IEOR 142A: Machine Learning and Data Analytics I, Fall 2024

Homework Assignment #2

Due October 9th, 2024 at 11:59pm Pacific time

IMPORTANT NOTE: As indicated in the slides of Lecture 1 (pgs. 36-38) and the syllabus (pgs. 6-8), please list any resources outside of the course materials that you find helpful in completing the assignment (e.g. peers you discuss with, materials from different classes, blog posts, AI Tools, etc.). Please also be mindful of all policies in the syllabus concerning academic integrity and the use of AI Tools, including that you need to write your own solutions individually.

Problem 1: (10 points) Let Y be a discrete random variable that takes values in the set $\{-2, -1, 1, 2\}$. Let θ, τ be two parameters such that $0 \leq \theta \leq 1$, $0 \leq \tau \leq 1$, and the probability distribution of Y is given by the following equations:

$$\begin{aligned} \mathbb{P}(Y = -2) &= (1-\theta)(1-\tau) \\ \mathbb{P}(Y = -1) &= \theta(1-\tau) \\ \mathbb{P}(Y = 1) &= \theta\tau \\ \mathbb{P}(Y = 2) &= (1-\theta)\tau \end{aligned}$$

(Although it is not required for this problem, you can check that the above equations imply that $\mathbb{P}(|Y| = 1) = \theta$ and $\mathbb{P}(\text{sign}(Y) = 1) = \tau$ where $|\cdot|$ is the absolute value and $\text{sign}(\cdot)$ is the sign function.)

The parameters θ, τ are unknown, but suppose that we are given the following dataset of five observed values that are assumed to be independent and identically distributed (i.i.d.) from the above distribution (given some values of the parameters):

$$y_1 = -1, y_2 = -1, y_3 = 1, y_4, = 1, y_5 = 2$$

Please answer the following:

a) (5 points) Given the above dataset $y_1, \ldots, y_5$, construct the likelihood function $L(\theta, \tau)$ and log-likelihood function $l(\theta, \tau) = \log(L(\theta, \tau))$.

b) (5 points) Find the maximum likelihood estimators (MLE) of θ and τ by maximizing the log-likelihood function $l(\theta, \tau)$.

a) $L(\theta, \tau)$ = product of probabilities

↳ $= P(Y = y_1) \cdot P(Y = y_2) \cdot P(Y = y_3) \cdot P(Y = y_4) \cdot P(Y = y_5)$

replace with values given $\Rightarrow P(Y = -1) \cdot P(Y = -1) \cdot P(Y = 1) \cdot P(Y = 1) \cdot P(Y = 2)$

replace with probabilities given $\Rightarrow = \theta(1-\tau) \cdot \theta(1-\tau) \cdot \theta\tau \cdot \theta\tau \cdot (1-\theta)\tau$

$$= [\theta(1-\tau)]^2 \cdot (\theta\tau)^2 \cdot (1-\theta)\tau$$

$$L(\theta,\tau) = \theta^4\tau^3(1-\tau)^2(1-\theta)$$

↓ log-likelihood $\log(L(\theta,\tau))$

$$\Rightarrow \quad \ell(\theta,\tau) = \log\left(\theta^4\tau^3(1-\tau)^2(1-\theta)\right)$$

expand using properties of log

$$\boxed{\ell(\theta,\tau) = 4\log(\theta) + 3\log(\tau) + 2\log(1-\tau) + \log(1-\theta)}$$

b) max log-likelihood

1) find partial derivatives with respect to θ and τ

$$\frac{\partial}{\partial\theta}(\ell(\theta,\tau)) = \frac{4}{\theta} - \frac{1}{1-\theta}$$

$$\frac{\partial}{\partial\tau}(\ell(\theta,\tau)) = \frac{3}{\tau} + \frac{2}{1-\tau}$$

2) save for θ and τ by setting derivative $= 0$

$$\frac{4}{\theta} - \frac{1}{1-\theta} = 0$$

$$\frac{4}{\theta} = \frac{1}{1-\theta}$$

$$4(1-\theta) = \theta$$

$$4 - 4\theta = \theta \rightarrow 4 = 5\theta \Rightarrow \boxed{\hat{\theta} = \frac{4}{5}}$$

