Review Jeopardy

Blue vs. Orange

How this works

- For the Jeopardy and Double Jeopardy rounds, each question has a dollar amount, D.

- Suppose that the proportion of students answering the question correctly is p.

- Then the amount earned by the class for that question is pD.

Jeopardy Round

$200

Transaction ID	Items
10001	**Bread**, **Juice**
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, **Bread**, Formula
10004	**Juice**, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, **Bread**
10006	Formula, **Juice**, Diapers, **Bread**

What is the *support* of the rule Bread —> Juice?

A) 33.3%
B) 20%
C) 5/6
D) 40%
E) I withdraw my support

$200

Transaction ID	Items
10001	**Bread**, **Juice**
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, **Bread**, Formula
10004	**Juice**, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, **Bread**
10006	Formula, **Juice**, Diapers, **Bread**

What is the *support* of the rule Bread —> Juice?

A) 33.3%
B) 20%
C) 5/6
D) 40%
E) I withdraw my support

$200

Transaction ID	Items
10001	**Bread**, **Juice**
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, **Bread**, Formula
10004	**Juice**, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, **Bread**
10006	Formula, **Juice**, Diapers, **Bread**

What is the *confidence* of the rule Bread —> Juice?

A) 50%
B) 75%
C) 25%
D) 40%
E) I vote no confidence.

$200

Transaction ID	Items
10001	**Bread**, **Juice**
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, **Bread**, Formula
10004	**Juice**, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, **Bread**
10006	Formula, **Juice**, Diapers, **Bread**

What is the *confidence* of the rule Bread —> Juice?

A) 50%
B) 75%
C) 25%
D) 40%
E) I vote no confidence.

$200

Transaction ID	Items
10001	**Bread**, **Juice**
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, **Bread**, Formula
10004	**Juice**, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, **Bread**
10006	Formula, **Juice**, Diapers, **Bread**

What is the *lift* of the rule Bread —> Juice?

A) 50%
B) 1
C) 0
D) 40%
E) pi

$200

Transaction ID	Items
10001	**Bread**, **Juice**
10002	Diapers, Beer, Eggs, Formula
10003	Milk, Diapers, **Bread**, Formula
10004	**Juice**, Diapers, Milk, Eggs
10005	Soda, Milk, Eggs, **Bread**
10006	Formula, **Juice**, Diapers, **Bread**

What is the *lift* of the rule Bread —> Juice?

A) 50%

B) 1

C) 0

D) 40%

E) pi

$400

What does it mean when the *lift* of the rule A—> B is greater than 1?

A) It means that the purchase of item A and item B are not independent events
B) It means that *the majority of* people who buy A also buy B, i.e. that the confidence is >50%
C) It means that the there is no association between item A and item B
D) It means that item A doesn't occur as frequently as item B
E) It means A told B but B didn't tell C and no one went up the tree

$400

What does it mean when the *lift* of the rule A—> B is greater than 1?

