**MISY467: Machine Learning for Business**

**Final Take-Home Exam**

**Question 1 (20 pts)**: Given the following confusion matrix, answer the subsequent questions.

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
|  | + - | |
| +  Actual  - | 4 | 2 |
| 1 | 3 |

1) What is the value of accuracy? What are the values of precision, recall, for both + and – class? What are the values of FPR (false positive rate), FNR (false negative rate) and TNR (true negative rate)?

Accuracy:

(4+3) / (4+2+1+3)=0.7

Precision:

+ class:

4/(4+1)=0.8

- class:(In the internet say this is called [Negative predictive value](https://en.wikipedia.org/wiki/Negative_predictive_value) (NPV))

3/(3+2)=0.6

Recall:

+:

4/(4+2)=2/3

-::(In the internet say this is called[True negative rate](https://en.wikipedia.org/wiki/True_negative_rate) (TNR))

3/(3+1)=0.75

FPR:

1/(3+1)=0.25

FNR:

2/(2+4)=1/3

-

2) If the cost of false positive prediction = $1 and cost of false negative prediction = $5, what is the total cost? How to adjust decision threshold to lower the total cost in this case?

Cost of Model

1+2\*5=11 Dollars

I would adjust the treshold so that its less likely that my model produces False negatives, because they are 5 times as expensive as False Positive so I would adjust my threshold that it doesn't produces as much false negatives before and therefor than is more likely to produces False Positives, but I would still save money because False Positives are cheaper.

3) Briefly describe a business application in which different costs are associated with different types of prediction error.

Imagine I do a test if I do a test if something is safe to eat for humans , if I test its safe to eat but in real its not safe to eat an the human might die from it on the other hand if I predict its not save to eat but its actually is, not much harm is done maybe I loose some money but I don't loose lives(and than get sued for it). So one the one hand the cost might be lives on the other hand the cost are financially.

**Question 2** (**20 pts)** Some basketball experts argue that one can predict if an rookie NBA player will become an all-star by just looking at two features: a) if he is among top 5 rookie selection, b) if the average score per game during the rookie year is above 15. To examine this interesting statement, you are going to build a decision tree about it. All three variables are binary and noted as **A** (target variable all-star, yes or no), **T5** (top 5 selection, yes or no) and **S** (Score >15, yes or no). The following table shows a data set that contains 10 data points.

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **T5** | **S** | **A** |
| 1 | Yes | Yes | Yes |
| 2 | Yes | No | Yes |
| 3 | No | No | Yes |
| 4 | Yes | Yes | Yes |
| 5 | Yes | No | No |
| 6 | No | Yes | Yes |
| 7 | No | No | No |
| 8 | No | Yes | No |
| 9 | Yes | Yes | Yes |
| 10 | No | Yes | No |

You need to build the decision tree manually (not using code) using Entropy as the criterion and show which feature you should use to split the root node and why. You have to show the details of your calculations.

Entropy of T5

P(No of T5)\*Entropy(2,3)+P(Yes of T5)\*Entropy(4,1)

Entropy(2,3)=0.97

Entropy(4,1)=0.722

5/10\*0.97+5/10\*0.722=0.846

Entropy of S

P(No of S)\*Entropy(2,2)+P(Yes of S)\*Entropy(4,2)

Entropy(4,2)=0.918

4/10\*1+6/10\*0.918=0.9508

So I have more information gain if I split first with T5(top 5 selection, yes or no), becaue the entropy goes more down when I split at t5 than with S

**Question 3** (**60 pts)**

In this question, you are given a dataset about the patients’ medical costs charged by insurance company, which has the following columns:

* age: age of primary beneficiary
* sex: insurance contractor gender, female, male
* bmi: body mass index (https://en.wikipedia.org/wiki/Body\_mass\_index)
* children: number of children covered by health insurance
* smoker: smoking or not
* region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
* charges: individual medical costs billed by health insurance

Your task is to build a model to predict the medical charges.

Requirements:

* you need to use at least three different models we covered in lectures and compare their performance using cross validation
* you need to fine tune at least two models for the best parameters (you choose what and how many parameters to tune)
* you need to use RMSE as the evaluation metric and explain your RMSE in this context (what does the RMSE value mean – is your final model good or bad?)