$$\frac{3}{\tau} - \frac{2}{1-\tau} = 0$$

$$\frac{3}{\tau} = \frac{2}{1-\tau}$$

$$3(1-\tau) = 2\tau$$

$$3 - 3\tau = 2\tau$$

$$3 = 5\tau \Rightarrow \boxed{\hat{\tau} = \frac{3}{5}}$$

IEOR 142A: Machine Learning and Data Analytics I

HW Assignment 2

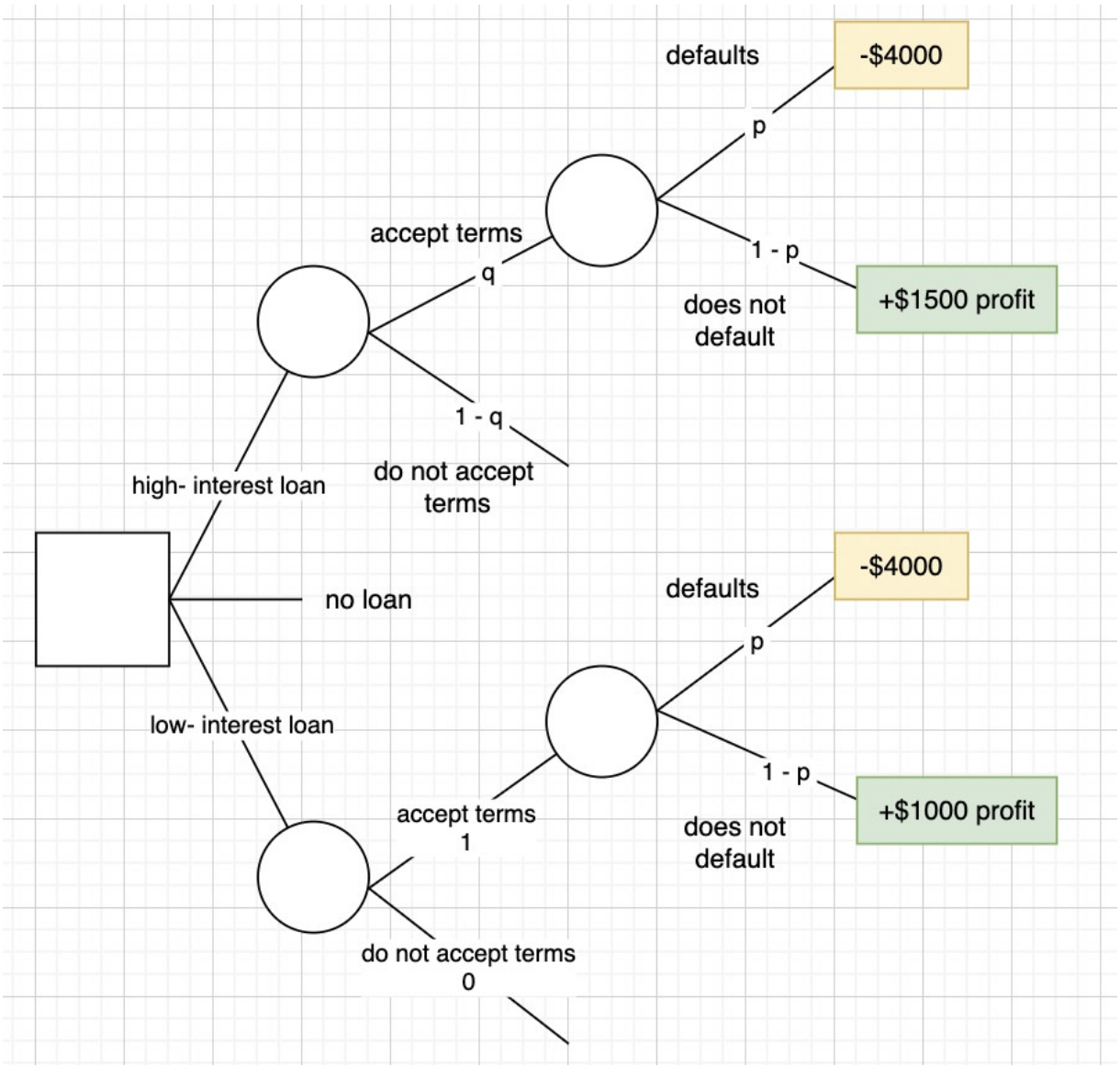
Problem 1: (10 points) Let Y be a discrete random variable that takes values in the set {−2, −1, 1, 2}.

work written on attached pdf above

Problem 2: (20 points) Let us consider an extension of the lending decision problem from class.

```
from IPython.display import display, Latex
```

a) (5 points) Create a decision tree diagram to model the previously described scenario. Use squares to denote decision nodes and circles to denote chance nodes representing random events. Each terminal node of the tree should have a corresponding profit value.



b) (5 points) Derive formulas for the expected profit under each of the three possible decisions for the lender: (i) fund with low interest, (ii) fund with high interest, and (iii) do not fund. Your formulas should depend on the probabilities p and q.

E[Profit | Low Interest]

$$E[\text{Profit}] = -4000\,p + 1000(1-p)$$

$$-5000\,p + 1000$$

E[Profit | High Interest]

$$E[\text{Profit}] = (q\,p(-4000)) + (q(1-p)1500) + ((1-q)(0))$$

$$(-5500\,p + 1500)q$$

$$-5500\,pq+1500\,q$$

E[Profit | Didn't Fund]

$$E[\text{Profit}]=0$$

c) (5 points) Suppose that q = 1/2. Segment the range of possible values of p, i.e., the interval [0, 1] into three subintervals corresponding to ranges of values where each of the three options are optimal decisions in order to maximize expected profit. Create a graph to visually display your answer.

$$\text{Formulas after knowing q} = \frac{1}{2}:$$

$$E[\text{Profit}]=-4000\,p+1000(1-p)$$

$$-5000\,p+1000$$

$$E[\text{Profit}]=(q\,p(-4000))+(q(1-p)1500)+((1-q)(0))$$

$$(-5500\,p+1500)\frac{1}{2}$$

$$-5500\,p\left(\frac{1}{2}\right)+1500\left(\frac{1}{2}\right)$$

$$-2750\,p+750$$

$$E[\text{Profit}]=0$$

Best course of action:

Making comparisons help segment the probability space into ranges where each option is the most profitable, allowing the lender to make data-driven decisions based on the borrower's risk profile.

1) Comparing low-interest vs high-interest

$$-5000\,p+1000>-2750\,p+750$$

$$-5000\,p+2750\,p>-1000+750$$

$$-2250\,p>-250$$

$$p<\frac{-250}{-2250}=\frac{1}{9}\approx 0.111$$

So, if p < 0.111, the low-interest option is better than the high-interest option.

2) Comparing high-interest vs no loaning

$$-2750\,p+750>0$$

$$p<\frac{-750}{2750}=\frac{3}{11}\approx 0.273$$

So, if p < 0.273, the high-interest option is better than not funding the loan.

