Question 1 − veep−1 − 169134.1.1

Veep is an American political satire comedy TV series that aired on HBO from 2012 to 2019. It ran for 7 seasons, with a total of 65 episodes. Veep has been praised as one of the funniest shows ever, with one of the highest number of jokes per minute (JPM). To see how funny Veep actually is, we obtain a dataset stored in the file `veep.csv` that contains information about the series and its individual episodes, such as the running time and number of jokes in each episode.

Assume we successfully loaded the dataset as a DataFrame named `veep`. The DataFrame contains many more rows than can be viewed on one page. For convenience, we display only the first and last 10 rows of the DataFrame below. Running times are measured in minutes.

	Season	Episode	Running time	Number of Jokes
0	All	NaN	1940.2	8227
1	1	NaN	244.8	1052
2	2	NaN	294.3	1264
3	3	NaN	296.3	1217
4	4	NaN	300.2	1277
5	5	NaN	297.1	1233
6	6	NaN	297.8	1289
7	7	NaN	209.7	895
8	1	1.0	30.5	111
9	1	2.0	29.9	126

	Season	Episode	Running time	Number of Jokes
63	6	8.0	29.3	115
64	6	9.0	29.5	143
65	6	10.0	31.1	141
66	7	1.0	30.4	118
67	7	2.0	28.2	120
68	7	3.0	30.4	120
69	7	4.0	29.6	133
70	7	5.0	29.3	134
71	7	6.0	30.7	124
72	7	7.0	31.1	146

Which of the following are true about the granularity and structure of the dataset? Select all that apply.

A the records are on the same level of granularity

B the dataset has missing values

C the dataset contains nested records (i.e. contains records that summarize information from other records in the dataset)

D the dataset stores rectangular data

Question 2 − veep−2 − 169135.1.1

(continues from Question 1)

We notice there are some `NaN` values in the "`Episode`" column. Which of the following is most appropriate for addressing these missing values? Select only one option.

A Impute them using the average of the rest of the values in the same column.

B Impute them using the mode of the rest of the values in the same column.

C Keep them as `NaN`.

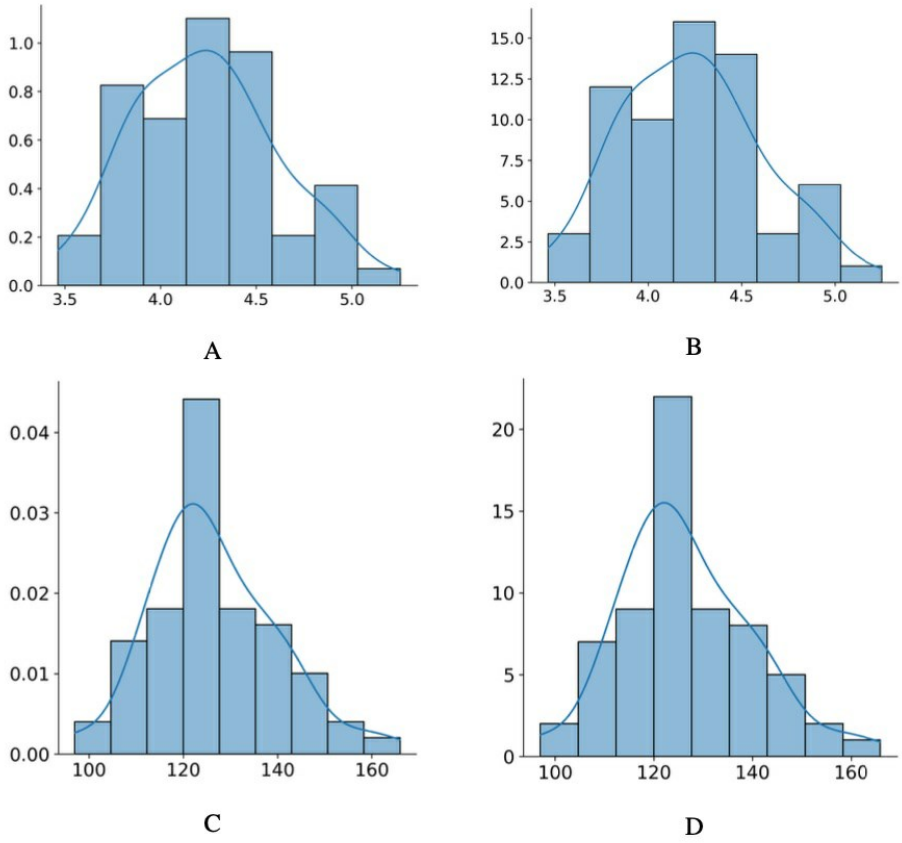
D Replace them with 0.