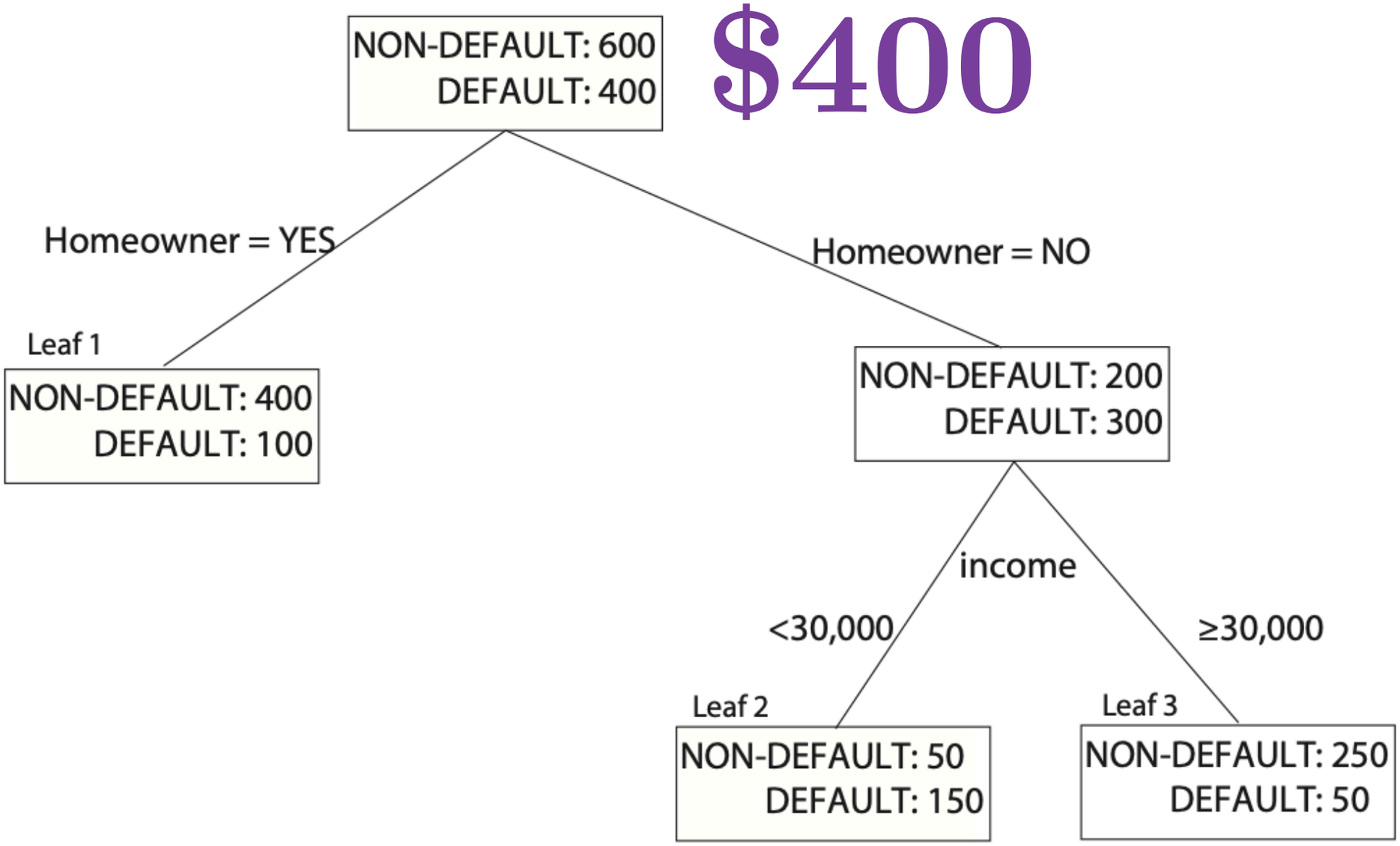
A) It means that the purchase of item A and item B are not independent events

B) It means that *the majority of* people who buy A also buy B, i.e. that the confidence is >50%

C) It means that the there is no association between item A and item B

D) It means that item A doesn't occur as frequently as item B

E) It means A told B but B didn't tell C and no one went up the tree



Assuming a cutoff probability of 0.50, what is the misclassification rate of the tree above?

