Final Exam
Bryce Bowles
Nov 30

Write the letter of the choice that most closely corresponds with each of the following terms. Each item will be used only once.
Group of answer choices

Question 1.

1. Leakage
    a. Using data for a predictive model that will not be available at the time of future predictions
2. Modeling
    a. The use of algorithms to "peak" into higher dimensions to discover patterns
3. data preprocessing
    a. The data mining task that usually consumes the most time in a project
4. Regression
    a. Predicting a continuous response
5. predictive analytics
    a. Using data to develop rules and models to support future decisions
    b. Building models using data to make predictions about future events
6. Testing data
    a. Data used to evaluate predictive models
7. descriptive analytics
    a. Discovering patterns in data via summarization and visualization to support decision making
8. Support vector machine
    a. A classification method that balances minimizing error while maximizing margin
9. No Free Lunch Theorem
    a. For any universally consistent classification method, there are distributions of data for which convergence to the Bayes optimal rules is arbitrarily slow
10. Mean absolute error
    a. Error due to using a model that is too simple to capture the patterns in the data
11. Margin
    a. The distance between sets of correctly-classified observations
12. Lift
    a. The number of times better that a model will perform in identifying positive values compared to selecting randomly from the population
13. ROC Curve (Receiver Operating Characteristic)
    a. A plot of the true positive rate versus false positive rate for a family of classification models
14. principal component analysis
    a. A framework that can be used to create two and three-dimensional visualizations of high-dimensional data

    b. A framework that can help visualize clusters from a cluster analysis and identify variables that define clusters
15. Bayes optimal rule
    a. A model that minimizes the probability of misclassification
16. Consistency
    a. A property of a method that, in the limit as more data are collected, converges to a Bayes optimal rule
17. Neural network
    a. A classification model consisting of layers of nodes where data are combined using weights and then transformed via activation functions
18. Precision
    a. The “hit rate” – the percentage of predicted positive values that are actually positive
19. Correlation
    a. A measure of linear relationships between two variables
20. Classification
    a. Predicting a binary or categorical response
21. Bias
    a. Error due to using a model that is too complex for the patterns in the data
22. Confusion matrix
    a. A table of actual and predicted values for a classification model
23. histogram
    a. A visualization for the relationship between one continuous and one categorical variable
24. k-means
    a. A method for clustering that alternately estimates centroids and assigns points to the nearest centroid
25. box-and-whisker plot
    a. A visualization for inspecting the distribution of a continuous attribute
26. Training data
    a. Data used to build predictive models
27. prescriptive analytics
    a. Using data to specify the actions that will help achieve an objective
28. visualization
    a. The beginning and end point of data mining visualization
29. Variance
    a. The average absolute deviation of predictions from actual values; used for evaluating the predictive ability of regression models
30. Random forest
    a. An ensemble method for classification comprised of decision Trees

2 - Question 25 pts

You are an analytics professional working for an e-commerce firm selling athletic shoes on the internet. The company collects data on products and customers to investigate the properties of shoes that sell the best and the demographics of customers who purchase the products. For each shoe design, you have data on profitability ("profitable" or "not"), number of reviews, median review, price, brand ("Nike", "Reebok", "Adidas", or "Other"), production cost, outsole cost, endorsement cost, length of laces, number of eyelets, and profit per shoe (in dollars).

Suppose you are asked to produce a classification model using historical data to predict whether a shoe will be profitable or not.

The data on shoe designs contains a mix of categorical and continuous data. There are missing values for many of the important attributes for about 10% of the designs. Which classification method would minimize the amount of preprocessing needed? Explain why.

- To minimize the amount of preprocessing needed, I would start with a classification tree because it naturally handles missing values. This would also provide variable importance measures.

3- Question 35 pts

In your data, there are 5,000 shoe designs that are not profitable and 300 designs that are. What challenge will this distribution pose in building a model? Describe two ways of overcoming the challenge.

