IEOR 142: Introduction to Machine Learning and Data Analytics, Fall 2019

Midterm Exam

October 2019

Name: ____________________

SID: ____________________

Instructions:

1. Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.
2. You are allowed one (double sided) 8.5 x 11 inch note sheet and a simple pocket calculator. The use of any other note sheets, textbook, computer, cell phone, other electronic device besides a simple pocket calculator, or other study aid is not permitted.
3. You will have until 5:00PM to turn in the exam.
4. Whenever a question asks for a numerical answer (such as 2.7), you may write your answer as an expression involving simple arithmetic operations (such as $2(1) + 1(0.7)$).
5. Good luck!

1 True/False and Multiple Choice Questions – 45 Points

Instructions: Please circle exactly one response for each of the following 15 questions. Each question is worth 3 points. There will be no partial credit for these questions.

1. The probability model underlying logistic regression states that $\Pr(Y = 1|X) = h(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$ where Y is the dependent variable, X is the vector of independent variables, $(\beta_0, \beta_1, \ldots, \beta_p)$ are the logistic regression coefficients, and $h(w) = \frac{1}{1+e^{-w}}$ is the logistic function.

   A. **True**

   B. False

2. Consider a linear regression model with a highly insignificant variable such that the p-value of the corresponding coefficient is greater than 0.50. Then, removing this variable from the model and re-training always results in a decrease in the training set R^2 value.

   A. True

   B. **False**

3. Consider a linear regression model with a highly insignificant variable such that the p-value of the corresponding coefficient is greater than 0.50. Then, removing this variable from the model and re-training always results in an increase in the test set OSR^2 value.

   A. True

   B. **False**

4. Consider a simple linear regression problem with a continuous dependent variable Y and a single independent variable X. Suppose that we have a training dataset of $n = 2$ observations $(x_1, y_1), (x_2, y_2)$ that satisfies $x_1 \neq x_2$ and $y_i = \beta_0 + \beta_1 x_i$ for $i = 1, 2$, where β_0, β_1 are the true coefficients for the model. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimates of β_0 and β_1, respectively, based on minimizing the RSS (residual sum of squared errors) on the training set. Then, it must be the case that $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$.

   A. **True**

   B. False

5. In order to train a boosting model (with trees as the base models), one of the required inputs to the algorithm is the number of splits in each of the base tree models, and this parameter should ideally be tuned with cross-validation.

   A. **True**

   B. False

6. Consider training a CART model for binary classification and suppose that we use either the error rate impurity function or the Gini index impurity function. Then, in both cases, the total impurity cost of the tree is guaranteed to strictly decrease after every additional split.

   A. True

   B. **False**

7. Consider using the bootstrap to asses the variability of the OSR^2 value of a previously trained Random Forests model on the test set, e.g., by constructing a confidence interval. Suppose that we set $B = 10,000$ for the number of bootstrap replications. Then, this procedure requires computing the OSR^2 value of the Random Forests model on 10,000 different bootstrapped datasets.

   A. **True**

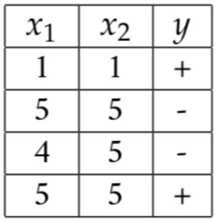
   B. False

8. The accuracy of a logistic regression model does not depend on the choice of the probability threshold value.

   A. True

   B. **False**

9. Consider the dataset below in Figure 1 for a binary classification problem with $p = 2$ features and where $+$ denotes a positive label and $-$ denotes a negative label.



x_1	x_2	y
1	1	+
5	5	-
4	5	-
5	5	+

Figure 1

   Then, it is possible for some classifier to achieve perfect 100% accuracy on this dataset.

   A. True

   B. **False**

10. After removing punctuation, the bag of words representation of "Paul likes to travel" is the same as that of "Paul likes to travel. Paul likes to travel."

   A. True

   B. **False**

11. It is always the case that nonparametric methods (like boosting and random forests) will outperform parametric methods (like linear regression) in terms of out of sample predictive performance.