3) Comparing low-interest vs no loaning

$$-5000\,p+1000>0$$

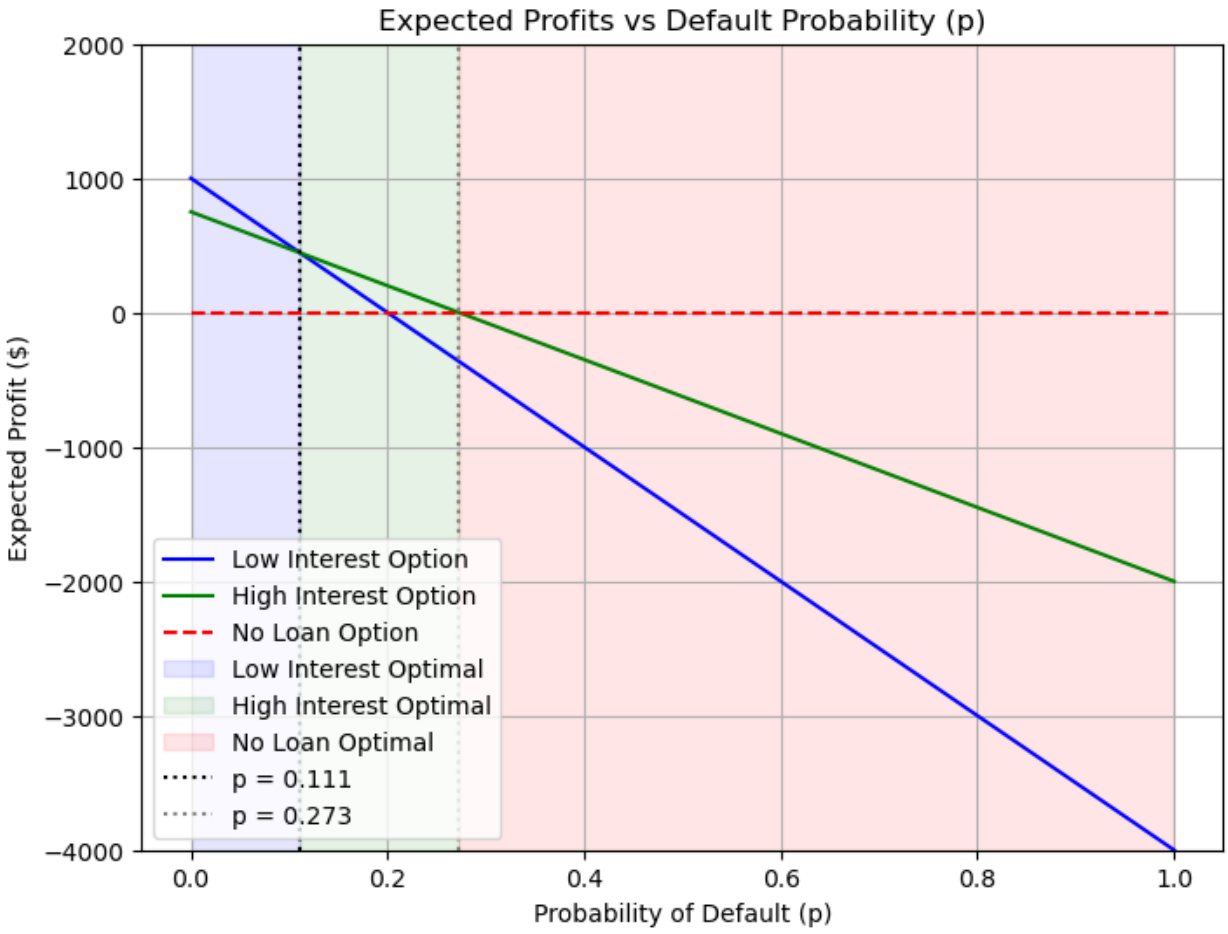
$$p<\frac{-1000}{5000}=\frac{-1}{5}\approx 0.2$$

So, if p < 0.2, the low-interest option is better than not funding the loan.

p=0.111 is the key threshold where the low-interest option beats the high-interest option. Although p=0.2 was calculated, it doesn't factor in the optimal decision when all three options are considered because the high-interest option is still better than no loan up to p=0.273.

```
import numpy as np
import matplotlib.pyplot as plt

# Define the probability range
p = np.linspace(0, 1, 500)

# Expected profits for each option
low_interest_profit = 1000 - 5000 * p
high_interest_profit = 750 - 2750 * p
no_loan_profit = np.zeros_like(p)

# Define shading areas based on the inequalities
p_shade_low_interest = np.linspace(0, 0.111, 100)
p_shade_high_interest = np.linspace(0.111, 0.273, 100)
p_shade_no_loan = np.linspace(0.273, 1, 100)

# Create the plot
plt.figure(figsize=(8, 6))

# Plot the expected profits
plt.plot(p, low_interest_profit, label='Low Interest Option',
color='blue')
plt.plot(p, high_interest_profit, label='High Interest Option',
color='green')
plt.plot(p, no_loan_profit, label='No Loan Option', color='red',
linestyle='--')

# Add shading for the inequalities
plt.fill_between(p_shade_low_interest, -4000, 2000, color='blue',
alpha=0.1, label='Low Interest Optimal')
plt.fill_between(p_shade_high_interest, -4000, 2000, color='green',
alpha=0.1, label='High Interest Optimal')
```

```
plt.fill_between(p_shade_no_loan, -4000, 2000, color='red', alpha=0.1,
label='No Loan Optimal')

# Highlight thresholds
plt.axvline(x=0.111, color='black', linestyle=':', label='p = 0.111')
plt.axvline(x=0.273, color='gray', linestyle=':', label='p = 0.273')

# Labels and legend
plt.title('Expected Profits vs Default Probability (p)')
plt.xlabel('Probability of Default (p)')
plt.ylabel('Expected Profit ($)')
plt.ylim(-4000, 2000)
plt.legend(loc='best')

# Show the plot with shading
plt.grid(True)
plt.show()
```



d) (5 points) Briefly discuss how one might estimate the probabilities p and q in practice, in a personalized way depending on features associated with the borrower. Your discussion should include what type of dataset(s) would need to be collected and what model(s) you would fit.

