1 Introduction

1.1 Name the three main types of analytics tasks in business intelligence, as well as the question each of them addresses.

Descriptive Analytics What happened?

(**Diagnostic Analytics** Why did it happen?)

Predictive Analytics What (and why) will happen?

Prescriptive Analytics How and why make it happen?

1.2 To which analytics task does the following belong to:

- optimization: prescriptive
- data mining: predictive
- data warehousing: descriptive
- forecasting: predictive
- simulation: prescriptive
- Online analytical processing: descriptive

1.3 Data science is the defined as intersection of which three competences?

(X) Domain expertise

(X) Statistics & Mathematics

(X) Computer Science & Hacking

1.4 What is the difference between OLTP (online transaction processing) and OLAP (online analytical processing)? In particular, consider their use, the way they organize their data, and the objective they are optimized for.

1. use: operational use (OLTP) vs decision making tool (OLAP)
2. data organization: transactional DB (OLTP) vs data warehouse (OLAP)
3. optimization objective: optimized for efficiency & consistency (OLTP) vs. accuracy & completeness (OLAP)

1.5 Name the three main Vs in big data and explain them.

volume the volume of data, e.g., more observations, more variables/features, . . .

variety the variety of data, e.g., data of many different forms, from variety of sources, variety over time, . . .

velocity the speed in which data arrives, e.g., streaming data

2 Descriptive Analytics Part 1

2.1 Which of the following data content types is considered structured?

( ) text in product reviews

(X) ratings from one to five stars in product reviews

(X) counts of the word “good” in a review

(X) product category, encoded in numbers

(X) product category, encoded as “books”,“films”,“shoes”,...

( ) product images

(X) transaction value

2.2 Indicate for which level of measurement (1) mode, (2) median, (3) arithmetic mean, (4) geometric mean are appropriate measures

- nominal: only mode
- ordinal: mode and median
- interval: mode, median, arithmetic mean
- ratio: mode, median, arithmetic mean and geometric mean

2.3 In data streams, data is

( ) available at once

(X) arrives sequentially

2.4 What are the characteristics of streaming data? Can you relate them to the big data “Vs”? What are the corresponding challenges?

- sequential arrival, i.e., volume increases over time
- volume: possibly infinite number of instances
  variety: changes over time
  velocity: instances arrive fast, analysis needs to be done quickly
- computational complexity (time and storage),
  change/drift (e.g., non-stationarity of distributions)

2.5 Consider the example of calculating the average sales per month, where data from the last five years is available. Using this example, explain why it is important to consider the temporal ordering of data instances.

- sales might vary over time (e.g., seasonality like christmas, long-term trends like popularity)
- we are interested in an accurate information of the **current** sales
- depending on its time of observation, the data is representative of current sales or not
- thus, e.g., for computing the current monthly average sales, we should exclude old data (e.g., from previous year/months)
- Note: it might be good to show the current month's sales in comparison to that of the previous month and the same month one year ago

2.6 Name and describe four data quality and usability metrics

See slides on "Data Quality and Usability Metrics"

- Data source reliability
  Do we have the right confidence and believe in this data source?
  Prefer original source rather than 2nd hand data

- Data content accuracy
  Correctness and appropriateness of data for objective of our analysis
  E.g., right data (e.g., scoring: customer residence/work address),
  Intention by original data provider correctly reflected?
  E.g., Is a zero a missing value or a value of zero, e.g. in income?

- Data accessibility
  Are the data accessible when needed?
  The type of data used during development of a prediction model
  should be available when deploying the model
  E.g., 3rd-party data (e.g., credit bureau/Facebook)

- Data security and data privacy
  Ensure that solely authorized people have read/write access to the data
  Ensure no data are lost due to, e.g., system failures or attacks

- Data richness / comprehensiveness
  Are all required data elements (variables) included in the data set?
  E.g., potentially confounding variables

- Data consistency
  Are the data accurately collected and combined/merged?
  E.g., when merging data (attribute values) about the same subject from different sources, this data needs to be consistently assigned to the same subject

- Data currency/data timeliness
  Are the data as recent/new as required?
  Best practice is to enter an observation as soon as it is made,
  in order to avoid incorrect remembering/encoding later,

