Chair of Decision Sciences & Systems
Department of Informatics
Technical University of Munich

Business Analytics: Endterm Exam

Module:	IN2028	**Date:**	Wednesday 17th February, 2021
Examiner:	Prof. Dr. Martin Bichler	**Exam:**	Endterm

First name:	Andreas	**Registration number, UIDs:**	03690388, E5076 / S5076
Last name:	Binder	**Last update:**	Monday 8th March, 2021

Σ (Endterm)	Grade (Endterm)	Grade intervall
58.0	**1.7**	[57.0; 61.5)

Notes:

- Please make sure that the total amount of credits stated above is correct.
- Solely the second correction (green color) is decisive.

Corrections:
The table below lists all corrections (image recognition and complaints during review) that are already considered in the calculation of your grade. If a problem or subproblem is listed multiple times, the correction with the highest number (column “Correction”) takes precedence.

Problem	Correction	credits	Annotations

Chair of Decision Sciences & Systems
Department of Informatics
Technical University of Munich

Personal sticker

S5076

Compliance to the code of conduct
I hereby assure that I solve and submit this exam myself under my own name by only using the allowed tools listed below.

Signature or full name if no pen input available

Business Analytics: Endterm Exam

Exam:	IN2028 / Endterm	**Date:**	Wednesday 17th February, 2021
Examiner:	Prof. Dr. Martin Bichler	**Time:**	08:00 – 09:30

For personal review only

Working instructions

- This exam consists of **22 pages** with a total of **6 problems**.
- The total amount of achievable credits in this exam is 90 credits.
- Allowed resources: open book.
- Subproblems marked by * can be solved without results of previous subproblems.
- **Answers are only accepted if the solution approach is documented.** Give a reason for each answer unless explicitly stated otherwise in the respective subproblem.
- Do not write in red or green colors.
- Any intermediate or final numbers in your calculations may be **rounded to four (4) digits**.

ONLINE SUBMISSION

Problem 1 Regression Analysis (15 credits)

You collected the following data about $n = 73$ different locations in Germany for a representative winter day in 2020:

Variable	Range	Explanation
snowfall	$\{0, 1\}$	occurrence of snowfall
temperature	$\mathbb{R}$	air temperature [°C]
humidity	$\mathbb{R}_{\geq 0}$	absolute humidity [g/m^3]
speed	$\mathbb{R}_{\geq 0}$	wind speed [km/h]
direction	$\{North, East, South, West\}$	wind direction

You would like to model *snowfall* by applying a logistic regression.

0 1 2 **2.0** a)* Briefly describe the data set:

1. Name the dependent and independent variables.
2. What scale of measurement does the variable *direction* belong to? How would you process the variable *direction* for the regression?

1. snowfall is the dependent one, temper. humidity, speed & direction the independent ones

2. it is nominal. I would create three dummy variables (North, East, South)

0 1 2 **0.5** b)* You decide to remove the variable *direction* since you consider it too unspecific. The remaining variables serve as an input to the logistic regression. Provide a formula to compute the likelihood function. The formula should include the variable names.

variables not explained

$$\prod \left(\frac{1}{1+e^{-\beta_0+\beta_1 T}}\right)^{SN_1} \times \left(1-\frac{1}{1+e^{-\beta_0+\beta_1 T}}\right)^{1-SN_1} \times \left(\frac{1}{1+e^{-\beta_0+\beta_2 S}}\right)^{SN_2} \times \left(1-\frac{1}{1+e^{-\beta_0+\beta_2 S}}\right)^{1-SN_2} \times \left(\frac{1}{1+e^{-\beta_0+\beta_3 H}}\right)^{SN_3} \times \left(1-\frac{1}{1+e^{-\beta_0+\beta_3 H}}\right)^{1-SN_3}$$

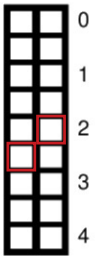
You apply a logistic regression and obtain the following output:

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.7088 -0.3151 -0.0666  0.3791  0.6938