1. There will be an imbalance in the Test and training dataset splits. You can adjust/assign weights

	Number	Weights
Not profitable	5,000	300
Profitable	300	5,000

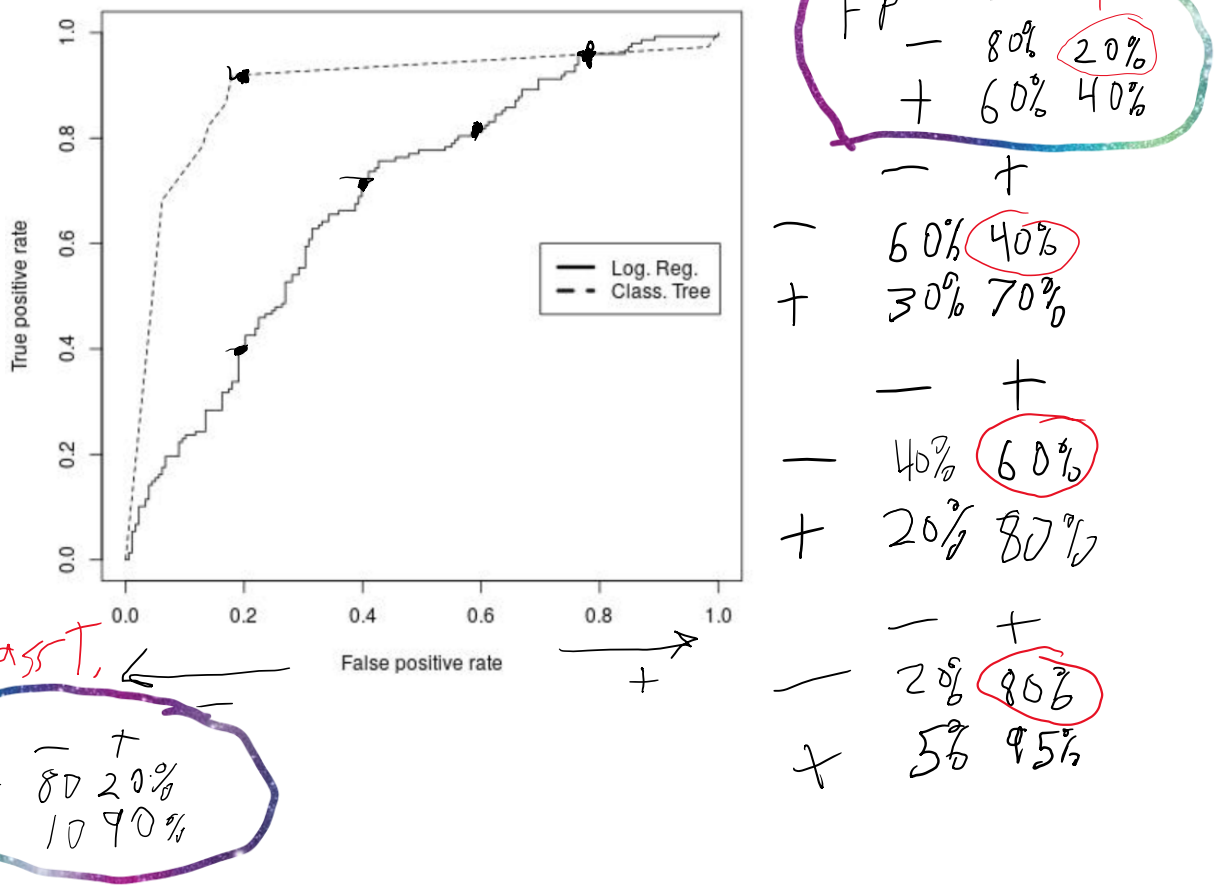
2. Downsample larger class. Take a 300 subsample of not profitable.

4 - Question 415 pts

1. Suppose you fit a logistic regression model and a classification tree model using training data and obtain the following ROC curves for testing data. Which model performs better for this data?

2. Based on a cost analysis, management determines that a 20% false positive rate is desired. What is the true positive rate for each model?

3. Write down a confusion matrix for each model in the previous question when the false positive rate is 20% (Note that the confusion matrix will be expressed in terms of percentages instead of counts.)



1. From the looks of the graph, the classification tree will "perform" better.
2. The true positive rates for each model are: Log. Reg. = 40% and Class. Tree = 90%
3.