   A. True

   B. **False**

12. Consider a binary classification problem where the test set has $N_{\text{pos}} > 0$ positive observations and $N_{\text{neg}} > 0$ negative observations. Suppose that we have previously trained a model on the training set, and that, on the test set, this model has a true positive rate value denoted by TPR and a false positive rate value denoted by FPR. Then a correct expression for the accuracy of this model on the test set is given by:

$$\text{Accuracy} = \frac{N_{\text{pos}} \cdot \text{TPR} + N_{\text{neg}}(1 - \text{FPR})}{N_{\text{pos}} + N_{\text{neg}}}$$

A. **True**

B. False

13. Which of the following actions has the *least* risk of increasing the likelihood of overfitting?

A. Increasing the number of trees/iterations when training a boosting model

B. Increasing the number of trees when training a random forests model while leaving the value of m (**`mtry`**) fixed

C. Decreasing the value of m (**`mtry`**) when training a random forests model while leaving the number of trees fixed

D. Introducing new independent variables in a linear regression model that are quadratic functions of the original set of independent variables

Answer: B

14. Which of the following statements are true regarding k-fold cross-validation?

1. Increasing the value of k results in more overall computation time for the cross-validation procedure
2. Using $k = n$ where n is the number of data points in the training set is the same as leave-one-out cross-validation (LOOCV).
3. Using $k = 1$ is the same the validation set method.

A. Only (1.) and (2.)

B. Only (1.) and (3.)

C. Only (2.) and (3.)

D. All three statements

Answer: A

15. Consider training a CART model for a classification problem on a training set of size $n = 6$ with $p = 2$ independent variables. Figure 2 below displays a scatter plot of the independent variables (X_1, X_2) along with 5 regions corresponding to the CART model that was trained. What is the most definitive (i.e., strongest) statement that can be made about the accuracy A of this CART model on the training set?

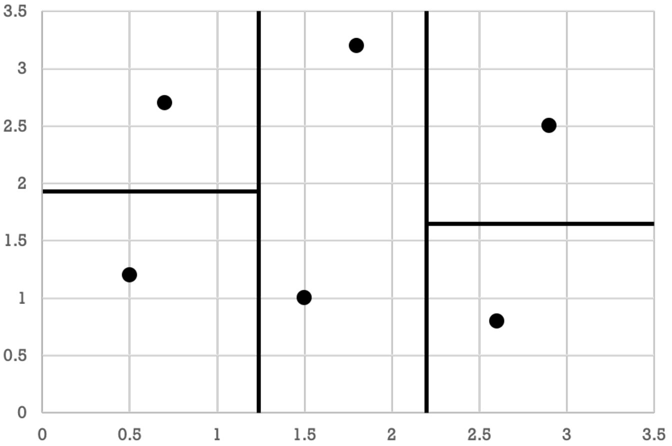


Figure 2

A. $0 \leq A \leq 1$

B. $4/6 \leq A \leq 1$

C. $5/6 \leq A \leq 1$

D. $A = 1$

Answer: C

2 Short Answer Questions – 55 Points

Instructions: Please provide justification and/or show your work for all questions, but please try to keep your responses brief. Your grade will depend on the clarity of your answers, the reasoning you have used, as well as the correctness of your answers.

The first two problems concern a dataset[1] of golf player statistics with 162 observations, each corresponding to a different top professional golfer who participated in the PGA tour in 2018. Various attributes[2] concerning player performance and winnings throughout the entire length of the 2018 season were collected and aggregated. Table 1 below describes these attributes in more detail. For clarity, the first 6 observations of the dataset are also included below. We are primarily interested in building models for predicting player success – in terms of monetary winnings – based on the four direct performance statistics/attributes that are provided. We are also interested in which performance statistics have the greatest impact on success.

Table 1: Description of the dataset.

Variable	**Description**
`PlayerName`	The player's name
`Winnings`	Total monetary winnings over the entire season, in millions of dollars (USD)
`AverageScore`	Average total point score per 18 hole round
`AveragePutts`	Average number of putts per hole
`AverageDrivingDist`	Average drive distance per hole, in yards
`DrivingAccuracy`	Percentage of shots where the drive shot successfully lands on the fairway area

```
> head(golf_data)
# A tibble: 6 x 6
  PlayerName     Winnings AverageScore AveragePutts AverageDrivingDist DrivingAccuracy
  <chr>             <dbl>        <dbl>        <dbl>              <dbl>           <dbl>
1 Aaron Baddeley    0.905         70.8         1.72               286.            57.7
2 Aaron Wise        1.05          70.7         1.73               303.            61.8
3 Abraham Ancer     3.17          70.6         1.75               293.            70.2
4 Adam Hadwin       2.22          70.5         1.73               291.            67.8
5 Adam Long         1.65          71.5         1.79               292             66.5
6 Adam Schenk       1.26          70.8         1.75               301.            61.3
```

[1] This dataset is a subset of a much more comprehensive dataset available at `https://www.kaggle.com/bradklassen/pga-tour-20102018-data`.