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.808268   0.334746  -2.415  0.01841 *
temperature -0.046228   0.010961  -4.218 7.37e-05 ***
humidity     0.056305   0.019977   2.818  0.00629 **
speed        0.005037   0.005824   0.865  0.39010
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 15.753 on 72 degrees of freedom
Residual deviance: 10.861 on 69 degrees of freedom
```

2.0 c)* Interpret the estimated model:

1. Which variables are statistically significant at 5%? Explain your reasoning.
2. Interpret the *intercept* and the coefficient for *temperature*. Refer to the odds and provide necessary calculation steps.

1. -0.5 argument not clear
2. -1 intercept interpretation missing



1. All variables except speed. The p-value is too low

2. Intercept

$$\frac{p(x+1)}{1-p(x+1)} = e^{\text{temp.}} \times \frac{p(x)}{1-p(x)}$$

If temp increases by 1 the odds increase by $e^{-0.046228}$

The intercept is at 0.446

2.5 d)* The McFadden R^2 for this model is 0.329482. Interpret this value. Explain the meaning of the McFadden R^2 and how it differs conceptually from the R^2 measure for OLS regression.

-0.5 mcFadden interpretation not full
-1p can not understand first sentence for the model performance due to bad writing

The value is fine, giving a good fit. The McFadden checks how much better the model performs than if all coefficients are 0.
R^2 measures how much of the variance is captured

You are advised to include air pressure as another covariate. Introducing the variable *pressure*, your regression yields the following results:

Residuals:

Min	1Q	Median	3Q	Max
-0.74349	-0.28478	-0.05542	0.34002	0.67652

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.798156	0.335005	-2.383	0.02000	*
temperature	-0.128528	0.084762	-1.516	0.13407	
humidity	0.056654	0.019986	2.835	0.00604	**
speed	0.004232	0.005884	0.719	0.47442	
pressure	0.083481	0.085255	0.979	0.33096	

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 15.753 on 72 degrees of freedom
Residual deviance: 10.710 on 68 degrees of freedom

The new McFadden R^2 is 0.330173.

3.5 e)* Analyze these results:

1. Examine the significance and the McFadden R^2. What is the impact of adding the variable *pressure*?
2. Explain these results. Which Gauss-Markov property is affected and how can you test for it?

-1 significance of variables not included

1. There seems to be a slight increase in the information that pressure added

2. The significance levels dropped which is cause caused by multicollinearity. I would use the variance inflation factor. and

Problem 2 Data Preparation (10 credits)

You have obtained a dataset with metadata about 2,731 videos uploaded to a popular streaming platform. Some videos get featured on the platform's main page which usually helps video creators to reach a much wider audience. Based on your available data, you want to use a general linear model in R, i. e. a logistic regression, and perform statistical inference to determine the main drivers behind the platform's decision whether or not to feature a video.

While looking at the data, you realize that there are several data preparation steps you will need to undertake before you can run your model. Below, you are given three resources: (A) a short description of the variables in your data, (C) a summary of each column's distribution, as well as (B) the correlation between the numerical variables.

