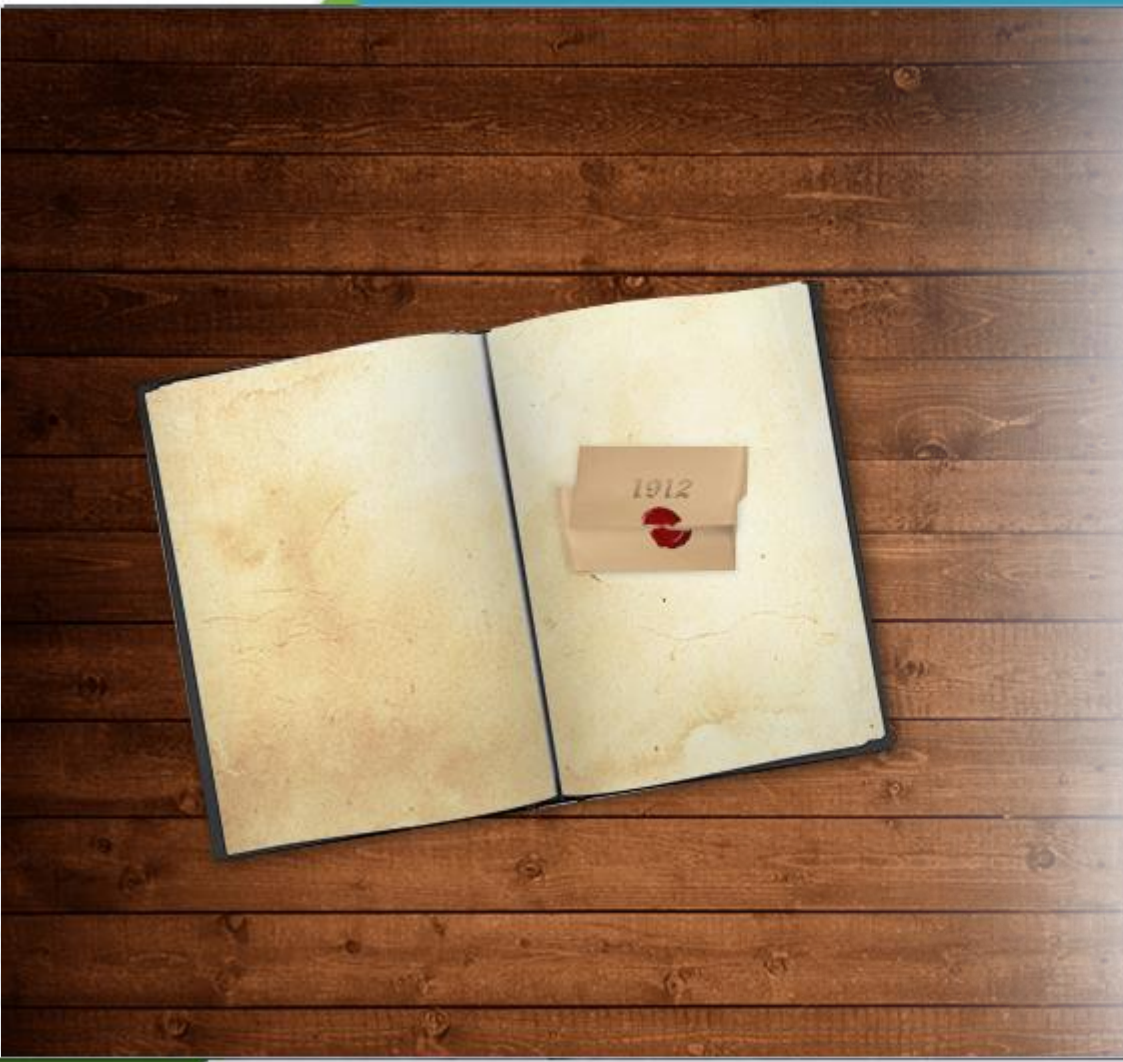
Cognizant

Welcome
to the course

Classification Model

Classification Model

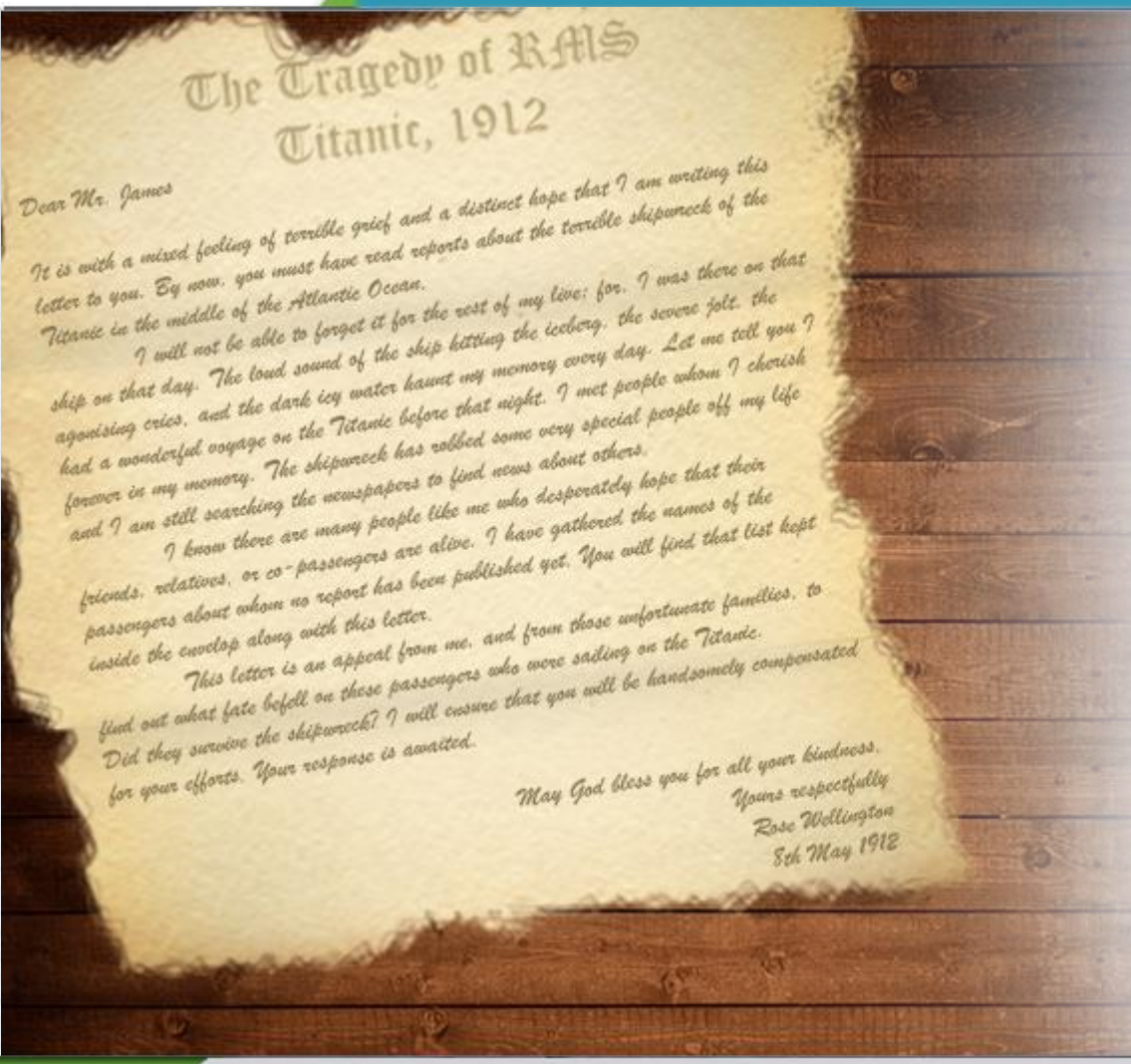


Did They Survive?

Nathan had no idea that he would find a mysterious letter inside an old copy of a detective story he has recently bought from a bookstore.

The letter is dated 1912, and its seal is broken.