[2] To understand some of the attributes better, note that a "putt" is a very short distance shot taken on the "green" near the hole, whereas a "drive" is the initial shot which is typically a very long distance shot.

1. (25 points) The dataset was split into a training set with 105 observations and a test set with 57 observations, and a linear regression model was built, using the training data, to predict **Winnings** based on the four direct player performance stats, namely **AverageScore**, **AveragePutts**, **AverageDrivingDist**, and **DrivingAccuracy**. The output from R is given below.

```
> summary(mod1)

Call:
lm(formula = Winnings ~ AverageScore + AveragePutts + AverageDrivingDist +
    DrivingAccuracy, data = golf_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4422 -0.6233 -0.0387  0.5000  4.9060

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        114.161145  14.961424   7.630 1.40e-11 ***
AverageScore        -1.745918   0.186456  -9.364 2.46e-15 ***
AveragePutts         2.192717   4.374250   0.501    0.617
AverageDrivingDist   0.026401   0.017066   1.547    0.125
DrivingAccuracy     -0.003636   0.029532  -0.123    0.902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.89 on 100 degrees of freedom
Multiple R-squared:  0.6087,Adjusted R-squared:  0.5931
F-statistic: 38.89 on 4 and 100 DF,  p-value: < 2.2e-16
```

The training data was further used to compute a correlation table, the output of which is given below.

```
> cor(golf_train[,c(2,5,6,7,8)])
                      Winnings AverageScore AveragePutts AverageDrivingDist DrivingAccuracy
Winnings            1.00000000   -0.7610713  -0.27987104         0.29221875     -0.01249549
AverageScore       -0.76107132    1.0000000   0.42866433        -0.16524349     -0.14614163
AveragePutts       -0.27987104    0.4286643   1.00000000         0.04029845      0.04488629
AverageDrivingDist  0.29221875   -0.1652435   0.04029845         1.00000000     -0.70911574
DrivingAccuracy    -0.01249549   -0.1461416   0.04488629        -0.70911574      1.00000000
```

Furthermore, variance inflation factors for the independent variables in the linear regression model were also computed.

```
> vif(mod1)
      AverageScore       AveragePutts AverageDrivingDist    DrivingAccuracy
          1.651558           1.393514           2.648311           2.629330
```

Please answer the following questions.

(a) (4 points) A particular golf player is considering adjusting his training strategy and expects that, in the 2019 season, his average score will be 70.50 points per round, he will average 1.76 putts per hole, his average driving distance will be 300 yards, and his driving accuracy will be 60%. The player also expects that there are no major differences in how player performance impacts winnings in the 2019 season versus the 2018 season. Use the R output on the previous pages to make a prediction for this player's total winnings in millions of dollars in the 2019 season.

Answer:

$$\begin{aligned}\hat{y} &= 114.161145 - 1.745918 * (70.50) + 2.192717 * (1.76) + 0.026401 * (300) - \\ &\quad - 0.003636 * (60) \\ &= 2.6341\end{aligned}$$

Grading:

- 2 points: correct coefficients and intercepts
- 2 points: correct number for variables
- One mistake correspond to 1 point deduction

(b) (4 points) Is there a high degree of multicollinearity present in the training set? On what have you based your answer?

Answer: No, there is not a high degree of multicollinearity present in the training set since the VIF values for all coefficients are relatively low, e.g., less than 5.

Grading:

- 1 points: correct answer
- 3 points: correctly justify answer using VIF values

(c) (4 points) Based on the R output on the previous pages, is there enough evidence to conclude that the true coefficient corresponding to `AverageScore` is not equal to 0? On what have you based your answer?

Answer: Yes, there is evidence to conclude that this coefficient is not equal to 0 since the p-value associated with that coefficient, 2.46e-15, is very small and hence significant. Hence we are able to reject the null hypothesis that this coefficient is equal to 0.

Grading:

- 1 points: correct answer
- 3 points: correctly justify answer using p-value

(d) (4 points) Based on the R output on the previous pages, is there enough evidence to conclude that the true coefficient corresponding to `AveragePutts` is not equal to 0? On what have you based your answer?

