 0.8174)
    No Information Rate : 0.608
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5355
 Mcnemar's Test P-Value : 0.1254

            Sensitivity : 0.6786
            Specificity : 0.8487
         Pos Pred Value : 0.7430
         Neg Pred Value : 0.8037
             Prevalence : 0.3920
         Detection Rate : 0.2660
   Detection Prevalence : 0.3580
      Balanced Accuracy : 0.7636

       'Positive' Class : 0
```



Do we know how to interpret these results?



Exercise:
first-hand glance at overfitting

INSEAD
The Business School for the World®

- Create a table of AUCs for the `rpart` method using various cps on both training and testing data
- Do you observe that while on training the AUC improves the lower cp you use?
  - Why? A: the tree becomes more elaborate.
- But what happens on testing data?
  - Do you observe that those elaborate trees preform worse - exactly because they too elaborately capture the nuances of the training data, which may not be present in testing.
  - That's overfitting!

STC(A) Summary

INSEAD
The Business School for the World®

- We have three models: logistic regression, CART1, CART2
- Which one would you use and why?

- What are the business implications?

- How can you improve the performance of the model?
  - STC(B) case: additional data on customer satisfaction, as measured by the NPS (net promoter score) -> Tutorial 2

Summary of Sessions 5-6

INSEAD
The Business School for the World®

- Large volumes of data about people/behavior increased the importance of an analytical task to predict an outcome of an event:
  - Will a customer churn? Default? Open email? [binary outcome]
  - Which item from a set will the customer choose? (iPhone model, bottle size, transit mode, job offer) [multinomial outcome]
- Predicting events ~ Classification: Which customers will churn? Which customers will buy iPhone X, etc.?
- We studied two Data Science methods for classification:
  - Logistic regression: build a linear model for utility and an exp transformation to predict the probability of an event
  - CART: build a decision-tree-like structure for describing pockets of data with similar properties wrp the occurrence of the event
- R code templates for both + some further "tricks"
  - data cleaning, custom functions, holdout, stepAIC, against overfitting

Next…

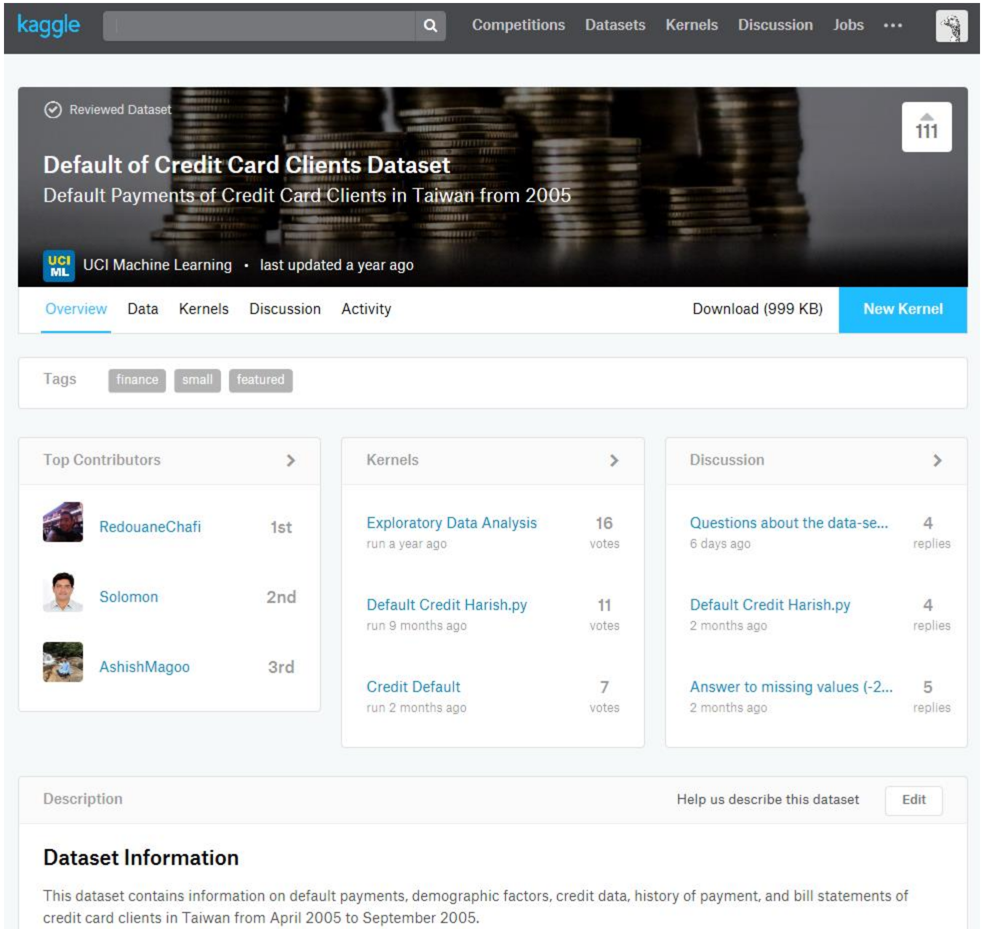
INSEAD
The Business School for the World®

- Tutorial 2: [next Mon, Feb 5, 19:15h, amphi 102]
  - mid-term R help, specifically on predicting events
  - First exposure to notebooks and *.rmd files
    - Make sure to follow instructions for Session 7, open GitHub account and "fork" the materials for the rest of the course
    - Sessions 7-12 will use the open/INSEADAnalaytics site a lot

- Assignment 2:
  - "Predicting credit defaults"
  - The data and the Assignment 2 comes from Kaggle…

Assignment 2 Kaggle source

INSEAD
The Business School for the World®

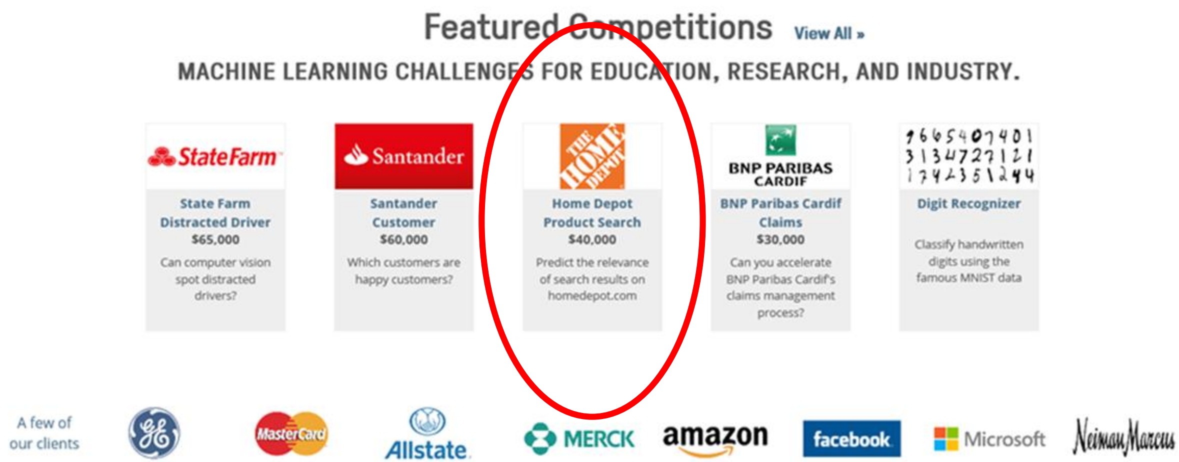


Assignment 2 is on the DSB course's open website (INSEADAnalytics):
http://inseaddataanalytics.github.io/INSEADAnalytics/SGP18J.html

What is Kaggle?

