

[Home](#) / [Dashboard](#) / [My Courses](#) / [AWS Certified Machine Learning Specialty](#) / [Modeling](#) / [Report](#)[← Back to the Course](#)

Level: Advanced

AWS Certified Machine Learning Specialty**Modeling**Completed on **Mon, 27 Jun 2022****1st**

Attempt

**11/15**

Marks Obtained

**73.33%**

Your Score

**0h 17m 2s**

Time Taken

**FAIL**

Result

Domain wise Quiz Performance ReportJoin us on **Slack community**

No.	Domain	Total Question	Correct	Incorrect	Unattempted
1	Modeling	15	11	4	0
Total	All Domains	15	11	4	0

Review the Answers

Filter By

[All Questions](#)**Question 1**

Incorrect

Domain: Modeling

You are a machine learning specialist at a large financial services firm. You are building a machine learning model to manage risk for your firm using data from your traders daily trading activity. You are in the stage of your model development where you need to provide jupyter notebooks to your development team that they can use in SageMaker Studio. Your developers need to use the Scala kernel based on the Almond Scala Kernel as their development environment in their jupyter notebooks.

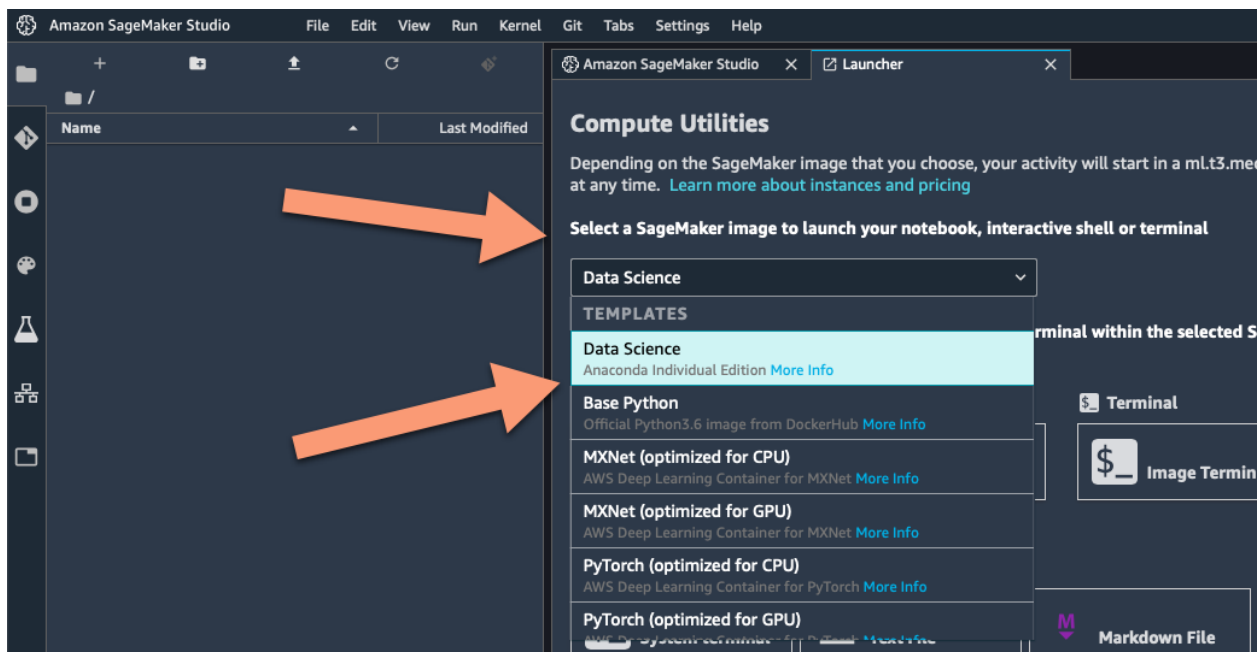
How can you provide the required development environment to your developers for their jupyter notebooks in the most efficient manner?

- ☐ A. Allow the developers to select the Scala kernel based on the Almond Scala Kernel from the list of included SageMaker images available in SageMaker studio. wrong
- B. Create a custom SageMaker image based on an AMI that includes the Scala kernel based on the Almond Scala Kernel and attach the image to your SageMaker domain.
- C. Create a custom SageMaker docker image using the Scala kernel based on the Almond Scala Kernel SageMaker custom image sample and attach the image to each user's profile.
- D. Create a custom SageMaker docker image using the Scala kernel based on the Almond Scala Kernel SageMaker custom image sample and attach the image to the SageMaker domain. right

Explanation:

Answer: D

Option A is incorrect. The default list of available images in SageMaker studio does not include a scala image.