Answer: No, there is not enough evidence to conclude that this coefficient is not equal to 0 since the p-value associated with that coefficient, 0.617, is too large. Indeed, at any reasonable significant level (e.g., 0.05 or 0.10) we would not be able to reject the null hypothesis that this coefficient is equal to 0.

Grading:

- 1 points: correct answer
- 3 points: correctly justify answer using p-value

(e) (4 points) Consider adding a new independent variable to the model called `AveragePuttsPerRound`, which is equal to the average number of putts per 18 hole round. (Recall that `AveragePutts` is the average number of putts per hole, and you may assume that each round consists of exactly 18 holes.) Is it possible for this new variable to improve the linear regression model for predicting `Winnings`? Explain your answer.

Answer: No, it is not possible for this new variable to improve the linear regression model since `AveragePuttsPerRound` is linearly related to the existing feature `AveragePutts`. In particular, `AveragePuttsPerRound` = 18*`AveragePutts`. Thus, any linear relationship between `Winnings` and `AveragePuttsPerRound` is already captured by the existing model. Put another way, if we were to add `AveragePuttsPerRound` to the model then there would be perfect multicollinearity between `AveragePuttsPerRound` and `AveragePutts` and so removing any one of them would not change the model.

Grading:

- 1 points: correct answer
- 3 points: correctly justify answer by mentioning that the relationship is captured by the existing variable

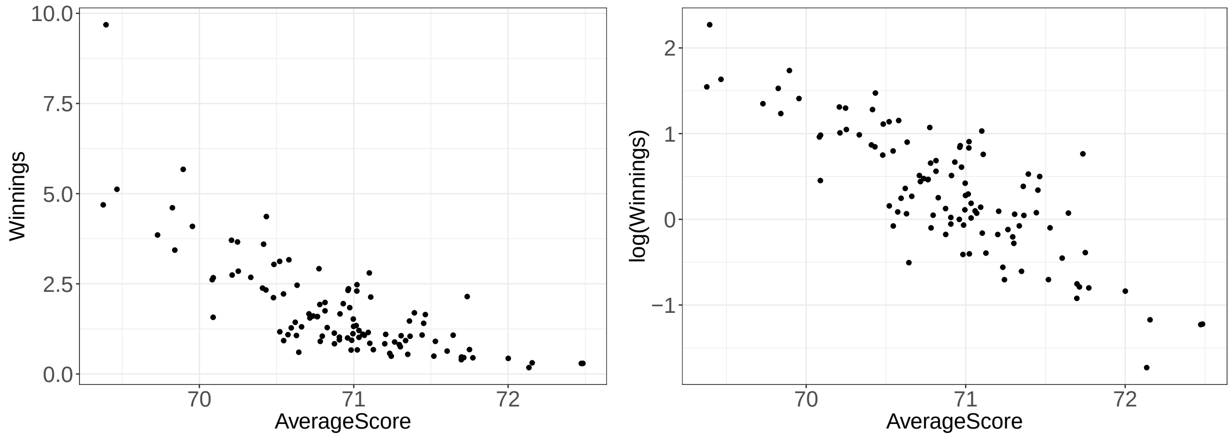


Figure 3

(f) (5 points) A data scientist working with the PGA Tour has determined that a simple linear regression model that only uses a single independent variable, **`AverageScore`**, would strike the best balance between interpretability and performance in this application domain. The data scientist is considering using one of two possible dependent variables: **`Winnings`** as before, or a logarithmic transformation **`log(Winnings)`**. Figure 3 shows scatter plots on the training data of these two possible dependent variables versus **`AverageScore`**. Based on Figure 3, which dependent variable choice would you recommend in order to get the best predictive performance? Explain your answer.

Answer: Based on Figure 3, logarithmic transformation `log(Winnings)` would get the best predictive performance since there is a strong linear relationship between `AverageScore` and `log(Winnings)`. On the other hand, the relationship between `AverageScore` and `Winnings` appears to be nonlinear.

Grading:

- 2 points: correct answer
- 3 points: correctly justify answer by mentioning linear relationship between **`log(Winnings)`** and **`AverageScore`** and/or that **`Winnings`** and **`AverageScore`** has a non-linear relationship (curve upwards)

2. (20 points) Next, a CART model was built to predict `Winnings` as a function of the four provided independent variables. The tree diagram corresponding to this model is shown in Figure 4 below.