In regards to dataset collection, it would be very insightful to have historical past data on the background of loan applicants and whether or not they defaulted or not. Some useful information include the borrower demographics (income, employment history, education level, credit FISCO score, etc), behavorial analysis (previous loan defauts, payment history), loan data (loaned amount, motive, terms, interest rate at the time), and even macro-level factors (inflation, unemployment rates).

When looking into classification models, logistical regression and decision trees are good ways to approach this problem. For logistic regression, it is very interpretable and ideal for binary outcomes like our case. Decision trees are useful in handling non-linear relationships and can segment data into different groups based on feature differences very clearly.

Problem 3: Framingham Heart Study

In this exercise, you are asked to build models using Framingham Heart Study data in order to predict CHD and to make recommendations to better prevent heart disease. There are 3,658 total observations in our data, with each observation representing the data from a particular study participant. There are 16 variables in the dataset, which are described in Table 1. You will be asked to predict TenYearCHD (whether the patient experiences coronary heart disease within 10 years of their first examination). As a consequence of your modeling efforts, you should be able to identify risk factors, which are the variables that increase the risk of CHD

Using all of the provided independent variables, build a logistic regression model to predict the probability that a patient will experience CHD within the next 10 years. Use dataset framingham train.csv to train your model. This training set has 2560 data points, which are randomly selected from the original framingham.csv dataset (around 70%). Use dataset framingham test.csv to test your model. This test set has the remaining 1098 data points.

```
from pandas.plotting import scatter_matrix
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import math
import random
import statsmodels.api as sm

#load csv data
f_test = pd.read_csv("framingham_test.csv")
f_train = pd.read_csv("framingham_train.csv")

f_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2560 entries, 0 to 2559
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   male             2560 non-null   int64
 1   age              2560 non-null   int64
 2   education        2560 non-null   object
```

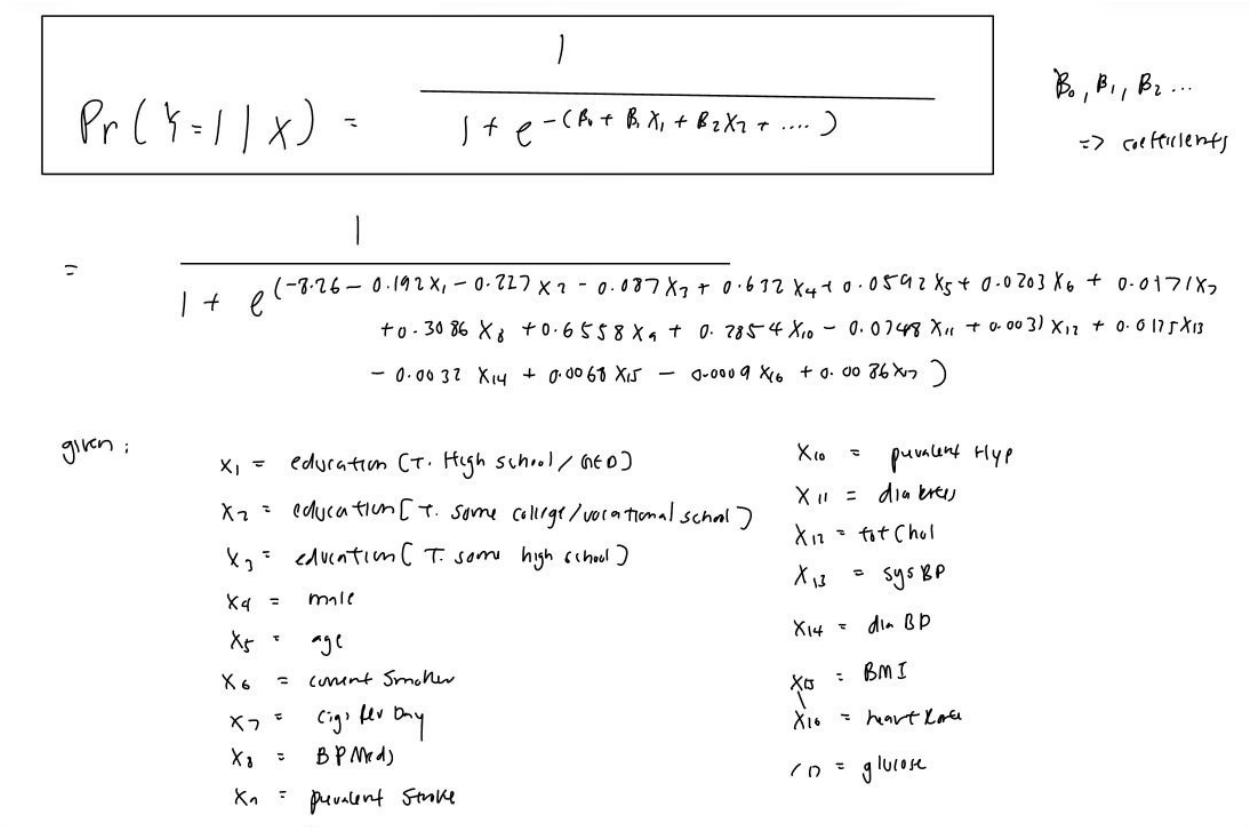
```
 3   currentSmoker    2560 non-null   int64
 4   cigsPerDay       2560 non-null   int64
 5   BPMeds           2560 non-null   int64
 6   prevalentStroke  2560 non-null   int64
 7   prevalentHyp     2560 non-null   int64
 8   diabetes         2560 non-null   int64
 9   totChol          2560 non-null   int64
 10  sysBP            2560 non-null   float64
 11  diaBP            2560 non-null   float64
 12  BMI              2560 non-null   float64
 13  heartRate        2560 non-null   int64
 14  glucose          2560 non-null   int64
 15  TenYearCHD       2560 non-null   int64
dtypes: float64(3), int64(12), object(1)
memory usage: 320.1+ KB
```