Option B is incorrect. SageMaker Studio images are built on Docker images, not AMIs.

Option C is incorrect. While you should create a custom SageMaker docker image using the Scala kernel based on the Almond Scala Kernel SageMaker custom image sample, attaching it to each user's profile is less efficient than attaching it to the SageMaker Studio domain.

Option D is CORRECT. The most efficient option is to create a custom SageMaker docker image using the Scala kernel based on the Almond Scala Kernel SageMaker custom image sample and attach it to the SageMaker Studio domain.

Reference:

Please see the [Amazon Sagemaker developer guide](#) titled [Bring your own SageMaker image](#).

Please refer to the GitHub repository titled [SageMaker Studio Custom Image Samples](#).

Ask our Experts

 [View Queries](#)



Question 2

Correct

Domain: Modeling

You work as a machine learning specialist for a city that wants to monitor air quality to address air pollution in their environment. You and your machine learning specialist team need to forecast the city air quality in parts per million of contaminants over the next week, taking into account weather, traffic conditions, and other pollutant contributors. You are building your model using daily data from the last year as your data source. Your team has decided to use SageMaker Studio to leverage its collaborative notebooks feature.

Which model and SageMaker Studio image will provide the best results for your team in the most efficient manner?

- A. Use the SageMaker Studio Base TensorFlow [tensorflow-2.3.0] image and the k-Nearest-Neighbors algorithm on the single time series consisting of the full year of data with a predictor_type of regressor.
- B. Use SageMaker Studio Base R [r-4.0.3] image and the Random Cut Forest algorithm on the single time series consisting of the full year of data.
- ☒ C. Use the SageMaker Studio Base Python [python-3.6] image and the Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of regressor. right
- D. Use the SageMaker Studio Base Scala [scala-2.13.3] image and the Linear Learner algorithm on the single time series consisting of the full year of data with a predictor_type of classifier.

Explanation:

Answer: C

Option A is incorrect. The problem of forecasting the air quality is a regression problem best solved by the Linear Learner built-in algorithm, not the k-Nearest Neighbors SageMaker built-in algorithm. Also, there is no Base TensorFlow image available in SageMaker Studio. You would have to create a custom image to use TensorFlow in SageMaker Studio. This would be far less efficient than using an image that is already available with SageMaker Studio.

Option B is incorrect. The problem of forecasting the air quality is a regression problem best solved by the Linear Learner built-in algorithm, not the Random Cut Forest SageMaker built-in

algorithm. Also, there is no Base R image available in SageMaker Studio. You would have to create a custom image to use R in SageMaker Studio. This would be far less efficient than using an image that is already available with SageMaker Studio.

Option C is CORRECT. The Linear Learner SageMaker built-in algorithm is the best choice from the options given to solve this regression problem. Also, the SageMaker Studio Base Python [python-3.6] image is a valid choice of an available SageMaker Studio image.

Option D is incorrect. While the Linear Learner SageMaker built-in algorithm is the best choice from the options given to solve this regression problem, there is no Base Scala image available in SageMaker Studio. You would have to create a custom image to use Scala in SageMaker Studio. This would be far less efficient than using an image that is already available with SageMaker Studio.

Reference:

Please see the **AWS SageMaker developer guide** titled **Available Amazon SageMaker Images**.

Please refer to the **AWS SageMaker developer guide** titled **Bring your own SageMaker image**.

Ask our Experts

 [View Queries](#)



Question 3

Incorrect

Domain: Modeling

You are a machine learning specialist working for a large retail clothing company. Your marketing department wants to leverage machine learning to understand product loyalty. Your machine learning team has decided to group the customers into categories based on which customers may churn, meaning abandon a given product for another, maybe a competitor's product, within the next 6 months. You have labeled data for customer product loyalty for the previous two years available to you for training your model.

Which machine learning model type should you and your team use to build your customer loyalty model?

- A. Linear regression using either the Linear Learner or XGBoost built-in SageMaker algorithms
- ☒ B. Clustering using either the K-Means or K-Nearest Neighbors algorithms wrong
- C. Classification using either the Linear Learner or XGBoost built-in SageMaker algorithms right
- D. Reinforcement learning using SageMaker RL

Explanation:

Answer: C

Option A is incorrect. Because you have labeled data and you want to group your customers into categories, classification is a better choice than regression.