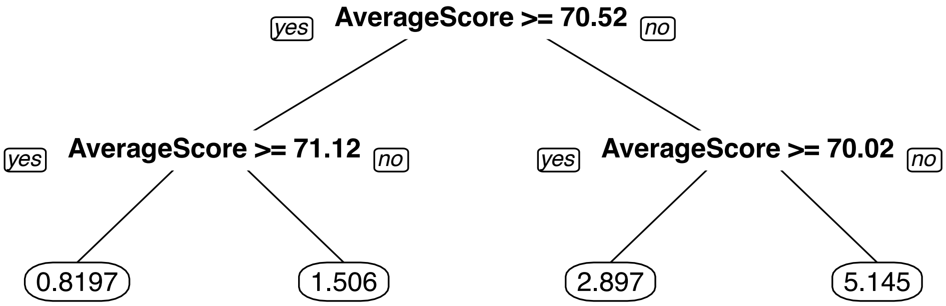


Figure 4

Note that the training set R^2 value of the above CART model is 0.704. Furthermore, the value of the `cp` parameter used when training the above CART model was set to `cp` $= 0.01$.

(a) (5 points) Consider a new CART model that results after some new split on one of the four leaf nodes (buckets) of the current model. Using the information above, what is the most definitive (i.e., strongest) statement you can make concerning the training set R^2 value of this new CART model? Explain your answer.

Answer: The training R^2 value of the new model will be in the interval $[0.704, 0.714]$. Indeed, adding a new split will cause the training set R^2 to go up or stay the same. On the other hand, since `cp` = 0.01, we know that the new model must have R^2 below $0.704 + 0.01 = 0.714$. Indeed, if the R^2 of the new model was strictly greater than 0.714, then according to the definition of `cp` an additional split beyond what is in the current tree diagram above would have been retained by the second phase of the CART algorithm in the training process.

Grading:

- 2 points: explain that the training R^2 will increase or stay the same
- 3 points: get the full answer (and reasoning) that the increase in R^2 is bounded so that the new R^2 must be on the interval $[0.704, 0.714]$

(b) (5 points) Consider a new CART model that results after removing the bottom right split "**`AverageScore`** $\geq$ 70.02". Using the information above, what is the most definitive (i.e., strongest) statement you can make concerning the training set R^2 value of this new CART model? Explain your answer.

Answer: The training R^2 value of the new model will be in the interval $[0, 0.694]$. Indeed, after removing any split, the training R^2 will decrease or stay the same, since training R^2 always increases or stays the same as we add splits during the first phase of the CART algorithm. Now, since the bottom right split mentioned above was actually included in the final tree, this means that the increase in R^2 based on adding the bottom right split must have been at least equal to the value of `cp` = 0.01. Thus, the R^2 value after removing the bottom right split must be in the interval $[0, 0.704 - 0.01] = [0, 0.694]$.

Grading:

- 2 points: explain that the training R^2 will decrease (or decrease or stay the same)
- 3 points: get the full answer (and reasoning) that the R^2 must actually decrease by at least 0.01

(c) (4 points) Provide a brief but precise explanation for why **AverageScore** is the only (out of four possible) independent variable that was selected by the CART algorithm at each split in the above tree.

Answer: Intuitively, based on all results we've seen so far, AverageScore is a lot more significant than the other independent variables. Thus, it is not surprising that AverageScore was the only variable selected. Precisely, AverageScore was selected at each iteration because the CART algorithm chooses the best split (in terms of decreasing the RSS impurity) among all possible splits among all 4 independent variables. For this data set, it happened to be the case that, for the first 3 splits, a split using AverageScore decreased the RSS impurity more than any other split using any other variable.

Grading:

- 2 points: intuitive explanation or mention the result of p-value
- 2 - 4 points: precise explanation about how CART selects splits
- (Full credit given if sufficient precise explanation is given without intuitive explanation.)
- -2 points: no explanation of how CART selects splits

(d) (6 points) Let $\hat{f}$(**AverageScore**) denote the prediction function corresponding to this CART model, i.e., the function that returns the predicted value of **Winnings** as a function of **AverageScore**. Draw the graph of the function $\hat{f}$(**AverageScore**) in Figure 5 below. (You only need to draw the graph for values of **AverageScore** between 69 and 72 and you do not have to be concerned with evenly spacing the ticks on the x and y axes.)