Classification Model

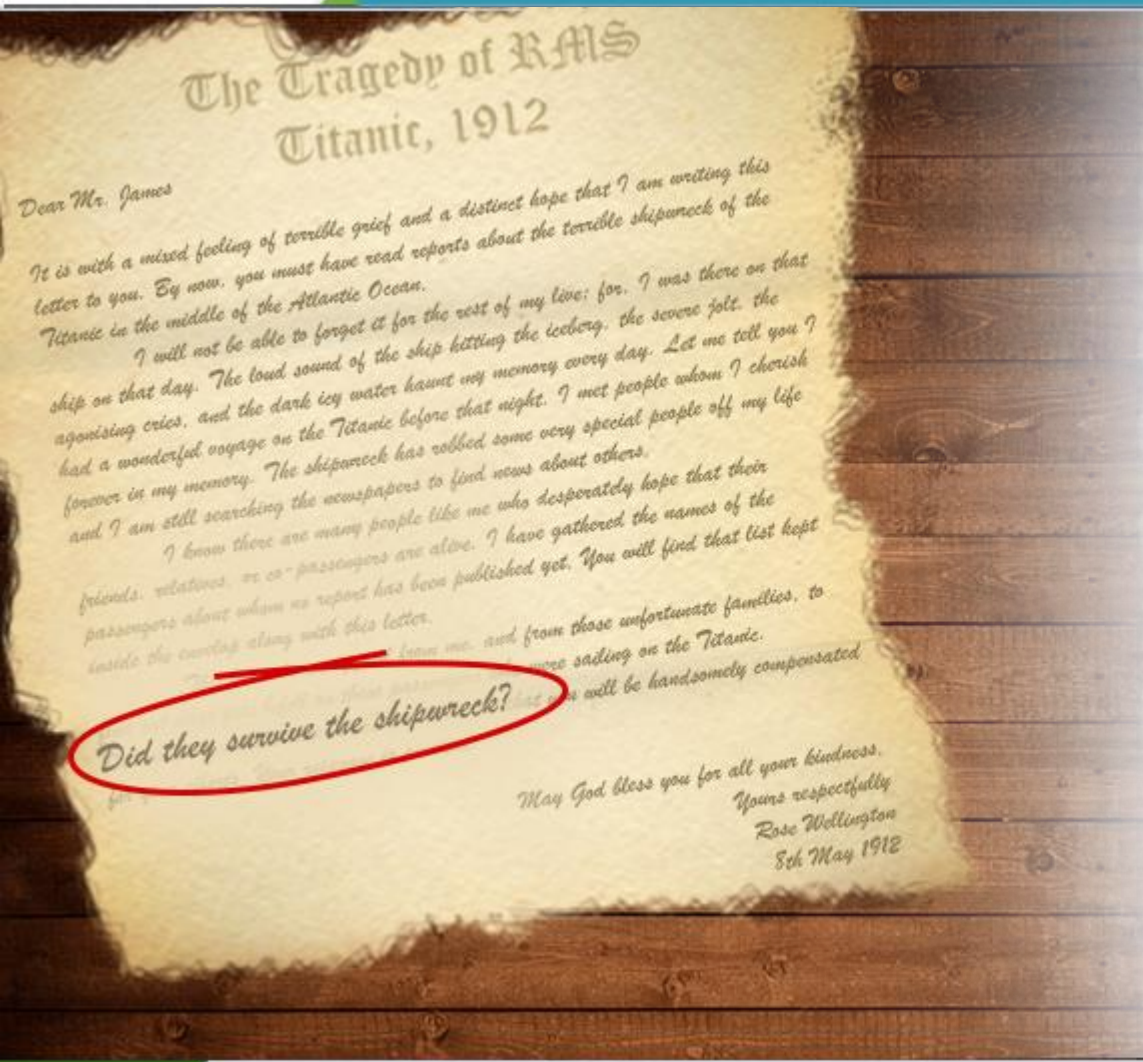


Did They Survive?

Being a fan of detective stories, Nathan could not resist himself from opening the letter.

As he opens it, he realizes it has something to do with the tragedy of the Titanic, the grand ship that sank on its maiden voyage in 1912.

