ALTKLAUSUR 18/19

One step in the Data Mining Process (supervised learning) is to partition the datainto di erent datasets. Consider you partition the data into two sets. Please, statethe name of those datasets and explainshortlyhow you use each of them.

Training Dataset: Train the model, estimate coefficients
Validation Dataset: Apply estimated model, evaluate performance

Explain shortly why we partition the data and relate it to the problem of overfitting in this context. State one possible reason for overfitting as well.

Partition to train model in training data set and evaluate it in validation dataset. If estimated model fits to the training data set too well/too closely, it also predicts "noise" (treats noise as a signal) -> overfitting.

Potential reason: too many predictors, or trying too many models

What does the following R command do?

s <- sample(row.names(housing.df), 5, prob = ifelse(housing.df$ROOMS>10,0.9, 0.01))

Draws random sample of 5 observations (rows), while giving more weight to the case "houses with more than 10 rooms" of being drawn.

"Prob" defines the probability weights. Here, the observation "houses with >10 rooms" is drawn with a probability vector of 90% & with a probability vector of 0.01 else

Consider you have 5 observations for variable x1 and x2 and the standard deviation sd1 for x1 and sd2 for x2 given in the following table:

x_1	x_2
2	49000
5	156000
10	99000
20	192000
16	39000
$sd_1 = 6.841053$ 7.46934	$sd_2 = 62867.06$ 66479.32

State in words or by equation how you normalize a variable

Substract the mean ($\bar{x}$) of each observation (xi) and divide by the standard deviation (sd).

1. Calculate mean: 1/5 * (2 + 5 + 10 + 20 + 16) = 10.6 for $\bar{x}1$
2. $\bar{x}2$ = 107000

Now: normalized value for first observation in x1:

$$\frac{2 - 10.6}{7.46994} = -1.151$$

Standard Deviation Formula:

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

State one reason why we normalize variables

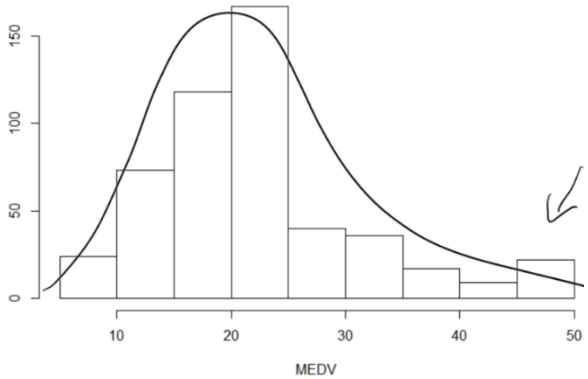
Variables have different scales: Large scales would dominate and skew results

What type of graph is obtained by the following command and what does this type of graph show?

hist(housing.df$MEDV, xlab = "MEDV")

Histogram of the median house value using the housing.df data frame. A histogram shows the distribution of a continuous variable.

Consider you have the following output graph:



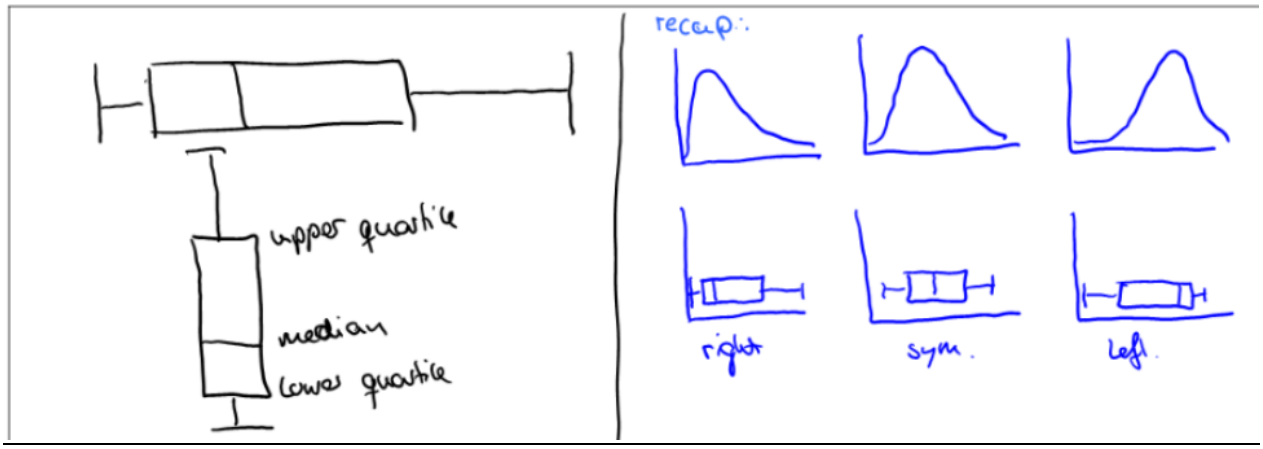
Is the distribution symmetric or skewed? If you state "skewed", also give the direction of the skewness

➔ The distribution is skewed to the right side

What would you in general do when you face a skewed distribution?

Transform to a log-scale

Draw a boxplot for a variable with a right-skewed distribution. Label the upperquartile, the median, and the lower quartile in your graph



Upper quartile: 75th percentile

Median: 50th percentile

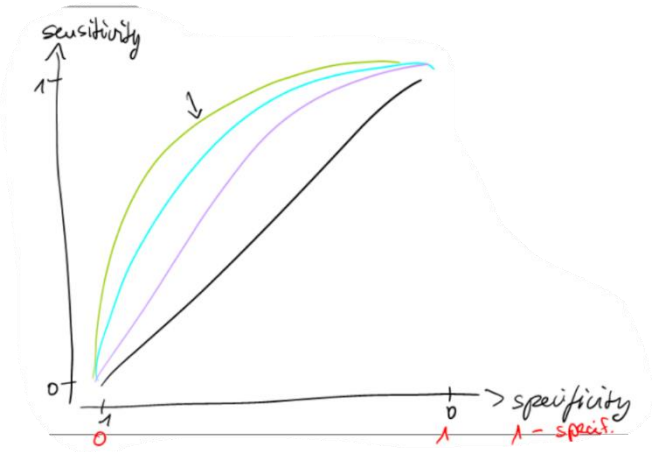
Lower Quartile: 25th percentile

<u>You estimate three different models and want to choose one of them based on their ROC curves.</u>

<u>Briefly explain what an ROC curve is.</u>

ROC curve shows specificity and sensitivity, i.e., how well a model classifies

<u>Explain how to use ROC curves in order to determine which model is best and illustrate your answer with a graph</u>



choose the curve that maximizes the AUC (area under the curve). Here: Green curve, hence the underlying model performs best.

<u>Provide an example for a situation where we might prefer a model which does not minimize the misclassification rate.</u>

Situations with asymmetric error costs, e.g. credit card fraud, tax fraud, etc.

<u>What is the difference between R^2 and R^2adj?</u>

The adjusted R^2 has a penalty for numbers of predictors, whereas R^2 continues to increase with increasing number of predictors which makes it unsuitable for measuring model fit

Consider the following output.

	ME	RMSE	MAE	MPE	MAPE
Test set	-40.1	1321	1012	-1.72	9.01

What does ME stand for?

➔ Mean Error

Was ME = -40.1 calculated in the training or validation sample?

It was calculated in the validation sample. Because for the training sample (where the model was training) the average error should be zero, otherwise the coefficients would be biased.

Write down the formula for the RMSE. [hint: If you cannot remember the formula, you may derive it from the name.]

