INSEAD

The Business School for the World®

Prof. Anton Ovchinnikov

Prof. Spyros Zoumpoulis

DSB Classes 7-8, February 6, 2018

- Advanced Classification. From .R to Notebooks. Dimensionality Reduction

Structure of the course

INSEAD
The Business School for the World®

- SESSIONS 1-2 (AO): Data analytics process; from Excel to R
  - Tutorial 1: Getting comfortable with R
- SESSIONS 3-4 (AO): Time Series Models
- SESSIONS 5-6 (AO): Intro to classification, logistic regression and machine learning
  - Tutorial 2: Midterm R help / classification
- SESSIONS 7-8 (SZ): Advanced Classification; From .R to Notebooks; Dimensionality reduction
- SESSIONS 9-10 (SZ): Clustering and Segmentation
  - Tutorial 3: Q&A on R for three main modules
- SESSIONS 11-12 (SZ): Catch-up and wrap-up; Guest speaker
  - "Tutorials 4,5": Hands-on help on projects
- SESSIONS 13-14 (AO+SZ): Project presentations

Plan for the day
Learning objectives

INSEAD
The Business School for the World®

- Advanced classification: metrics and methods
  - Regularization. Advanced tree methods.

- From .R scripts to Notebooks
  - New way/process for doing and communicating analytics with reproducible, publication-quality output

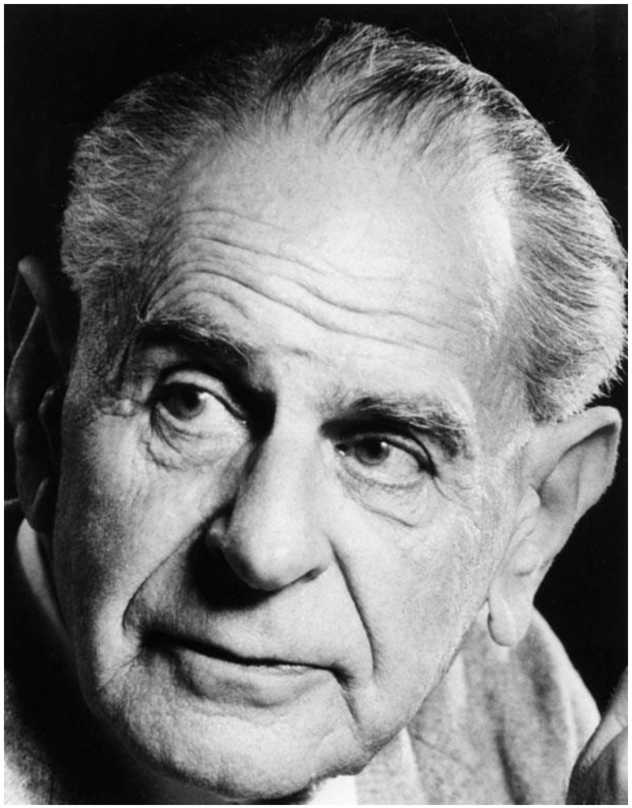
- Derived attributes and dimensionality reduction
  - Generate (a small number of) new manageable/interpretable attributes that capture most of the information in the data

Overfitting & Regularization

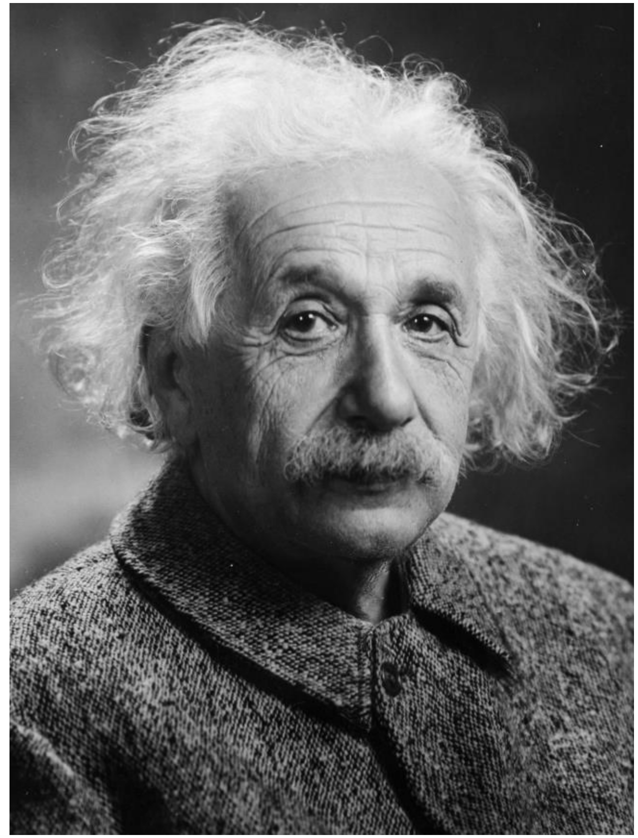
INSEAD
The Business School for the World®