Question 3 − Veep−3 − 169136.2.0

(continues from Question 1)

Jokes per Minute (JPM) is defined as the number of jokes per minute of running time, i.e. dividing column "`Number of Jokes`" by "`Running time`". We want to plot the distribution of the "`JPM`" variable. The plots below show a few attempts at visualizing this distribution. Some of the plots visualize the incorrect variable, and some do not display valid kernel density estimates.

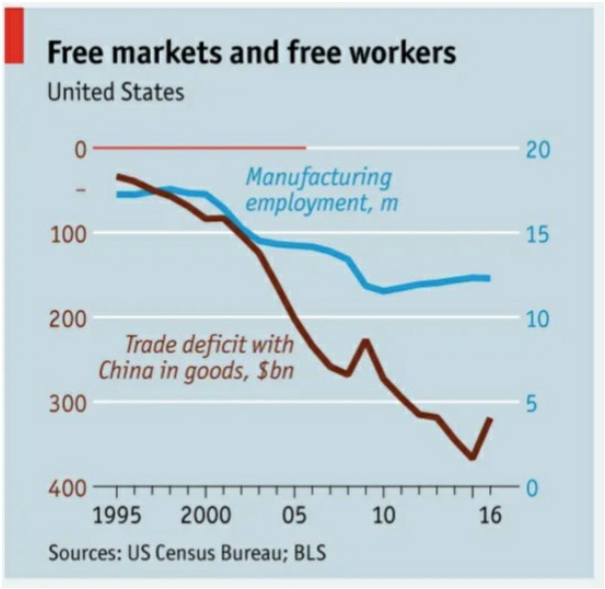
Based on the DataFrame shown in Question 1, which one of the following is most likely the correct plot?



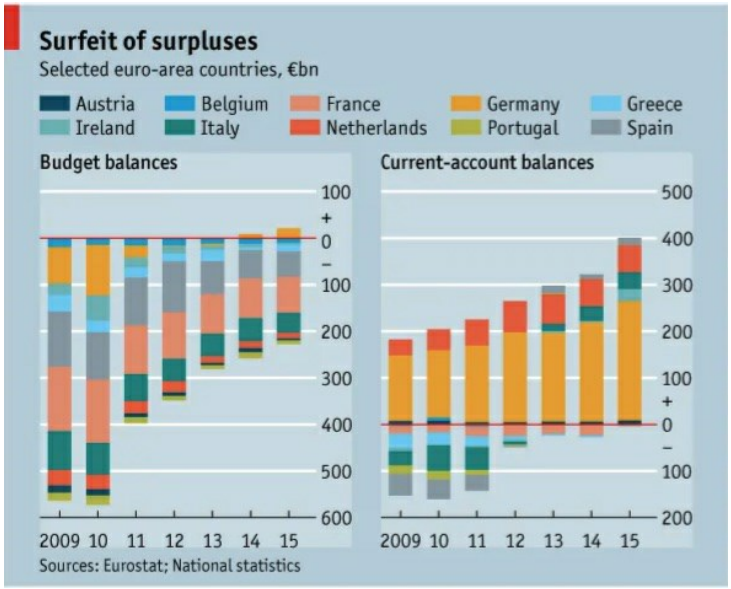
- **A** A
- **B** B
- **C** C
- **D** D

Question 4 – vis – 169137.1.1

Economist is one of the media publishers that take data visualisation seriously. Every week they publish around 40 charts across print, their website and apps. Sometimes they get it wrong. For each of the two plots below mention which design choices are (1) misleading and/or (2) confusing and/or (3) failing to make a point. Provide (in short) ways to improve the visualizations.