Option B is incorrect. Clustering is done using unlabeled data. You have labeled data. So you will get better results using a classification algorithm.

Option C is CORRECT. Since your data is labeled, you will get the best results using a classification algorithm when attempting to classify your customers.

Option D is incorrect. Reinforcement algorithms are used for solving problems such as supply chain management, HVAC systems, industrial robotics, game artificial intelligence, dialog systems, and autonomous vehicles. They attempt to learn a strategy. They are not well suited to grouping data elements into categories.

Reference:

Please see the [Amazon SageMaker developer guide](#) titled [Use Amazon SageMaker Built-in Algorithms](#).

Please refer to the [article](#) titled [Regression vs Classification in Machine Learning: What is The Difference?](#).

Please review the [article](#) titled [How to Use Unlabeled Data in Machine Learning](#).

Please refer to the [Amazon SageMaker developer guide](#) titled [Use reinforcement learning with Amazon SageMaker](#).

Ask our Experts

 [View Queries](#)



Question 4

Correct

Domain: Modeling

Your machine learning team is building and planning to operationalize a machine learning model that uses deep learning to recognize and classify images of potential security threats to important government officials using stills taken from live security video. However, when your team runs their mini-batch training of the neural network, the training accuracy oscillates over your training epochs.

What is the most probable cause of your training accuracy problem?

A. The validation error has stopped decreasing.

B. The mini-batch size is too small.

C. The mini-batch size is too large.

☒ D. The learning rate is very high. right

Explanation:

Answer: D

Option A is incorrect. In deep learning training, choosing the correct number of epochs is important. The validation error is used to determine how many epochs to run through. When the learning rate stops decreasing, you should stop running training epochs. The point when your validation error stops decreasing has no correlation to the oscillation of the accuracy of your training epochs.

Option B is incorrect. A small mini-batch is used to prevent your training process from stopping at local minima. Having a small mini-batch size won't cause oscillation in your training epoch accuracy.

Option C is incorrect. A large mini-batch size is used to allow for highly computational demanding matrix multiplication in your training calculations. Having a large mini-batch size won't cause oscillation in your training epoch accuracy.

Option D is CORRECT. A very high learning rate tends to cause oscillation in your training accuracy. A high learning rate causes your weight updates to be too large, and you will overestimate your goal and oscillate around the true goal.

Reference:

Please see the [Amazon SageMaker developer guide](#) titled **DeepAR Forecasting Algorithm**.

Please refer to the [Machine Learning Mastery article](#) titled **How to Configure the Learning Rate When Training Deep Learning Neural Networks**.

Please review the [article](#) titled **Hyperparameters in Machine /Deep Learning**.

Please refer to the [Towards Data Science article](#) titled **Hyper-parameter Tuning Techniques in Deep Learning**.

Ask our Experts

 [View Queries](#)



Question 5

Correct

Domain: Modeling

You are a machine learning specialist on a software development team in a real estate company. Your management team has asked your team to build a logistic regression model that your company wishes to use to predict whether or not a person will buy a given listing based on multiple attributes of the sale, the property, and the customer profile. Your team lead has assigned your team to find the optimal model with an ideal classification threshold.

Which model evaluation technique should your team use to discover how different classification thresholds will affect the model's performance?

- A. Rand Index
- B. Root Mean Square Error (RMSE)
- C. Mean Absolute Error (MAE)
- ☒ D. Receiver Operating Characteristic (ROC) curve right

Explanation:

Answer: D

Option A is incorrect. The Rand Index evaluation technique is used for optimizing unsupervised models, not supervised models like logistic regression models.

Option B is incorrect. The RMSE evaluation technique is used for regression problems where you are solving for a continuous variable. In this use case, you are solving a binary classification: will or will not buy.

Option C is incorrect. The MAE evaluation technique is also used for regression problems where you are solving for a continuous variable. In this use case, you are solving a binary classification: will or will not buy.

Option D is CORRECT. The Receiver Operating Characteristic curve evaluation technique is used for regression problems where you are solving for a binary variable. In this use case, you are solving a binary classification: will or will not buy.

Reference:

Please see the [Google Machine Learning Crash Course](#) article titled **Classification: ROC Curve and AUC**.