- What happened when in Assignment 2, you made a `rpart` CART tree with very small cp?
- Fundamental tradeoff of learning with data
  - Models that are too simple are not accurate on the training set, and also don't generalize well on the test set
  - Models that are too accurate on training set are too complex, and therefore don't generalize well on the test set

Overfitting & Regularization



Karl Popper



Albert Einstein

Overfitting & Regularization

- Need to fine-tune the model so that is strikes a good balance between accuracy and simplicity
- Cross-validation does this fine-tuning
  - Break the data into training data, validation data, test data
  - Train model using training data
  - Test on validation data to fine-tune parameters, and iterate
  - "When happy," test (once) on test data to simulate how model would do in the real world

Overfitting & Regularization

INSEAD
The Business School for the World®

- Regularization: set of techniques to reduce overfitting
  - For logistic regression:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} - \log likelihood(\beta, data) + \lambda \left(\frac{1-\alpha}{2} \sum_i \beta_i^2 + \alpha \sum_i |\beta_i| \right)$$

measures **fit**

controls trade off between maximizing fit and minimizing complexity

measures **complexity**

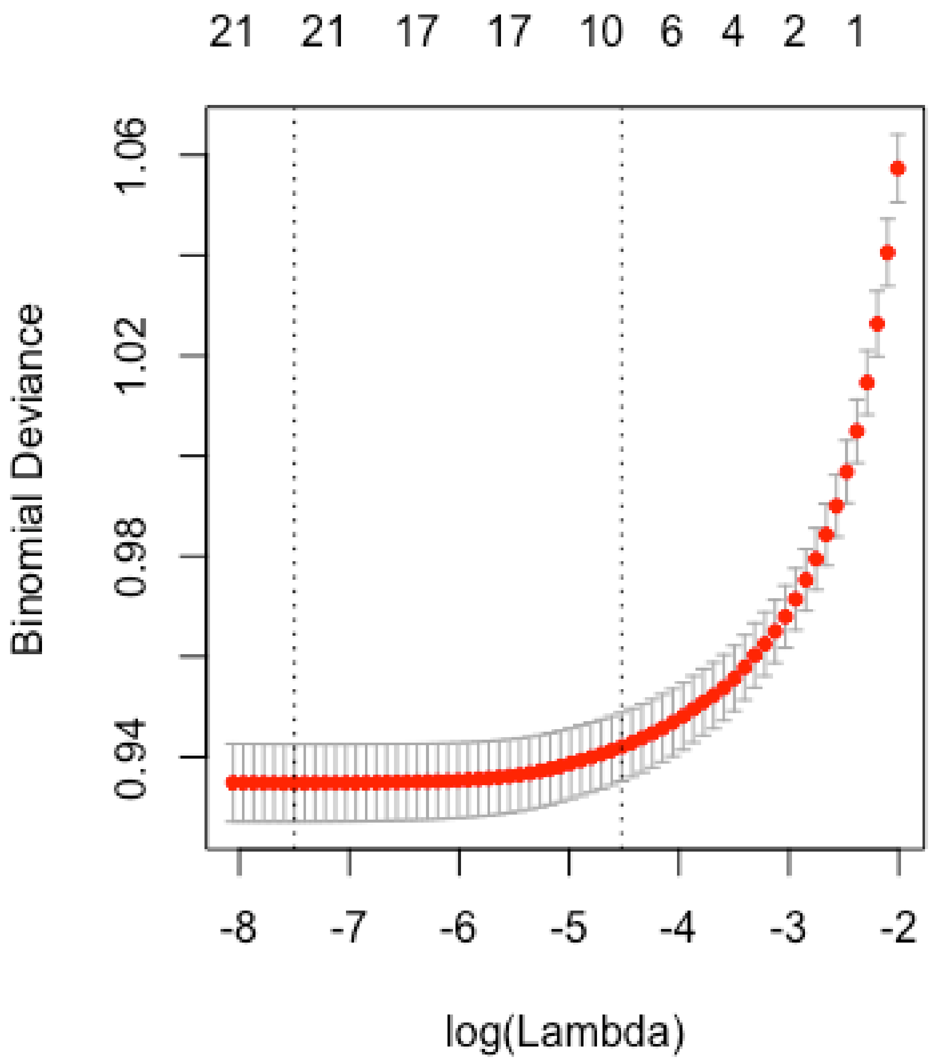
- $\alpha = 1$: penalize sum of absolute values of coefficients. Lasso regression
- $\alpha = 0$: penalize sum of squares of coefficients. Ridge regression

```
Package: glmnet
cv.out <- cv.glmnet(as.matrix(estimation_data[,independent_variables]),estimation_data[,dependent_variable],alpha=1,
                 family="binomial" )
#family= "binomial" => logistic regression
#alpha=1: Lasso
lambda <- cv.out$lambda.1se #choose value of λ
log_reg_coefficients <- as.matrix(coef(cv.out,s=lambda)) #extract the estimated coefficients
```