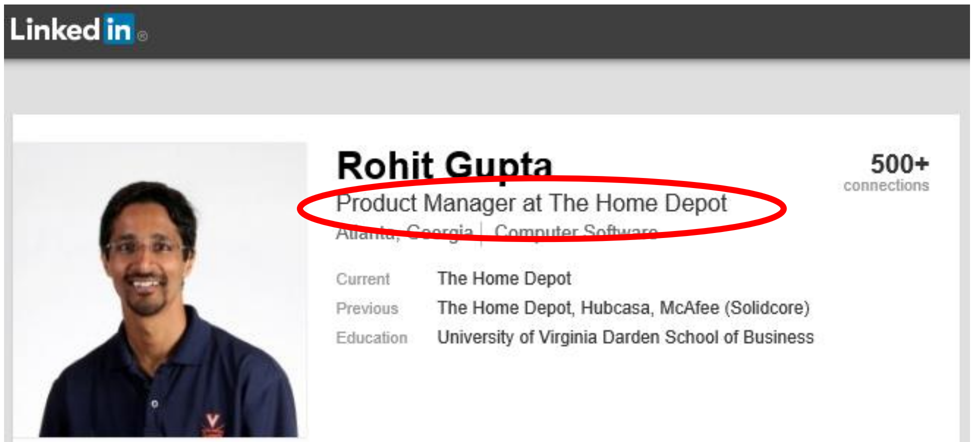
INSEAD
The Business School for the World®

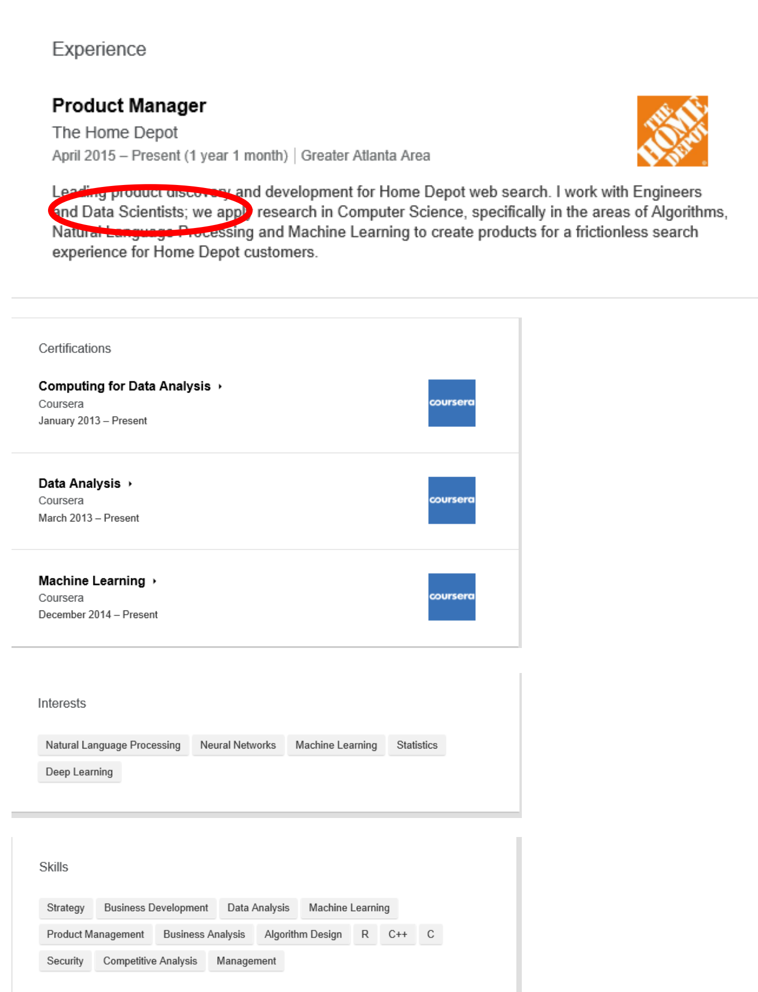
- Kaggle is the world's largest community of data scientists
- Kagglers compete with each other to solve complex data science problems in **free, public competitions**
- Top data scientists/competitors are invited to work on the most interesting and sensitive business problems from some of the world's biggest companies through **Masters competitions** ["Netflix challenge"]



New "Breed" of Managers Focused on Data Science