- Data granularity
  Are the variables and data values defined in the adequate (e.g., lowest)
  level of detail for their intended use?
  E.g., customer income encoded as numerical value (vs. discretization)

- Data validity
  Are the actual and expected data values aligned?
  E.g., age wrongly defined as range between 0 and 99

- Data relevancy
  Are the variables included in the data set relevant for objective of our analysis?
  Note: Inclusion of variables of lesser relevance depends on the analysis/algorithm
  E.g., some similarity/distance calculations might be difficult with large numbers of variables (see "curse of dimensionality")
  E.g., when segmenting customers into groups using euclidean distance measure,
  the number of variables should be small (e.g., us Principal Component Analysis)

2.7 Review Application Case 2.2 and the questions therein.

See figure 1 below, or Figures 2.4–2.6 in [Sharda et al., 2018, p. 95,97,99].

2.8 What is class imbalance? What is the problem of measuring a classifier's accuracy under extreme class imbalance? Which counter-measures could be used in preprocessing?

- Class imbalance corresponds to a situation with few observations (data instances) of one class, compared to many observations of the other class
  Example from Application Case 2.2: Few students quitting their studies (minority class), compared to a large majority who continues

- It favours a classifier that predicts all instances as belonging to the majority class (e.g., all continue), although that classifier is not useful for identifying minority class instances

- Re-weighting or Re-sampling of instances based on their class: Selecting few instances of the majority class (or weighting them lower), selecting all instances from the minority class (weighting them higher)

2.9 Describe the approach provided in Application Case 2.2 for assessing the impact (importance) of a variable in the classification model.

Two copies of the classification model are trained, (1) one with the variable included, and (2) one without the variable. The difference in performance between the two copies of the model indicates the impact of including the variable.

2.10 Name the four steps of data preprocessing. Name one example for each.

- Data consolidation, e.g., for calculating the turnover, selecting from a data source the data of all customer transactions from the previous month with the relevant attributes ordered quantity and price

- Data cleaning, e.g., removing or imputing missing values, removing outliers, e.g., transactions with unknown quantity or negative price

- Data transformation, e.g., create attribute transaction value as product $quantity \cdot price$

- Data reduction, e.g., reduce dimensions or volume

References

[Sharda et al., 2018] Sharda, R., Delen, D., and Turban, E. (2018). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. Pearson, 4 edition.