.



Grading instruction

figure 1 flaws (Number of points: 2)

figure 2 flaws (Number of points: 2)

left plot: one y-axis is flipped (0 is on top), x-axis is weirdly scaled, what is the red line?

right plot: too much clutter, too much color, no info to gain

check the actual blog of economist on these:

https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

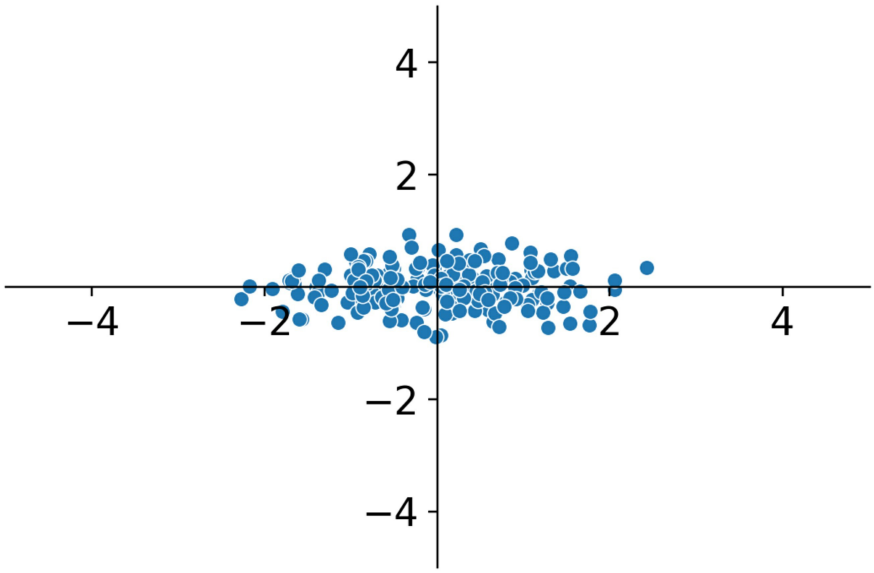
Question 5 – MC–Q2–PCAvsLasso – 169118.2.0

Both PCA and Lasso (L1) Regression can be used for feature selection. Which of the following statements are true? Select all that apply.

- **A** Lasso selects a subset of the original features
- **B** PCA and Lasso both allow you to explictly specify how many features are chosen
- **C** PCA produces features that are linear combinations of the original features
- **D** PCA selects a subset of the original features

Question 6 − Q6−PCA−choose−matrix − 169125.2.1

Consider a dataset containing two features, x_1 and x_2. Below we see the plot that visualizes the data (on the x axis is x_1 and on the y axis is x_2). Based on this plot and if we were to apply SVD on the data, which of the following matrices is most likely to represent the Σ matrix (aka the matrix of singular values)?



A $\begin{vmatrix} 0.5 & 0.45 \\ 0.45 & 0.75 \end{vmatrix}$

B $\begin{vmatrix} 0.75 & 0.45 \\ 0.45 & 0.5 \end{vmatrix}$

C $\begin{vmatrix} 10 & 0 \\ 0 & 8 \end{vmatrix}$

D $\begin{vmatrix} 13 & 0 \\ 0 & 5 \end{vmatrix}$

Question 7 − Q6−PCA−equivalent − 169126.1.1

In the context of Singular Value Decomposition (SVD), which of the following expressions is equivalent to $U\Sigma$?

A ΣU

B XV

C XV^T

D VX

Question 8 − Q6−PCA−match − 169123.2.2

We are still talking about PCA. Identify the words from the left column that correspond to those in the second column. Drag the relevant word(s) from the right column to the word(s) in the left column that you believe there is a match."

Variance	Maximization
Reconstruction Error	Minimization
SVD	$A = U\Sigma V^T$

Question 9 − Q6−PCA−true−false − 169124.2.1

Given a set of n-points in a `d`-dimensional space, using PCA to reduce the dataset to `k < d` dimensions will always lead to loss of information. We know that the rank of `n`-by-`d` data matrix is `d`.

