Questionnaire (9th chapter)

Total de pontos 0/29

O e-mail do participante (**breno.xavier@ufpe.br**) foi registrado durante o envio deste formulário.

✗ What is a continuous variable? * .../1

It is a type of numerical variable that can take on any value within a range of possible values, typically represented by real numbers

✗ What is a categorical variable? * .../1

It is a type of variable that represents data with a limited number of possible values, typically expressed as labels or categories

✗ Provide two of the words that are used for the possible values of a categorical variable. *.../1

labels and categories

✗ What is a "dense layer"? * .../1

It refers to a type of neural network layer where each neuron in the layer is connected to every neuron in the previous layer, and each connection is assigned a weight

✗ How do entity embeddings reduce memory usage and speed up neural networks? *.../1

Entity embeddings help neural networks to be faster and use less memory by converting categorical variables into smaller and simpler vectors, which means that the network needs to learn fewer things.

✗ What kinds of datasets are entity embeddings especially useful for? * .../1

Is useful for datasets with high cardinality categorical variables, meaning those with a large number of unique categories

✗ What are the two main families of machine learning algorithms? * .../1

supervised learning and unsupervised learning

✗ Why do some categorical columns need a special ordering in their classes? How do you do this in Pandas? *.../1

se the Categorical data type with the ordered=True parameter and specify the desired order in the categories parameter.

✗ Summarize what a decision tree algorithm does. * .../1

It recursively based on decision rules, optimizing each split for information gain or Gini index, resulting in a tree structure for classification or regression

✗ Why is a date different from a regular categorical or continuous variable, and how can you preprocess it to allow it to be used in a model? *.../1

A date is different from a regular categorical or continuous variable because it has a temporal aspect; it can be preprocessed by extracting features such as day, month, and year, or converting it to a numerical value like Unix time

✗ Should you pick a random validation set in the bulldozer competition? If no, what kind of validation set should you pick? *.../1

No, you should not pick a random validation set in the bulldozer competition, but rather a time-based split since the data is ordered by date

✗ What is pickle and what is it useful for? * .../1

Pickle is a Python module that allows for the serialization and deserialization of Python objects, making it useful for saving and loading models or data structures

✗ How are `mse`, `samples`, and `values` calculated in the decision tree drawn in this chapter? *.../1

In a decision tree, mse measures the mean squared error, samples is the number of samples in the node, and values is the predicted value for the node

✗ How do we deal with outliers, before building a decision tree? * .../1

outliers can be removed or transformed using techniques such as winsorizing, trimming, or capping

✗ How do we handle categorical variables in a decision tree? * .../1

Categorical variables can be handled in a decision tree by using one-hot encoding, label encoding, or ordinal encoding

✗ What is bagging? * .../1

it is an ensemble learning technique that combines multiple models trained on bootstrap samples of the data to reduce variance and improve generalization

✗ What is the difference between `max_samples` and `max_features` when creating a random forest? *.../1

max_samples specifies the maximum number of samples used in each bootstrap sample, while max_features specifies the maximum number of features considered at each split in each tree

✗ If you increase `n_estimators` to a very high value, can that lead to overfitting? Why or why not? *.../1

Increasing n_estimators to a very high value can reduce overfitting by averaging the predictions of many trees, but can also increase computation time and memory usage

✗ In the section "Creating a Random Forest", just after <<max_features>>, why did `preds.mean(0)` give the same result as our random forest? *.../1

preds.mean(0) gave the same result as the random forest because it takes the mean of the predictions across all trees, which is equivalent to the sum of the votes divided by the number of trees

✗ What is "out-of-bag-error"? * .../1

the estimation of the error rate of a random forest using only the samples that were not used to train each individual tree

✗ Make a list of reasons why a model's validation set error might be worse than the OOB error. How could you test your hypotheses? *.../1

Different distribution of data, data leakage, or a smaller sample size; hypotheses could be tested by comparing the error on different subsets of the data or using cross-validation

✗ What's the purpose of removing unimportant variables? * .../1

It can reduce overfitting, improve model interpretability, and speed up training and inference

✗ What's a good type of plot for showing tree interpreter results? * .../1

bar plot or waterfall plot, which displays the contribution of each feature to the final prediction

✗ What is the "extrapolation problem"? * .../1

The problem occurs when a model is used to make predictions outside of the range of the training data, which can lead to unreliable or erroneous result

✗ How can you tell if your test or validation set is distributed in a different way than your training set? *.../1

You can test if your distributed differently than your training set by comparing the distribution of feature vlues or target values using summary statistics or visualization

✗ Why do we ensure `saleElapsed` is a continuous variable, even although it has less than 9,000 distinct values? *.../1

should be a continuous variable to preserve the temporal relationship between dates and to allow for interpolation or extrapolation between date

✗ What is "boosting"? * .../1

it is ensemble learning technique that combines multiple weak models trained sequentially on weighted versions of the data, with each model focused on the samples that were misclassified by the previous models

✗ How could we use embeddings with a random forest? Would we expect this to help? *.../1

can be used with a random forest by replacing categorical variables with their corresponding embeddings and using them as continuous features in the model; this could help capture nonlinear relationships and interactions between categorical variables

✗ Why might we not always use a neural net for tabular modeling? * .../1

because it may not be the most efficient or interpretable model for the given dataset or task, and simpler models such as decision trees or linear models may suffice