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2}$$

Suppose we are interested in the determinants of prices of used cars. Explain what each of those ten lines of code is doing (lines 9 and 10 belong together and are one line of command)

```
1  selected.var <- c(3, 4, 7, 8, 9, 10, 12, 13, 14, 17, 18)
2  Fuel_Type <- as.data.frame(model.matrix(~ 0 + Fuel_Type, data=car.df))
3  car.df <- cbind(car.df[,-4], Fuel_Type[,])
4  car.df$Fuel_TypePetrol <- NULL
5  set.seed(1)
6  train.index <- sample(c(1:1000), 600)
7  train.df <- car.df[train.index, selected.var]
8  valid.df <- car.df[-train.index, selected.var]
9  search <- regsubsets(Price ~ .^2, data = train.df[c(1:4, 7:10)],
10         nbest = 1, nvmax = 20, method = "exhaustive", really.big=T)
11 summary(search)
```

1. Select the variables we want to use (here: columns 3, 4, 7, 8, 9, …)
2. Transform the variable Fuel_Type into a set of dummy variables. From the dataset car.df
3. Replace the 4th column (Fuel Type) with the newly created set of dummy variables for fuel type
4. Delete the dummy variable for Petrol, perhaps to avoid multicollinearity
5. Fix a starting point for the “pseudo” random draw (for reproducibility reasons, makes code reproducible)
6. We randomly sample a vector of 600 numbers out of 1000 and store them under train.index
7. Based on the generated 600 numbers, we select the rows from the car.df dataset, that are going into the training dataset. Use only the selected variables (columns) from line 1#

8. For validation.df select only those rows that are NOT in the train.index
9. We perform an exhaustive search with interactions. Outcome variable is Price, X variables are all restricted to (1:4 and 7:10). Only one best model for models from 1 to 20. Really.big=TRUE, to state that the computational power is high, therefore to account for that.
10. Summary of the previously estimated results.

ALTKLAUSUR 19/20

Suppose you have data on the price of pizzas, their size, the quality of ingredients used, and the distance to the town center of pizza restaurants. You want to predict pizza prices for new restaurants in this city. While exploring the data, you find that the variable size contains several missing values.

Why could this be a problem, especially when the number of missing values is high?

Not all algorithms can handle missing values. Most algorithms drop them by default. If they are dropped, the information from the other variables of the observation with the missing value in one variable is lost. The higher the number of missing values, the more observations are lost.

State two possible ways to handle missing values

- Omission (only practical if number of missings is small)
- Imputation (Replace missings with reasonable substitutes [e.g., mean, median]
    - Advantage: We can keep the record and its non-missing information

[Extrapolation, Interpolation?]

Suppose you detect that for one observation the value of the variable size is very distant to the other values of this variable (it could be a very high or very low value). How is such a value called? What could you do about this problem?

How is it called? -> Outlier

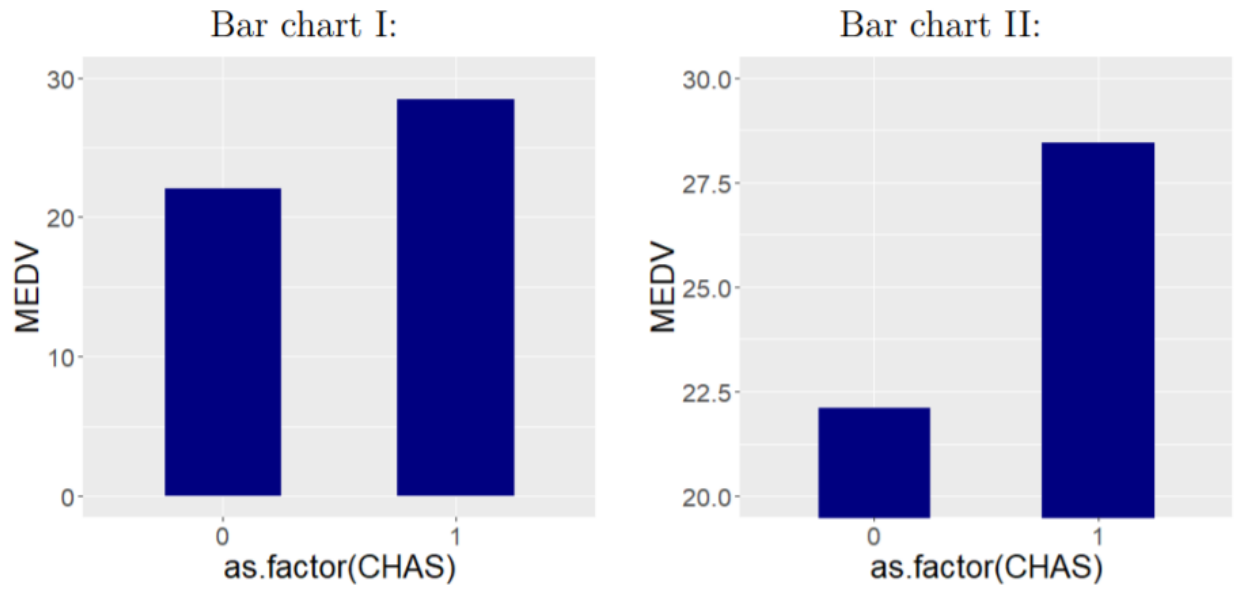
What could be done about this problem? ->

1. Check whether the value is an error -> correct if possible
2. Can be treated as missing value (impute with mean or median)

State two ways how you could detect whether there are values of a variable that are very distant to the other values of this variable.

- 1. Use summary statistics and compare minimum/maximum value to values of 1^{st} / 3^{rd} quartile or mean
- Order variables and compare values
- 2. Graphically with boxplots (the dots above or under the lines are defined as outliers) or scatterplots

Data Vizualisation: Consider the following two bar charts of the median value of houses (MEDV) by the two categories: bounding to the Charles River (CHAS=1) or not (CHAS=0)



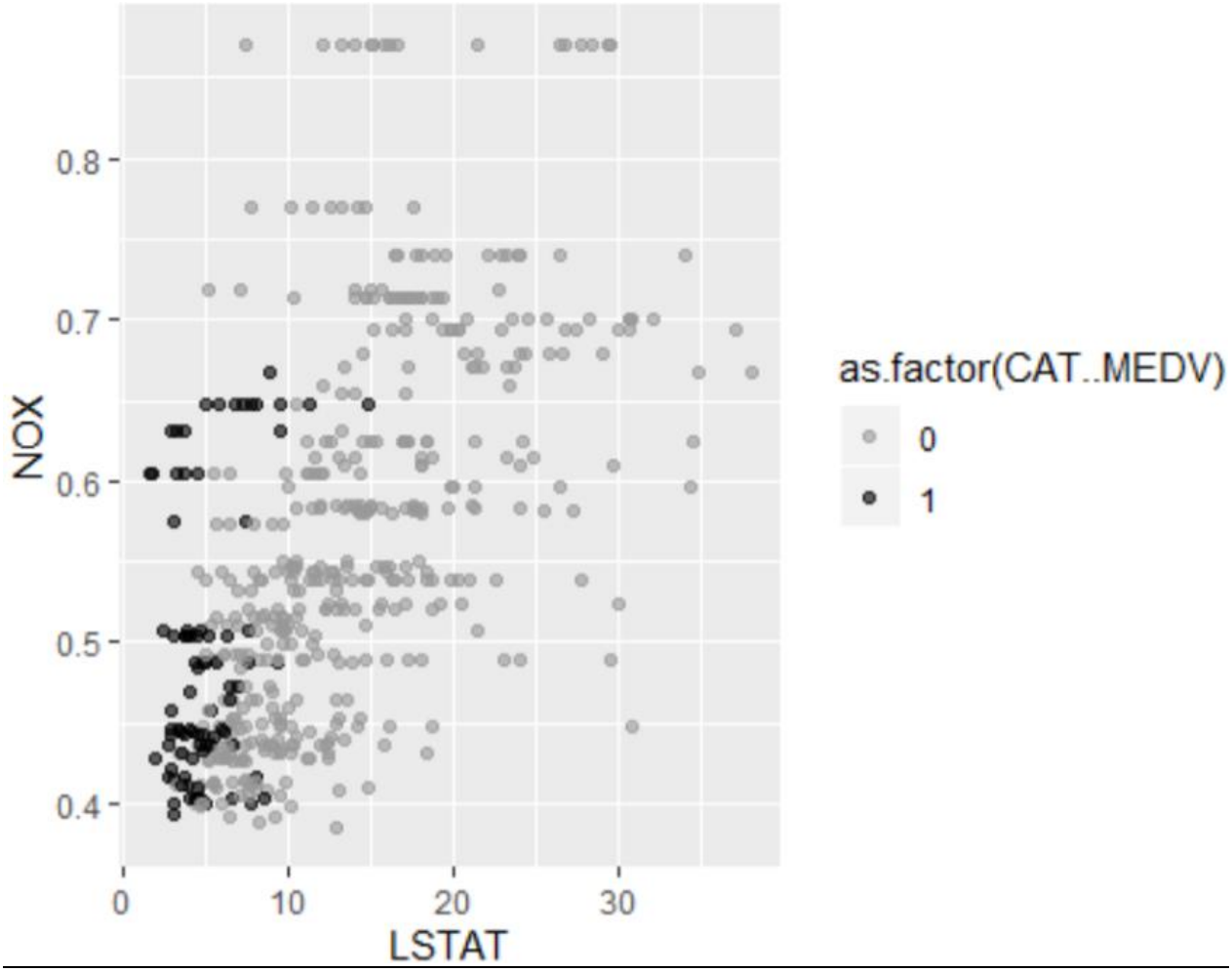
Bar chart I and bar chart II plot the same variables. What is the difference between the two bar charts? How would you assess this difference?