A True

B False

Question 10 − xregr1 − 169141.1.1

The ACT (American College Testing) is a standardized test used for college admissions in the United States. ACT scores are used by colleges and universities as one of several factors in the admissions process. Higher scores generally increase a student's chances of being admitted to their desired colleges and may also be used to qualify for scholarships. A high school counselor was interested in finding out how well student grade point averages (GPA) predict ACT scores. The next 3 questions refer to the same problem.

For this purpose he decides to collect data from school past students. He gets a list of telephones and decides to call every 100th student on the list and ask if they are interested in participating in the research study. What are some forms of error that we may encounter with our sampling strategy from above? Select all that apply.

A chance error

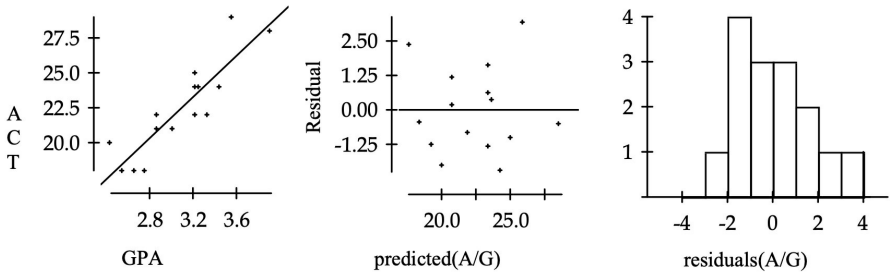
B selection bias

C non-response bias

Question 11 – xregr2 – 169142.2.1

(continues from previous question)

After collecting the sample, the counselor runs the regression model that predicts ACT from GPA. Below we see 3 plots of this model: (a) the scatterplot between GPA and ACT, (b) the plot of the residuals against the predicted value of the model for ACT and (c) a histogram of the residuals. We also show the results of the regression model below the plots.



Dependent variable is: **ACT**
No Selector
R squared = 78.1% R squared (adjusted) = 76.4%
s = 1.630 with 15 - 2 = 13 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	123.041	1	123.041	46.3
Residual	34.5589	13	2.65838	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-0.427035	3.382	-0.126	0.9014
GPA	7.39697	1.087	6.80	≤ 0.0001

Comment on how good is the fit of the model by referring to as much information above as possible.

Grading instruction

statistical significance (Number of points: 1)

straight line (Number of points: 1)

residual plot (Number of points: 1)

1) Accoding to the first plot, the relationship between ACT and GPA is approximatly linear,
so higher GPA, higher ACT.
2) Residuals (plots 2 and 3) appear to be small and approx. normally distributed. No real pattern
3) The co-efficient seems to be statistically significant ($p<0.01$) and R-squared is 78.1%

Question 12 – xregr3 – 169143.3.0

(continues from previous question)

Interpret the coefficient and the intercept of the model in the context of the problem. Comment on how useful/practical they are to explain the problem to people that do not understand regression specifics.

Grading instruction

coefficient (Number of points: 1)

intercept (Number of points: 1)

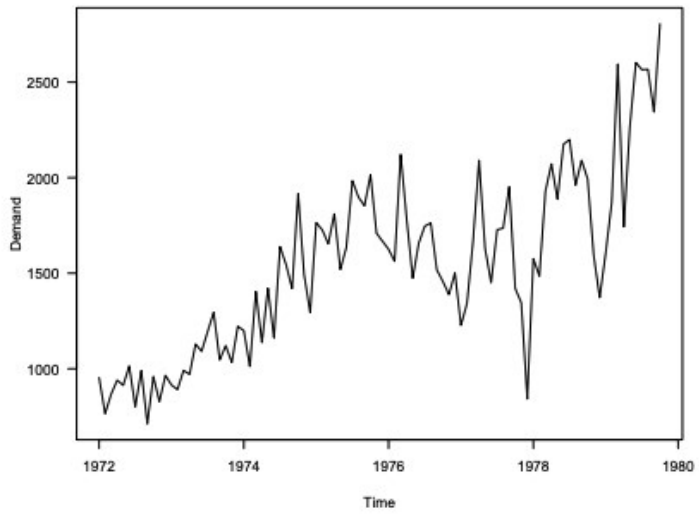
Coefficient for GPA 7.4 = For one point increase on GPA, that leads to 7.4 points increase in ACT scores. Useful to explain the relationship between the two, esp. since they are measured in different units/scales.