Overfitting & Regularization

INSEAD
The Business School for the World®

> plot(cv.out)



- λ that minimizes mean cross-validated error:

> log(cv.out$lambda.min)

[1] -7.498859

- largestλ s.t. error is within 1 standard error of the minimum:

> log(cv.out$lambda.1se)

[1] -4.52178

Emphasizes simplicity (even) more

Important classification metric: Profit Curve

INSEAD
The Business School for the World®

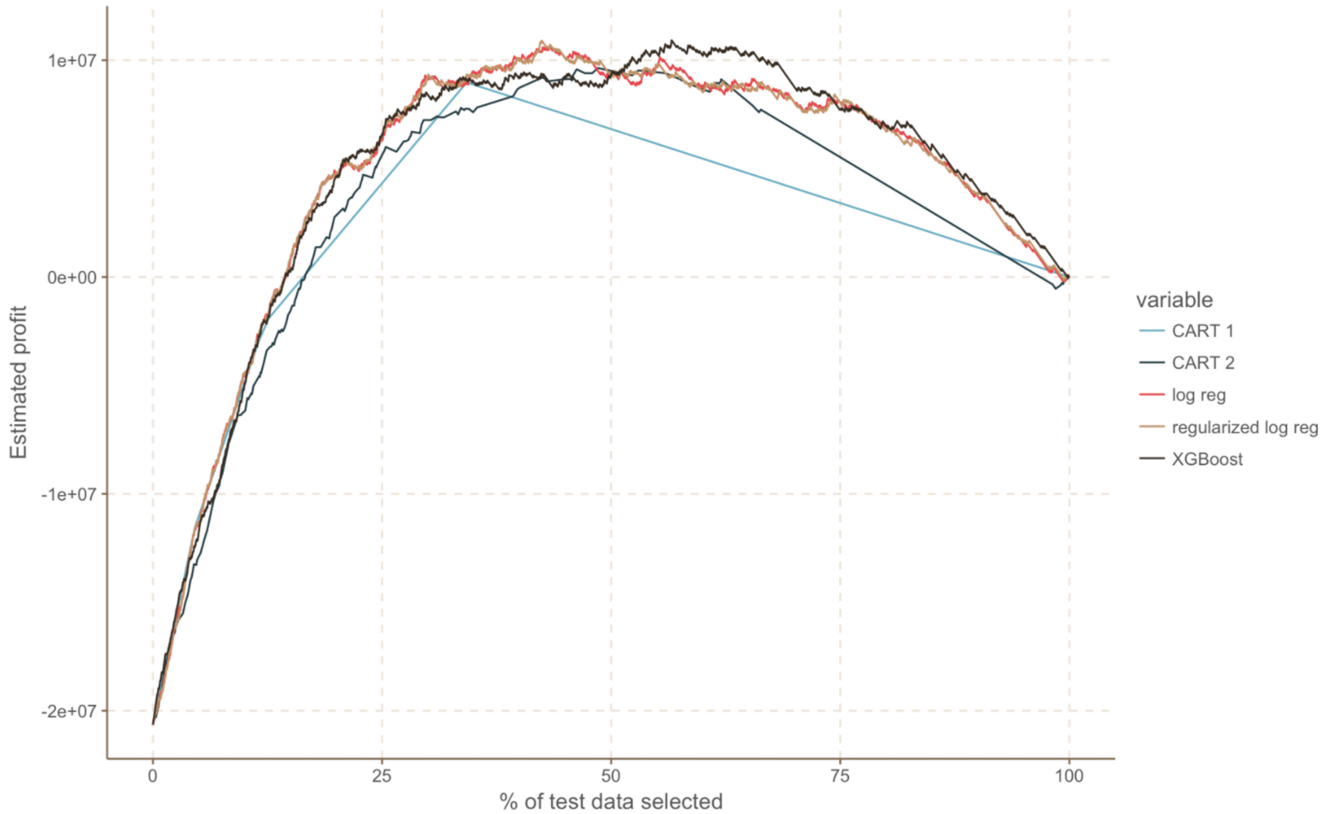
- Measure business profit if we only select the top cases in terms of the probability of "response"
- For this, we need to define values and costs of correct classifications and misclassifications

	Predicted: default	Predicted: no default
Actual: default	$0	-$100000
Actual: no default	$0	$20000

Profit = # of 1's correctly predicted * value of capturing a 1
+# of 0's correctly predicted * value of capturing a 0
+# of 1's incorrectly predicted as 0 * cost of missing a 1
+# of 0's incorrectly predicted as 1 * cost of missing a 0

Important classification metric: Profit Curve

- Given a classifier, rank instances in the test data from highest predicted probability of belonging to class 1 (= default) to lowest
- Can put the cutoff for giving vs. not giving credit at any rank
- As I move the cutoff, calculate the corresponding profit…



