1. What are the steps that you have followed in your last project to prepare the dataset?

Answer: The choice of data entirely depends on the problem you’re trying to solve but, in my case, I have followed these 5 steps to prepare the dataset;

Step1: Gathering the data

The quantity of the data is important but not as important as the quality of it.

Step2: Handling missing data

This is one of the hardest step and handling missing data in the wrong way can cause disasters.

Step3: Taking data further with the feature extraction

Feature extraction can be a turning point. It is what makes a dataset unique. Getting insight by making relations between features is important thing.

Step4: Deciding which key factors are important

AI is able to decide which features truly affect the output and which doesn’t. On the downside, The more data you give your model, it costs you money (computer power) & time. Both not always available. So, Giving your program a little help isn’t always a bad idea. If you’re sure that a certain feature is completely unrelated to the output, you should just disregard it altogether.

Step5: Splitting the data into training & testing sets

The data is 80–20 percent training & testing sets respectively. The 20% for the test set has engineered in a way that they’re not just randomly cut out of the dataset.

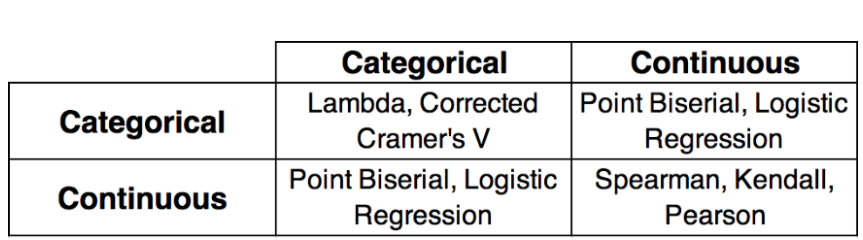
1. In your last project what steps were involved in model selection procedure?

Answer: As considering the data-rich situation, the approach is selected where we randomly divide the dataset into three parts: training set, validation set, and, test set. Where training set was used to fit the models; the validation set was used to estimate the prediction error for model selection; and finally, the test set was used for assessment of the generalization error of the final chosen model.

We used k-fold cross-validation that splits the training dataset into k folds, where each example appears in a test set only once. At last, we have finalized the model based on test data score.

1. If I give you 2 columns of any dataset, what will be the steps will be involved to check the relationship between those 2 columns?

Answer: The first step will be to check the columns contain which kind of data type such as; Continuous or categorical. So that we can check correlation between categorical and numerical variables.

**

1. Can you please explain 5 diff kind of strategies at least to handle missing values in dataset?

Answer:

1. Deleting Rows with missing values: Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.
2. Impute missing values with mean/median: Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.
3. Impute missing values for categorical variable: When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large then it can be replaced with a new category.
4. Missing values imputation using k-NN:: The k nearest neighbours is an algorithm that is used for simple classification. The algorithm uses ‘feature similarity’ to predict the values of any new data points.
5. Imputation using Deep Learning Library (Datawig): This method works very well with categorical and non-numerical features. It is a library that learns Machine Learning models using Deep Neural Networks to impute missing values in a dataframe. It also supports both CPU and GPU for training.
6. What kind of diff. issues you have faced wrt your raw data? At least mention 5 issues.

Answer:

1. Getting data from multiple sources
2. Unlocking value out of Unstructured Text Data
3. Setting up the infrastructure and velocity of data
4. Adapting to different tools to collect unstructured data
5. Building a robust strategy before collecting data
6. What is your strategy to handle categorical dataset? Explain with example.

Answer: Categorical features have a lot to say about the dataset thus it should be converted to numerical to make it into a machine-readable format.

Two major types of categorical features are

* **Nominal** – These are variables which are not related to each other in any order such as colour (black, blue, green).
* **Ordinal** – These are variables where a certain order can be found between them such as student grades (A, B, C, D, Fail).

Encoding Categorical Variables is main approach to handle categorical dataset.