As mentioned, Bar chart II and I plot the same variables. There seems to be a 6 point difference between houses bounding to Charles River (CHAS=1) and houses not bounding to the Charles river (CHAS=0). However, the Bar Chart II makes it seem like this difference is much larger than it is in reality. This "illusion" happens due to the fact that the Y-values of the two bar charts are differently scaled. While Bart I starts at zero, bar chart II starts at ~18.

Bar chart I starts at a zero baseline whereas bar chart II starts with a baseline at 20 (Non-zero baseline). Looking at bar charts, we compare the relative end points. If the baseline is different from zero, we get a false visual comparison.

Suppose you get the following graph where NOX is the nitric oxide concentration in the tract, LSTAT is the percent of the lower status of the population, and CAT..MEDV is a binary variable that equals 1 if the

house has a value of more than $30.000.



How is the type of graph called?

Scatterplot

State a reason to use color coding in this regard.

To display multivariate relationships (relationship of three variables). Here: between LSTAT, NOX and whether or not the house has a value >30k

Explain in detail what information you can deduce from this graph

Houses with higher values (CAT..MEDV) are on average more likely to be in a low status and low NOX neighborhood (LSTAT). Houses with higher values have a maximum LSTAT of about 15% and a maximum NOX of around 0.67

The number of houses with a value above $30.000 is lower than the number of houses with a lower value.

Performance Evaluation

What is the leading question in performance evaluation?

How well does the model predict new data? (not: How well does it predict the data it was trained with?)

Explain what each line of the following code does

```
1  bank.df <- read.csv("UniversalBank.csv")
2  set.seed(2)
3  train.index <- sample(c(1:dim(bank.df)[1]), dim(bank.df)[1]*0.6)
4  train.df <- bank.df[train.index, ]
5  valid.df <- bank.df[-train.index, ]
6  logit.reg <- glm(Personal.Loan ~ ., data = train.df, family = "binomial")
7  logit.reg.pred <- predict(logit.reg, valid.df[, -8], type = "response")
8  roc.df <- data.frame(actual = valid.df$Personal.Loan, predicted = logit.reg.pred)
9  r <- roc(roc.df$actual, ref=1, roc.df$predicted)
10 plot.roc(r)
```

1. Imports the csv file UniversalBank.csv as data frame called bank.df
2. sets a seed to enable replication of draw
3. draws a random sample of 60% of the bank.df data(rows) and stores it in ”train.index”
4. creates data frame with training data using the random sample ”train.index”, stores it as ”train.df”
5. creates data frame with validation data using the remaining data from ”bank.df” that is not part of ”train.index”, stores it as ”valid.df”
6. logistic regression of personal loan (binary) on all other variables in the data set, stores it as ”logit.reg”
7. predicts probabilities with model estimated in 6) in validation data, stores it as ”logit.reg.pred”
8. creates a data frame named ”roc.df” with the predictions from 7) and the true values from the validation data
9. defines an element ”r” with actual and predicted values
10. draws the ROC based on 9)

You type the following command into R and get the following output: confusion matrix

```
confusionMatrix(as.factor(ifelse(pred > 0.8, 1, 0)), as.factor(valid.df$Personal.Loan))
```

```
          Reference
Prediction    0    1
         0 1803   88
         1    6  103
```

What is the cut-off value in this case and what does it mean?

The cut-off value is 0.8. Observations with predicted values/propensities equal to or larger than 0.8 are classified as ”1”, all other as ”0”.

Based on the confusion matrix calculate the sensitivity and the specificity showing your calculations.

The code uses the default option for the definition of the positive class. Such the positive class is ”0”:

sensitivity = % of “positive” correctly classified = 103/(103+88)=0.539

specificity = % of “negative” correctly classified = 1803/(1803+6)=0.997

You estimated three models and obtained the following ROC curves. Which model do you choose and why?

Choose model 1 because it has the highest ROC curve (and thus also the highest AUC).

Linear regression: Your data on used cars contains the following variables:

Variable	Description
price	price of the car (Euro)
metallic	= 1 if metallic color, = 0 else
weight	weight of the car (kg)
warranty	remaining warranty (years)

You are interested in the relationship between car prices and the other variables.

Write down the estimation model

$$price_i = \beta_0 + \beta_1 metallic_i + \beta_2 weight_i + \beta_3 warranty_i + \varepsilon_i$$

R produces the following output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -33136       1560     -21   <2e-16 ***
metallic         572        163       4    5e-04 ***
weight            40          2      28   <2e-16 ***
warranty         185         25       7    5e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Are metallic-colored cars sold at higher or lower prices than “ordinary” cars?

Buyers tend to pay more for metallic cars as the coefficient for metallic is 572 which is greater than zero and “non-metallic” being the reference/omitted category. Since this coefficient is statistically significant, its plausible to assume that this is the case.

Your supervisor is surprised about $\hat{\beta}_{weight} = 40$ being both positive and significant. Interpret this coefficient. Give an explanation for $\hat{\beta}_{weight} > 0$

= 40 means that, on average, buyers are willing to pay 40 Euros for an increase in the car’s weight by one unit. Possible explanation: more weight is correlated with bigger engines/more horse power.

Assume one unit is 100kg. A 1500kg car increases the cars price by 600 on average.

Your supervisor tells you that the firm is considering to offer car warranty extensions for 200 Euro per additional year. Will there be a large demand for this kind of warranty extension? Explain based on your estimation results.

No. On average, buyers are only willing to pay $\hat{\beta}_{warranty} = 185$ Euros for an extra year of warranty, which is less than 200 Euros

Logistic Regressions: Using data on online sales, you are interested in predicting the variable

$$buyer = \begin{cases} = 1 & \text{if visitor bought something} \\ = 0 & \text{if visitor didn't buy anything} \end{cases}$$

Why should you use logistic instead of linear regression if y = buyer

- Buyer is caterogical/a factor. Linear regression should be used for numeric outcomes.
- Alternative solution: Logistic regression makes sure that P(y) ∈ [0, 1] which is not guaranteed if a linear regression is used.

How do we interpret logistic regression coefficients?

βˆ is relative change in the odds if the predictor variable increases by one unit

One unit increase, increases the dependent variable by one unit %

Are the odds larger or smaller than 1 if most visitors leave the online shop without buying anything? Explain using the mathematical definition of the odds.

Letting $p = \Pr(buyer = 1)$ **(1P)**, "most visitors leave without buying anything" translates into $p < (1 - p)$ **(1P)**. Consequently, the $odds = \frac{p}{1-p}$ **(1P)** must be smaller than 1 **(1P)**.

DEUTSCH

What does the command table() do?

Gibt die einzelnen Ausprägungen der Variable Color, bzw. Der Autofarbe, an: Im Datensatz sind zb 283 blaue Autos und 191 schwarze Autos enthalten

Please explain why we could decide to reduce the number of categories and state how you would do this in this example (only in words, no R code required)

Für die Variable Farbe macht es keinen Sinn, eine numerische Variable in R zu hinterlegen. Hierfür muss man nämlich mit Factors arbeiten. Man hätte also 9 Dummys statt 10, um Multikollinearität zu verhindern.