Please refer to the [Towards Data Science](#) article titled **Understanding AUC - ROC Curve**.

Please review the [Data Institute](#) article titled **Choosing the Right Metric for Evaluating Machine Learning Models — Part 1**.

Ask our Experts

 [View Queries](#)



Question 6

Correct

Domain: Modeling

You are a machine learning specialist at a pharmaceutical drug manufacturer. Your team has the task of building a deep learning model to be used for drug discovery by combining data from various

sources. Your team will use the deep learning model to predict novel candidate biomolecules for new disease targets such as COVID 19. You and your team have created a deep learning neural network disease target prediction model that performs well on the training data. However, it performs poorly on your test data.

Which of the methods listed should your team use to correct your testing problem? (SELECT THREE)

- ☐ A. Increase dropout right
- ☐ B. Decrease regularization
- ☐ C. Increase regularization right
- ☐ D. Decrease dropout
- ☐ E. Decrease feature combinations right
- ☐ F. Increase feature combinations

Explanation:

Answers: A, C and E

Option A is CORRECT. When your model performs well in training but poorly in testing, it quite often means the model is overfitted or not generalized enough. Increasing dropout, or the probability of a node to be turned off, helps generalization.

Option B is incorrect. You need to increase regularization to avoid overfitting, a common problem when your training data is “memorized” by your neural network. Decreasing regularization is used when your model is underfitting.

Option C is CORRECT. You need to increase regularization to avoid overfitting, a common problem when your training data is “memorized” by your neural network. Increasing regularization helps generalization.

Option D is incorrect. Decreasing dropout will not help generalization. It will have the opposite effect.

Option E is CORRECT. Decreasing feature combinations and combining the strength of multiple complementary features to yield more powerful features (reducing the number of features) will help with generalization.

Option F is incorrect. Increasing feature combinations is used when your model is underfitting. Decreasing feature combinations, combining the strength of multiple complementary features to yield more powerful features (reducing the number of features) will help with generalization.

Reference:

Please see the [Amazon Machine Learning developer guide](#) titled **Model Fit: Underfitting vs. Overfitting**.

Please refer to the [Machine Learning Mastery article](#) titled **A Gentle Introduction to Dropout for Regularizing Deep Neural Networks**.

Ask our Experts

[+ View Queries](#)

Question 7

Correct

Domain: Modeling

You work as a machine learning specialist for a publishing company. The company has labeled a historical dataset of publication sales data. Using this labeled data, you need to predict how many copies of a given publication should be produced each month.

Which machine learning algorithm type should you use to generate your predictions?

- ☒ A. Linear regression right
- ☐ B. Principal component analysis
- ☐ C. Random Cut Forest
- ☐ D. Logistic regression

Explanation:

Answer: A

Option A is CORRECT. You are trying to solve a “how many” question, and your data is labeled. These two factors lead to the choice of linear regression as the best option from those given.

Option B is incorrect. The **principal** component analysis is used for dimensionality reduction, not for solving predictions of “how many” problems. Also, it is an unsupervised algorithm. We have labeled data. So we should use a supervised algorithm.

Option C is incorrect. The random cut forest is used primarily as an unsupervised algorithm for detecting anomalous data points within a data set. Since we have labeled data, we will use a supervised algorithm.

Option D is incorrect. Logistic regression is used to solve “yes/no” or binary predictions, not “how many” predictions.

Reference:

Please see the **Amazon Machine Learning developer guide** titled **Regression Model Insights**.

Please refer to the **Amazon SageMaker developer guide** titled **Random Cut Forest (RCF) Algorithm**.

Please refer to the **Amazon SageMaker developer guide** titled **Principal Component Analysis (PCA) Algorithm**.

Ask our Experts

[+ View Queries](#)

Question 8

Correct

Domain: Modeling

You are the technical leader of your machine learning team where you are responsible for the quality of your team's code. Your team generates a very high level of code per week, thousands of lines of Java code used in machine learning data ingestion, transformation, as well as EMR cluster jobs. You need a way to improve the pull request review process. So you have decided to use CodeGuru Reviewer to leverage its program analysis and machine learning to detect defects that your developers may fail to find while also leveraging its suggestions for improving the team's Java code. To this end, you want to use CodeGuru to create code reviews. You are using GitHub Enterprise Server for your code repository.