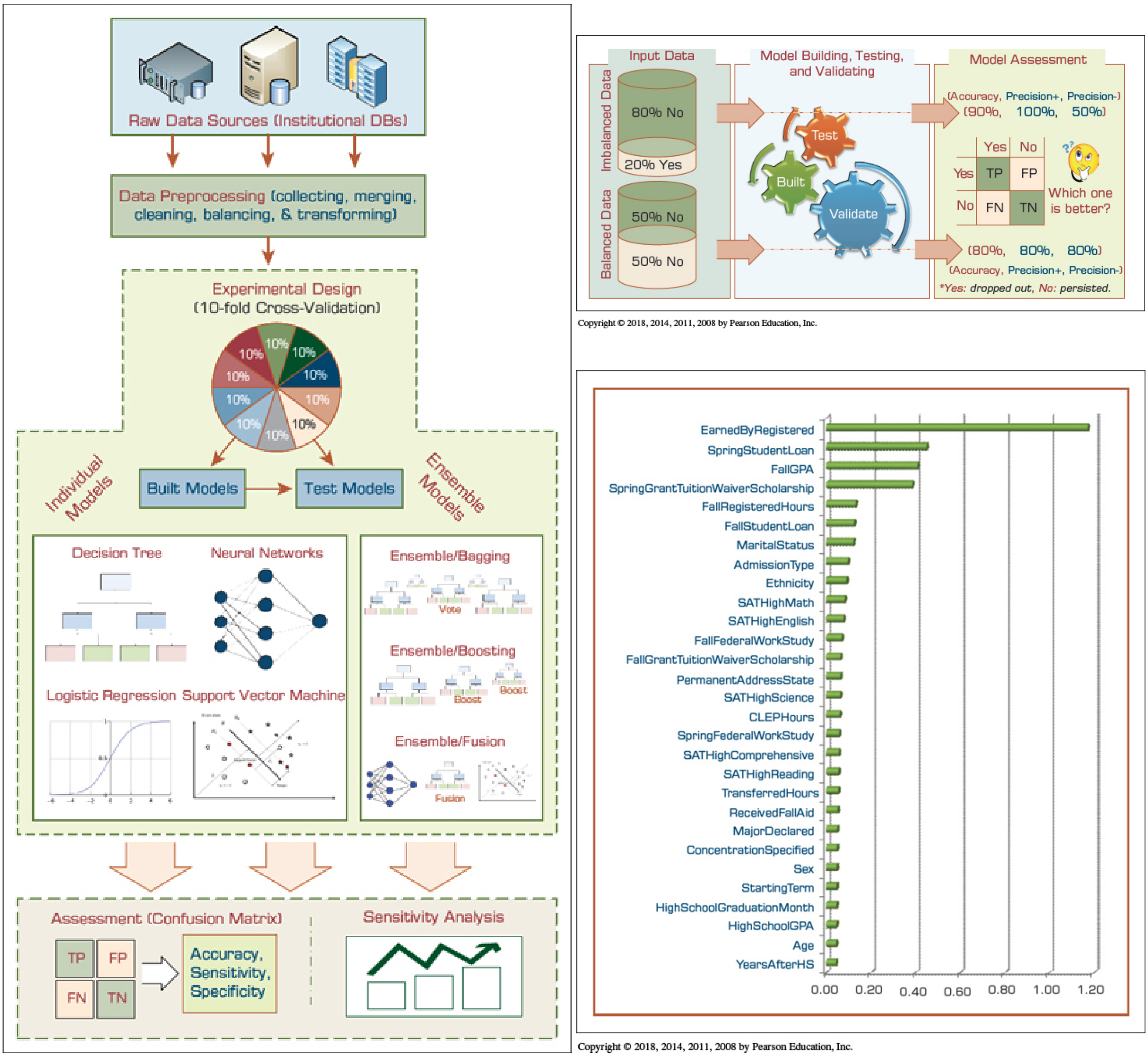

Figure 1: Preprocessing Steps and Sensitivity Analysis of Variables (Source: [Sharda et al., 2018, p. 95,97,99])

3 Descriptive Analytics Part 2

3.1 Probability Density (PDF) and Cumulative Distribution (CDF) Functions

You are working for a Dutch-based, international group of retailers named SPQR, who operates stores across 40 countries world-wide. From a previous analysis of historical data, the company knows that the monthly aggregated sales (in Euros) of its product X are distributed according to a Gaussian normal distribution with mean $\mu = 230,000$ and standard deviation $\sigma = 20,000$. For the following questions, assume stationarity, i.e. that the distribution does not change.

1. SPQR asks you to provide an estimate of the probability that the sales will exceed $200,000$ next month.
   Hint: Use the function $pnorm(q, mean, sd)$ in R.

   ```
   pnorm(q=200000,mean=230000,sd=20000) # prob. of X<=200,000
   1-pnorm(q=200000,mean=230000,sd=20000) # prob. of X>200,000
   ```

2. SPQR asks you to provide an estimate of the probability that the sales will be between $200,000$ and $260,000$ next month.
   Hint: Use the function $pnorm(q, mean, sd)$ in R.

   ```
   c200=pnorm(q=200000,mean=230000,sd=20000) # prob. of X<=200,000
   c260=pnorm(q=260000,mean=230000,sd=20000) # prob. of X<=260,000
   c260-c200 # probability of 200,000<X<=260,000
   ```

3. Using the probability density function $dnorm(x, mean, sd)$, provide estimates of the density (i.e., for the probability that the next month's sales are in an infinitesimal interval around these X values) for (a) $X = 230,000$ and (b) $X = 240,000$. Which one is greater?