Feature Engineering

INSEAD
The Business School for the World®

Your data may have more information than what is contained in your existing variables

- Spend lots of time thinking of ways to combine your variables into new ones!
- "Engineering" good features may be more important than using a better method
- Requires contextual knowledge of the business
  - Can not be outsourced
  - Can not be automated

Feature Engineering

INSEAD
The Business School for the World®

Example for credit card default case:

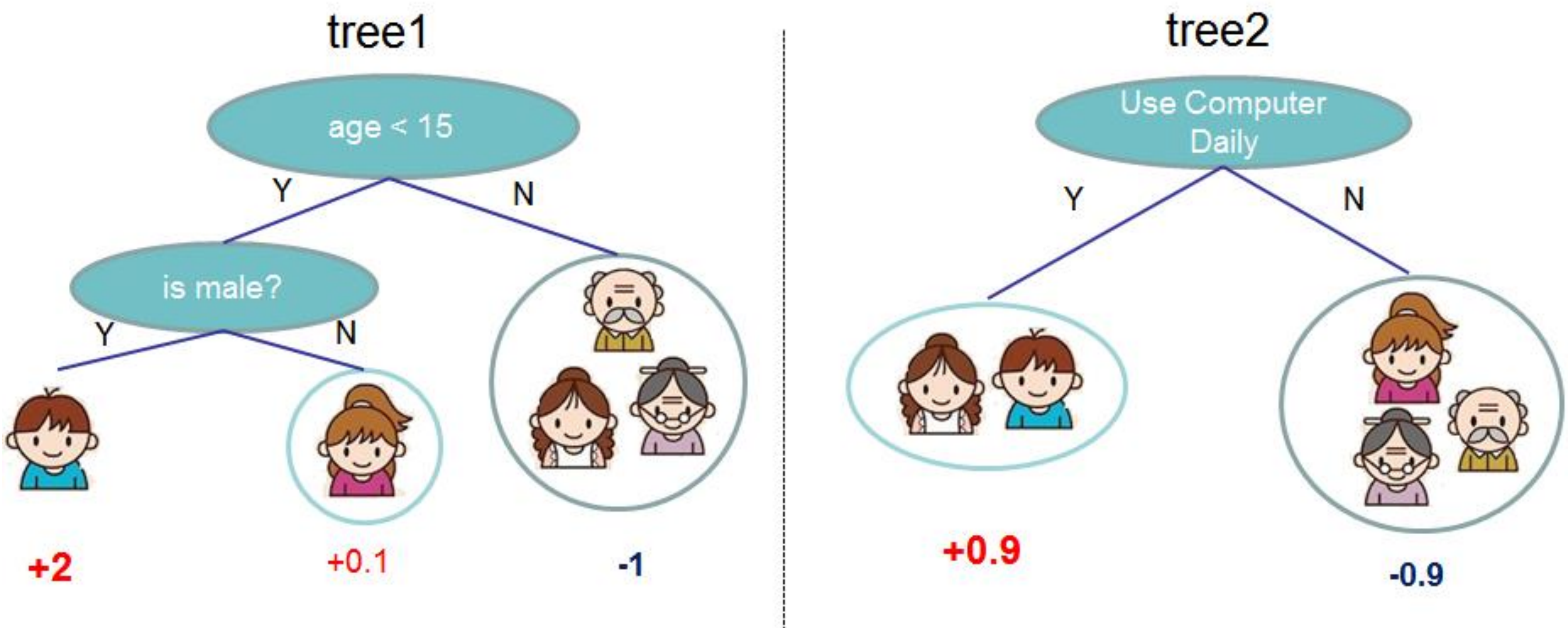
```
tmpx = t(apply(ProjectData[,7:12], 1,
                    function(r) matrix(c(sum(r==-2), sum(r==-1), sum(r==0),sum(r > 0)), nrow=1)))
                    #apply: apply the function to an array of values
                    # argument "1": apply the function over rows
ProjectData = cbind(ProjectData[,2:5], #cbind: combine a set of columns
                     tmpx,
                     apply(ProjectData[,13:18], 1, function(r) median(r[!is.na(r)])),
                     apply(ProjectData[,19:24]/ProjectData[,13:18], 1, function(r)
                              ifelse(sum(!is.na(r) & !is.infinite(r)), mean(r[!is.na(r) & !is.infinite(r)]),0)),
                     ProjectData[,25])

dependent_variable = 11
independent_variables = c(1:10) # use all the new attributes
```

Tree Ensemble Methods

INSEAD
The Business School for the World®

- Main idea: put a set of CARTs together, output a combination (e.g., mode, mean) of the respective outputs the CARTs

Does someone like computer games?



f() = 2 + 0.9= 2.9 f()= -1 - 0.9 = -1.9

Source: http://xgboost.readthedocs.io/en/latest/model.html#

Tree Ensemble Methods

INSEAD
The Business School for the World®

Both **random forests** and **boosted trees** generate multiple random samples from the training set (with replacement), and train a different CART for each sample of the data. This is called bagging.