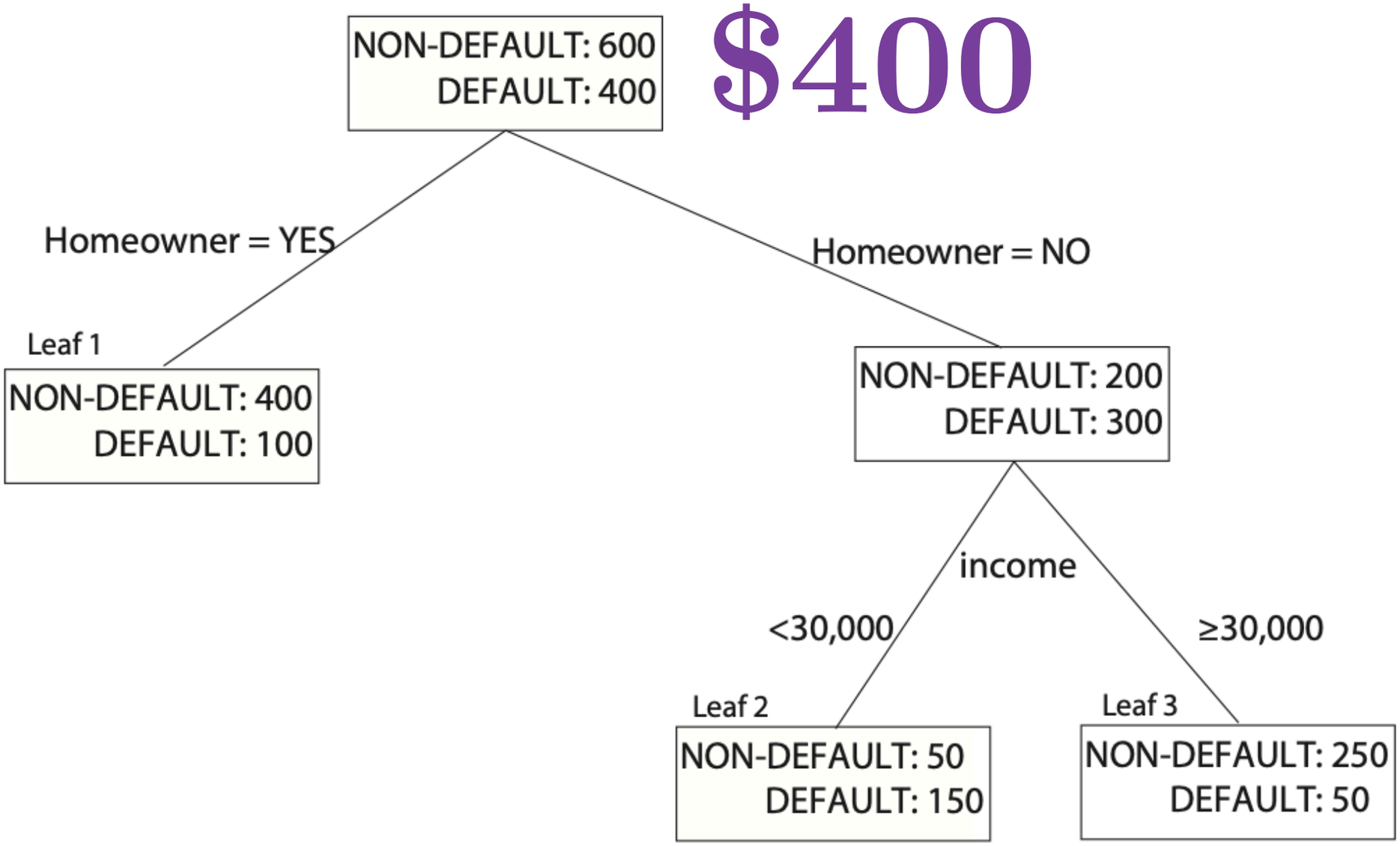
A) impossible to tell

B) 20%

C) 2%

D) 10%

E) 0. This 🌳 is 💯



Assuming a cutoff probability of 0.50, what is the misclassification rate of the tree above?