i) What is the fitted logistic regression model? Do not simply copy the results of your code, but instead state the equation used by the model to make predictions. Use all features from Table 1 to build your model.

```
import statsmodels.formula.api as smf

# Fit the logistic regression model

logreg = smf.logit(formula = 'TenYearCHD ~ male + age + education +
currentSmoker + cigsPerDay + BPMeds + prevalentStroke + prevalentHyp +
diabetes + totChol + sysBP + diaBP + BMI + heartRate + glucose',
                   data = f_train).fit()

print(logreg.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.379592
         Iterations 7
                           Logit Regression Results

==============================================================================
========
Dep. Variable:              TenYearCHD   No. Observations:
2560
Model:                           Logit   Df Residuals:
2542
Method:                            MLE   Df Model:
17
Date:                Tue, 08 Oct 2024   Pseudo R-squ.:
0.1147
Time:                        20:20:57   Log-Likelihood:
-971.75
converged:                       True   LL-Null:
```

```
-1097.6
Covariance Type:            nonrobust   LLR p-value:
9.215e-44
==============================================================================
=========================================
                                                  coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
-----------------------------------------
Intercept                                      -8.2638      0.839
-9.854      0.000      -9.908      -6.620
education[T.High school/GED]                   -0.1922      0.205
-0.938      0.348      -0.594       0.210
education[T.Some college/vocational school]    -0.2265      0.231
-0.982      0.326      -0.679       0.226
education[T.Some high school]                  -0.0871      0.193
-0.451      0.652      -0.465       0.291
male                                            0.6322      0.132
4.793      0.000       0.374       0.891
age                                             0.0592      0.008
7.393      0.000       0.044       0.075
currentSmoker                                   0.0203      0.187
0.109      0.914      -0.345       0.386
cigsPerDay                                      0.0171      0.007
2.360      0.018       0.003       0.031
BPMeds                                          0.3086      0.295
1.047      0.295      -0.269       0.886
prevalentStroke                                 0.6558      0.532
1.233      0.218      -0.387       1.698
prevalentHyp                                    0.2854      0.165
1.734      0.083      -0.037       0.608
diabetes                                       -0.0748      0.379
-0.198      0.843      -0.817       0.667
totChol                                         0.0031      0.001
2.399      0.016       0.001       0.006
sysBP                                           0.0125      0.004
2.775      0.006       0.004       0.021
diaBP                                          -0.0032      0.008
-0.420      0.674      -0.018       0.012
BMI                                             0.0068      0.015
0.450      0.653      -0.023       0.036
heartRate                                      -0.0009      0.005
-0.189      0.850      -0.011       0.009
glucose                                         0.0086      0.003
3.236      0.001       0.003       0.014
==============================================================================
=========================================
```

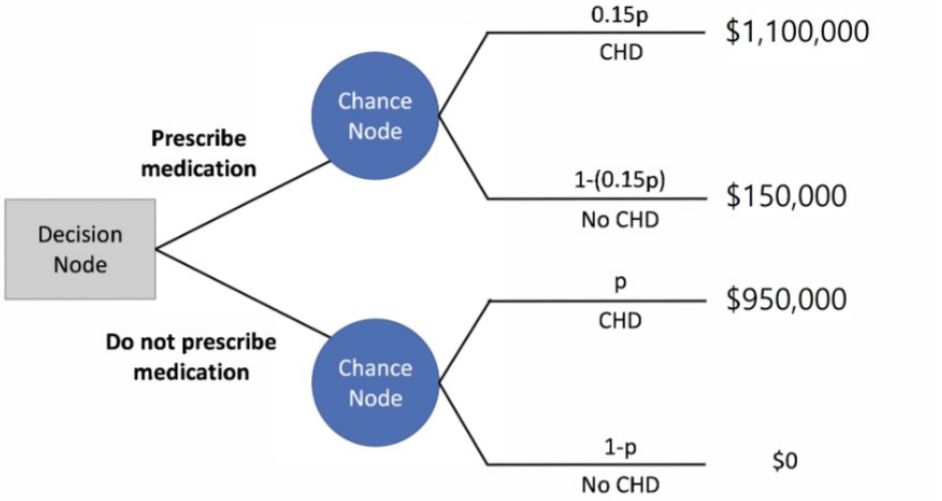


ii) What are the most important risk factors for 10-year CHD risk identified by the model? In other words, those variables that appear to have a significant increasing effect on 10-year CHD risk. Pick one of these variables and describe its impact on a patient's predicted odds of developing CHD in the next 10 years.

Based on the model, the coefficients for the features male, prevalentStroke, and BPMeds are likely the top 3 most important risk factors for 10-year CHD risk. For every 1-unit increase in the male feature, the odds of CHD increase by 63.22%. For every 1-unit increase in the prevalentStroke feature, the odds of CHD increase by 65.58%. For every 1-unit increase in BPMeds feature, the odds of CHD increase by 30.86%.