Classification Model

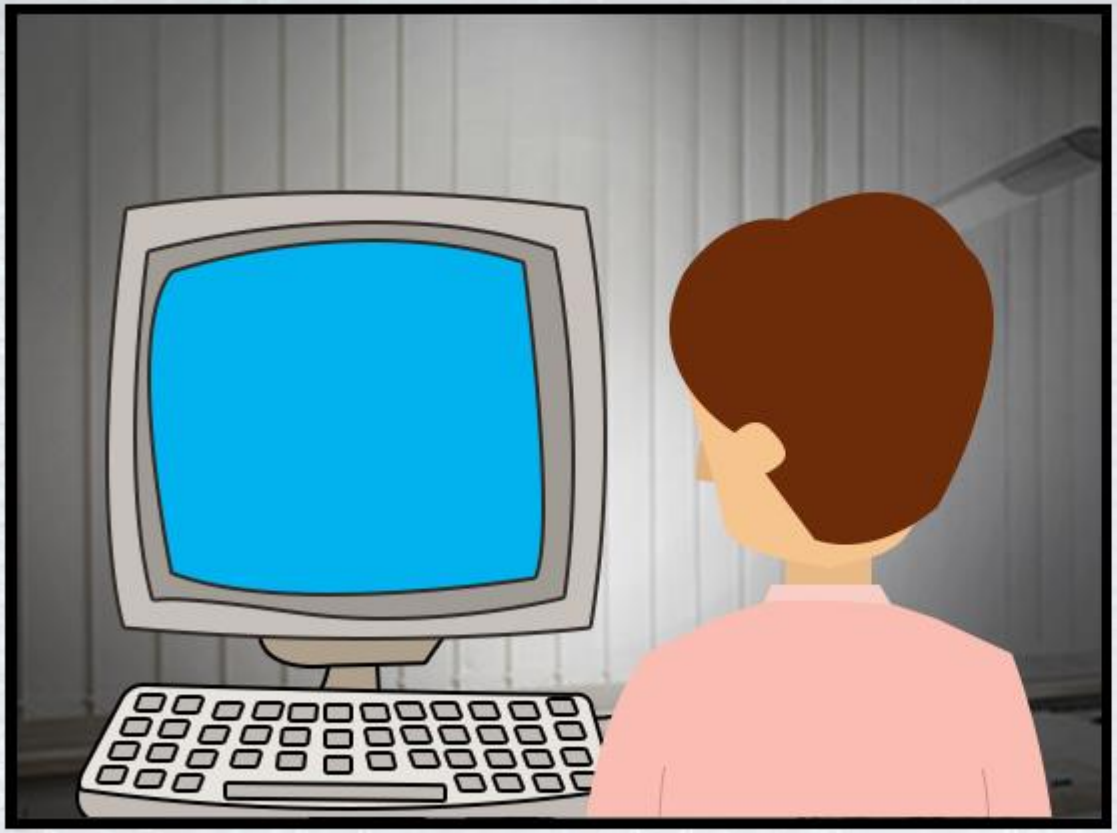


Did They Survive?

The letter was written to a private detective by someone called Rose Wellington. It contains a list of passengers who were onboard the Titanic on that fateful night. The lady requested the detective to find out how many of them survived.

It seems the book belonged to that detective!