When your developers commit code to your GitHub repository, you want the CodeGuru Reviewer code review to run. What do you need to do to make this workflow possible?

- A. Perform your own input validation for your developers code.
- B. Select the type of regression SageMaker built-in algorithm to use in the machine learning defect detection.
- C. Select the type of clustering SageMaker built-in algorithm to use in the machine learning defect detection.
- ☒ D. Create a connection to your repository using AWS CodeStar connections to connect the CodeGuru Reviewer service to your GitHub repository. right

Explanation:**Answer: D**

Option A is incorrect. CodeGuru Reviewer identifies problems with input validation for you using its machine learning capabilities.

Option B is incorrect. You don't need to select the machine learning algorithm used by CodeGuru Reviewer.

Option C is incorrect. You don't need to select the machine learning algorithm used by CodeGuru Reviewer.

Option D is CORRECT. If you use GitHub Enterprise Server as your code repository, you are required to create a connection to your repository using AWS CodeStar connections to connect the CodeGuru Reviewer service to your GitHub repository.

Reference:

Please see the **Amazon CodeGuru Reviewer user guide** titled **What is Amazon CodeGuru Reviewer?**.

Please see the **Amazon CodeGuru Reviewer user guide** titled **Create a repository for your source code**.

Please see the **Developer Tools console user guide** titled **What are connections?**.

Ask our Experts

 [View Queries](#)



Question 9

Incorrect

Domain: Modeling

You are a machine learning specialist at a large bank. Your machine learning team has recently been assigned the task of detecting fraud in the bank's web and mobile applications. Your management team is excited about using machine learning for fraud detection. But they have limited money in the yearly budget for this work.

You have decided to use the Amazon Fraud Detector service to deliver your fraud detection layer in your web and mobile architecture. When building your Fraud Detector model, which model type should you choose?

- ☒ A. ONLINE_FRAUD_DETECTOR wrong
- ☐ B. ONLINE_FRAUD_INSIGHTS right
- ☐ C. FRAUD_INSIGHTS
- ☐ D. FRAUD_DECTOR

Explanation:

Answer: B

Option A is incorrect. The model type available for the Fraud Detector service is the ONLINE_FRAUD_INSIGHTS model.

Option B is CORRECT. The model available in the Fraud Detector service is the ONLINE_FRAUD_INSIGHTS model.

Option C is incorrect. The model type available for the Fraud Detector service is the ONLINE_FRAUD_INSIGHTS model.

Option D is incorrect. The model type available for the Fraud Detector service is the ONLINE_FRAUD_INSIGHTS model.

Reference:

Please see the [Amazon Fraud Detector user guide](#) titled [How Amazon Fraud Detector works](#).

Please see the [Amazon Fraud Detector welcome page](#) titled [CreateModel](#).

Please see the [Amazon Fraud Detector FAQs](#).

Ask our Experts

 [View Queries](#)



Question 10

Incorrect

Domain: Modeling

Your machine learning team is part of the research department of a hedge fund firm. Your team has been assigned a project to forecast the price movement of several stocks in the NASDAQ index. You have decided to use historical related time series in your model to improve the accuracy of your model. Your management team has asked that your team produces the model quickly and at a low administrative overhead. So your team lead has decided to use the Amazon Forecast service.

Which Amazon Forecast algorithm would be the best choice for your stock price movement forecasting problem?

- A. Prophet
- ☒ B. DeepAR+ wrong
- C. ARIMA
- D. CNN-QR right

Explanation:

Answer: D

Option A is incorrect. The Amazon Forecast Prophet algorithm does not accept related time series data without future values.

Option B is incorrect. The Amazon Forecast DeepAR+ algorithm does not accept related time series data without future values.

Option C is incorrect. The Amazon Forecast ARIMA algorithm does not accept related time series data without future values.

Option D is CORRECT. The Amazon Forecast CNN-QR algorithm is the only Forecast algorithm that accepts related time series data without future values.

Reference:

Please see the [Amazon Forecast developer guide](#) titled [Getting Started](#).

Please see the **Amazon Forecast developer guide** titled **Choosing an Amazon Forecast Algorithm**.

Please see the **Amazon Forecast developer guide** titled **Using Related Time Series Datasets**.

Ask our Experts

 [View Queries](#)



Question 11

Correct

Domain: Modeling

You work as a machine learning specialist for an analytics consulting firm that produces machine learning models for businesses that wish to understand the effects of social media on their product sales. Your latest assignment is to build a model that predicts whether a user will click-through an advertisement on a set of social media apps. For this problem, you have hundreds of millions of observations with hundreds of features. Which type of algorithm should you use to meet your business problem?