Picking the male feature, it seems like the sex of the individual played a pretty impactful role in determining the odds of CHD. This could be inferred as CHD being a sex-related disease where men are more likely to get it than women. From understanding this fact, we would be more careful in watching men and their developments over the next 10 years than women overall as their odds are relatively higher.

Figure 1: Decision tree for prescribing the approved medication to prevent CHD. The leaf nodes represent cost values.



iii) Suppose that you wish to determine the optimal strategy for assigning which new patients receive the medication. Given your colleague's analysis of the costs and benefits associated with the recently approved treatment, identify a threshold value of p, call it $\bar{p}$, such that it is optimal to prescribe the medication to a patient if and only if their 10-year CHD risk exceeds $\bar{p}$.

E[Cost | Prescribe]

$$E[\text{Cost}] = 0.15\,p(1{,}100{,}000) + (1 - 0.15\,p)(150{,}000)$$

$$¿\, 165000\,p + (150000 - 22500\,p)$$

E[Cost | Do not Prescribe]

$$E[\text{Cost}] = p(950{,}000) + (1 - p)0$$

Comparing cost if prescribe vs do not prescribe:

$$165000\,p + (150000 - 22500\,p) > p(950{,}000)$$

$$150000 > (950000 + 22500 - 165000)$$

$$\frac{150000}{807500} < p \approx 0.18576 < p$$

So, if p > 0.18576, prescribing is better cost-wise than not prescribing.

iv) Describe the test set performance of the logistic regression model, using the threshold identified in part (iii) to separate patients into those who are at high risk for CHD (risk exceeding the threshold $\bar{p}$) and those who are at low risk for CHD (risk below the threshold $\bar{p}$). State the model's accuracy, True Positive Rate (TPR), and False Positive Rate (FPR), and briefly describe these three metrics in a way that is accessible to a non-technical audience.

```
y_prob = logreg.predict(f_test)
y_prob
```

```
0       0.069521
1       0.203809
2       0.063341
3       0.055820
4       0.275753
          ...
1093    0.059773
1094    0.046459
1095    0.171108
1096    0.153250
1097    0.035033
Length: 1098, dtype: float64
```

```
y_prob = logreg.predict(f_test)
y_pred = pd.Series([1 if x > 0.18576 else 0 for x in y_prob],
index=y_prob.index)
```

If the probability is of getting CHD is higher than 18.58%, then it would be better to get prescribed than not prescribed for cost. So, to determine this for the test set, we could compare the values of y_prob with 18.58%.

```
from sklearn.metrics import confusion_matrix


cm = confusion_matrix(f_test['TenYearCHD'], y_pred)
print ("Confusion Matrix : \n", cm)
```

```
Confusion Matrix :
 [[713 221]
 [ 71  93]]
```

```
Accuracy= (cm.ravel()[0]+cm.ravel()[3])/sum(cm.ravel())
TPR = (cm.ravel()[3])/(cm.ravel()[2] + cm.ravel()[3])
FPR = (cm.ravel()[1])/(cm.ravel()[0]+cm.ravel()[1])
print(Accuracy)
print(TPR)
print(FPR)
```

```
0.7340619307832422
0.5670731707317073
0.2366167023554604
```

Accuracy: Overall, how correct our risk predictions are for everyone we assess.

True Positive Rate (TPR): Our ability to correctly identify individuals who are truly at high risk for CHD.

False Positive Rate (FPR): Our tendency to incorrectly classify individuals who are not at risk as being at risk for CHD.

v) Patients are prescribed the medication using the strategy implied by the model, use the test set data to provide an estimate(s) for the expected economic cost per patient. You should first report your estimate assuming that the CHD outcomes in the test set are not affected by the treatment decision. Is this assumption reasonable? You should then adjust your estimate in a way that takes into account the fact that the treatment decision impacts a patient's risk of developing CHD. (Hint: keep in mind that this dataset was collected before the option of prescribing the medication was even considered.)

TN = 713 FP = 221 FN = 71 TP = 93

Cost for if Prescribed Med Prediction:

Cost of treatment if patient develops CHD (TP): $1,100,000

Cost of treatment if patient doesn't develop CHD (FP): $150,000

Cost for if not Prescribed Med Prediction:

Cost if patient develops CHD (FN): $950,000

Cost if patient doesn't develop CHD (TN): $0

The overall expected economic cost per patient is a weighted sum based on the number of TPs, FPs, FNs, and TNs in the test set:

Expected Cost:

$$\frac{(93*1100000+71*950000+221*150000+0*713)}{1098} \approx 184790.52823$$

This is the expected cost per patient under the assumption that the CHD outcomes are not influenced by treatment decisions.

However this assumption is unrealistic. Because the medication is supposed to reduce the risk of developing CHD, not addressing this value underestimates its potential benefits. For example, having such medication prescribed as a preventive measure could lower the overall probability of a high-risk patient actually developing CHD, which could reduce their healthcare cost in the future.

In the first part, we assumed the medication has no effect on reducing the risk of CHD. In other words, we use the test set's observed CHD outcomes as if the treatment didn't exist.