- Random Forests
  - The samples are completely random. No adaptiveness.
  - Use fully grown CARTs (each with low bias, high variance). Reduce variance by bagging together many uncorrelated trees.
  - Final prediction is the simple average

- Boosted trees
  - Based on weak learners (each with high bias, low variance)
  - But adaptive: instances that had been modeled poorly by the overall system before have larger probability of being picked now → higher weight
  - Final prediction is a weighted average

Tree Ensemble Methods

INSEAD
The Business School for the World®

- Random Forests

Package: randomForest

```
model_forest <- randomForest(x=estimation_data[,independent_variables],
                             y=estimation_data[,dependent_variable],
                             importance=TRUE, proximity=TRUE, type="classification")
```

- Boosted trees

Package: xgboost

```
model_xgboost <- xgboost(data = as.matrix(estimation_data[,independent_variables]),
                         label = estimation_data[,dependent_variable],
                         eta = 0.3, max_depth = 10, nrounds=10, objective = "binary:logistic", verbose = 0)
#objective= "binary:logistic" => logistic regression for classification
#eta: step size of each boosting step. max.depth: maximum depth of tree.
#nrounds: the max number of iterations
```

How to then retrieve predicted probabilities (and therefore also classes)?

```
validation_Probability_class1<-   predict(model,newdata=as.matrix(validation_data[,independent_variables]),
                                  type=  "prob" )
```

(A) Process for Classification

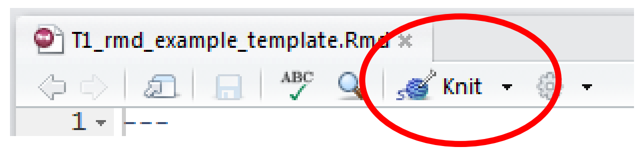
INSEAD
The Business School for the World®

1. Split the data
2. Set up the dependent variable
3. Simple Analysis
4. Classification and Interpretation
5. Validation accuracy
   - Use various classification metrics you know
6. Test accuracy

From R to Notebooks

- You traditional approach for "using" analytics has been two-step:
  - "do" analytics (e.g., plot a graph in Excel)
  - "communicate" analytics (e.g., copy-paste the graph into a PowerPoint presentation / Word file report, etc.)

- With coding (and R) there is a better way: "notebooks"
  - "knit" the R markdown (*.Rmd) file



  - This will create a *.html report (a webpage) with the analysis outputs, graphs, text. Can also create a PDF report
  - Main advantage of this approach: ALL IN ONE PLACE
    - When the new data is available (e.g., next quarter's sales numbers come in), creating an updated report will take you… 1 click
  - Along with sharing tools (GitHub): reusable, replicable, easy to share, all-in-one-place way of doing analytics and communicating them with publication-quality output

Derived Attributes and Dimensionality Reduction

INSEAD

The Business School for the World®

- What is dimensionality reduction?
  - Generate (a small number of) new attributes that are (linear) combinations of the original ones, and capture most of the information in the original data
  - Often used as the first step in data analytics

- Why do dimensionality reduction?
  - Computational and statistical reasons: with thousands of features, very expensive and hard to estimate a good model
  - Managerial reason: the new attributes are interpretable and actionable

- The key idea of dimensionality reduction
  - Transform the original variables into a smaller set of **factors**
  - Understand and interpret the factors
  - Use the factors for subsequent analysis

Dimensionality Reduction: Key Questions

INSEAD

The Business School for the World®

1. How many factors do we need?
2. How would you name the factors? What do they mean?
3. How interpretable and actionable are the factors we found?

(A) Process for Dimensionality Reduction

INSEAD

The Business School for the World®

1. Confirm the data is metric
2. Scale the data
3. Check correlations
4. Choose number of factors
5. Interpret the factors
6. Save factor scores

Applying Dimensionality Reduction: Evaluation of MBA Applications

INSEAD
The Business School for the World®

Variables available:

1. GPA
2. GMAT score
3. Scholarships, fellowships won
4. Evidence of communications skills
5. Prior job experience
6. Organizational experience
7. Other extra curricular achievements

Which variables are correlated? What do these groups of variables capture?