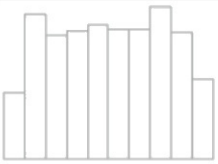
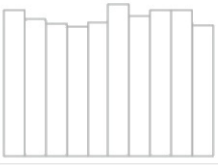
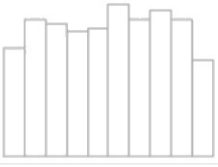
Resource A Your dataset has the following columns (see table below for types and distributions):

1. **video_id:** A unique identifier of each video.
2. **views:** The number of views the video has received on the platform.
3. **rating:** The star-rating (from 1 star '*' to 5 stars '*****').
4. **earnings:** earnings in US$ that were paid out to the video creator as part of a revenue-sharing agreement with the platform.
5. **account_gender**: the gender of the video creator
6. **account_age**: the age (in years) of the video creator
7. **sentiment_english**: A score between 0.0 (very negative comments) and 1.0 (very positive comments) that describes the sentiment of user comments posted below the video in English language.
8. **sentiment_other**: A score between 0.0 (very negative comments) and 1.0 (very positive comments) that describes the sentiment of user comments posted below the video in languages other than English.
9. **featured**: Whether or not the video was featured on the home page or not.

Resource B The *correlation matrix* between numerical features is given below:

	views	earnings	account_age	sentiment_english	sentiment_other
views	1.00000000	0.76582721	NA	0.01824006	0.01940797
earnings	0.76582721	1.00000000	NA	0.01947576	0.02048991
account_age	NA	NA	1	NA	NA
sentiment_english	0.01824006	0.01947576	NA	1.00000000	0.99860600
sentiment_other	0.01940797	0.02048991	NA	0.99860600	1.00000000

Resource C Summary of each variable's distribution

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	video_id [character]	1. 00Etx6TK 2. 016AmFYV 3. 01nrJJBC 4. 025OtRki 5. 02Kf4xNZ 6. 02VPhDIb 7. 03Xe0Uza 8. 06PdjMvm 9. 07yFfne3 10. 082cBmZJ [2721 others]	1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 2721 (99.6%)		2731 (100%)	0 (0%)
2	views [numeric]	Mean (sd) : 9300.9 (28084) min < med < max: 26 < 2824 < 799669 IQR (CV) : 7080 (3)	2351 distinct values		2731 (100%)	0 (0%)
3	rating [character]	1. * 2. ** 3. *** 4. **** 5. *****	543 (19.9%) 538 (19.7%) 533 (19.5%) 561 (20.5%) 556 (20.4%)		2731 (100%)	0 (0%)
4	earnings [numeric]	Mean (sd) : 194.3 (679.2) min < med < max: 0 < 39.7 < 19567 IQR (CV) : 147.5 (3.5)	2230 distinct values		2731 (100%)	0 (0%)
5	account_gender [character]	1. F 2. M	1236 (51.8%) 1152 (48.2%)		2388 (87.44%)	343 (12.56%)
6	account_age [integer]	Mean (sd) : 40.9 (13.2) min < med < max: 18 < 41 < 64 IQR (CV) : 23 (0.3)	47 distinct values		2388 (87.44%)	343 (12.56%)
7	sentiment_english [numeric]	Mean (sd) : 0.5 (0.3) min < med < max: 0 < 0.5 < 1 IQR (CV) : 0.5 (0.6)	2731 distinct values		2731 (100%)	0 (0%)
8	sentiment_other [numeric]	Mean (sd) : 0.5 (0.3) min < med < max: 0 < 0.5 < 1 IQR (CV) : 0.5 (0.5)	2731 distinct values		2731 (100%)	0 (0%)
9	featured [character]	1. No 2. Yes	2599 (95.2%) 132 (4.8%)		2731 (100%)	0 (0%)

8.0

Based on the resources and the goal above, identify five problems of the data set in its current form. For each, quickly describe why it is problematic in the given context, and name a possible data preparation strategy to overcome it. *After performing the steps you describe, the resulting data set should be suitable to fit your model given the following specification in R:*