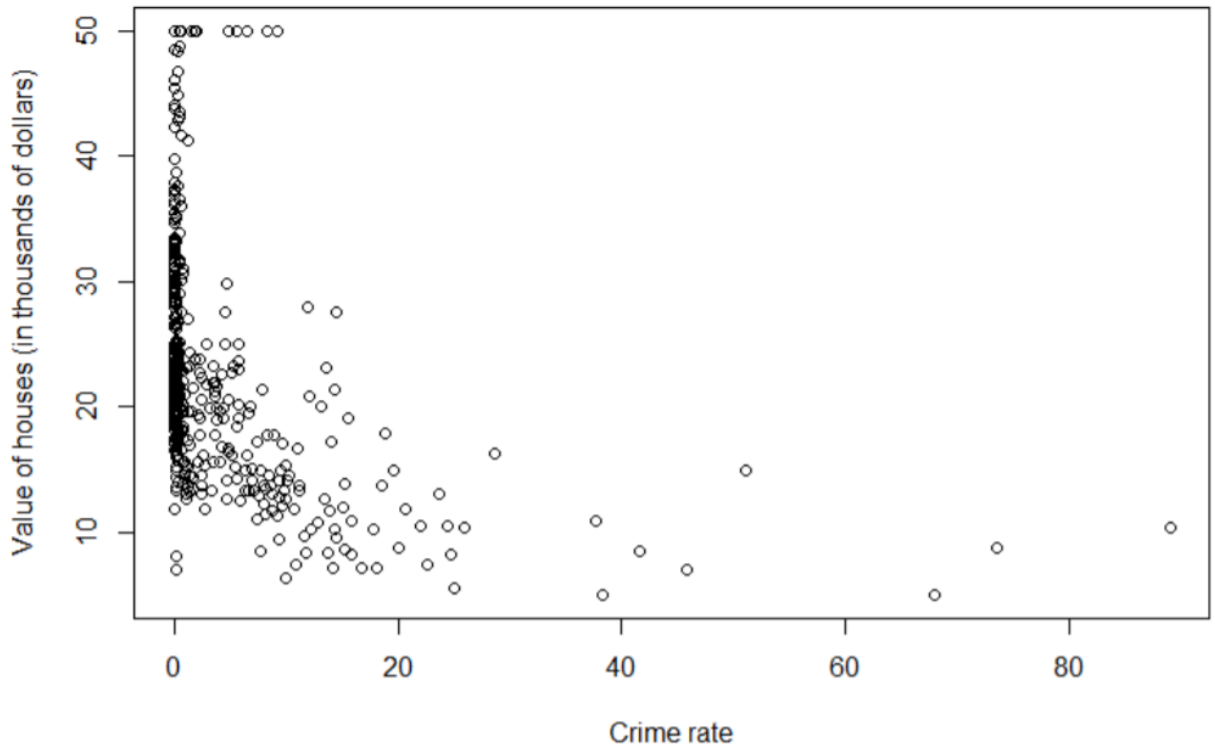
Man könnte die Anzahl an Variablen reduzieren, um der Gefahr des Overfittings entgegen zu wirken. Für kategoriale Variablen kann man dummies einfügen. Hier könnte man die 10 Ausprägungen zusammenfassen in helle oder dunkle Farben. Somit könnte man zB eine Dummy Variable für helle Farben einfügen, die 1 wird, wenn das Auto schwarz, grau, grün oder blau ist. Und 0 wenn das Auto weiß, beige, silber, geld o.ä. ist.

Andernfalls könnte man einfach die Farben rauswerfen, für die man im Vergleich zu den anderen Farben nicht viele Beobachtungen hat. Für beige, violette oder gelbe Autos macht eine Regressionsanalyse z.B. wenig Sinn, da schlicht nicht genügend Beobachtungen vorhanden sind.

Boxplot interpretation

- Man erkennt, dass Häuser die näher am River liegen im durchschnitt mehr wert sind als Häuser die nicht am River liegen. Man erkennt dies daran, dass einerseits die gesamte Verteilung des rechten Boxplots (inkl. Median, Minimum, Maximum) nach oben versetzt ist im Vergleich zu Häusern, die nicht in der Nähe vom River liegen.
- Zusätzlich liegt bei den Häusern, die näher am River liegen, eine höhere Varianz vor (erkennbar am größeren IQR).
- Linker Boxplot ist symmetrisch verteilt (wenn man die Ausreißer nicht hinzuzählt), während rechter Boxplot rechtsschief ist
- Das günstigste Haus ist…. Wert
- Der rechte Boxplot weist zudem eine erkennbare rechtsschiefe auf. Zählt man beim linken Boxplot die Ausreißer mit, scheint auch dieser rechtsschief zu sein.

Scatterplot



Type of variable: numeric

Man erkennt eine negative Korrelation zwischen Hauswert und Kriminalrate. Häuser in Gegenden mit hoher Kriminalrate (X-Achse) sind tendenziell weniger wert (Y-Achse) als Häuser in Gegenden mit geringer Kriminalrate. Mit steigender Kriminalrate nimmt der Wert der Häuser ab

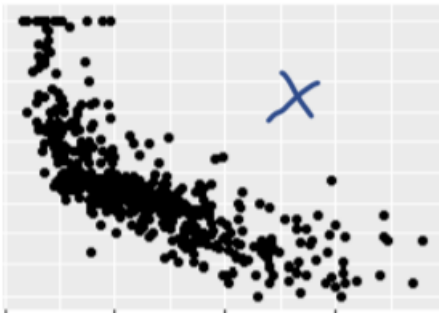
How could you modify the graph to obtain a better visualization of the value of housing units in areas with very low crime rates? (2 points)

Was kann man grafisch tun, wenn man eine bessere Visualisierung der Werte haben möchte, wenn Viele Punkte sind aufeinander gehäuft sind?

- ➔ Man könnte Crime rate zb Normalisieren bzw standardisieren, oder die X-AChse (crime rate) in eine logarithmierte Achse umändern

State what Corr is. What range of values can “Corr” take?

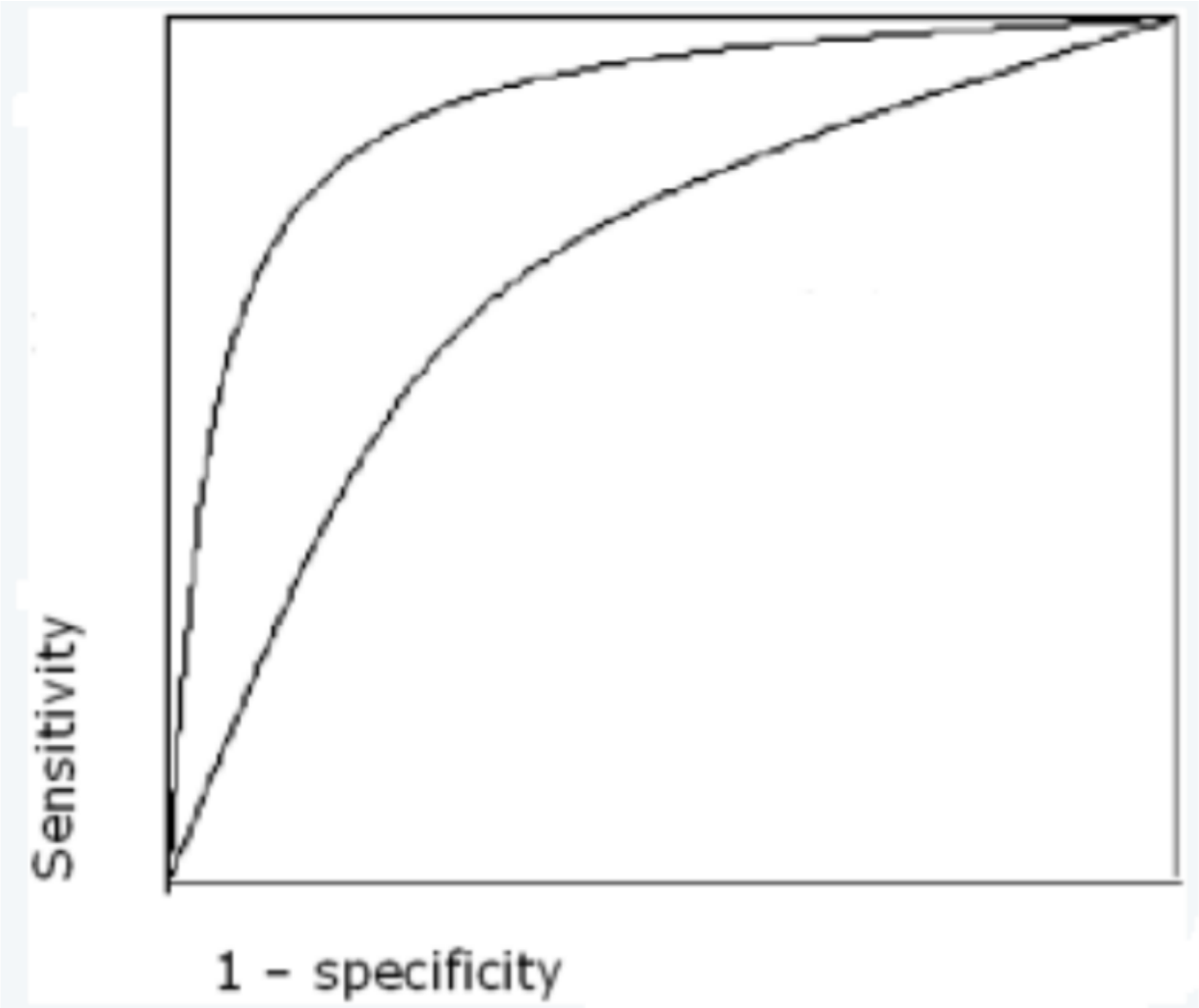
Korrelation zeigt Informationen über die Beziehung zwischen zwei Variablen an. Korrelation kann zwischen -1 und +1 liegen.