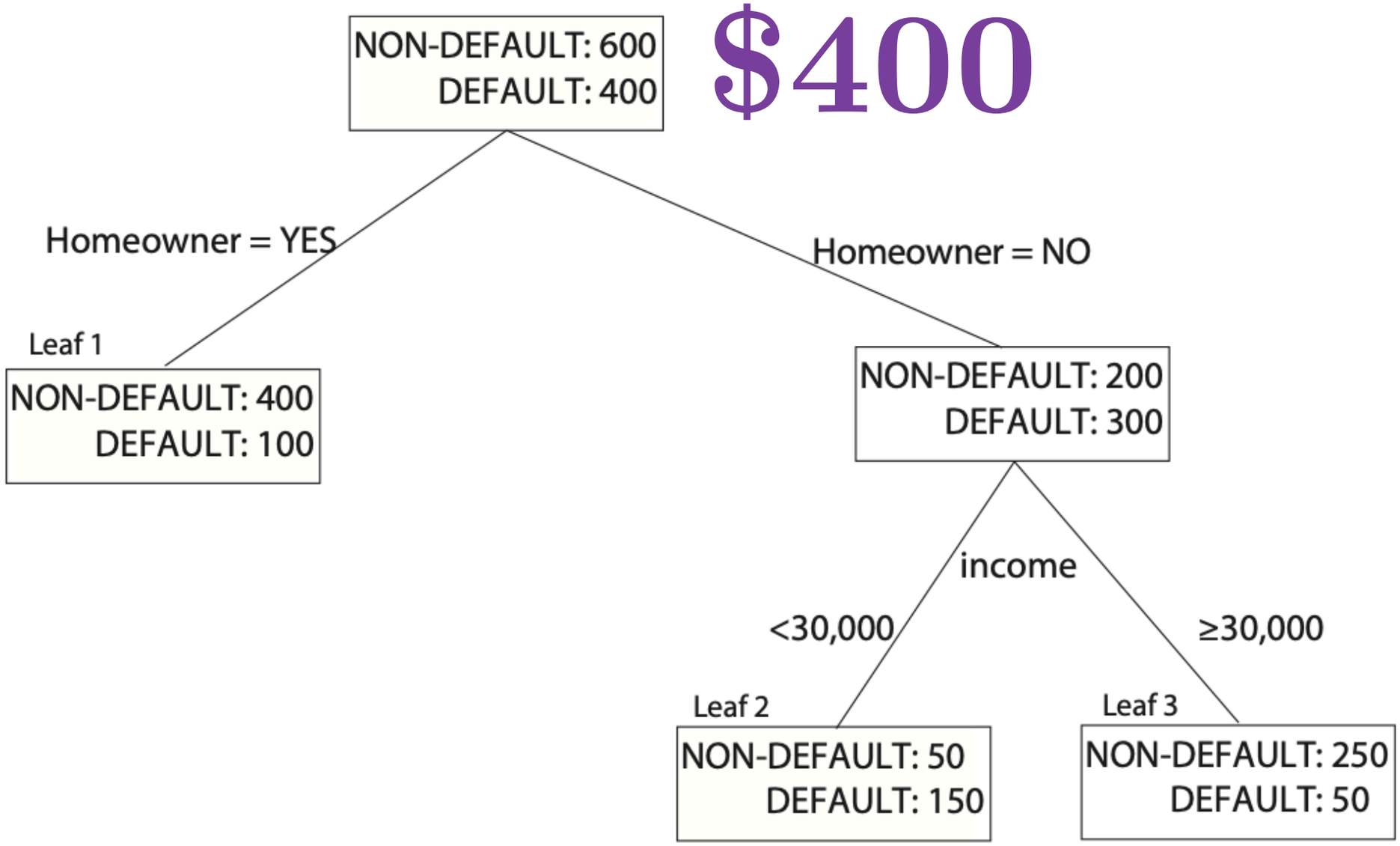
A) impossible to tell

B) 20%

C) 2%

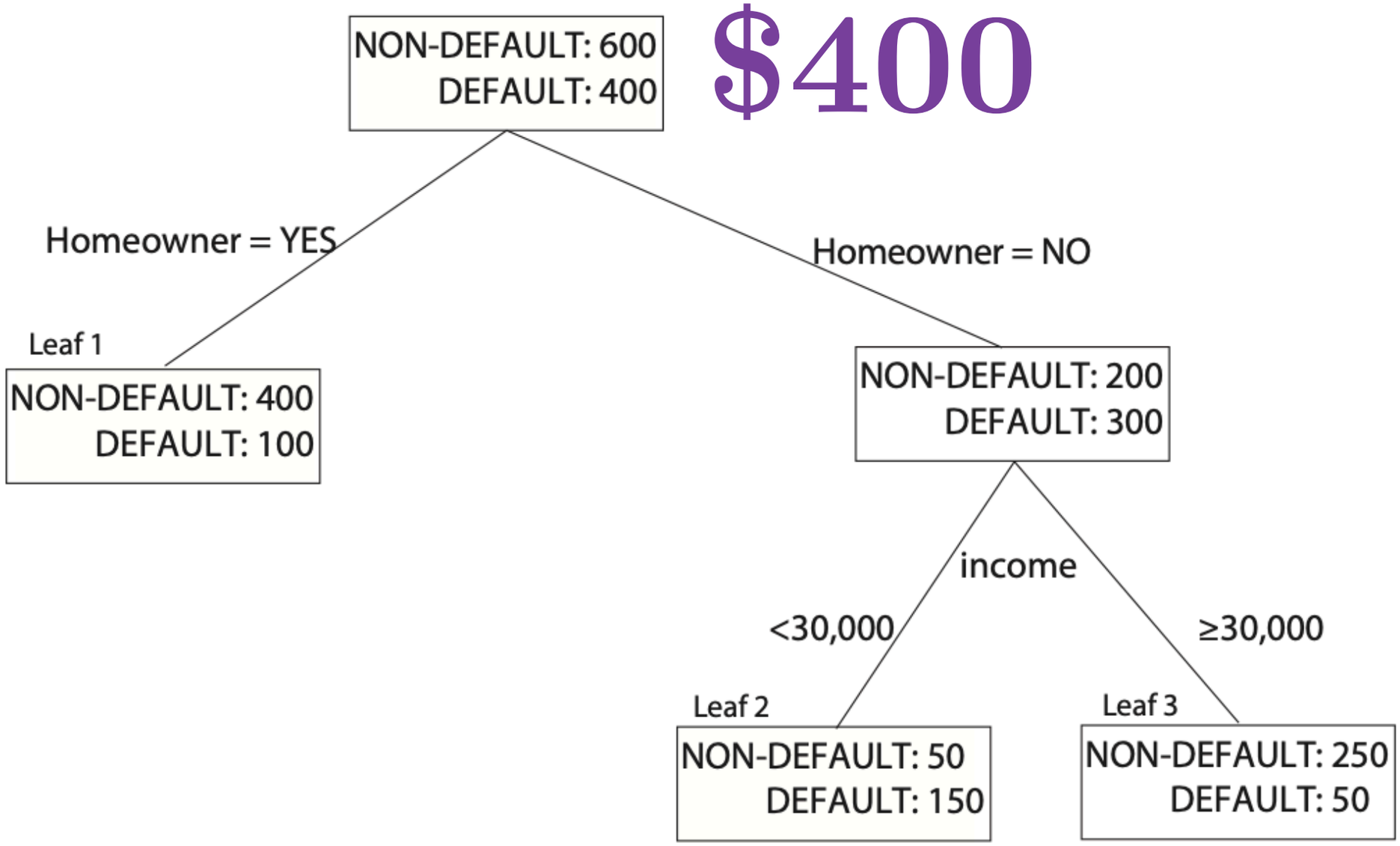
D) 10%

E) 0. This 🌳 is 💯



What is the predicted probability that a homeowner with an income of \$25,000 will default on their credit card payments?

A) impossible to tell
B) 20%
C) 75%
D) 0%
E) 💯



What is the predicted probability that a homeowner with an income of $25,000 will default on their credit card payments?

A) impossible to tell

B) 20%

C) 75%

D) 0%

E) 💯

$600

If your goal is to determine what items might be affected by a price change in item A, you'd be looking for association rules where item A was the ...

A) Antecedent
B) Consequent
C) Antequent
D) Consequedent
E) I was told this review would be fun

$600

If your goal is to determine what items might be affected by a price change in item A, you'd be looking for association rules where item A was the ...

A) Antecedent
B) Consequent
C) Antequent
D) Consequedent
E) I was told this review would be fun

$600

A 50-degree polynomial with all 2-factor interactions is likely to be a model with

A) High bias
B) High variance
C) High training error
D) Low validation error
E) my best interests at heart.

$600

A 50-degree polynomial with all 2-factor interactions is likely to be a model with

A) High bias

B) High variance

C) High training error

D) Low validation error

E) my best interests at heart.

$800

Even when run with the same input data and the same number of clusters, k-means may give the user a different answer each time it is run because...

A) At each iteration of the algorithm, the centroids of each cluster are randomly placed.
B) It often starts with a random initial set of centroids. The initial centroids eventually determine the final solution.
C) k-means has to downsample the data and so the different answers are due to the different random samples of data being used.
D) You have to put the k-means centroids into hierarchical clustering, which is randomized.
E) That's soooo random.

$800

Even when run with the same input data and the same number of clusters, k-means may give the user a different answer each time it is run because...