```
model <- glm(featured ~ ., data=video_data_clean, family='binomial')
```

0 1 2 3 4 5 6 7 8 9 10

1. The two sentiment features have a high correlation which can cause multicoll. I would exclude one of those ✓ 2 ✓

2. I would leave out the ID before fitting because this feature contains no useful info. ✓ 2 ✓

3. You have to deal with missing values in Acc. Gender & Acc Age. for gender, I would remove ✓ 2 ✓ NAs & for ~~Acc~~ Acc. age I would impute this via the mean

4. You should convert character columns to factor columns 1 → R does it automatically (except for label)

5. You have to create dummy variables for coding. One hot encoding for stars 1-4. Leave one out to aft multicoll.

better convert to factor or numeric to reflect ordinality (+1P for problem)

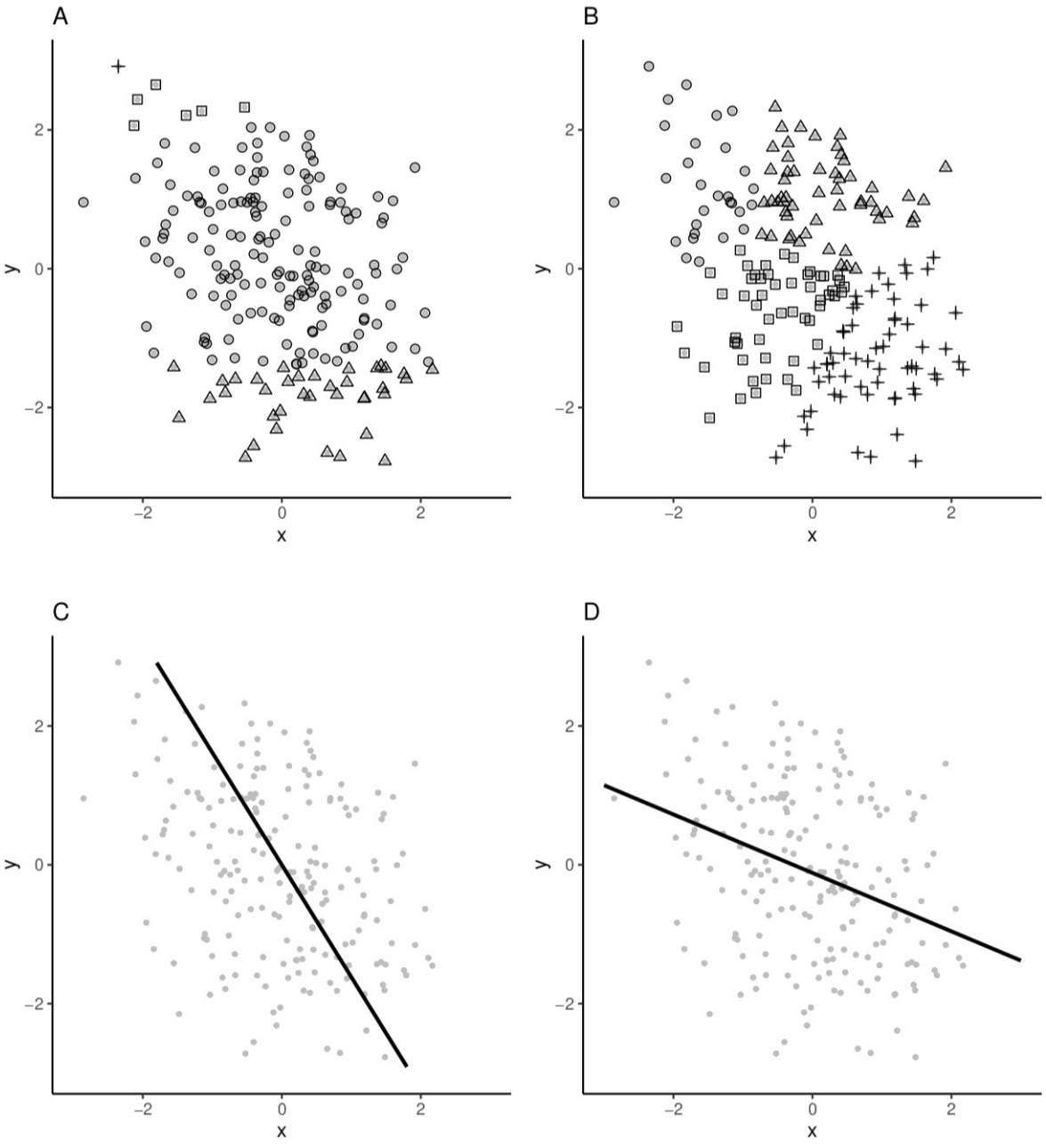
Problem 3 Identify Models (10 credits)

You have been hired by a client to help them understand a problem in one of their business processes. In your analysis, you have collected a dataset of 200 problem instances and fitted several statistical models on the data:

- linear regression (OLS),
- k-means,
- principal components analysis,
- decision tree.

You are about to present your results to the clients and have printed handouts with visualizations of your models, but suddenly you realize that your assistant has forgotten to label each visualization with its title.

For each printout (A,B,C,D) below, (a) name which of the models is shown, (b) quickly summarize the visualization (What aspect(s) of the model are shown?) (c) justify your choice. (Why is it this model and not one of the others?)



5.5

B: K-Means: One can see four separated clusters ✓ ~~2~~ 1.5

A: Decision Tree: ✓ 1

C: PCA: the line captures the maximal variance ✓ ~~2~~

D: Linear Regression: it fails to draw a meaningful line ✓ 1

→ line tries to minimize $(y-\hat{y})^2$ and hence a little flatter in comparison to PCA which minimizes orthogonal residuals.

Problem 4 Evaluation (18 credits)

Your cybersecurity company has developed two new email spam filters. The managers would like to have an analysis on both software systems (A and B). You have received a small but representative trial of experimental results:

True class	Probability of Classifier A	Probability of Classifier B
+	0.87	0.81
-	0.21	0.25
-	0.78	0.79
+	0.53	0.68
+	0.94	0.91
-	0.68	0.89
-	0.37	0.54
+	0.83	0.88
-	0.02	0.17
+	0.67	0.77

Each row represents an email, which is either spam (positive label) or no spam (negative label). The probabilities indicate the likelihood of being classified as spam by the respective software. The default cutoff value for both systems is 0.76.

0 1 2