X achse: MEDV

Y Achse: LSTAT

Man erkennt eine negative Beziehung zwischen MEDV und LSTAT. Wenn LSTAT steigt, sinkt MEDV tendenziell. D.h. Häuser in Gegenden mit hoher Anzahl an sozioökonomisch schwachen Personen sind weniger wert als Häuser in Gegenden mit geringer Anzahl an sozioökonomisch schwachen



Man sieht zwei Receiver Operating Curves. ROC-Kurve ist ein Tool zur Evaluierung der Performance eines Klassifizierungsmodells. Die ROC illustriert die Sensitivität und Spezifität bei sinkendem Cut-off value (von 1 bis 0). Wird benutzt um den optimalen Cut-off Value zu identifizieren. Spiegelt den Trade-Off zwischen Sensitivität und Spezifität wider. Auf der X-Achse ist die Specificity bzw die False Positive Rate abgetragen. Die sensitivity gibt die Prozentzahl korrekt Klassifizierter 1en bzw Responder wieder. Die Specificity gibt die Prozentzahl korrekt klassifizierter negatives wieder.

Mit der ROC wird der optimale Cut-Off value identifiziert, bei dem ein Modell die beste Performance liefert.

AUC ist der Bereich unter der ROC. Je größer dieser Bereich, d.h. je näher dieser Wert an 1 ist, desto besser perform das Modell. Je höher die AUC, desto höher ist gleichzeitig die specificity und sensitivity und desto besser performt das Modell. Generell (ohne Beachtung asymmetrischer Kosten) wird immer eine ROC gesucht, die sich am weitesten links oben von der Diagonalen (Naïve Rule) verläuft. Je weiter links oben die ROC von der Naïve Rule verläuft, desto besser performt das Modell im Vergleich zu einer Zufallsziehung.

Briefly explain three different alternatives for selecting subsets of predictors. State the procedure for variable selection you would pick if you were not restricted by computational power.

Forward Selection: ist eine schrittweise Regression, die mit einem leeren Modell anfängt und die Variablen sukzessiv einfügt. Dabei wird nacheinander die Variable eingefügt, die das Modell am besten verbessert (zB die am meisten im Vergleich zu den anderen Variablen zum adjustierten R^2 beiträgt). Wenn das nächste Einfügen nicht mehr zu einem signifikanten Anstieg des adj. R^2 führt, hört dieser Algorithmus auf. (erfordert weniger Rechenleistung)

Exhaustive Search: Probiert alle möglichen Kombinationen von den X-Variablen und schaut, wie sie performen (mithilfe des adjustierten R^2). Für jede Anzahl an beliebigen X-Variablen kann dadurch die optimale Kombination identifiziert werden.

Backward Elimination: Hier werden alle Prädiktoren in das Modell aufgenommen. Dann werden sukzessiv (nach und nach) die Variablen rausgeschmissen, die am wenigsten (nicht signifikant) zum adj. R^2 beitragen.

- ➔ Generell ist Exhaustive Search am besten (wenn Rechenpower keine Beschränkung ist), da hierbei sämtliche Kombinationen von Variablen durchprobiert werden vom Algorithmus, bis der perfekte Fit für jede Anzahl an Variablen gefunden wird. Dies erfordert jedoch ab einer bestimmten Menge von Variablen Unmengen an Rechenpower, weshalb andere Verfahren wie Forward Selection / Backward Selection bevorzugt werden

R Code interpretation LINEARES MODELL

Car.lm <- lm(Price ~., data = train.df): Lineare Regression der abhängigen Variable Price auf sämtliche unabhängige Variablen aus dem Trainingsdatensatz und speichert dies unter Car.lm ab.

Car.lm.null <- lm(Price ~1, data =train.df: Schätzt das Modell Price = beta0 + e_i aus dem Trainingsdatensatz. Dieses Schätzmodell wird dann in car.lm.null gespeichert. In diesem Fall ist die Regression gleich mean(Y), da nur das Intercept gefitted wird.

Car.lm.step <- step(car.lm.null, scope = list (lower=car.lm.null, upper=car.lm), direction = "forward"): Führt eine Fordward Selection durch, in der minimum das Intercept und maximum die gesamten Variablen enthalten sein sollen und speichert das resultierende Modell unter Car.lm.step ab

Summary(car.lm.step): Zeigt die Schätzungen der Koeffizienten mit enstprechenden Signifikanztests an.

Car.lm.pred <- predict(car.lm.step, valid.df): Wendet das Regressionsmodell (car.lm.step) auf dem Validierungsdatensatz an um Preise vorherzusagen und speichert es unter Car.lm.pred ab

Accuracy(car.lm.step.pred, valid.df$Price): Misst die Genauigkeit der Vorhersagen mithilfe gängiger Genauigkeitsmaße wie RMSE, ME, MAE etc

Suppose one of the explanatory variables was categorical and could have three possible outcomes. Suppose you have transformed this variable into three binary/dummy variables and received an error message during the analysis. How many binary variables should you use instead? Explain why

Der Fehlerterm ist entstanden, weil multikollinearität zwischen den Variablen besteht. Man hätte hier zwei statt drei dummy variablen verwenden sollen, um Multikollinearität zu verhindern. Wenn

Warum kann es zu Problemen kommen, wenn Beobachtungen fehlen?

Wenn viele Beobachtungen fehlen, sinkt auch Verlässlichkeit eines Regressionsmodells. Einige Funktionen von R funktionieren nicht wenn NAs vorliegen, z.B. bei der Berechnung des Means

2 possible ways to handle missing values

1. Omitting der fehlenden Beobachtungen, also das Weglassen fehlender Beobachtungen
2. Gründe für fehlende Werte suchen. Ggf. Extrapolieren oder Imputation mithilfe des Means z.B.

Suppose you detect that for one observation the value of the variable size is very distant to the other values of this variable (it could be a very high or very low value). How is such a value called? What could you do about this problem?

Dies ist ein sogenannter Outlier. Man könnte hier den Grund für den Ausreißer suchen und korrigieren falls ein Fehler z.B. vorliegt. Andernfalls kann man den Ausreißer auch weglassen und stattdessen den Median oder Mittelwert aus dem Datensatz einsetzen (Imputation).