- ☐ A. Neural network with a small number of hidden layers
- ☐ B. Logistic regression
- ☐ C. Clustering
- ☒ D. Neural network with a large number of hidden layers right

Explanation:

Correct Answer: D

Option A is incorrect. A neural network with a small number of hidden layers will not perform well with hundreds of millions of observations with hundreds of features.

Option B is incorrect. A logistic regression algorithm will not perform well with hundreds of millions of observations with hundreds of features.

Option C is incorrect. Clustering is not the best choice of algorithm for a problem where you are trying to solve for a binary target, click-through or not.

Option D is correct. For a problem where you have hundreds of millions of observations with hundreds of features, a neural network with a large number of hidden layers will perform the best.

Reference:

Please see the Machine Learning Yearning by Andrew Ng, chapter 4: **Scale Drives Machine Learning Progress** (<https://github.com/ajaymache/machine-learning-yearning>), and the Towards Data Science article titled **Machine Learning vs. Deep Learning** (<https://towardsdatascience.com/machine-learning-vs-deep-learning-62137a1c9842>),

the Statistics Solutions article titled **What is Logistic Regression?** (<https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-logistic-regression/>), the Investopedia article titled **Neural Network** (<https://www.investopedia.com/terms/n/neuralnetwork.asp>), the Amazon SageMaker developer guide titled **Linear Learner Algorithm** (<https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html>)

Ask our Experts

 [View Queries](#)



Question 12

Correct

Domain: Modeling

You work as a machine learning specialist for a security trading trading firm where you are responsible for building a machine learning model that can predict the price movement of a given stock throughout the trading day. You have produced several models and you now need to select the best model for your machine learning problem. You are using scikit-learn to implement your evaluation process. Which validation technique should you use to determine the best model?

- A. k-Fold Cross-Validation (k-Fold CV) using the scikit-learn KFold method
- B. Leave-one-out Cross-Validation (LOOCV) using the scikit-learn LeaveOneOut method
- ☒ C. Time Series Cross-Validation using the scikit-learn TimeSeriesSplit method right
- D. Bayesian optimization using the scikit-learn BayesianSearchCV method

Explanation:

Correct Answer: C

Option A is incorrect. k-Fold Cross-Validation is the most used model validation technique. However, you are working with time series data (you are predicting price movement over time). Therefore, Time Series Cross-Validation is a better choice.

Option B is incorrect. Leave-one-out Cross-Validation does not inherently handle time series data. Time Series Cross-Validation is a better choice.

Option C is correct. Using the TimeSeriesSplit scikit-learn method for your cross-validation will give you the best results. You are predicting price movement over time.

Option D is incorrect. Bayesian optimization is used for optimizing hyperparameters.

Reference:

Please see the Machine Learning Mastery article titled **A Gentle Introduction to Model Selection for Machine Learning** (<https://machinelearningmastery.com/a-gentle-introduction-to-model->

[selection-for-machine-learning/](#)), and the Machine Learning Mastery article titled **A Gentle Introduction to k-fold Cross-Validation** (<https://machinelearningmastery.com/k-fold-cross-validation/>), the Towards Data Science article titled **Validating your Machine Learning Model** (<https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7>), the Towards Data Science article titled **Using the latest advancements in deep learning to predict stock price movements** (<https://towardsdatascience.com/aifortrading-2edd6fac689d>), the scikit-optimize page titled **skopt.BayesSearchCV** (<https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html>)

Ask our Experts

 [View Queries](#)