1. How do you define a model in terms of machine learning or in your own word?

Answer: Machine learning is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.

1. What do you understand by k fold validation & in what situation you have used k fold cross validation?

Answer: We used k-fold cross-validation that splits the training dataset into k folds, where each example appears in a test set only once. At last, we have finalized the model based on test data score in model selection procedure.

1. What is meaning of bootstrap sampling? explain me in your own word.

Answer: Bootstrap Sampling is a method that involves drawing of sample data repeatedly with replacement from a data source to estimate a population parameter.

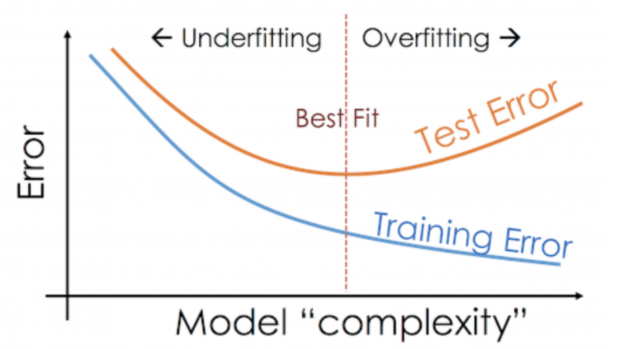
Bootstrap sampling is used in a machine learning ensemble algorithm called bootstrap aggregating (also called bagging). It helps in avoiding overfitting and improves the stability of machine learning algorithms.

In bagging, a certain number of equally sized subsets of a dataset are extracted with replacement. Then, a machine learning algorithm is applied to each of these subsets and the outputs are ensembled.

1. What do you understand by underfitting & overfitting of model with example?

Answer: The situation where any given model is performing too well on the training data but the performance drops significantly over the test set is called an overfitting model.

For example, non-parametric models like decision trees, KNN, and other tree-based algorithms are very prone to overfitting. These models can learn very complex relations which can result in overfitting.



On the other hand, if the model is performing poorly over the test and the train set, then we call that an underfitting model. An example of this situation would be building a linear regression model over non-linear data.

1. What is diff between cross validation and bootstrapping?

Answer: Bootstrapping is a technique that helps in many situations like validation of a predictive model performance, ensemble methods, estimation of bias and variance of the model. It works by sampling with replacement from the original data, and take the “not chosen” data points as test cases. We can make this several times and calculate the average score as estimation of our model performance.

In addition, Bootstrapping helps in ensemble methods as we may build a model (like a Decision tree) using each bootstrap data set and “bag” these models in an ensemble (like Random Forest) and take the majority voting for all of these models as our resulting classification.

On the other hand, cross validation is a technique for validating the model performance, and it’s done by split the training data into k parts. We take k-1 parts as our training set and use the “held out” part as our test set. We repeat that k times differently (we hold out different part every time). Finally we take the average of the k scores as our performance estimation.

Cross validation can suffer bias or variance. if we increase the number of splits (k), the variance will increase and bias will decrease. On contrast, if we decrease (k), the bias will increase and variance will decrease. Generally 10-fold CV is used but of course it depends on the size of the training data.

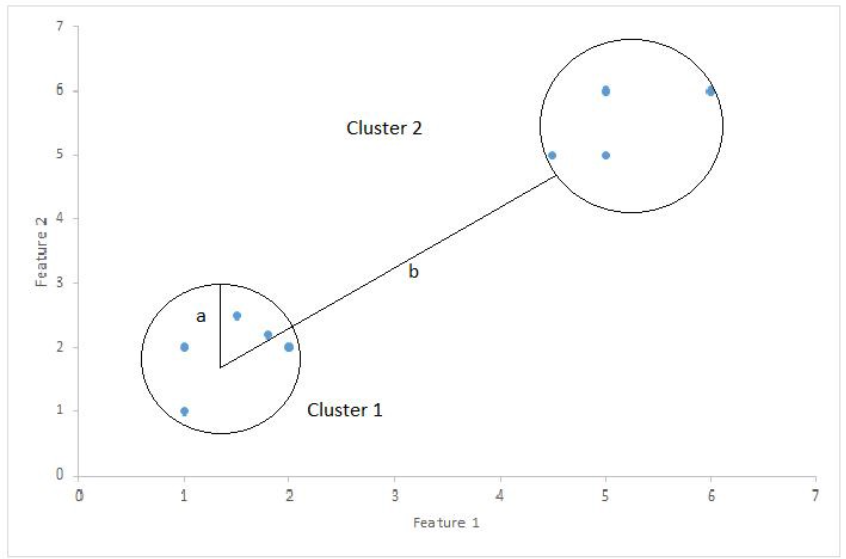
1. What do you understand by silhouette coefficient?