Constant -0.42 = ACT score if your GPA is 0. Not useful since we do not have data on that range. Also, a negative ACT does not make sense.

Question 13 – time1 – 169138.1.0

The question concerns a time series data set giving the monthly demand for repair parts large/heavy equipment in Iowa during the period 1972–1979.

A graph of the series is shown below.



This graph indicates that it is appropriate to work with the logarithms of the original values. Explain why this is and indicate which competing transformations (if any) could have been tried.

Grading instruction

Criterion 1 (Number of points: 1)

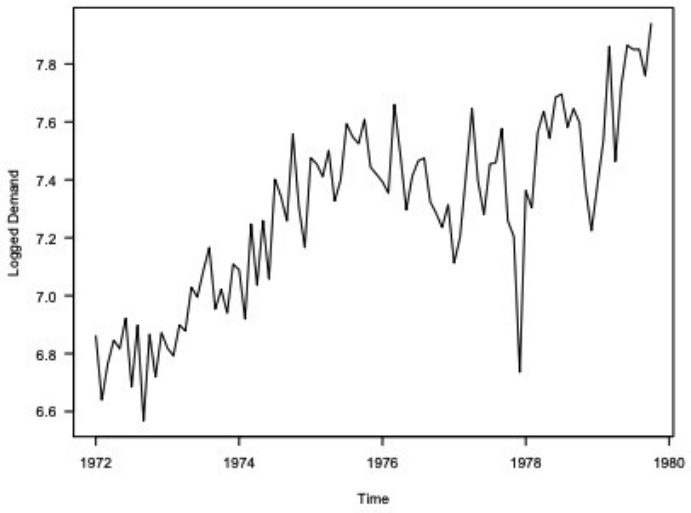
we see non-stationarity and a slight increase in the trend. log-scale will fix some of these issues

other transformations: box-cox, sqrt (probably), other answers were taken correct provided you justified them on the basis of the plot

Question 14 – time2 – 169139.1.1

(continues from previous question)

A plot of logged values is shown below. On the basis of this plot it was decided to work with a differenced (first-order) version of the series. Explain why this was.



Grading instruction

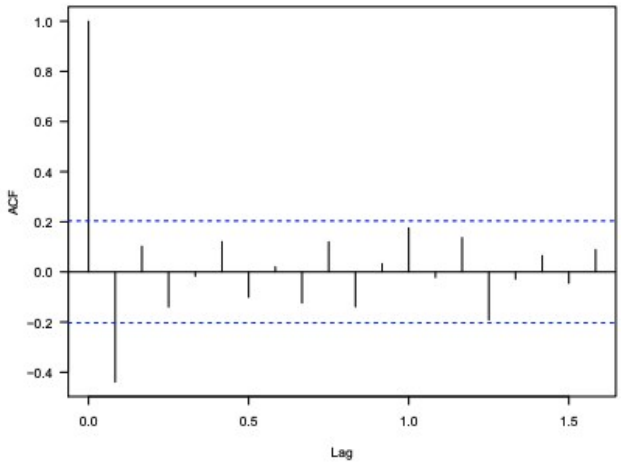
Criterion 1 (Number of points: 1)

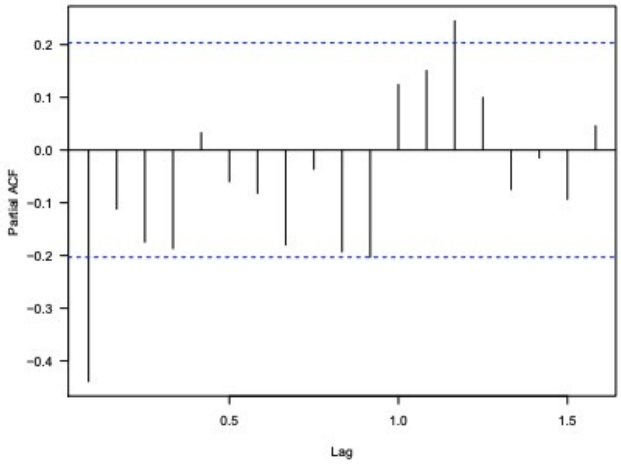
even with the logged values, we see the trend changing, therefore taking first-order differences will smoothen this