1.0 a)* Calculate the *accuracy* of both software systems. Provide formula and calculations.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Acc_A = \frac{3 + 4}{3 + 4 + 1 + 2} = 0.7 \checkmark$$

$$Acc_B = \frac{3 + 3}{3 + 3 + 4} = 0.6 \times$$

3.0 b)* Provide formulas and calculate *recall* and *specificity* for both software systems. Explain the results and characterize both software systems accordingly.

$Recall = \frac{TP}{TP + FN}$ $Specif. = \frac{TN}{TN + FP}$

$Rec._A = \frac{3}{5} = 0.6$ ✓ $Spec_A = \frac{4}{5} = 0.8$ ✓

$Rec_B = \frac{2}{5} = 0.4$ f $Spec_B = \frac{3}{5} = 0.6$ ✓

Based on this sample, Classifie A performs generally better although both classifier at minimizing False positives than min. False Negatives 1

0 1 2 3 4

2.5 c)* Many customers prefer that a high share of the emails classified as spam are indeed spam emails. Which software would you recommend? Name and calculate a meaningful metric.

We want a software with high Specificity/thus few False Pos. f I would recommend Class. A (✓)

because of Precision

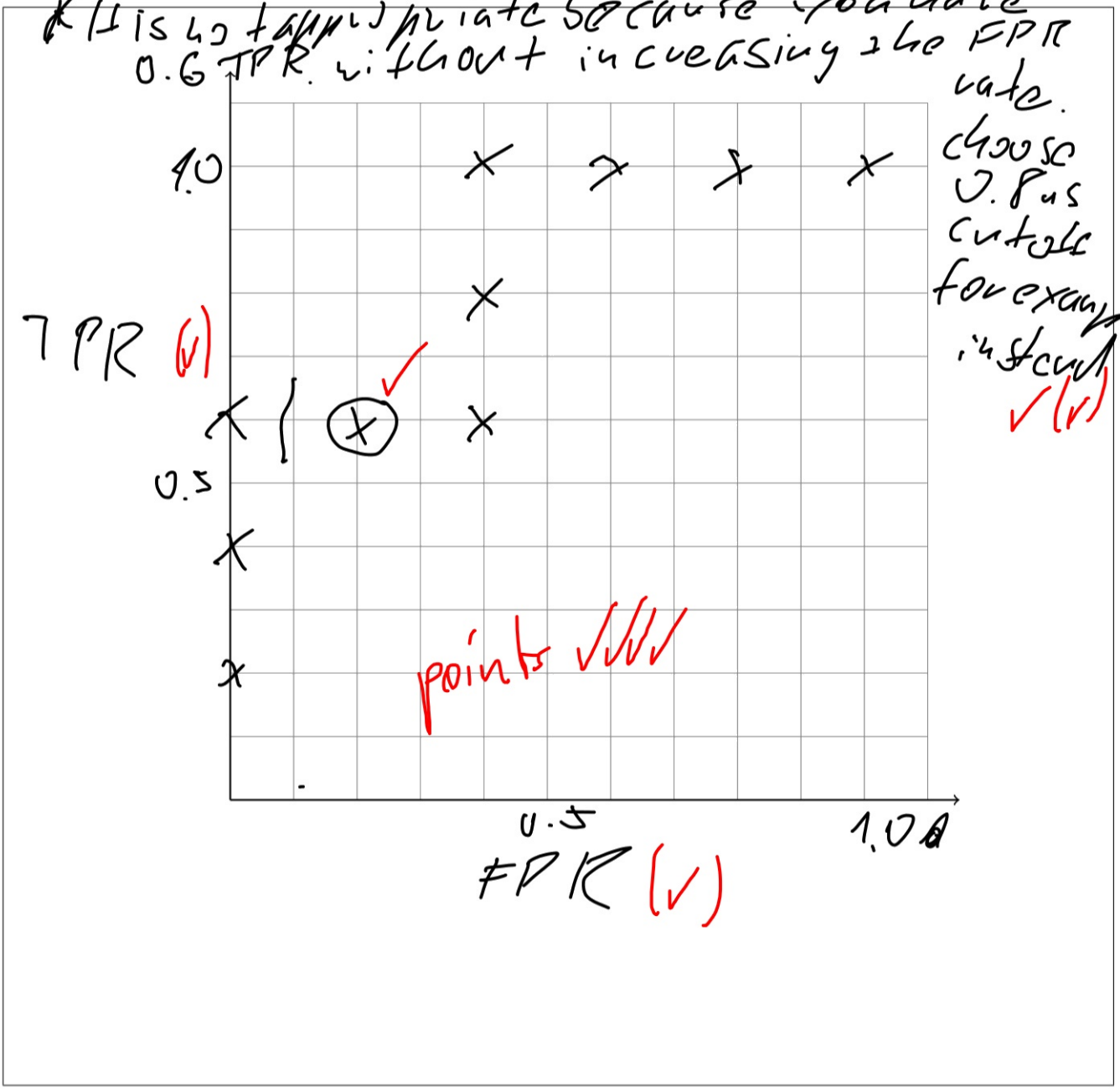
0 1 2

Note: Table identical as above (depicted here for ease of readability).

True class	Probability of Classifier A	Probability of Classifier B
+	0.87	0.81
-	0.21	0.25
-	0.78	0.79
+	0.53	0.68
+	0.94	0.91
-	0.68	0.89
-	0.37	0.54
+	0.83	0.88
-	0.02	0.17
+	0.67	0.77

0 1 2 3 4 5 6 7

7.5 d)* Draw the ROC curve of software A. Name the axes and write down the (x,y)-coordinates for every point. Mark the point corresponding to the default cutoff value of 0.76. Would you argue that this cutoff value is appropriate? Explain your reasoning.



2.0 e)* Do you agree with the following statement? Explain your reasons.

"The precision can be easily computed from the ROC curve. For a particular cutoff value, it equals the slope of the line connecting the origin and the point on the ROC curve corresponding to the cutoff value."

0 1 2

No, for point (0.4, 0.6) for example we would get 1.25 ~~but~~
The max value is 1. You find the Recall values on the y-axis

Page empty

Problem 5 Decision Trees (17 credits)

A laundry manager wants to optimize her energy and detergent usage. Therefore, she tested the washing results for various settings and reported her results in Table 5.1.

Temperature	Detergent	Dirt	Result
30	low	low	good
30	low	high	bad
30	low	medium	bad
30	high	high	bad
40	low	low	good
40	low	medium	bad
40	high	high	bad
40	low	medium	good
60	high	medium	good
60	high	medium	good

Table 5.1: Laundry results.

4.0 a)* Of the first three attributes, find that one that would be at the root of a decision tree using the **information gain** criterion. For the numeric attribute *Temperature* consider binary splits.