INSEAD
The Business School for the World®

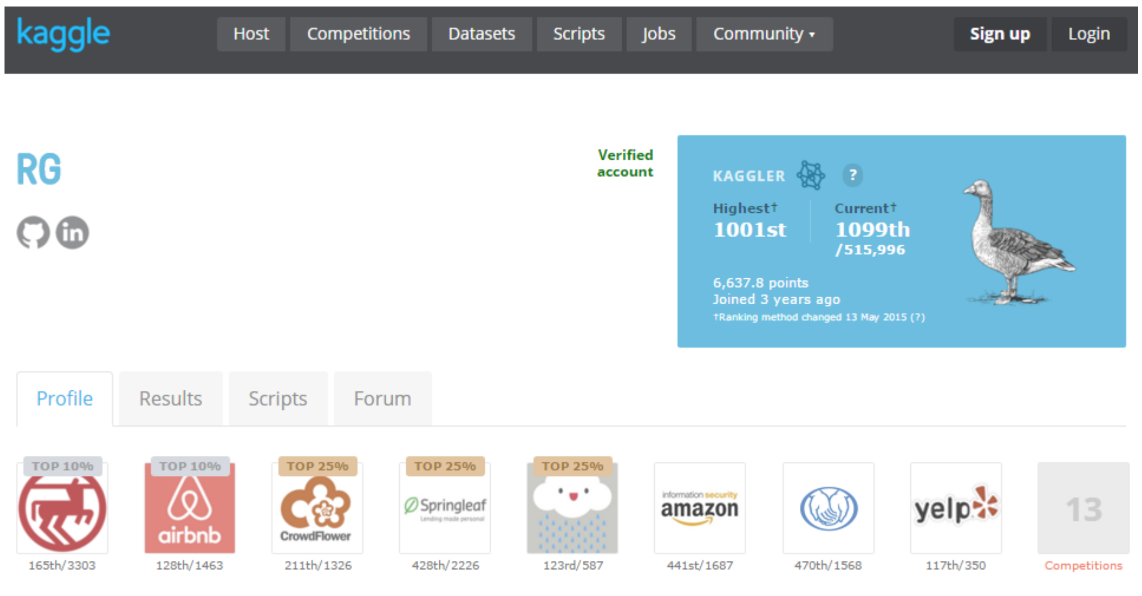




Kaggle Competitor

INSEAD
The Business School for the World®

- [As of end of 2015] Rohit has finished in the top 10% twice and in the top 25% three times

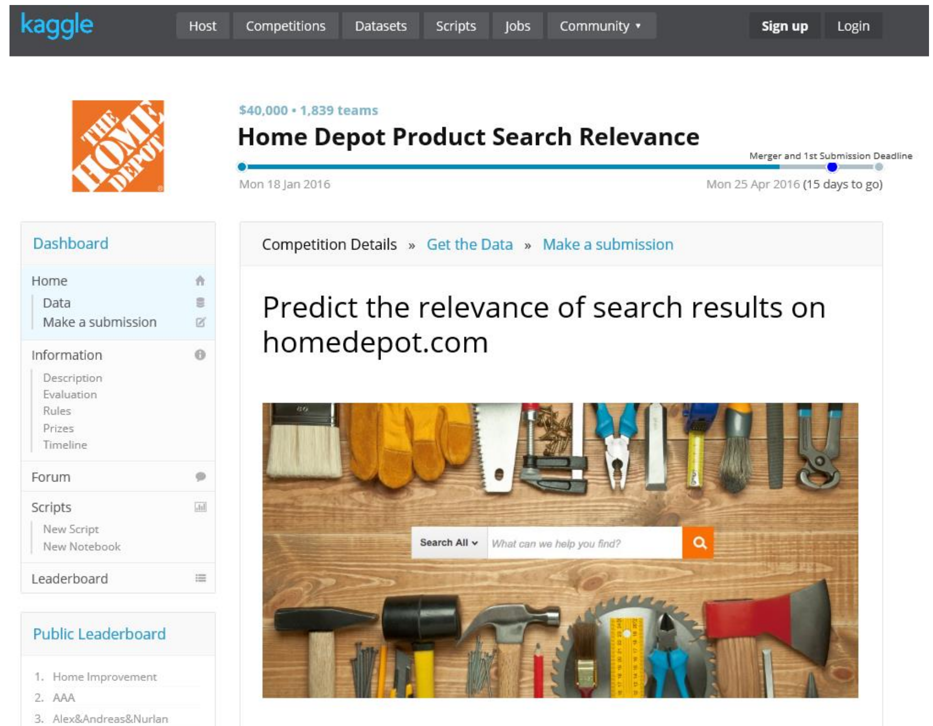


Kaggle Host

INSEAD
The Business School for the World®

HomeDepot website search relevancy project:

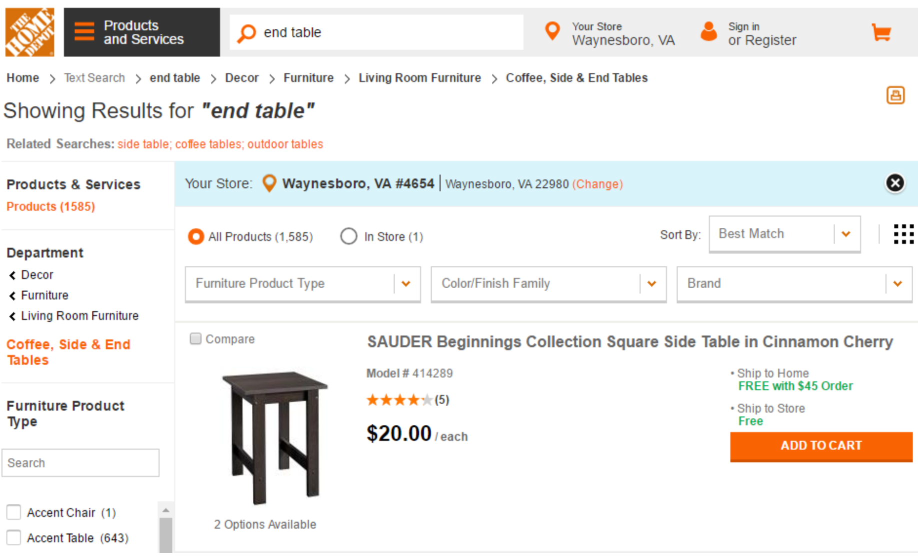
- Search relevancy is an implicit measure Home Depot uses to gauge how quickly they can get customers