Step 1: Confirm data is metric

INSEAD
The Business School for the World®

	Variables	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
1	1	3	580	2	3.5	5	3.8	4
2	2	3.2	570	2	3.8	6	3.8	3.8
3	3	3.7	690	3	3.3	3	3.2	3.6
4	4	3.9	760	3	3.8	5	3.9	3.2
5	5	2.8	480	2	3.2	6	3.8	3.8
6	6	3.4	520	2.5	2.6	2	2.5	2.4
7	7	3.6	670	3	3.7	4	3.5	2.9
8	8	3.6	760	3	3.9	5	3.3	3.2

Step 2: Scale the data

INSEAD
The Business School for the World®

Before standardization

	Variables	min	X25.percent	median	mean	X75.percent	max	std
1	GPA	2.5	2.8	3.45	3.31	3.62	3.9	0.47
2	GMAT	380	480	575	583.5	682.5	760	119.44
3	Fellow	1	2	2.8	2.45	3	3.8	0.91
4	Comm	2	3.18	3.4	3.34	3.73	3.9	0.49
5	Job.Ex	2	3	5	4.25	5.25	6	1.52
6	Organze	1	3.05	3.4	3.2	3.8	3.9	0.73
7	Extra	2.4	2.88	3.4	3.3	3.8	4	0.52

Step 2: Scale the data

Standardization….

```
ProjectDatafactor_scaled=apply(ProjectDataFactor,2, function(r) { #"2" applies the function over columns
 if (sd(r)!=0) {
  res=(r-mean(r))/sd(r)
  } else {
   res=0*r; res
   }
 })
```

Step 2: Scale the data

INSEAD
The Business School for the World®

After standardization

	Variables	min	X25.percent	median	mean	X75.percent	max	std
1	GPA	-1.72	-1.08	0.31	0	0.68	1.27	1
2	GMAT	-1.7	-0.87	-0.07	0	0.83	1.48	1
3	Fellow	-1.6	-0.5	0.39	0	0.61	1.49	1
4	Comm	-2.73	-0.33	0.13	0	0.8	1.16	1
5	Job.Ex	-1.48	-0.82	0.49	0	0.66	1.15	1
6	Organze	-2.99	-0.2	0.27	0	0.82	0.95	1
7	Extra	-1.75	-0.83	0.19	0	0.97	1.36	1

Step 3: Check correlations

INSEAD
The Business School for the World®

	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
GPA	1.00	0.90	0.92	0.56	0.15	-0.03	0.01
GMAT	0.90	1.00	0.86	0.78	0.33	0.19	0.16
Fellow	0.92	0.86	1.00	0.59	0.18	0.01	0.02
Comm	0.56	0.78	0.59	1.00	0.60	0.47	0.39
Job.Ex	0.15	0.33	0.18	0.60	1.00	0.80	0.77
Organze	-0.03	0.19	0.01	0.47	0.80	1.00	0.61
Extra	0.01	0.16	0.02	0.39	0.77	0.61	1.00

Step 3: Check correlations

INSEAD
The Business School for the World®

	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
GPA	1.00	0.90	0.92	0.56	0.15	-0.03	0.01
GMAT	0.90	1.00	0.86	0.78	0.33	0.19	0.16
Fellow	0.92	0.86	1.00	0.59	0.18	0.01	0.02
Comm	0.56	0.78	0.59	1.00	0.60	0.47	0.39
Job.Ex	0.15	0.33	0.18	0.60	1.00	0.80	0.77
Organze	-0.03	0.19	0.01	0.47	0.80	1.00	0.61
Extra	0.01	0.16	0.02	0.39	0.77	0.61	1.00

Step 4: Choose the number of factors

INSEAD

The Business School for the World®

We use Principal Component Analysis

Package: psych

```
UnRotated_Results<-principal(ProjectDataFactor, nfactors=ncol(ProjectDataFactor),
                             rotate="none", score=TRUE)
```

- Factors are linear combinations of the original raw attributes…
- …so that they capture as much of the variability in the data as possible
- Factors are uncorrelated, and as many as the variables
- Each factor has an associated "eigenvalue" – which corresponds to the amount of variance captured by that factor
- First factor has the highest eigenvalue and explains most of the variance, then the second, …, and so on

Step 4: Choose the number of factors

INSEAD
The Business School for the World®

Package: FactoMineR

```
Variance_Explained_Table_results<-PCA(ProjectDataFactor, graph=FALSE)
Variance_Explained_Table<-Variance_Explained_Table_results$eig
```

	Eigenvalue	Pct of explained variance	Cumulative pct of explained variance
Component 1	3.74	53.48	53.48
Component 2	2.27	32.40	85.88
Component 3	0.42	6.07	91.95
Component 4	0.29	4.11	96.06
Component 5	0.14	1.99	98.05
Component 6	0.10	1.41	99.46
Component 7	0.04	0.54	100.00

```
> Variance_Explained_Table[1,1]/sum(Variance_Explained_Table[,1])
[1] 0.5347987
```