Note: For this exercise, we will again use $0 \cdot \log(0) = 0$.

info(R) = info([5,5]) = 1

info(T=35) = info([2,2],[4,2]) = 4/10 × 1 + 6/10 × 0.918 = 0.951

gain(T=35) = 0.049 ✗✗

info(T=50) = info([2,0],[3,5]) = 0.2 × 0 + 0.8 × 0.954 = 0.7632

gain(T=50) = 0.237 ✓

info(De) = info([3,3],[2,2]) = 0.6 × 1 + 0.4 × 1 = 1

gain(De) = 0 ✓

info(Di) = info([2,0],[0,3],[3,2]) = = 0.2 × 0 + 0.3 × 0 + 0.5 × 0.971 = 0.486

gain(Di) = 0.514 ✓

Choose Dirt because of the highest gain

round to 4 digits -0.5p

0.0 b)* Assume there is an additional data point with a missing value:

Temperature	Detergent	Dirt	Result
60	high	?	good

However, your colleague had access to all data (including the new observation) and calculated a **gain ratio** of 0.2326 for the attribute *Dirt*. What is the missing value marked by the question-mark?

~~gain ratio = gain~~
based on the table, I would suspect that dirt = medium

X

2.0 c)* You build a decision tree with pruning activated. Nevertheless, it is still large and difficult to interpret. What can you do to make it smaller?

you can do postpruning which tries to minimize the error you made at leafs

X postpruning already done

2.0 d)* What is the reason for pruning and how does the pruning method subtree replacement work?

you combat overfitting ✓ which is the danger of ~~memorizing~~ generalizing the data, and thus loss of generalizability. You estimate the error one makes at a leaf and compare it to the node ~~one up~~ that comes before. If the error is higher at the leaf than at the one before, you ✓ prune that

0 1 2 **0.0** e)* Generally, why is it not a good idea to discretize numerical attributes into many bins?

you loose the possibility of more splitpoints

Problem 6 Inference and Causality (20 credits)

The following scenario is based on real events, however numbers and other details are adjusted for academic purposes.

The Federal Reserve Act divided Mississippi between the 6th (Atlanta) and 8th (St. Louis) Districts. During the Great Depression, these districts' policies differed. Atlanta championed monetary activism and the extension of aid to support banks. St. Louis denied such an aid for banks. During the banking crisis that started at the end of 1930 and lasted for 2 years, Atlanta lent money to banks in need while St. Louis did not. In a study you want to address the following question: "Did providing liquidity (or credibly committing to do so) reduce rates of bank suspension and liquidation?"

3.0 a)* What type of experiment/study does the above described scenario allow for? Provide 2-3 arguments.

Natural experiment
- the banks have to belong to the region of aid, thus does not allow for random assignment

0 1 2 3