Answer: Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.



Silhouette Score = (b-a)/max(a,b)

Where,

a= average intra-cluster distance i.e the average distance between each point within a cluster.

b= average inter-cluster distance i.e the average distance between all clusters.

1. What is the advantage of using ROC Score?

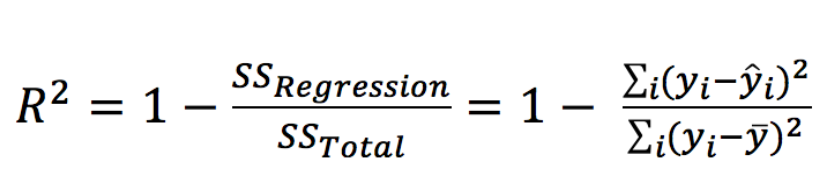
Answer:

* A simple graphical representation of the diagnostic accuracy of a test: the closer the apex of the curve toward the upper left corner, the greater the discriminatory ability of the test.
* Allows a simple graphical comparison between diagnostic tests
* Allows a simple method of determining the optimal cut-off values, based on what the practitioner thinks is a clinically appropriate (and diagnostically valuable) trade-off between sensitivity and false positive rate.
* Also, allows a more complex (and more exact) measure of the accuracy of a test, which is the AUC
  + The AUC in turn can be used as a simple numeric rating of diagnostic test accuracy, which simplifies comparison between diagnostic tests.
  + The AUC is non-parametric, which means it is unaffected by abnormal distributions in the population

1. Explain me complete approach to evaluate your regression model

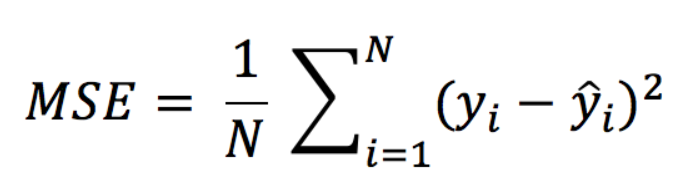
Answer: There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square: R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

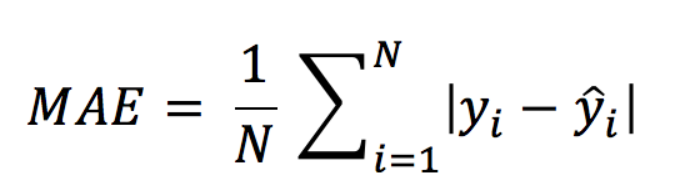


2. Mean Square Error(MSE)/Root Mean Square Error(RMSE): While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and makes it easier for interpretation.



3. Mean Absolute Error(MAE): Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.



1. Give me example of lazy learner and eagar learner algorithms example.

Answer:

**Lazy learner:**

1. Just store Data set **without** learning from it
2. Start classifying data when it receive **Test data**
3. So it takes less time learning and more time classifying data

**Eager learner:**

1. When it receive data set it starts classifying (learning)
2. Then it does not wait for test data to learn
3. So it takes long time learning and less time classifying data

In supervised learning Some examples are :

**Lazy** : K - Nearest Neighbour, Case - Based Reasoning

**Eager** : Decision Tree, Naive Bayes, Artificial Neural Networks

1. What do you understand by holdout method?

Answer: Holdout Method is the simplest sort of method to evaluate a classifier. In this method, the data set (a collection of data items or examples) is separated into two sets, called the Training set and Test set.