to the right products. Currently, human raters evaluate the impact of potential changes to their search algorithms, which is a slow and subjective process. By removing or minimizing human input in search relevance evaluation, Home Depot hopes to increase the number of iterations their team can perform on the current search algorithms

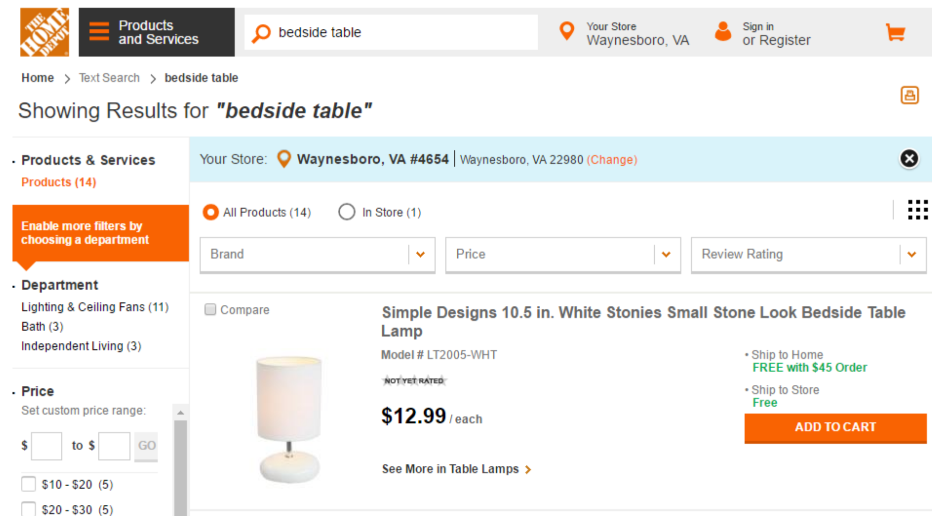


Implications for Home Depot

INSEAD
The Business School for the World®

Web sales accounted for $3.37 billion of Home Depot's $67.54 billion in 2015 Q1-Q3 sales.