A) At each iteration of the algorithm, the centroids of each cluster are randomly placed.

B) It often starts with a random initial set of centroids. The initial centroids eventually determine the final solution.

C) k-means has to downsample the data and so the different answers are due to the different random samples of data being used.

D) You have to put the k-means centroids into hierarchical clustering, which is randomized.

E) That's soooo random.

$800

Which of the following is a drawback of hierarchical clustering?

A) It is randomly initialized and will give you different results with each function call.
B) The dendrogram it creates doesn't let you visualize how the data points and clusters relate to each other.
C) It requires computation and storage of the pairwise distance matrix between observations, which is *huge* for large datasets.
D) You have to pick initial seed points.
E) You have to minimize a global objective function (SSE) which makes this method too slow in practice.

$800

Which of the following is a drawback of hierarchical clustering?

A) It is randomly initialized and will give you different results with each function call.

B) The dendrogram it creates doesn't let you visualize how the data points and clusters relate to each other.

C) It requires computation and storage of the pairwise distance matrix between observations, which is *huge* for large datasets.

D) You have to pick initial seed points.

E) You have to minimize a global objective function (SSE) which makes this method too slow in practice.

$1000

Which of the following imputation examples sound *reasonable?*

A) Income is missing for 40% of observations, so I fill in those values with the median income.

B) I have binary variable X which is '1' for 80% of the population. It is missing for 2% of my observations, so I set those values to '1'.

C) I'm building a daily time series model with 2 years of data but I'm missing a sporadic 5% of observations for temperature. I impute these values with the mean of all observations.

D) I am missing 20% of the data in a column called "total_taxes_past_3_years" so I fill in the value 0 for those observations.