From the given situation, it states that the medication reduces the probability of developing CHD in high-risk patients by 85%. If their current 10-year risk (probability) of developing CHD is p without taking the medication, then their 10-year risk (probability) with the medicine would instead be

$(0.15 * p)$

For True Positives (TP): Instead of having 93 patients expected to develop CHD, the amount will be reduced to

$$93 * 0.15 = 13.95 \approx 14\, individuals.$$

For False Positives (FP): There will be an additional number of patients who will get misidentified from high-risk (resulting in be prescribed the medicine) to then turning out to NOT have CHD 10 years later due to taking the medicine:

$$93 - 13.95 = 79.05 \approx 79 individuals$$

Adjusted Expected Cost:

$$\frac{(13.95 * 1100000 + 79.05 * 150000 + 71 * 950000 + 221 * 150000 + 0 * 713)}{1098} \approx 116395.71949$$

vi) Consider a simple baseline model that predicts none of the patients are at high risk for CHD and therefore does not recommend treatment for any of the patients. Describe the test set performance of the baseline model in terms of accuracy, TPR, and FPR, as well as expected economic cost per patient.

```
# baseline model where it predicts none of the patients are at high
risk for CHD and does not recommend treatment

baseline_no = np.sum(f_test['TenYearCHD'] == 0)  # no CHD developed
baseline_yes = np.sum(f_test['TenYearCHD'] == 1)   # CHD developed

print(pd.Series({'0': baseline_no, '1': baseline_yes}))
```

```
0    934
1    164
dtype: int64
```

```
ACC = baseline_no/(baseline_no + baseline_yes)
TPR = 0
FPR = 0
ACC, TPR, FPR
```

```
(0.8506375227686703, 0, 0)
```

TN = 934 FP = 0 FN = 164 TP = 0

Cost for if Prescribed Med Prediction:

Cost of treatment if patient develops CHD (TP): $1,100,000

Cost of treatment if patient doesn't develop CHD (FP): $150,000

Cost for if not Prescribed Med Prediction:

Cost if patient develops CHD (FN): $950,000

Cost if patient doesn't develop CHD (TN): $0

Expected Cost:

$$\frac{(0*1100000+164*950000+0*150000+0*713)}{1098} \approx 141894.35337$$

vii) Use an example to explain how to use the model in a real clinical setting. Suppose a new patient arrives, and the physician accesses the patient's electronic medical records and retrieves the following about the patient: Male, age 40, Some high school education, currently a smoker with an aerage of 2 cigarettes per day. Currently not on blood pressure medication, had a stroke and is hypertensive. No diabetes; total Cholesterol at 180. Systolic/diastolic blood pressure at 140/100, BMI at 28, heart rate at 80, glucose level at 100.

What is the predicted probability that this patient will experience CHD in the next ten years? Based on your calculated $\bar{p}$ threshold from part (iii) from the decision tree, should the physician prescribe the preventive medication for this patient?

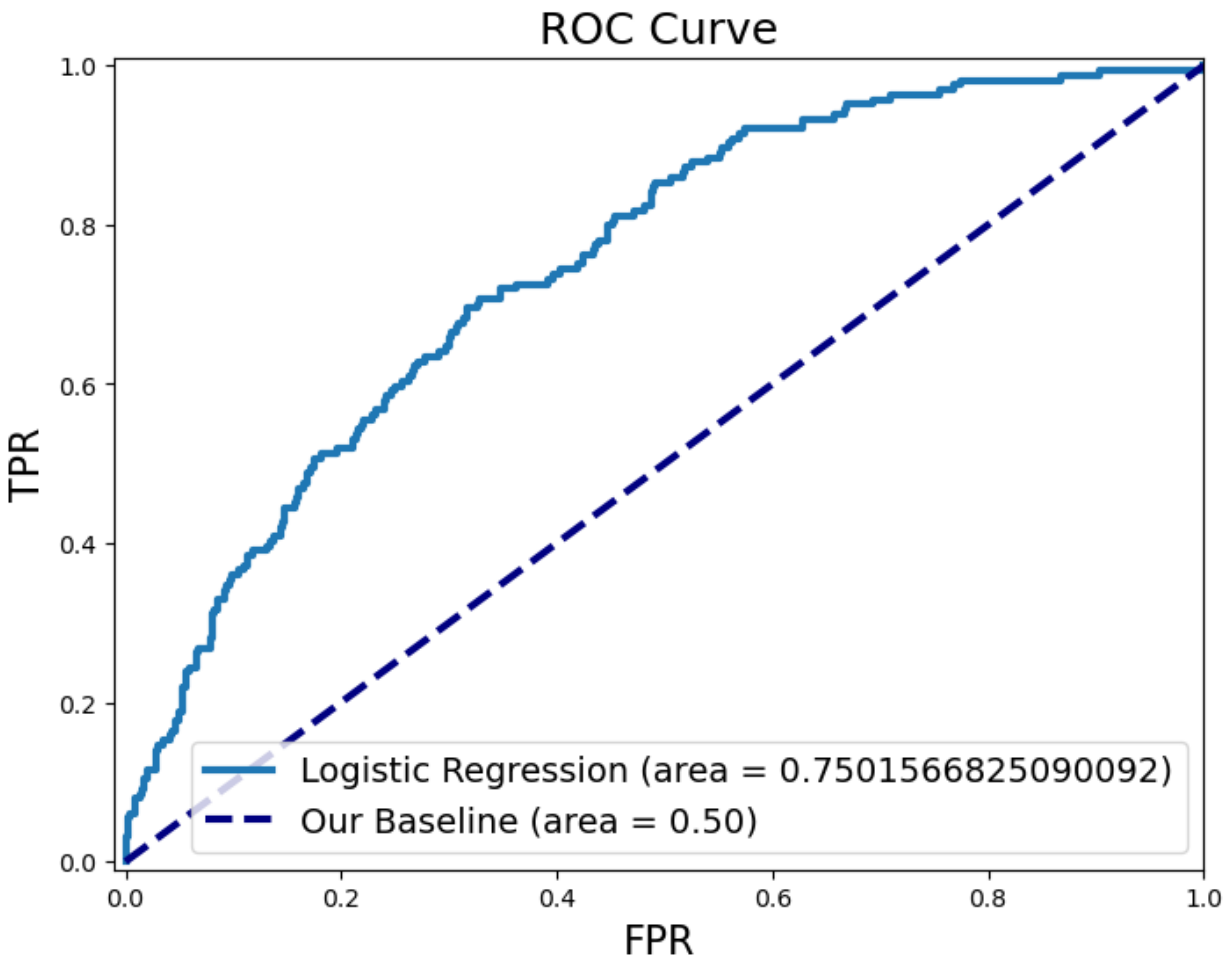
```
patient_data = {
    'male': [1],
    'age': [40],
    'education': ['Some high school'],
    'currentSmoker': [1],
    'cigsPerDay': [2],
    'BPMeds': [0],
    'prevalentStroke': [1],
    'prevalentHyp': [1],
    'diabetes': [0],
    'totChol': [180],
    'sysBP': [140],
    'diaBP': [100],
    'BMI': [28],
    'heartRate': [80],
    'glucose': [100]
}


patient_df = pd.DataFrame(patient_data)

predicted_p = logreg.predict(patient_df)[0]

print(f"Predicted probability of CHD: {predicted_p:}")
```

```
Predicted probability of CHD: 0.19885768512679286
```