State two ways how you could detect whether there are values of a variable that are very distant to the other values of this variable.

Analytisch/Rechnerisch: Mithilfe von summary statistics könnte z.B. Minimum- und Maximumwerte der Variable mit dem Mittelwert oder den 25% und 75% Quantil vergleichen

Grafisch: Auf Boxplot oder Scatterplot kann man zB gut Outlier erkennen

Man könnte die Werte in einer Tabelle (in R) nach Größe ordnen und vergleichen

Bar chart I and bar chart II plot the same variables. What is the difference between the two bar charts? How would you assess this difference?

Der Unterschied ist, dass Bar Chart I bei 0 beginnt und Bar chart II nicht, obwohl dieseben Werte abgebildet werden. Dies führt zu einer optischen Täuschung

Balkendiagramme müssen eine Null-Basislinie haben! Wenn unsere Augen Balkendiagramme interpretieren, vergleichen wir die relativen Höhen der Balken. Wenn die Basislinien nicht bei Null beginnen, wird dieser visuelle Vergleich verzerrt und der Unterschied zwischen den Balken auf eine Weise überbetont, die nicht ehrlich ist. Mit einer Null-Basislinie kann man den Unterschied zwischen zwei Balken deutlich besser und realistischer einschätzen.

State a reason to use color coding in this regard.

Ermöglicht eine multivariate Analyse der Daten. Durch das Color Coding wird es einfacher, Korrelationen und Trends zwischen den drei Variablen zu erkennen

Information from the Scatterplot

Man erkennt, dass sich teure Häuser (jene über 30,000$) tendenziell eher in Gegenden mit geringer Prozentzahl an sozioökonomisch schwacher Population und in Gegenden mit geringer NOX Belastung befinden.

Man erkennt eine negative Beziehung zwischen Preis des Hauses und Anzahl sozioökonomisch schwacher Population. Je größer der Anteil der sozioökonomisch schwächeren Bevölkerung in einer Population, desto günstiger der Hauspreis

Houses with higher values have a maximum LSTAT of about 15% and a maximumNOX of around 0.65.The number of houses with a value above $30.000 is lower than the number ofhouses with a lower value

What is the leading question in performance evaluation?

Wie gut ist das Modell darin, neue Werte im Validierungsdatensatz vorherzusagen?

How well does the model predict new data? (not: How well does it predict thedata it was trained with?)

Explain what each line of the following code does // INTERPRETATION CODE LOGIT

1. Bank.df <- read.csv("UniversalBank.csv"): importiert die CSV-Datei "UniversalBank" als Dataframe und speichert diese unter Bank.df ab
2. Set.seed(2): setzt den Zufallszahlengenerator auf einen bestimmten Startwert. Ermöglicht reproduzierbare Ergebnisse. Interessant zB für Debugging, oder um den Code nachvollziehbarer für andere zu machen
3. Train.index <- sample(c(1:dim(bank.df)[1]), dim(bank.df)[1]*0.6): Zieht aus dem gesamten Datensatz bank.df 60% Zeilennummern zufällig und speichert diese im Vektor Train.index ab
4. Train.df <- bank.df[train.index,]: Erstellt den Trainingsdatensatz: Fügt die zuvor zufällig gezogenen Zeilen dem Data Frame Train.df hinzu, inkl. Aller Variablen.
5. Valid.df <- bank.df[-train.index,]: Erstellt den Validierungsdatensatz: Die restlichen Zeilen aus dem bank.df Datensatz (die nicht im Trainingsdatensatz gelandet sind) werden nun dem Validierungsdatensatz hinzugefügt.
6. Logit.reg <- glm(Personal.Loan ~., data=train.df, family = "binomial"): Schätzt ein Logit-Modell. Versucht den Effekt aller im Trainingsdatensatz enthaltenen Variablen auf die Wahrscheinlichkeit, ob der Kunde den Kredit annehmen wird (Y=1) oder nicht (Y=0), zu bestimmen.
7. Logit.reg.pred <- predict(logit.reg, valid.df[, -8], type = "response"): Anhand des geschätzten Logit-Modells (Logit.reg) werden nun Vorhersagen auf dem Validierungsdatensatz getroffen, ausgenommen ist hierbei jedoch die 8. Spalte (bzw. 7. Erklärende Variable). Dies wird unter Logit.reg.pred gespeichert
8. Roc.df <- data.frame(actual = valid.df$Personal.Loan, predicted = Logit.reg.pred): Erstellt einen Dataframe, in dem die vorhergesagten Werte (Wahrscheinlichkeiten) und die tatsächlichen Werte (ob Kredit angenommen wurde oder nicht) abgetragen.
9. r <- roc(roc.df$actual, ref=1, roc.df$predicted): Erstellt eine ROC-Kurve. Refernzklasse ist hierbei 1, also dass der Kredit angenommen wird, als zweites wird die Variable mit den vorhergesagten Wahrscheinlichkeiten angegeben.
10. plot.roc(r): Plottet die ROC-Kurve

You type the following command into R and get the following output:

Ist eine Tabelle die dafür genutzt wird, die Performance eines classification Modells zu evaluieren auf Basis eines Validierungsdatensatz, für den die tatsächlichen Werte bekannt sind.

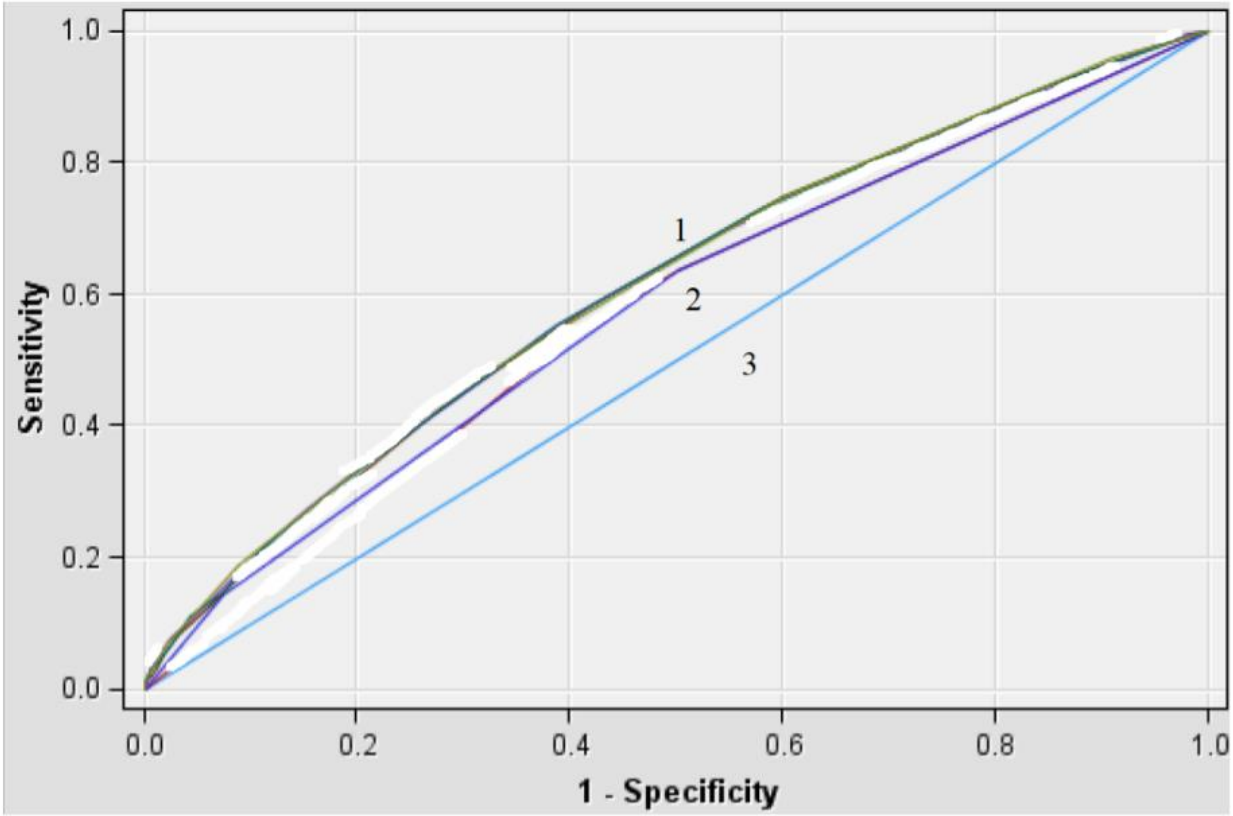
Cut-Off value ist hierbei 0.8. Dh dass ab einem vorhergesagten Wert von 0.8 diese Beobachtung als 1 (also Kunde wird Kredit annehmen) klassifiziert wird und alle Beobachtungen unter 0.8 werden als 0 klassifiziert.