BANK SUSPENSIONS AND LIQUIDATIONS

Begin July 1		End June 30	Percentage of Banks Suspending: All (1)	Suspending: 6th Atlanta (2)	Suspending: 8th St. Louis (3)	Percentage of Banks Liquidating: All (4)	Liquidating: 6th Atlanta (5)	Liquidating: 8th St. Louis (6)
				Federal Reserve District			Federal Reserve District	
1929	to	1930	4.8	7.1	3.0	4.5	7.1	2.4
1930	to	1931	28.9	14.2	39.5	13.6	7.1	18.6
1931	to	1932	13.2	14.9	11.8	8.0	7.9	8.1
1932	to	1933	7.7	7.5	7.9	7.3	6.5	7.9
1933	to	1934	.9	.0	1.7	.9	.0	1.7
1929	to	1934[a]	49.8	38.7	59.2	30.9	26.8	34.4

SOURCE.—*Rand McNally Bankers Directory* and National Archives and Records Administration Record Group 82. See Section II and Richardson (2006, 2007*a*, 2007*b*, 2008) for details.

[a] The last row indicates the percentage of banks operating on July 1, 1929, that either suspended or liquidated by June 30, 1933.

0 1 2 3 4 5 6 7 8 9

6.0 b)* You are provided with the data in the table above.

1. What technique can you apply to identify causal effects in this setting? Explain the technique and argue briefly why you choose it (3-4 sentences). (6P)
2. Why do you need that technique and cannot simply make a causality statement based on the data of Atlanta only? (3P)

1. Differences in Differences: You want to account for the differences in years. We suggest the common trends assumptions. Also you can

fixed effects to differentiate between liquidating & suspending

2. we would get omitted variables bias to the error term. This would increase the error and would decrease overall significance

+ explanation of technique

0 1 2 3 4 5 6 7 8

5.0 c)*

Fictional

Assume you have data with multiple attributes for each bank in Atlanta in 1929. That year, some of the banks performed a program to potentially increase financial stability. You know which banks have performed the program and you know which banks failed in 1931.

Now, you want to make a statement of whether the program helped the banks to avoid suspension/liquidation.

1. Name one appropriate technique to help you identify causal effects in this setting. (2P)
2. Explain in 2-3 sentences why it is appropriate. (3P) +
3. Describe in 2-3 sentences how you would apply it. (3P)

1. I would use Propensity score Matching ✓
2. First, you have to identify the ~~one~~ banks that used the program, & match them, ~~using~~ then would you continue on; + Observational study design, no time related data
3. Including a log. regression to identify how & what the prog. Would match similar banks ✓

via propensity score. If I get satisfying matches, I would then run a model to identify the outcome

Page empty

Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.

Page empty