Log. Reg.	-	+
-	80%	20%
+	60%	40%

Class. Tree	-	+
-	80%	20%
+	10%	90%

5 - **Question 55 pts**

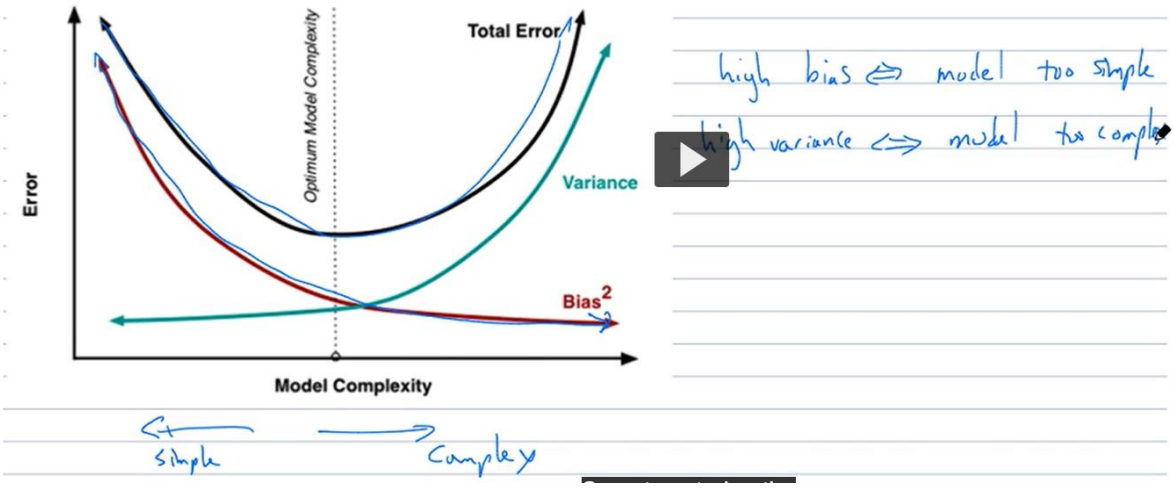
A software vendor contacts you and claims that their implementation of a classification algorithm outperforms all other algorithms by at least 5% on any dataset. They provide you with a quote to implement the software with your firm's systems for $200,000. A co-worker estimates that a 5% increase in accuracy would result in increases in revenue of $100,000 per year and recommends purchasing the software. What is your response?

- That's not possible; no one algorithm works best for all datasets. My initial impression be to ask my co-worker how they calculated the return. Then I would ask the vendor for more details on the algorithm and datasets it was being trained, validated, and tested on so I could analyze the model and predictions myself. What are the maintenance costs per year? Algorithms always need to be updated over time to prevent accumulating unintended bias. What is our budget? How long will it take to implement it? If all goes well, it seems as if the algorithm will pay for itself but it seems too good to be true.

6 - Question 65 pts

Your group decides to deploy a support vector machine with the linear kernel that uses a subset of five predictors to predict whether a shoe design will be profitable. You suspect that a more complex model can capture more subtle relationships between available data and profitability. Describe two ways to increase the complexity of the model.

1. Decrease bias – capture more patterns / more observations
2. Increase variance – add more variables or attributes
3. Use Gaussian kernel or radial basis function kernel instead of the linear kernel for SVM with a large C (to penalize error over margin) and small sigma



7 - Question 75 pts

A neural network model is trained by tuning the size and decay parameters. The best set of parameters is chosen based on the AUC. The output is below. Which combination of parameters is selected as the best?

```
> my_nn
Neural Network

234 samples
  7 predictor
  2 classes: 'profitable', 'not'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 234, 234, 234, 234, 234, 234, ...
Resampling results across tuning parameters:

  size  decay  ROC        Sens       Spec
  1     0e+00  0.5000000  1.0000000  0.0000000
  1     1e-04  0.5000000  1.0000000  0.0000000
  1     1e-01  0.6636565  0.8150772  0.3464436
  3     0e+00  0.5000000  1.0000000  0.0000000
  3     1e-04  0.5000000  1.0000000  0.0000000
  3     1e-01  0.6802752  0.7655241  0.3994653
  5     0e+00  0.5000000  1.0000000  0.0000000
  5     1e-04  0.5000000  1.0000000  0.0000000
  5     1e-01  0.6798116  0.7566192  0.3998660
```

- Size 3 and decay .1 (1e-01) have the greatest ROC indicating it is the best.