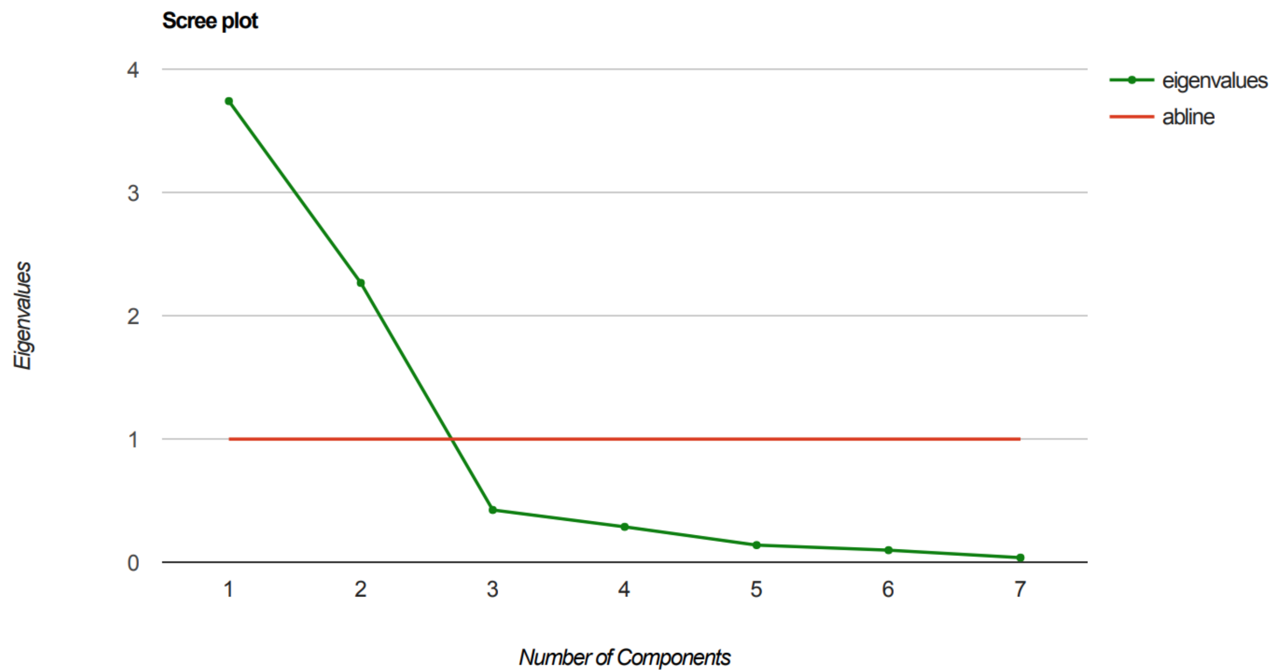
??

Step 4: Choose the number of factors

INSEAD
The Business School for the World®

We want to capture as much of the variance as possible, with as few factors as possible. How to choose the factors? Three criteria to use:

- Select all factors with eigenvalue > 1
- Select factors with highest eigenvalues up to exceeding a threshold (e.g. 65%) in cumulative % of explained variance
- Select factors up to the "elbow" of the scree plot



Step 5: Interpret the factors

INSEAD
The Business School for the World®

To interpret the factors, we want them to use only a few, non-overlapping original attributes

- Factor "rotations" transform the estimated factors into new ones that satisfy that, while capturing the same information

Step 5: Interpret the factors

INSEAD
The Business School for the World®

Package: psych

```
Rotated_Results<-principal(ProjectDataFactor, nfactors=max(factors_selected),
                           rotate="varimax", score=TRUE)
Rotated_Factors<-round(Rotated_Results$loadings,2)
```

	Component 1	Component 2
GPA	0.96	-0.05
GMAT	0.95	0.19
Fellow	0.95	-0.01
Comm	0.70	0.54
Job.Ex	0.19	0.93
Organze	0.01	0.89
Extra	0.01	0.86

To better visualize and interpret: suppress loadings with small values

```
Rotated_Factors_thres <- Rotated_Factors
Rotated_Factors_thres[abs(Rotated_Factors_thres) < MIN_VALUE]<-NA
```

	Component 1	Component 2
GPA	0.96	
GMAT	0.95	
Fellow	0.95	
Comm	0.70	0.54
Job.Ex		0.93
Organze		0.89
Extra		0.86

Step 5: Interpret the factors

INSEAD
The Business School for the World®

What factor loads “look good"? Three technical quality criteria:

1. For each factor (column) only a few loadings are large (in absolute value)
2. For each raw attribute (row) only a few loadings are large (in absolute value)
3. Any pair of factors (columns) should have different "patterns" of loading

Step 6: Save factor scores

INSEAD The Business School for the World®

Replace the original data with a new dataset where each observation (row) is described using the derived factors

- For each row, estimate the **factor scores**: how the observation "scores" for each of the selected factors