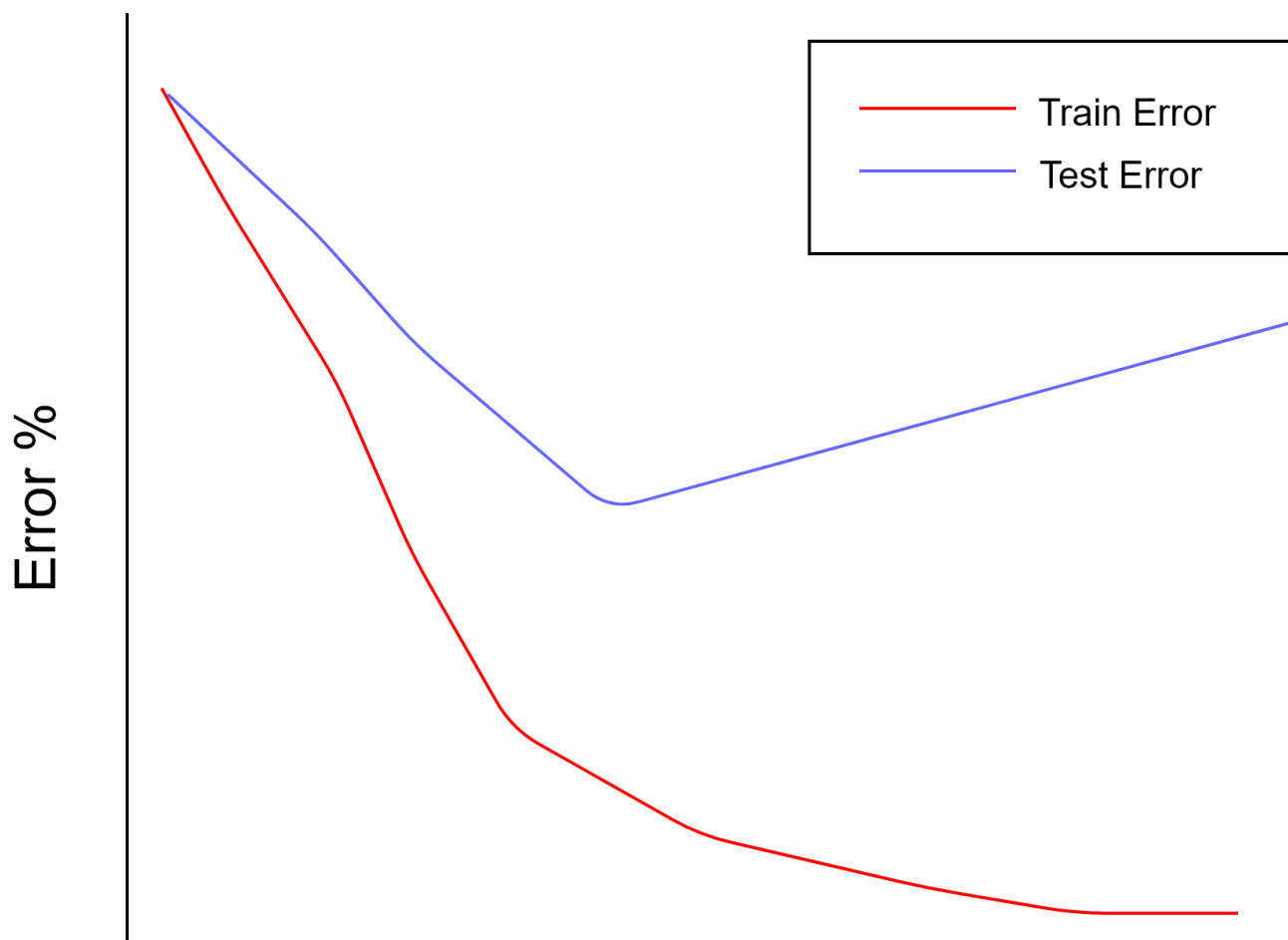
Question 13

Correct

Domain: Modeling

0

You work as a machine learning specialist for a software company that produces an auction website where users can buy art work through an auction process. Your task as a machine learning specialist for the company is to produce a machine learning model that estimates the value of the various art products put up for auction. You have decided to use a deep learning model to produce your estimates. Your data source has many features, such as artist, artist selling history, category, category selling history, rarity of product, etc. Some of these features have outliers. In your training you have realized that you have an overfitting problem. You have graphed your training error and testing error:



You need to address your model overfitting. You have decided to use regularization to address your overfitting problem. Which regularization technique best fits your situation?

- ☒ A. Lasso regularization right
- ☐ B. Ridge regularization
- ☐ C. Dropout
- ☐ D. Early stopping

Explanation:

Correct Answer: A

Option A is correct. Lasso regularization (L1) is the best choice here because you have outliers in your feature data. L1 regularization handles outliers well.

Option B is incorrect. Ridge regularization (L2) does not handle outliers as well as L1 regularization.

Option C is incorrect. Using dropout with outliers is more effort than using L1.

Option D is incorrect. Early stopping will not address your outlier problem.