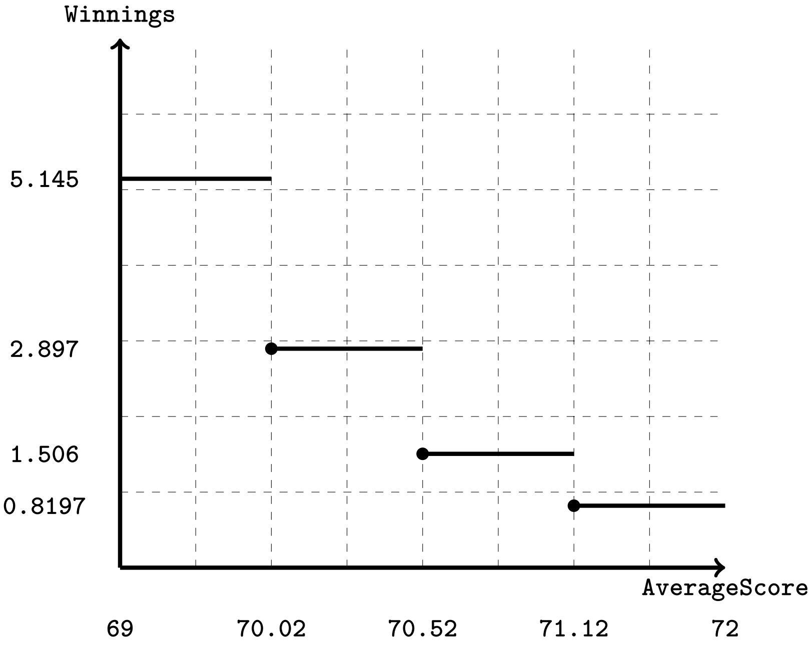


Figure 5

Grading:

- 1 points: correct label on x-axis

- 1 points: correct label on y-axis
- 4 points: correct function (stair-case shape)
- -1 point if filled dot (for inclusion) is not correct
- -1 if making the plot continuous

3. (10 points) Consider again the golfer from Q1 part (a) who is considering adjusting his training strategy and expects that, in the 2019 season, his average score will be 70.50 points per round, he will average 1.76 putts per hole, his average driving distance will be 300 yards, and his driving accuracy will be 60%. A friend of yours offers you a bet, whereby if you agree to the bet then you have to pay your friend \$100 right now. If the golfer mentioned above earns over \$2.5 million dollars in the 2019 season, then your friend will pay you back \$150. Otherwise, your friend keeps the \$100 and does not pay you back anything.

   Do you currently have enough information to decide if you should take this bet or not? If yes, then please mention if you will take the bet or not and describe how you used the information in the previous problems to make your decision. If no, then please precisely describe what additional information you need and, if applicable, what additional model(s) you would build on the training data and how you would use the results of those model(s) to make your decision.

   Answer: No, we do not have enough information. We need to estimate the probability p that this golfer earns over \$2.5 million. If we construct a decision tree for the bet (see Figure 6), we know that the break-even threshold should be $50p - 100(1-p) = 0 \Rightarrow p = 2/3$. However, we do not have information to obtain the value of p for this particular golfer from the current model. To estimate the probability p, we will need to train a classification model such as logistic regression that allows us to directly obtain an estimate of p as a function of the four features. Using the current dataset, this can be achieved by creating an additional binary column "over 2.5 m" and building a classification model using that column as the dependent variable.

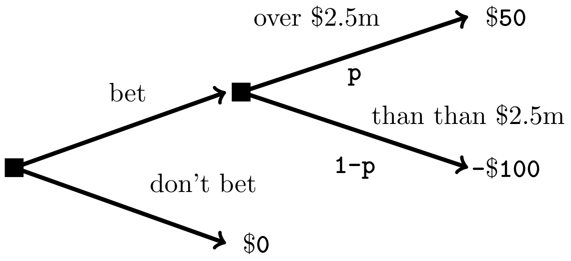


Figure 6

Grading:

- 2 points: correct answer (we do not have enough information)
- 4 points: provide justification for the betting threshold $p = 2/3$ (-2 points if p not provided)
- 4 points: describe what model(s) are needed for estimating p