8 - Question 810 pts

The output from R for the logistic regression model is included below.

1. Which variables appear to be important for prediction?
2. The plan for the model is to predict which shoes will be profitable before they are displayed on the website. Based on the variables included in the model below, do you anticipate any problems with predicting profitability for future designs?

```
> summary(my_lr)

Call:
glm(formula = profitability ~ ., family = binomial("logit"), data = my_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6155  -0.8900  -0.4162   1.0513   2.0431

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       43.57822   72.65511   0.600   0.5486
production_cost    1.12826    0.88738   1.271   0.2036
model_year         0.15175    0.15661   0.969   0.3326
number_of_views   -0.07927    0.13009  -0.609   0.5423
length_of_laces   -2.37098    1.42282  -1.666   0.0956 .
number_of_eyelets -2.65700    1.39018  -1.911   0.0560 .
brand_Nike       -18.90854 1167.47815  -0.016   0.9871
brand_Reebok      -1.37563    1.47334  -0.934   0.3505
brand_Adidas      -2.28434    1.39349  -1.639   0.1012
number_of_reviews -3.34920    1.40931  -2.376   0.0175 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296.48  on 233  degrees of freedom
Residual deviance: 242.67  on 220  degrees of freedom
AIC: 270.67

Number of Fisher Scoring iterations: 16
```

1. number_of_reviews is the more important variable while number_of_eyelets and length_of_laces are only marginally significant with a greater p value than typically needed to be considered significant, but we’ll still consider them important for prediction. The model also has a reduction from null to residual deviance indicating it is capturing the patterns in the data.

2. The variables are not statistically significant.

    a. If it is using data to predict that will not be available at the time of future predictions such as number of reviews or number or views, this may cause some degree of leakage. I recommend removing the variable and rerunning the model and variable importance.

9 - Question 910 pts
You learn that the historical data for the predictive model is for shoes that were successful in a test marketing at a physical retail location prior to being sold on the website. The goal is to deploy the predictive model prior to test marketing. What is a challenge posed by the data that you have? What is a way that the challenge can be overcome?

- If it is using data to predict that will not be available at the time of future predictions such as number of reviews or number or views, this may cause some degree of leakage and bias. I recommend removing the variables and rerunning the model and variable importance. The potential pitfalls are examples of leakage that can result in an overestimation of model performance. I also suggest using a training, validation and test set to provide an unbiased evaluation of final model fit on the training set.

10 - Question 1010 pts (Module 10_optimization)
In our modeling of Lending Club investments, we considered several strategies for creating a portfolio of loans to invest in, including: (1) choose the smallest loans until the budget is met, (2) choose the loans with the highest predicted return until the budget is met, and (3) use optimization to determine a mix of loans that maximizes predicted return subject to the budget constraint.

1. Which method produced the portfolio with the highest predicted return?
    a. Optimization delivered the best performing portfolio
        i. (1) choose the smallest loans until the budget (of $1,000,000) is met
            1. Predicted: 148,382.7
            2. Actual: 186,545.7
        ii. (2) choose the loans with the highest predicted return until the budget is met
            1. Predicted: 155,765.9
            2. Actual: 223,306.7
        iii. (3) use optimization to determine a mix of loans that maximizes predicted return subject to the budget constraint
            1. Predicted: 156,180.1
            2. Actual: 223,584.9
2. Why is there a discrepancy between the predicted and actual returns of the portfolios, and what can be done to improve agreement between the optimization model output and the actual return of a portfolio?
    a. There is a discrepancy between the predicted and actual returns of the portfolios since the predictions were inaccurate. The ability of the optimization model to select a good portfolio depends on having accurate predictions from the classification and regression models. To improve the optimization, improve the classification and regression models. Experiment with the regression model to make it generate better predictions on profitability before optimizing on this portfolio. One of the ways you can do this is to choose different variables or parameters and calculate variable importance to see which are statistically significant. To minimize error in the regression model, you can try different methods such as using a LASSO approach to place a penalty on the coefficients and enforce sparsity or regularization. Other methods include stepwise, forward and backward regression. Ideally, you're looking for a larger R squared (or goodness of fit) and a smaller MAE and RMSE.