Specificity: True Negative Rate:

1803 / 1812 = 0.9950

Sensitivity: True Positive Rate:

103 / 191 = 0.5393



Which model do you choose and why?

Model 1 weil bei diesem Modell die Area under the Curve am größten ist. Somit weist Model 1 die beste Performance auf.

Variable	Description
price	price of the car (Euro)
metallic	= 1 if metallic color, = 0 else
weight	weight of the car (kg)
warranty	remaining warranty (years)

Price = beta_0 + beta_1*metallic + beta_2*weight + beta_3*warranty + e

Are metallic-colored cars sold at higher or lower prices than “ordinary” cars?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -33136       1560     -21   <2e-16 ***
metallic         572        163       4    5e-04 ***
weight            40          2      28   <2e-16 ***
warranty         185         25       7    5e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Käufer tendieren dazu, mehr für metallic Autos zu zahlen, da colored Autos sind ceteris paribus 572€ teurer als nicht metallic-Autos.

Your supervisor is surprised about βˆ weight = 40 being both positive and significant. Interpret this coefficient. Give an explanation for βˆ weight > 0.

Pro Einheit Gewicht (in KG), erhöht sich der Preis des Autos c.p. um 40 Geldeinheiten. Dies erscheint plausibel, denn schwerere Autos haben häufig größere Motoren mit mehr Leistung, was sich häufig in einem höheren Preis niederschlägt. SUVs, die häufig teurer sind also andere Karosserieformen wie Limousine oder Kleinwagen, wiegen häufig auch mehr.

Your supervisor tells you that the firm is considering to offer car warranty extensions for 200 Euro per additional year. Will there be a large demand for this kind of warranty extension? Explain based on your estimation results.

Nein, da Kunden aktuell 185 € bereit sind zu zahlen für ein Jahr zusätzlicher Garantie. Somit würde der Preis von 200€ über der Zahlungsbereitschaft Kunden liegen und die Nachfrage nach einer solchen Versicherung wäre gering.

Why should you use logistic instead of linear regression if y= buyer?

Da es sich bei diesem problem um ein Klassifizierungsproblem handelt (Buyer ist kategorial) Eine lineare Regression würde einen kontinuierlichen Output herausgeben, der in diesem Fall schlechter zu interpretieren wäre. Eine logistische Regression hingegen gibt diskrete outputs aus, die in diesem Fall (Buyer = 1, oder non-buyer=0) angebrachter wären.

How do we interpret logistic regression coefficients?

Logistische Regressionskoeffizienten geben die relative Veränderung in den log(odds) wieder wenn die Prädiktorvariable um eine Einheit steigt. Somit ist es nicht trivial, genaue Aussagen über den Effekt einer Variable zu treffen. Anhand des Vorzeichens kann man jedoch schnell erkennen, ob eine Variable einen positiven oder negativen Effekt auf die Wkt. Für Y=1 hat.

Are the odds larger or smaller than 1 if most visitors leave the online shop without buying anything? Explain using the mathematical definition of the odds.

Letting p= Pr(buyer= 1), “most visitors leave without buying anything" translates into p < (1p) . Consequently, the odds=p/(1-p)must be smaller than 1.

Probeklausur 1819

One step in the Data Mining Process (supervised learning) is to partition the data into different datasets. Consider you partition the data into two sets. Please, state the name of those datasets and explain shortly how you use each of them. (4 points)

Trainingsdaten: Hier wird das Modell entwickelt und trainiert. Typischerweise ist dies die größte Gruppe

Validierungsdaten: In dieser Gruppe wird die Performance des Modells gemessen. Das in den Trainingsdaten entwickelte Modell schätzt mithilfe der Validierungsdaten nun neue Werte, welche mit den tatsächlichen Werten verglichen werden. Ggf. Finden Korrekturen und Optimierungen des Modells statt.

Explain shortly why we partition the data and relate it to the problem of overfitting in this context. State one posible reason for overfitting as well

Ziel ist es, ein Modell zu finden, welches möglichst gut neue Werte schätzen kann, nicht bereits bekannte Werte. Hierfür wird ein Datensatz in mind. 2 Gruppen (trainings- und validierungsdatensatz) aufgeteilt. Im Datensatz wird das Modell entwickelt und trainiert und im Validierungsdatensatz wird die Performance des Modells getestet. Diese Partitionierung erfolgt, da das entwickelte Modell sich sonst zu sehr an die bereits bekannten Daten anpassen würde (Gefahr des Overfittings). Ziel ist es aber, ein generalisierendes Modell zu finden, welches möglichst gut neue (unbekannte) Werte schätzt. Ein Grund für Overfitting ist z.B. das Anwenden von zu vielen Prädiktoren. Ein komplexes Model könnte zwar einen exzellten Fit zu dem bereits bestehenden Datensatz haben, aber in neuen Daten dafür umso schlechter performen.

What does the following R command do?

```
s <- sample(row.names(housing.df), 5, prob = ifelse(housing.df$ROOMS>10,
0.9, 0.01)) (3 points)
```

Aus dem Datensatz housing.df werden 5 Zeilen zufällig gezogen. Dabei werden Häuser mit mehr als 10 Zimmern übergewichtet in ihrer Wahrscheinlichkeit, gezogen zu werden. Jene Häuser werden nämlich mit einer Wkt. Von 90% gezogen, und 10% andernfalls.

State in words or by equation how you normalize a variable

Normalisieren:= Anpassung von Werten, die auf unterschiedlichen Skalen gemessen wurden, an eine gemeinsame Skala.

Formel: zi = (x_i / mean(x)) / sd(x)

Subtrahieren des Mittelwerts von jeder Beobachtung und anschließende Division durch die Standardabweichung.

State one reason why we normalize variables.

Wenn man Variablen mit unterschiedlicher Maßeinteilung hat. Große Maßeinheiten würden dominieren und die Ergebnisse verzerren. Z.B. wenn man Daten über Geld in 1000€ Schritten hat und Größe in cm,

kann dies die Vergleichbarkeit der Daten problematisch machen. Durch das Normalisieren werden die Daten auf eine gemeinsame vergleichbare Basis bzw auf eine gemeinsame Skala (zwischen 0 und 1) gebracht.

What type of graph is obtained by the following command and what does this type of graph show?

hist(housing.df$MEDV, xlab = "MEDV")

- ➔ Erstellt ein Histogramm über die Verteilung des Medianwerts der Häuser in dem Datensatz
- ➔ Histogramme werden genutzt um kontinuierliche Variablen darzustellen

What would you in general do when you face a skewed distribution?

In eine logarithmische Darstellung transformieren

Briefly explain what an ROC curve is

Ist ein Tool zur Evaluierung der Performance eines Modells. Die ROC-Curve illustriert die Sensitivität und Spezifität bei sinkendem Cut-off value (von 1 bis 0). Wird benutzt um den optimalen Cut-off Value zu identifizieren. Spiegelt den Trade-Off zwischen Sensitivität und Spezifität wider. Auf der X-Achse ist die Specificity bzw die False Positive Rate abgetragen. Die sensitivity gibt die Prozentzahl korrekt Klassifizierter 1en bzw Responder wieder. Die Specificity gibt die Prozentzahl korrekt klassifizierter negatives wieder.

Mit der ROC wird der optimale Cut-Off value identifiziert, bei dem ein Modell die beste Performance liefert.