   ```
   d230=dnorm(x=230000,mean=230000,sd=20000) # dens. of X=230,000
   d240=dnorm(x=240000,mean=230000,sd=20000) # dens. of X=240,000
   d230>d240 # density at X=230,000 is greater
   ```

3.2 Using R for calculating selected statistics:

For the following three samples

- $X_1 = \{-2, -1, -.5, 0, 0.25, 0.25, 1, 2\}$
- $X_2 = \{0, 0.1, 0.3, 0.7, 1.5, 3, 7\}$
- $X_3 = \{-2, -1.25, -1, -.8, -.5, 0, 0, 0\}$

Use R to calculate for the following statistics

- Mode $modeval(X)$ (see hint below)
- Median: $median(X)$
- 1st quantile: $quantile(x = X, probs = .25)$
- 3rd quantile: $quantile(x = X, probs = .75)$
- Arithmetic Mean: $mean(X)$
- Minimum, Maximum: $min(X), max(X)$
- Range: $max(X) - min(X)$
- Interquartile Range IQR: $quantile(x = X, probs = .75) - quantile(x = X, probs = .25)$
- (Unbiased) Variance: $var(X)$
- (Unbiased) Standard Deviation: $sd(X)$ or $var(X)^{.5}$
- (Sample) Skewness: $skewness(X)$

Hint: Load and use the library moments and use the following function definition for the mode:

```
library(moments);
# defining a function modeval(X) to calculate the mode of X:
modeval = function(X){
  return(as.numeric(names(sort(-table(X))[1])));
}

X1=c(-2,-1,-.5,0,0.25,0.25,1,2)
X2=c(0, 0.1, 0.3, 0.7, 1.5, 3,7)
X3=c(-2,-1.25,-1,-.8,-.5,0,0,0)
X=X1 # repeat below for X2 and X3 or use loop
modeval(X)
median(X)
quantile(x=X,probs=.25)
quantile(x=X,probs=.75)
mean(X)
min(X)
max(X)
max(X)-min(X) # Range
quantile(x=X,probs=.75)-quantile(x=X,probs=.25) # IQR
var(X)
sd(X) # or var(X)^.5
skewness(X)
```

X_1: (nearly) symmetric, similar to Gaussian
X_2: positive skewed, swayed left/tail on right side, mean greater median.
X_3: bi-modal, with mode at 0 (one cluster) and a second, Gaussian cluster centred at -1.

3.3 Kullback-Leibler Divergence

The SPQR group is interested in analysing the relative importance (i.e., share of total sales) of its different product categories (for simplicity, we consider three main categories A,B,C) to its total sales (world-wide, and in selected countries). These shares are given in Table 1. You are tasked to compare, how similar the distribution of sales' shares is in the Netherlands Q_1 (resp., Italy Q_2) to the aggregated (world-wide) share distribution (P), by calculating the Kullback-Leibler Divergence between P and Q_1 (resp., Q_2).

	A	**B**	**C**
World-Wide, P	0.4	0.3	0.3
Netherlands Q_1	0.4	0.35	0.25
Italy Q_2	0.5	0.3	0.2

Table 1: Shares of product categories in sales.

Note: You can use the following function definition for the KL-Divergence in R:

```
# definition of a function kldiv(p,q) to calculate the KL-Div(p||q)
kldiv = function(p,q){
  frac = p/q;
  plogfrac = p*log(frac);
  plogfrac[p == 0] = 0;
  plogfrac[(p != 0)&(q == 0)] = Inf;
  return(as.numeric(sum(plogfrac)));
}

tabnames=list(c("WW","NL","IT"),c("A","B","C"));
tabdata=c(0.4, 0.4, 0.5, 0.3, 0.35, 0.3, 0.3, 0.25, 0.2);
sales<-matrix(nrow=3,ncol=3,byrow=FALSE,data=tabdata,dimnames=tabnames) # orig. table
kldiv_ww_nl=kldiv(sales[1,],sales[2,]);
kldiv_ww_it=kldiv(sales[1,],sales[3,]);
print(paste("KLDiv. Sales WW vs NL: ",kldiv_ww_nl))
print(paste("KLDiv. Sales WW vs IT: ",kldiv_ww_it))
```

3.3.1 Independence

For the SPQR group in the example above, the sales' shares by product category and country are given in Table 2. If one interprets country and product as random variables X and Y, how would this table look like if country X and product Y were independent of each other?

Product Y**:**	**A**	**B**	**C**	**Total**
Country X**:**				
Netherlands Q_1	0.24	0.21	0.15	0.6
Italy Q_2	0.2	0.12	0.08	0.4
Total	0.44	0.33	0.23	1.0

Table 2: Sales' shares by product category and country.