A classifier performs function of assigning data items in a given collection to a target category or class.

Example –E-mails in our inbox being classified into spam and non-spam.

Classifier should be evaluated to find out, it’s accuracy, error rate, and error estimates. It can be done using various methods. One of most primitive methods in evaluation of classifier is ‘Holdout Method’.

In the holdout method, data set is partitioned, such that – maximum data belongs to training set and remaining data belongs to test set.

1. What is diff between predictive modelling and descriptive modelling.

Answer:



1. How you have derived a feature for model building in your last project?

Answer: The great features that describe the structures inherent in your data.

Better features means flexibility and Better features means simpler models.

Tabular data is described in terms of observations or instances (rows) that are made up of variables or attributes (columns). An attribute could be a feature.

The idea of a feature, separate from an attribute, makes more sense in the context of a problem. A feature is an attribute that is useful or meaningful to your problem. It is an important part of an observation for learning about the structure of the problem that is being modeled.

I use “meaningful” to discriminate attributes from features. Some might not. I think there is no such thing as a non-meaningful feature. If a feature has no impact on the problem, it is not part of the problem.

1. Explain 5 different encoding techniques.

Answer: Since most machine learning models only accept numerical variables, preprocessing the categorical variables becomes a necessary step. We need to convert these categorical variables to numbers such that the model is able to understand and extract valuable information.

1. Label Encoding or Ordinal Encoding: We use this categorical data encoding technique when the categorical feature is ordinal. In this case, retaining the order is important. Hence encoding should reflect the sequence. In Label encoding, each label is converted into an integer value. We will create a variable that contains the categories representing the education qualification of a person.
2. One hot Encoding: We use this categorical data encoding technique when the features are nominal(do not have any order). In one hot encoding, for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category. These newly created binary features are known as Dummy variables.
3. Dummy Encoding: Dummy coding scheme is similar to one-hot encoding. This categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). In the case of one-hot encoding, for N categories in a variable, it uses N binary variables. The dummy encoding is a small improvement over one-hot-encoding. Dummy encoding uses N-1 features to represent N labels/categories.
4. Binary Encoding: Binary encoding is a combination of Hash encoding and one-hot encoding. In this encoding scheme, the categorical feature is first converted into numerical using an ordinal encoder. Then the numbers are transformed in the binary number. After that binary value is split into different columns. Binary encoding works really well when there are a high number of categories. For example the cities in a country where a company supplies its products.
5. Target Encoding: In target encoding, we calculate the mean of the target variable for each category and replace the category variable with the mean value. In the case of the categorical target variables, the posterior probability of the target replaces each category.
6. How do you define some features are not important for ML model? What strategy will you follow

Answer: Unnecessary features decrease training speed, decrease model interpretability, and, most importantly, decrease generalization performance on the test set.

The FeatureSelector library can be used to select important features.

The most common feature selection methods:

* Features with a high percentage of missing values: The first method for finding features to remove is straightforward: find features with a fraction of missing values above a specified threshold.
* Collinear (highly correlated) features: Collinear features are features that are highly correlated with one another. In machine learning, these lead to decreased generalization performance on the test set due to high variance and less model interpretability.
* Features with zero importance in a tree-based model: It finds features that have zero importance according to a gradient boosting machine (GBM) learning model.
* Features with low importance: The function identify\_low\_importance finds the lowest importance features that do not contribute to a specified total importance. For example, the call below finds the least important features that are not required for achieving 99% of the total importance:
* Features with a single unique value: A feature with only one unique value cannot be useful for machine learning because this feature has zero variance. For example, a tree-based model can never make a split on a feature with only one value (since there are no groups to divide the observations into).