Package: psych

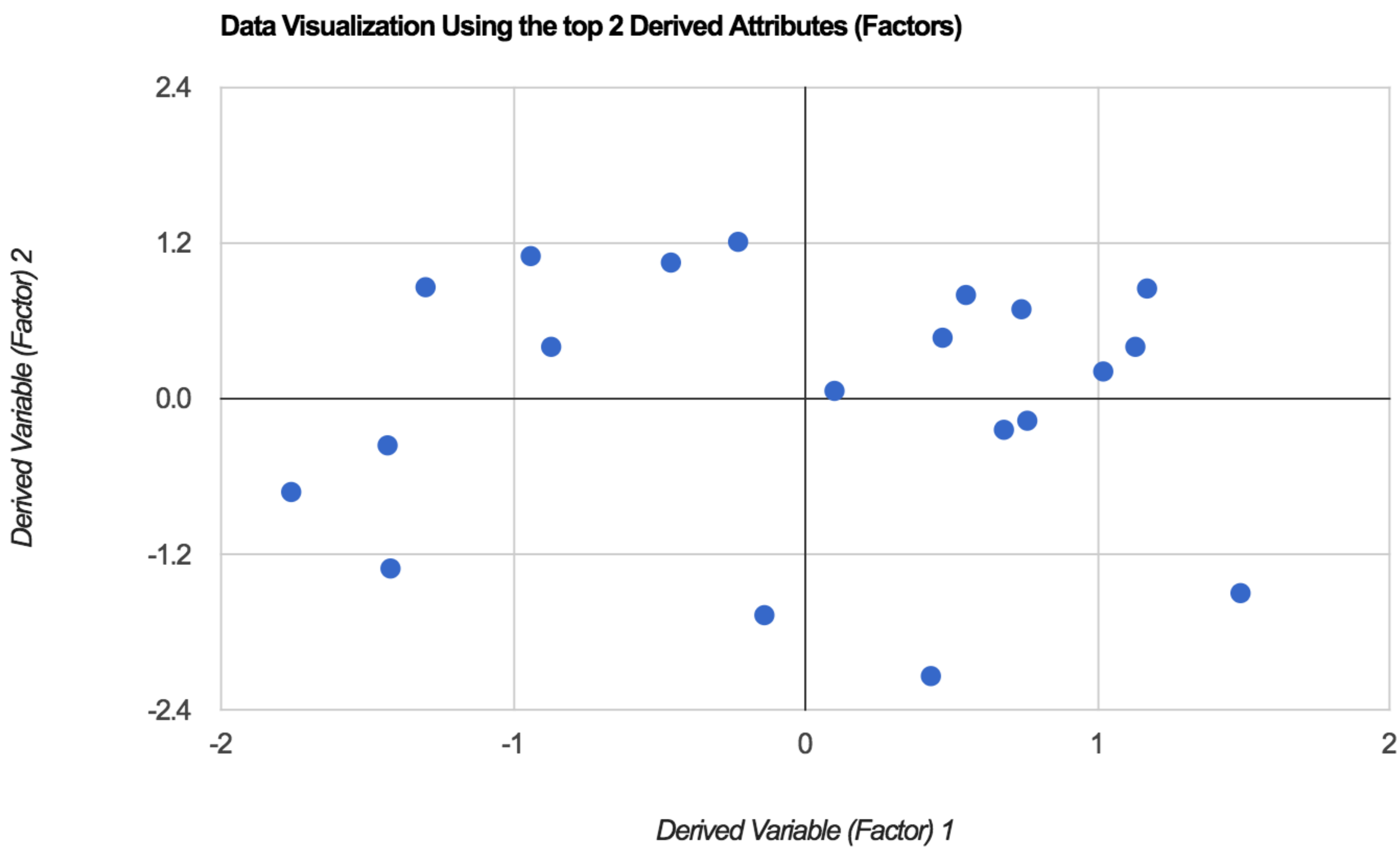
```
NEW_ProjectData <- round(Rotated_Results$scores[,1:factors_selected],2)
```

	Derived Variable (Factor) 1	Derived Variable (Factor) 2
observation 01	-0.46	1.05
observation 02	-0.23	1.21
observation 03	0.68	-0.24
observation 04	1.13	0.40
observation 05	-0.94	1.10
observation 06	-0.14	-1.67
observation 07	0.76	-0.17
observation 08	1.02	0.21
observation 09	-1.76	-0.72
observation 10	0.43	-2.14

Step 6: Save factor scores

INSEAD
The Business School for the World®

Then continue the analysis (e.g., make decision, or do clustering, etc.) with the new attributes



Summary of Sessions 7-8

- Advanced classification:
  - Regularization, profit curve, more methods (regularized regression, XGBoost, …), a process for classification
- Feature engineering
- From R scripts to Notebooks
  - New way/process for doing and communicating analytics with reproducible, publication-quality output
- Derived attributes and dimensionality reduction
  - A process for dimensionality reduction using Principal Component Analysis
  - Then continue analysis on the new attributes (next time: clustering and segmentation)

Next…

INSEAD

The Business School for the World®

- Sessions 9-10: [Fri, Feb 9 Amphi 307]
  - Cluster Analysis and Segmentation
  - BOR – work on the market segmentation process for the Boats (A) case

- Assignment 3 (due Feb 13):
  - Complete the market segmentation process for the Boats (A) case

- Proposal for final project (due Feb 14)


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