Question 15 − time3 − 169140.1.1

(continues from previous question)

The following plots show the ACF and PACF for the differenced version of the logged series. On the basis of these plots, choose a model to be fitted to the logged data values. Explain why you think the model you have chosen is appropriate.





AR(p) and MA(q) can be identified by observing striking spikes in the graphs above.
AR(1) would be a good model given the spike at lag 1 in the acf plot.

However, the plot doesn t die off after any lag, so we cannot call for a MA(q) model - it seems to be periodical in the pacf plot from where we could get our q value, there is no spike followed by decrease, but we can rather observe a pattern

Spikes at ACF mean, seasonal patterns

Grading instruction

Criterion 1 (Number of points: 2)

Question 16 – LR–1 – 169127.2.1

Consider the following binary classification problem with 4 data points. Suppose that we use a logistic regression model to predict the probability that y = 1 given x1 and x2:

x1	x2	y
0	1	1
1	0	0
-1	0	1
0	-1	0

After fitting a logistic regression model (**without** regularization) on features x1 and x2 (**without** an intercept) and class labels y, we find that the predicted probabilities for the 4 data points returned by `sklearn` are given below.

```
array([[2.69411751e-05, 9.99973059e-01],
       [9.99973059e-01, 2.69411751e-05],
       [2.69411751e-05, 9.99973059e-01],
       [9.99973059e-01, 2.69411751e-05]])
```

Is the data linearly separable? Given your answer to this question, what would be the values of the parameters of the logistic regression model (say b1 and b2 for x1 and x2 respectively)? Justify your answer.

Hint: Think about the values that cross-entropy loss will take.

Grading instruction

Criterion 1 (Number of points: 1)

linearly separable yes

Criterion 2 (Number of points: 2)

argumentation correct

Criterion 3 (Number of points: 2)

details about the loss etc.

A dataset is linearly separable if a line can be drawn in terms of the features xi that perfectly separates the two classes y. A sketch of the data shows that this dataset is linearly separable.

Tricky answer: When a dataset is linearly separable, the weights of an unregularized logistic regression model diverge to infinity.

Think about this way:

when x1=0, x2=1, then y^ should approach 0.

that happens when y^ = 1 / (1+e^-b1) and that gives us b1 —> - infinity

Similarly, we find that b2 —> + infinity

Question 17 – LR–2 – 169128.1.2

(continues from previous question)

We train a new logistic regression model (without an intercept) with regularization and find the optimal model parameters to be `b1=-2/3` and `b2=2/3`.

Suppose that we observe a new data point `xnew = [2,1]` (i.e. x1=2 and x1=1). What is the probability that our model believes `xnew` belongs to class 0?

Note: You may leave your final answer as an expression in terms of `e`.

Grading instruction

Criterion 1 (Number of points: 1)

correct formula

Criterion 2 (Number of points: 1)

result

P(y^=1|xnew) = σ ([2,1] x [-2/3,2/3]) = σ [-2/3] = 1 / (1+e^(2/3))

P(y^=0|xnew) = 1 - 1/(1+e^(2/3))

Question 18 – LR–3 – 169129.1.0

(continues from previous question but can be answered independently)

We decide to use a new dataset of 6 training points. For the remainder of this question, we will only consider these 6 new training points. We train a new logistic regression model to find the model's predicted probabilities, displayed in the table below.

Data point #	x1	x2	y	predicted probability
1	1	1	1	0.65
2	0.5	2.5	1	0.90
3	0.75	0.5	1	0.75
4	0	0	1	0.85
5	0.5	1.5	0	0.70
6	1	0	0	0.45

Which data point incurs the highest cross-entropy loss?