Kaggle: Home Depot Data

INSEAD

The Business School for the World®

Training data:

	A	B	C	D	E
1	id	product_uid	product_title	search_term	relevance
2	2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3
3	3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.5
4	9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 Tugboat Wood and Concrete Coating	deck over	3
5	16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)	rain shower head	2.33
6	17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)	shower only faucet	2.67
7	18	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Microwave in Stainless Steel with Sensor Cooking	convection otr	3
8	20	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Microwave in Stainless Steel with Sensor Cooking	microwave over stove	2.67
9	21	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Microwave in Stainless Steel with Sensor Cooking	microwaves	3
10	23	100007	Lithonia Lighting Quantum 2-Light Black LED Emergency Fixture Unit	emergency light	2.67
11	27	100009	House of Fara 3/4 in. x 3 in. x 8 ft. MDF Fluted Casing	mdf 3/4	3
12	34	100010	Valley View Industries Metal Stakes (4-Pack)	steele stake	2.67

Testing data:

	A	B	C	D	E
1	id	product_uid	product_title	search_term	relevance
2	1	100001	Simpson Strong-Tie 12-Gauge Angle	90 degree bracket	?
3	4	100001	Simpson Strong-Tie 12-Gauge Angle	metal l brackets	?
4	5	100001	Simpson Strong-Tie 12-Gauge Angle	simpson sku able	?
5	6	100001	Simpson Strong-Tie 12-Gauge Angle	simpson strong ties	?
6	7	100001	Simpson Strong-Tie 12-Gauge Angle	simpson strong tie hcc668	?
7	8	100001	Simpson Strong-Tie 12-Gauge Angle	wood connectors	?
8	10	100003	STERLING Ensemble 33-1/4 in. x 60 in. x 75-1/4 in. Bath and Shower Kit with Right-Hand Drain in White	bath and shower kit	?
9	11	100003	STERLING Ensemble 33-1/4 in. x 60 in. x 75-1/4 in. Bath and Shower Kit with Right-Hand Drain in White	bath drain kit	?
10	12	100003	STERLING Ensemble 33-1/4 in. x 60 in. x 75-1/4 in. Bath and Shower Kit with Right-Hand Drain in White	one piece tub shower	?
11	13	100004	Grape Solar 265-Watt Polycrystalline Solar Panel (4-Pack)	solar panel	?
12	14	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in Chrome (Valve Not Included)	1 handle shower delta trim kit	?

Next… [cont.]

INSEAD
The Business School for the World®

- … you will also have a different professor – Spyros Zoumpoulis
- I will still be available for help with projects, and will see you at the projects' presentations in Sessions 13-14

- "Final" remarks:
  - You've learned multiple powerful techniques/tools for "beyond Excel" data analyses. Become leaders of data-driven decision making in your organizations – use those tools and your knowledge!

INSEAD

The Business School for the World®

Next… [cont.]

INSEAD
The Business School for the World®

- … you will also have a different professor – Spyros Zoumpoulis
- I will still be available for help with projects, and will see you at the projects' presentations in Sessions 13-14

- "Final" remarks:
  - You've learned multiple powerful techniques/tools for "beyond Excel" data analyses. Become leaders of data-driven decision making in your organizations – **use those tools and your knowledge!**
  - You've also learned that there are lots of things you don't know: when in need, seek for help, hire experts …
  - … and become their bosses

- It was my pleasure teaching [and learning] with you, lets stay connected!
- *https://www.**linkedin**.com/in/**antonovchinnikov***


INSEAD
The Business School
for the World®
Europe
Asia
Middle East