Reference:

Please see the Neptune blog titled **Fighting Overfitting with L1 or L2 Regularization – Which One Is Better?** (<https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization>), and the Machine

Learning Mastery article titled **How to Avoid Overfitting in Deep Learning Neural Networks** (<https://machinelearningmastery.com/introduction-to-regularization-to-reduce-overfitting-and-improve-generalization-error/>), the Data Driven Investor article titled **L1 and L2 Regularization** (<https://medium.datadriveninvestor.com/l1-l2-regularization-7f1b4fe948f2>)

Ask our Experts

 View Queries



Question 14

Correct

Domain: Modeling

You work as a machine learning specialist for an analytics firm that produces machine learning models for clients that want to purchase data analysis on things like estimates for efficacy of advertising campaigns. You are currently working on an estimator for the effectiveness of a proposed direct mailing campaign. You have gathered your data, performed feature engineering and chosen the XGBoost algorithm for your model. Now you are ready to tune your hyperparameters for your model training. Which configuration strategy will give you the best model performance?

- A. Large learning rate, small number of estimators, without early stopping.
- B. Large learning rate, large number of estimators, with early stopping.
- ☒ C. Small learning rate, large number of estimators, with early stopping. right
- D. Small learning rate, large number of estimators, without early stopping.

Explanation:

Correct Answer: C

Option A is incorrect. A large learning rate with a small number of estimators without using early stopping will cause your model to oscillate.

Option B is incorrect. A large number of estimators with early stopping is a good set of configurations, but a large rate will cause your model to oscillate.

Option C is correct. Using a small learning rate, a large number of estimators, with early stopping will allow your model to find the correct number of estimators.

Option D is incorrect. Using a small learning rate, a large number of estimators without early stopping will make your training run very long and probably overfit your data.

Reference:

Please see the Kaggle article titled **XGBoost** (<https://www.kaggle.com/alexisbcook/xgboost>), and the Amazon SageMaker developer guide titled **XGBoost Hyperparameters** (https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost_hyperparameters.html)

Ask our Experts

[+ View Queries](#)

Question 15

Correct

Domain: Modeling

You work as a machine learning specialist for a retail marketing firm. You are responsible for the machine learning models used for product marketing. Your latest assignment has you building a model to predict whether or not a particular marketing campaign will benefit from social media advertising. You have gathered your social media and product marketing data and selected your model algorithm. You are now in the process of evaluating your model. You are using the scikit-learn `model_selection` package in a pipeline using cross validation to select the best performing model. When setting the parameters for your scoring your cross validation runs, which scoring technique should you use?

- ☒ A. `f1` right
- B. `adjusted_mutual_info_score`
- C. `rand_score`
- D. `completeness_score`

Explanation:

Correct Answer: A

Option A is correct. The `f1` scoring metric is used for binary targets. Your target is binary: *predict whether or not* a particular marketing campaign will benefit from social media advertising.

Option B is incorrect. The `adjusted_mutual_info_score` metric is used in clustering problems. You are not solving a clustering problem.

Option C is incorrect. The `rand_score` metric is used in clustering problems. You are not solving a clustering problem.

Option D is incorrect. The `completeness_score` metric is used in clustering problems. You are not solving a clustering problem.

Reference:

Please see the Kaggle article titled **Cross-Validation** (<https://www.kaggle.com/alexisbcook/cross-validation>), the Scikit-learn page titled **3.3. Metrics and scoring: quantifying the quality of predictions** (https://scikit-learn.org/stable/modules/model_evaluation.html), the Scikit-learn page titled **`sklearn.metrics.adjusted_mutual_info_score`** (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_mutual_info_score.html#sklearn.m

etrics.adjusted_mutual_info_score), the Scikit-learn page titled **sklearn.metrics.rand_score** (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.rand_score.html#sklearn.metrics.rand_score), the Scikit-learn page titled **sklearn.metrics.completeness_score** (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.completeness_score.html#sklearn.metrics.completeness_score), the Scikit-learn page titled **sklearn.metrics.f1_score** (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score)

Ask our Experts

 [View Queries](#)



Finish Review

Certification

Cloud Certification

Java Certification

PM Certification

Big Data Certification

Support

Contact Us

Help Topics

Company


Become Our Instructor

Support

Discussions

Blog

Business

 **Join us on Slack!**

Join our open **Slack community** and get your queries answered instantly! Our experts are online to answer your questions!