Explain how to use ROC curves in order to determine which model is best and illustrate your answer with a graph

Man kann die ROC-Curve für mehrere Modelle gleichzeitig verwenden. Das Modell mit der ROC-Curve, die am weitesten oben links von der Naive-Rule-Gerade verläuft, performt am besten, da hier die AUC am größten ist.

Provide an example for a situation where we might prefer a model which does not minimize the misclassification rate.

Oft bei asymmetrischen Kosten. Z.B. bei Corona-Tests, wo ein False-Negative-Test deutlich problematischer und mehr Schaden anrichten würde, als ein False-Positive-Test. Hier wird eher versucht, die False Positive Rate zu minimieren, auch wenn dies nicht die Missklassifizierungsrate minimieren würde. ZB bei Corona-Test (Wo False-Negative schlimmer ist als False-Positive)

What is the difference between R^2 and R^2 adj?

R^2 steigt stets mit steigender Anzahl an Variablen an, dies führt schnell zu Overfitting des Modells. Adjustiertes R^2 ist so formuliert, dass das Einfügen zu vieler Variablen bestraft wird.

Was ist der ME?

- ➔ Mean Error: ist ein Genauigkeitsmaß. Hier heben sich positive und negative Fehler gegenseitig auf. Daraus kann man schließen ob man eine systematisch über- oder unterschätzung hat. Wenn ME negative, hat man Überschätzung der Werte

Wird für Validierungsdaten oder Trainingsdaten der ME berechnet?

➔ ME wird in Validierungsdaten berechnet. Mit dem ME berechnet man, wie sehr das geschätzte Modell von tatsächlichen Werten abweicht. In den Trainingsdaten sollte der ME = 0 sein, da sonst die Koeffizienten verzerrt sind.

Suppose we are interested in the determinants of prices of used cars. Explain what each of those ten lines of code is doing (lines 9 and 10 belong together and are one line of command) // Exhaustive Search Code

```
1  selected.var <- c(3, 4, 7, 8, 9, 10, 12, 13, 14, 17, 18)
2  Fuel_Type <- as.data.frame(model.matrix(~ 0 + Fuel_Type, data=car.df))
3  car.df <- cbind(car.df[,-4], Fuel_Type[,])
4  car.df$Fuel_TypePetrol <- NULL
5  set.seed(1)
6  train.index <- sample(c(1:1000), 600)
7  train.df <- car.df[train.index, selected.var]
8  valid.df <- car.df[-train.index, selected.var]
9  search <- regsubsets(Price ~ .^2, data = train.df[c(1:4, 7:10)],
10         nbest = 1, nvmax = 20, method = "exhaustive", really.big=T)
11 summary(search)
```

1. Speichert einen Vektor mit den Zahlen in der Klammer unter dem Namen selected.var ab
2. Transformiert die Variable Fuel_Type aus dem Datensatz car.df in einen eigenen Dataframe und speichert diesen unter Fuel_Type an -> Dummy Variable wird erstellt
3. Die Variable Fuel Type wird aus dem Datensatz Car.df entfernt und die in 2. Erstellen Dummy-Variablen für Fuel-Type werden stattdessen eingefügt.
4. Setzt alle Werte aus der Kategorie Benzin auf Null (Um Multikollinearität zu vermeiden)
5. setzt den Zufallszahlengenerator auf einen bestimmten Startwert. Ermöglicht reproduzierbare Ergebnisse. Interessant zB für Debugging, oder um den Code nachvollziehbarer für andere zu machen
6. Zieht aus dem Intervall 1 bis 1000 zufällig 600 Zahlen und speichert diese unter train.index (Für Trainingsdatensatz)
7. Aus dem Datensatz car.df werden 600 Zeilen (entsprechend den 600 zuvor gezogenen zufälligen Zahlen) inklusive der im ersten Schritt festgelegten Variablen im Validierungsdatensatz valid.df gespeichert
8. Die restlichen Daten des car.df Datensatzes, die nicht im Trainingsdatensatz gelandet sind, werden in den Validierungsdatensatz eingefügt (inkl.der im erstne Schritt festgelegten Variablen)
9. Führt eine Exhaustive Search auf Basis der ersten 4 Zeilen und der 7.-10. Spalte im Trainingsdatensatz durch. Probiert alle möglichen Kombinationen von den X-Variablen und schaut, wie sie performen (mithilfe des adjustierten R^2). Nur das beste Modell für Modells von 1 bis 20. Really.Big=TRUE um R mitzuteilen, dass viel Rechenleistung für diesen Command benötigt wird. Speichert ab in search.
10. Präsentiert die Ergebnisse der Exhaustive Search

Schätzmodell

Price = b0 + b1*automatic + b2*parking_assistant + b3*doors + b4*cd_player + e

7619.99 + 539.4*2 = 8698.79€ ist der durchschnittliche Preis

-> wenn Schätzung für parking_assistant statistisch significant wäre, würden Käufer die Einparkhilfe mehr schätzen, da der Koeffizient höher ist als für Automatik (1210.04 > 621.48). Es lässt sich zumindest zu einem 90% Signifikanzniveau sagen, dass Kunden bereit sind, 621.48 c.p. mehr für ein Auto mit Automatik zu zahlen.

Sollte doors als numerische Variable behandelt werden?

Nein, als kategoriale, da numerische Variablen kontinuerlich sein können, was bei Anzahl der Türen wenig sinn ergeben würde. Hier müsste man z.B. Dummy Variablen für jede Anzahl an Türen einfügen.

Der Leiter der Verkaufsabteilung ist begeistert…

Er scheint begeistert zu sein, weil die Kunden überaus viel Geld bereit sind zu zahlen für ein Auto mit cd_player. Demnach sind Kunden für ein Auto mit CD Player dazu bereit, c.p. 4103.08€ mehr zu zahlen als für ein Auto ohne CD-Player. Dies wird dadurch unterstützt, dass der Koeffizient statistisch significant ist.

Der Leiter der Verkaufsabteilung könnte zB bei seinen Autos ohne CD Player einen CD-player nachrüsten, wenn dies durchschnittlich günstiger als 4103.08 ist, dann könnte man seine Begeisterung teilen.

Logit Model: was gibt b1 an?

Gibt die relative Änderung in den log(odds) wieder, wenn Alter des aktuellen Smartphones des Kunden um einen Monat steigt. Die Interpretation hierbei ist nicht trivial. Stattdessen kann man anhand des Vorzeichens schnell erkennen, ob ein positiver oder negativer Effekt auf die Wahrscheinlichkeit “Buyer” zu sein besteht

Kleiner oder größer Null?

Wahrscheinlich größer Null, da je Älter das Smartphone eines Individuums, desto größer ist tendenziell die Wahrscheinlichkeit, dass dieser sich ein neues Smartphone wünscht, aufgrund alternder Technologie o.ä..

Ältere Kunden mehr oder weniger geneigt…

Wenn sich das Alter des Kunden erhöht, verringert sich tendenziell die Wahrscheinlichkeit, dass dieser sich das neue Smartphone kauft. Dieser Effekt ist jedoch nicht statistisch significant, weswegen man keine feste Aussage über den Effekt des Alters auf die Wahrscheinlichkeit für einen Kauf treffen kann.

Decile-wise lift chart

Aus der Grafik kann man ableiten, dass wenn man die Werbung an die vom Modell am wahrscheinlichsten identifizierten obersten 10% der Kunden verschickt, dass diese fast 7 mal so

wahrscheinlich sind, das Smartphone zu kaufen, als wenn man die Werbung zufällig an 10% der Kunden verschickt. Ja ich würde dem Unternehmen empfehlen, mit diesem Modell zu arbeiten, da dadurch immense Kosten eingespart werden können für Werbungen, auf die nicht reagiert wird.

Nennen sie eine Variable, die zusätzlich aufgenommen werden könnte:

Vielleicht Einkommen? Falls Daten vorhanden. Intuitiv kann man argumentieren, dass jemand mit höherem EInkommen tendenziell dazu neigt, neue Technologien (inkl.neuer Smartphones) auszuprobieren im Vergleich zu jemandem, der nicht viel Geld zur Verfügung hat und dementsprechend andere Dinge finanzieren muss