Caluclate the vector of row-sums and the vector of column-sums, multiply these two vectors to get the table:

```
tabnames=list(c("NL","IT"),c("A","B","C"));
tabdata=c(0.24, 0.2, 0.21, 0.12, 0.15, 0.08);
sales<-matrix(nrow=2,ncol=3,byrow=FALSE,data=tabdata,dimnames=tabnames) # orig. table
tabdata2=matrix(rowSums(sales)) %o% t(matrix(colSums(sales)));
sales_if_indep<-matrix(nrow=2,ncol=3,data=tabdata2,dimnames=tabnames) # solution tab.
print(sales_if_indep)
```

3.3.2 Relationship between Variables

The SPQR group wants you to analyse a possible relationship between the sales of two of its products, X, and Y, with their sales data given in Table 3 below: Use R to calculate the covariance $cov(X, Y)$, Pearson's linear correlation coefficient $cor(X, Y, method = “pearson'')$, and Spearman's rank correlation coefficient $cor(X, Y, method = “spearman'')$. Furthermore, use lm to fit a linear model with X being the explanatory variable, and Y being the dependent (or response) variable.

Product	**1**	**2**	**3**	**4**	**5**
X	0.1	0.4	0.5	0.8	0.9
Y	0.9	0.3	0.2	0.21	0.1

Table 3: Sales data for products X and Y

```
Xvec=c(0.1, 0.4, 0.5, 0.8, 0.9);
Yvec=c(0.9, 0.3, 0.2, 0.21, 0.1);
sales<-data.frame(X=Xvec,Y=Yvec) # create a data frame
cov(sales$X,sales$Y)
cor(sales$X,sales$Y,method="pearson")
cor(sales$X,sales$Y,method="spearman")
linmod<-lm(sales$Y~sales$X) # fit linear model
linmod$coefficients # show its intercept and slope
```

3.4 Streaming or Time Series Data

The SPQR group's data warehouse provides (aggregated) sales records (in millions of Euros) of several quarters (see Table 4). You are called to settle the dispute between two colleagues, one who argues that all data should be used to calculate the average sales of this product, the other arguing to use just the most recent data. Use the arithmetic mean to demonstrate the effect of (a) using all data vs. (b) using a windowing approach that divides the data into three chunks, one for each year.

Year	**2015**				**2016**				**2017**			
Quarter	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Quantity	0.3	0.4	0.3	0.35	0.55	0.6	0.5	0.6	0.4	0.5	0.45	0.4

Table 4: Sales data (in millions of Euros) for products X over time

```
quantity=c(0.3, 0.4, 0.3, 0.35, 0.55, 0.6, 0.5, 0.6, 0.4, 0.5, 0.45, 0.4);
quarter=c(1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4);
year=c(2015, 2015, 2015, 2015, 2016, 2016, 2016, 2016, 2017, 2017, 2017, 2017);
sales<-data.frame(quantity=quantity,quarter=quarter,year=year)
mean(sales$quantity)
mean(sales$quantity[sales$year==2015])
mean(sales$quantity[sales$year==2016])
mean(sales$quantity[sales$year==2017])
```

3.5 Indicate which of the following statements are correct

( ) Data (or Information) Visualisation and Visual Analytics are precisely the same.

(X) Information Visualisation is foremost retrospective and descriptive, answering the questions such as what has happened or is happening.

( ) Information Visualisation is foremost predictive, future-oriented, answering the questions such as what will happen.

(X) Predictive Analytics is foremost predictive, future-oriented, answering the questions such as what will happen.

(X) Visual Analytics is the combination of information visualisation and predictive analytics.

3.6 Indicate which of the following statements about narrative visualization are correct

( ) The Martini Glass schema corresponds to (a) starting with a broad view that allows the user to interactively select among various aspects potential points of interest, and (b) then successively narrowing the scope down (e.g., by focusing on a single KPI).

(X) Quite the opposite! The Martini Glass schema starts with a narrow focus point and allows initially little interaction. Then, the scope is subsequently widened and more and more interaction is allowed.

(X) The Drill-Down-Story starts with a "map" that outlines potential points of interest, and allows the user to interactively select and explore aspects in depth.

( ) The Drill-Down-Story starts with one focus point, and leads the user through a step-by-step analysis, thereby "drilling down" to the underlying facts.

(X) The Interactive Slideshow is a an approach that structures the information into several subsequent steps ("tabs"), and allows the user to navigate through them interactively at their own speed.