So, if p > 0.18576, prescribing is better cost-wise than not prescribing. Since the probability we calculated is greater than 0.18576, we should prescribe the patient medicine.

b) 15 points) Show the ROC curve for your logistic regression model on the test set and describe how this curve may be helpful to decision-makers looking to further study the medication you have considered so far in this homework as well as other possible medications for preventing CHD. Describe one interesting observation implied by examining the ROC curve. What is the area under the curve (AUC) for your model in the test set?

```
y_train = f_train['TenYearCHD']
X_train = f_train.drop(['TenYearCHD'], axis=1)

y_test = f_test['TenYearCHD']
X_test = f_test.drop(['TenYearCHD'], axis=1)

import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc

fpr, tpr, _ = roc_curve(y_test, y_prob)
roc_auc = auc(fpr, tpr)
roc_auc

plt.figure(figsize=(8, 6))
plt.title('ROC Curve', fontsize=18)
plt.xlabel('FPR', fontsize=16)
plt.ylabel('TPR', fontsize=16)
plt.xlim([-0.01, 1.00])
plt.ylim([-0.01, 1.01])
plt.plot(fpr, tpr, lw=3, label='Logistic Regression (area =
{:})'.format(roc_auc))
plt.plot([0, 1], [0, 1], color='navy', lw=3, linestyle='--',
label='Our Baseline (area = 0.50)')
plt.legend(loc='lower right', fontsize=14)
plt.show()
```



The ROC curve helps us assess the model's ability to predict CHD risk by showing the trade-off between True Positive Rate (aka sensitivity) and False Positive Rate. A model with a higher curve indicates better performance, and here, it is visualized that our logistic regression model outperforms baseline model with an AUC of around 0.75. This indicates it has a 75% chance of correctly identifying at-risk patients. Through this, we are able to evaluate the model's effectiveness in targeting high-risk individuals for preventive medication and help them set appropriate thresholds for treatment. One interesting observation I have made is that as FPR increases, the improvement of TPR also begins to trail off/become stagnant, inferring that our model performs very well initially, but flattens.

c) (10 points) Rather than explicitly dictating which patients should receive the medication, let us consider letting patients decide for themselves. Suppose that if a patient has health insurance, the treatment costs for CHD (including the proposed medication) will be covered by their insurance company. However, a patient will still incur an equivalent cost of $600,000 for decreased quality of life if they develop CHD. Disregarding other factors such as side effects of the medication, if there were no insurance co-payment then it should be clear that every patient would always choose to receive the medication because it would cost them nothing and it would lower their risk of CHD. Thus let us consider setting a co-payment value C-- the amount that each patient would have to pay in order to receive the medication – in order to provide an incentive for some patients to forego the treatment while others would choose to receive the treatment. What value of C should the insurance company charge as a co-payment for the medication in order that the patients would "self select" in a manner that is consistent with the previously examined "optimal strategy" discussed in part (a) above?

Observing if patient chooses to RECEIVE medication:

Expected cost =

$$C + 600000 * (0.15 * p)$$

** note: it is 0.15*p because taking the medication will decrease the probability of developing CHD

Observing if patient chooses to NOT RECEIVE medication:

Expected cost =

$$p * 600000$$

Solving Optimal Value of C:

Set

$$C + 600000 * (0.15 * p) = p * 600000$$

and solve for C

$$C + 90000\, p = 600000\, p$$

$$C = 510000\, p$$

given that p = 0.18576 was our calculated threshold:

$$C = 510000 * 0.18576 = 94737.6$$

d) (5 points) Are there any aspects of the analysis performed thus far that raise ethical concerns? If so, suggest at least one way that this analysis could be changed to address such concerns.

Overall, over-reliance on algorithmic predictions for medical decisions that handle life and death may be controversial as these medical conditions are immensely complex and require real human judgment from doctors (people who have been in school for this for 8+ years). However, it would be more ethical if such judgment is paired with the model's insights to make inferences.

There may be some bias in our model if it was trained on data that underrepresents marginalized communities, potentially leading to inaccurate risk predictions for minority

populations. We should expand areas/regions/races/sexualities where we gather data and ensure it is randomly sampled if possible to avoid these issues. Alongside bias, the issue of data privacy has been a huge topic recently. Data privacy and informed consent must be ensured, and patients should know how their data is used for this to be an ethical analysis.