$1000

Which of the following imputation examples sound *reasonable?*

A) Income is missing for 40% of observations, so I fill in those values with the median income.

B) I have binary variable X which is '1' for 80% of the population. It is missing for 2% of my observations, so I set those values to '1'.

C) I'm building a daily time series model with 2 years of data but I'm missing a sporadic 5% of observations for temperature. I impute these values with the mean of all observations.

D) I am missing 20% of the data in a column called "total_taxes_past_3_years" so I fill in the value 0 for those observations.

Double Jeopardy Round

$400

Which statement accurately describes the k-Nearest Neighbor method?

A) An unsupervised technique to partition your data into k clusters
B) A supervised predictive technique that uses the k closest training observations to make a prediction for a test observation
C) A heuristic method of finding the k closest clusters to a given cluster
D) A special cross validation technique that uses one's k nearest validation observations to determine an error rate
E) A social distancing tactic where ones nearest neighbors stay at least k feet apart

$400

Which statement accurately describes the k-Nearest Neighbor method?

A) An unsupervised technique to partition your data into k clusters

B) A supervised predictive technique that uses the k closest training observations to make a prediction for a test observation

C) A heuristic method of finding the k closest clusters to a given cluster

D) A special cross validation technique that uses one's k nearest validation observations to determine an error rate

E) A social distancing tactic where ones nearest neighbors stay at least k feet apart

$400

Which of the following statements about the k-means algorithm are correct?

A) The k-means algorithm is sensitive to outliers

B) For different initializations, the k-means algorithm is expected to give the same clustering results.

C) The k-means algorithm will find the global minimum of it's objective function, which sums the squared distances of each point and its cluster centroid.

D) The output from k-means includes a dendrogram which allows you to see the evolution of the algorithm and how points were combined together

E) k-means j because it's opposite day.

$400

Which of the following statements about the k-means algorithm are correct?

A) The k-means algorithm is sensitive to outliers

B) For different initializations, the k-means algorithm is expected to give the same clustering results.

C) The k-means algorithm will find the global minimum of it's objective function, which sums the squared distances of each point and its cluster centroid.

D) The output from k-means includes a dendrogram which allows you to see the evolution of the algorithm and how points were combined together

E) k-means j because it's opposite day.

$800

Which of the following is an *advantage* of using decision trees over logistic regression?

A) A decision tree will always be more accurate than a logistic regression
B) Logistic regression has assumptions that require verification, decision trees do not.
C) Decision trees can handle observations that have missing values on input variables
D) Decision trees create response surface that is continuous (no holes or jumps) while logistic regression does not.
E) Both B and C
F) Trees turn carbon dioxide into oxygen. Let's see LR do that.

$800

Which of the following is an *advantage* of using decision trees over logistic regression?

A) A decision tree will always be more accurate than a logistic regression
B) Logistic regression has assumptions that require verification, decision trees do not.
C) Decision trees can handle observations that have missing values on input variables
D) Decision trees create response surface that is continuous (no holes or jumps) while logistic regression does not.
E) Both B and C
F) Trees turn carbon dioxide into oxygen. Let's see LR do that.

$1200

Which of the following is a *disadvantage* of using the k-Nearest Neighbor Method?

A) k-Nearest Neighbor models have rigid assumptions that require verification
B) k-Nearest Neighbor models do not offer any insight into variable importance
C) k-Nearest Neighbor models cannot use categorical variables as input
D) There is no good way to determine k for a k-Nearest Neighbor model.
E) 🤷‍♀️

$1200

Which of the following is a *disadvantage* of using the k-Nearest Neighbor Method?

A) k-Nearest Neighbor models have rigid assumptions that require verification

B) k-Nearest Neighbor models do not offer any insight into variable importance

C) k-Nearest Neighbor models cannot use categorical variables as input

D) There is no good way to determine k for a k-Nearest Neighbor model.

E) 🤷‍♀️

$1200

For which of the following tasks might clustering be a suitable approach?

A) Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.
B) Given historical weather records, predict if tomorrow's weather will be sunny or rainy.
C) Given information about connections on a social media platform, find out which users are most influential.
D) Given a database of information about your users, put them into groups for targeted advertising.
E) Laundry.

$1200

For which of the following tasks might clustering be a suitable approach?

A) Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

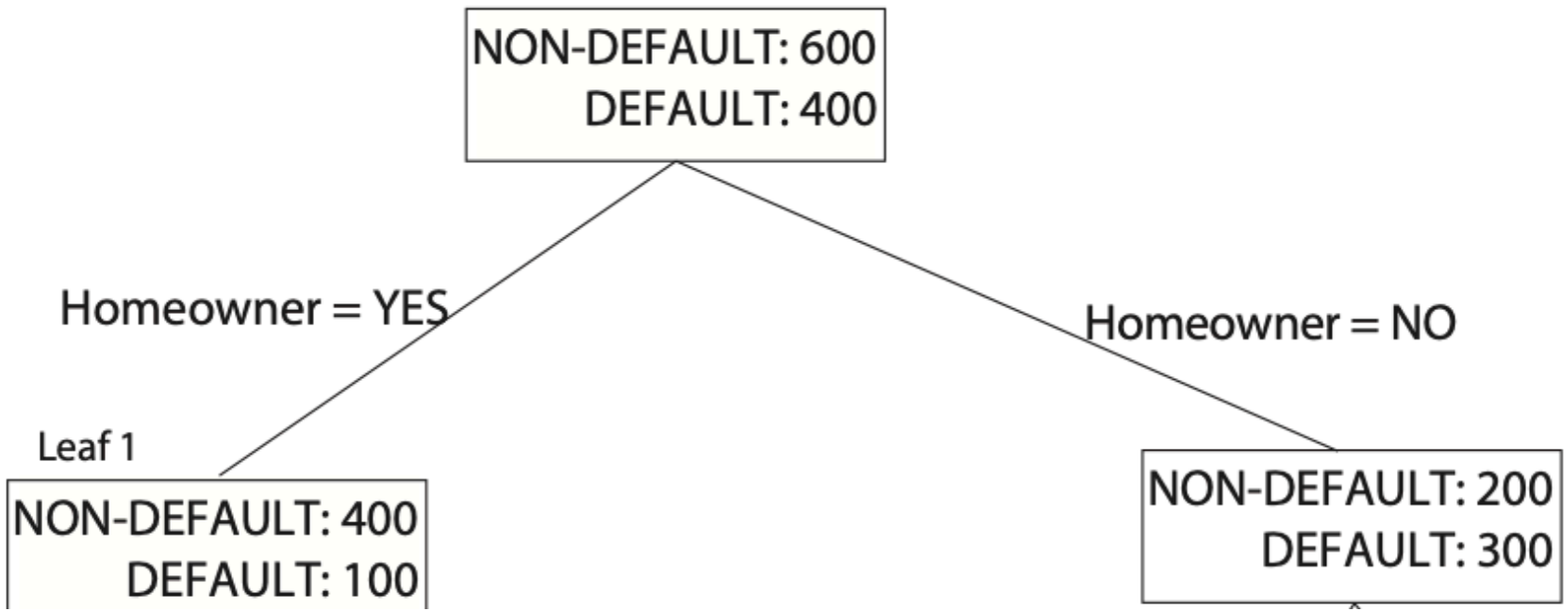
B) Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

C) Given information about connections on a social media platform, find out which users are most influential.

D) Given a database of information about your users, put them into groups for targeted advertising.

E) Laundry.

$1600



What is the *gain* in misclassification rate associated with the above split?

