Please upload to 2VU under the Problem Set 4 assignment.

Due: November 9th for section 61 and November 10th for section 60

*Conceptual*

1. **Tell me what a classifier is and when one would use a classifier (like logistic regression) as opposed to a model that generates point estimates (like linear regression).** *(5 pts)*
2. **Give me at least one reason why proportions (or probabilities) are potentially misleading when using conditional means/binning to predict a discrete outcome.** *(2 pts)*
3. **Why is it important to know how a data team discretizes (bins) their continuous predictors?** *(8 pts; Full points will articulate what the pros/cons are of the three methods discussed in the async)*
4. **Pick sensitivity or specificity and describe a situation/goal where you would prioritize it over the other.** *(5 pts)*

*Scenario*

A bank has a growing concern regarding folks defaulting on their loans. They contract out to your organization, asking you to identify what makes someone a good or bad loanee. You are provided with a dataset of 700 customers that have recently been approved of loans and whether they defaulted or not (defaulted cases are 1 and cases that did not default are 0).

1. **What is the goal of this model?** *(2 pts)*

Your data team suggests running a logistic regression with a 90% training/10% testing split-sample cross validation (standard industry split).

1. **Does a logistic regression analysis make sense here? Why?** *(3 pts)*

Your data team ran their analysis and show you the following results table:

![Table

Description automatically generated]()

1. **Which predictors (if any) are risk factors (they increase the odds of defaulting)? How do you know?** *(3 pts)*
2. **Which predictors (if any) are protective factors (they decrease the odds of defaulting)? How do you know?** *(3 pts)*

You turn the page of the team’s report and see these:

|  |  |
| --- | --- |
| **Metric** | **Estimate** |
| Accuracy | 0.84 |
| Sensitivity | 0.57 |
| Specificity | 0.94 |
| ROC-AUC | 0.87 |

Chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | ***Defaulted?*** | |
| **Prediction** | ***No*** | ***Yes*** |
| **No** | 0.69 | 0.11 |
| **Yes** | 0.04 | 0.15 |

1. **How well did this model perform overall? Is there anything you are unsatisfied with (how are the false positive and false negative rates)?** *(10 pts)*
2. **Did your team “accomplish” its goal? Would you use this model for its designed purpose? Why/Why not?** *(4 pts)*
3. **Bonus: Explain what threshold is in a classification task (like logistic regression) and what changing it does to model performance.** *(5 pts)*

**I would also greatly appreciate your feedback on:**

**How long this took to complete**

This took me X minutes to complete.

**How much you felt it complimented the async and sync**

This did/did not compliment the async and sync sessions. I felt it was facilitating/hindering my understanding of the concepts.

**How comparable it feels to assignments in other classes**

This feels more/equally/less challenging than assignments I have in my other classes.