Hint: You do not need to actually compute the cross-entropy loss

A Point 1

B Point 2

C Point 3

D Point 4

E Point 5

F Point 6

The prediction $\hat{y}=0.70$ for Point5 is furthest from the corresponding observed value $y = 0$.

Question 19 – LR–4 – 169130.1.0

(continues from previous question)

What are the range(s) of classification thresholds T that maximize(s) accuracy? Select all that apply. In the following question, justify your answer shortly.

A [0, 0.45]

B (0.65, 0.70]

C (0.75,0.85]

D (0.45, 0.65]

E (0.70, 0.75]

F (0.85,0.90]

Any threshold between .45 and .65 results in 4 true positives and 1 true negative.

Any threshold between .70 and .75 results in 3 true positives and 2 true negatives.

In either case, the accuracy of the model is TP +TN = (TP+TN+FP+FN) = 5/ 6

Question 20 – LR–45 – 169131.1.0

Justify shortly your answer to the previous question

Grading instruction

Criterion 1 (Number of points: 2)

Question 21 – LR–6 – 169132.2.0

(continues from previous question but can be answered independently)

Suppose we are interested in evaluating our model's performance with an ROC curve. To construct an ROC curve, we first need to find our model's TPR and FPR at various classification thresholds T. The TPR and FPR for the majority of thresholds have been computed for you in the table below; please select the correct values for the remaining entries (A,B,C,D) and shortly justify your answer.

T	TPR	FPR
0.95	0	0
0.9	1/4	0
0.85	1/2	0
0.75	A	B
0.7	C	D
0.65	1	0.5
0.45	1	1

For convenience, we copy the data table from the previous question below.

Data point #	x1	x2	y	predicted probability
1	1	1	1	0.65
2	0.5	2.5	1	0.90
3	0.75	0.5	1	0.75
4	0	0	1	0.85
5	0.5	1.5	0	0.70
6	1	0	0	0.45

Grading instruction

A (Number of points: 1)

B (Number of points: 1)

C (Number of points: 1)

D (Number of points: 1)

When the threshold is 0.75, there are 3 true positives, 2 true negatives, and 1 false negative.
The resulting metrics are TPR = 3/(3+1)=3/4 and FPR = 0/(0+2)=0

When the threshold is 0.7, there are 3 true positives, 1 true negative, 1 false positive, and 1 false negative.
The resulting metrics are TPR= 3/(3+1)=3/4 and FPR= 1/(1+1)=1/2 = 0.5

Hence,

A=3/4, B=0, C=3/4, D=1/2

Question 22 – MC–Q1–DT – 169117.1.0

What strategies can help reduce overfitting in decision trees? Select all that apply.

- **A** pruning
- **B** enforce a minimum number of samples in leaf nodes
- **C** make sure each leaf node is one pure class
- **D** enforce a maximum depth for the tree

Question 23 – MC–Q3–overfit – 169119.1.1

You train a classifier on 10,000 training points and obtain a training accuracy of 99%. However, when you check the accuracy on the test set that is only 67%. Which of the following, done in isolation, has a good chance of improving your performance on the test set? Select all that apply.

- **A** set your regularization parameter (λ) to 0
- **B** use validation to tune the hyperparameters
- **C** train on more diverse data
- **D** train on less diverse data

Question 24 – MC–Q4–Ridge – 169120.1.2

Which of the following is NOT a potential benefit of using Ridge (L2) regression?

- **A** It can reduce the variance of the model.
- **B** It can improve the interpretability of the model.
- **C** It can help to reduce overfitting.

Question 25 – bagboost – 169145.1.1

For the following scenarios, argue whether bagging **or** boosting would be your preferred choice and justify your choice shortly.

A. Noisy data

B. A very big dataset

C. Inbalanced data

Grading instruction

Noisy data (Number of points: 1)

Big data (Number of points: 1)

Inbalanced data (Number of points: 1)

Noisy data: Bagging would be preferred since with the multiple splits of the data we can avoid too much noise. Boosting could amplify this noise.

Very big dataset: Bagging due to the ability to run things in parallel. With boosting we have to use the whole dataset

Inbalanced data: Boosting will do a good job - it can focus on the minority samples compared to bagging