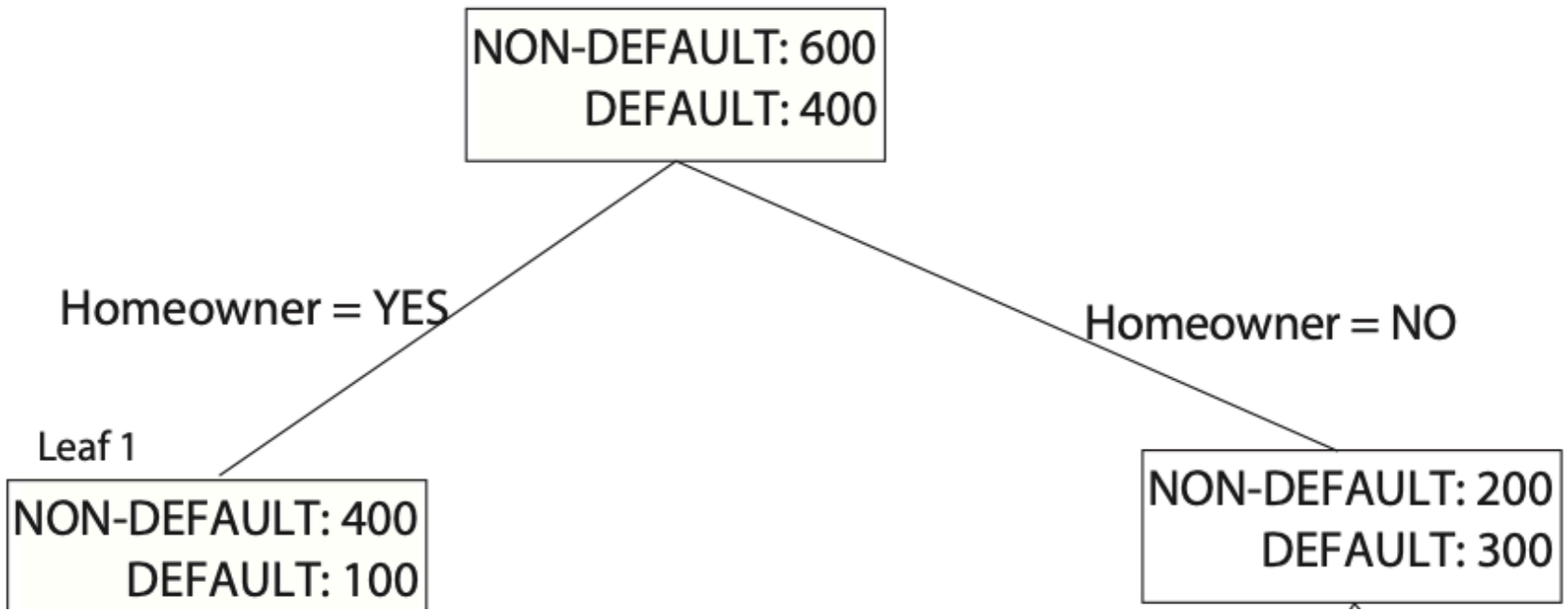
A) gain of 10%

B) gain of 30%

C) gain of 20%

D)  . final answer.

$1600



What is the *gain* in misclassification rate associated with the above split?

A) gain of 10% (40%)-[10%+20%]=10%

B) gain of 30%

C) gain of 20%

D)  . final answer.

$2000

What does it mean if a model has high variance?

A) It overfits the training data
B) It underfits the training data
C) The variance of the predicted values is higher than the variance of the input variables
D) If we trained the model with a different training set, the model parameters might be wildly different
E) Both A and D

$2000

What does it mean if a model has high variance?

A) It overfits the training data
B) It underfits the training data
C) The variance of the predicted values is higher than the variance of the input variables
D) If we trained the model with a different training set, the model parameters might be wildly different
E) Both A and D

$2000

The association rule A—>B has a confidence of 50% and a lift of 4.

What are the confidence and lift of the rule B—>A?

A) Confidence is 50%, lift is 4
B) Confidence is 50%, unable to know lift
C) Unable to know confidence, lift is 4
D) Confidence is 1/50%=2 and lift is 1/4=25%
E) Please stop.

$2000

The association rule A—>B has a confidence of 50% and a lift of 4.

What are the confidence and lift of the rule B—>A?

A) Confidence is 50%, lift is 4

B) Confidence is 50%, unable to know lift

C) Unable to know confidence, lift is 4

D) Confidence is 1/50%=2 and lift is 1/4=25%

E) Please stop.

Final Jeopardy

Category: Bias-Variance Tradeoff

How this works.

- This is the only place you can lose money
- You'll select a wager, W
- Some proportion, p, of you will get the question right.

IF p>0.80 then Payout = pW
ELSE Payout = - (1-p) W

Choose your Wager

(A) 0

(B) 2000

(C) 4000

(D) 5000

(E) 8000

Final Jeopardy Question

Which one of the following statements is `TRUE`?

A) For k-Nearest Neighbors, large of values of k correspond to higher variance models whereas small values of k correspond to higher bias models.

B) The purpose of pruning a decision tree is to reduce bias and increase variance.

C) If my model performs with a high level of accuracy on the training data but performs substantially worse on a validation dataset, then I should suspect my model suffers from high variance.

D) High variance is analogous to underfitting and high bias is analogous to overfitting.

E) A clustering solution with high bias will be easy to explain.

Final Jeopardy Question

Which one of the following statements is `TRUE`?

A) For k-Nearest Neighbors, large of values of k correspond to higher variance models whereas small values of k correspond to higher bias models.

B) The purpose of pruning a decision tree is to reduce bias and increase variance.

C) If my model performs with a high level of accuracy on the training data but performs substantially worse on a validation dataset, then I should suspect my model suffers from high variance.

D) High variance is analogous to underfitting and high bias is analogous to overfitting.

E) A clustering solution with high bias will be easy to explain.