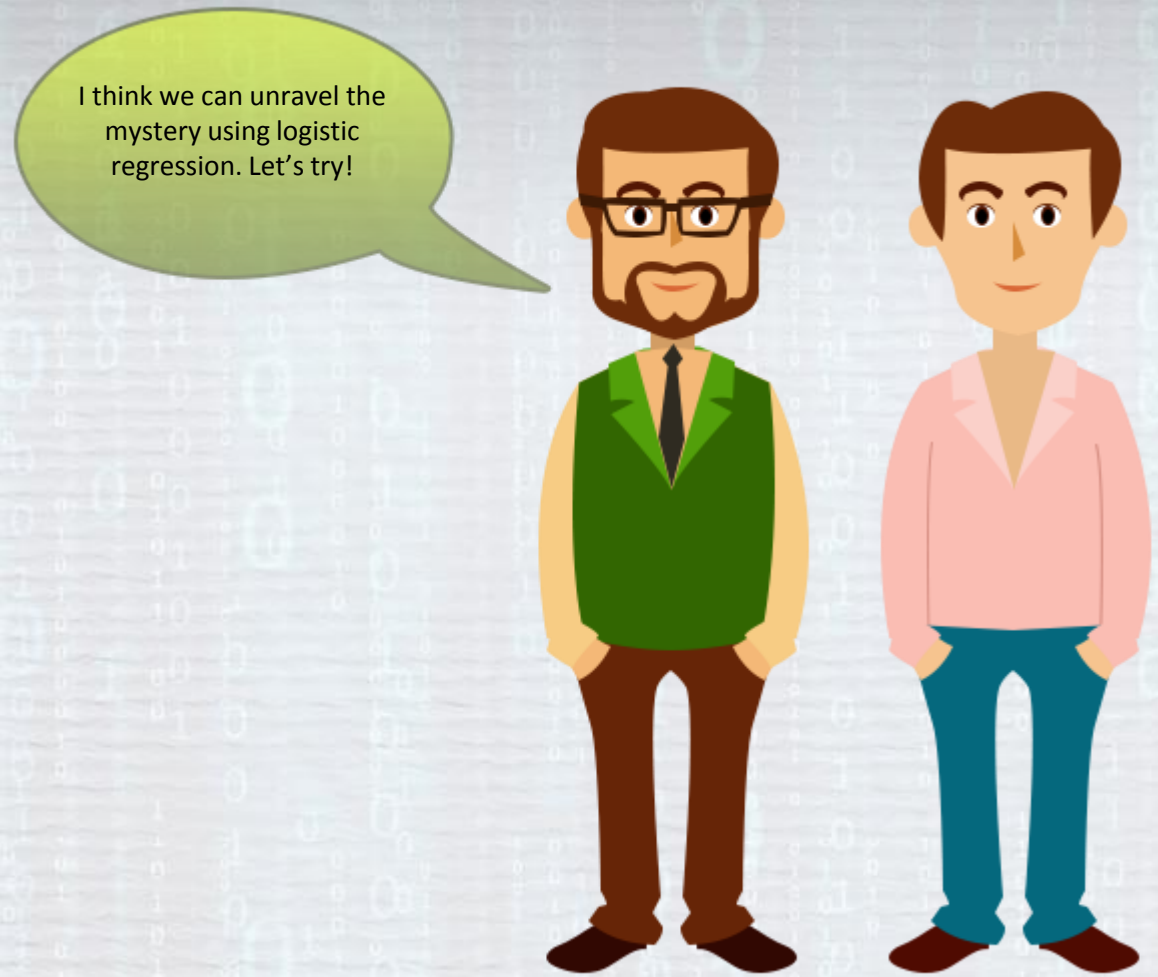
Classification Model



Did They Survive?

Nathan does some research online. He finds a list of passengers who were onboard the Titanic, and whose survival statuses are known. Unfortunately, the list does not contain the names he found inside the book.

Classification Model



Did They Survive?

Few days later, Nathan is discussing about the letter with his friend Scott, who is a data scientist.

Scott tells him that they can predict whether those passengers survived or not, using logistic regression.

Classification Model

Objectives

By the end of this module, you will be able to:

- Define logistic regression.
- Explain how to analyze data for logistic regression.
- Identify the input, parameters, and output details of the supervised learning model.
- Explain data selection strategy for logistic regression.
- Describe error handling and model selection.

Classification Model

Predicting Survival Status

As they start unraveling the Titanic mystery, Scott explains to Nathan “Since we are trying to predict whether the passengers survived or not, we will use a Boolean variable to indicate the survival status of the passengers. A Boolean variable has two outcomes: true and false.”

In this case, the variables indicate the following:

- 1 indicates that the passenger survived the shipwreck.
- 0 indicates that the passenger did not survive the shipwreck.

This type of problem is solved using a Logistic Regression Model, which is based on conditional probability.
In the Titanic case, the outcome should be measured against the threshold of 0.5. Thus, any value more than 0.5 will be considered as 1 (survived), and less than 0.5 will be considered as 0 (did not survive).

Conditional Probability

Classification Model

Predicting Survival Status

Conditional Probability

Conditional Probability

In probability theory, a conditional probability measures the probability of an event given that—by assumption, presumption, assertion, or evidence—another event has occurred.

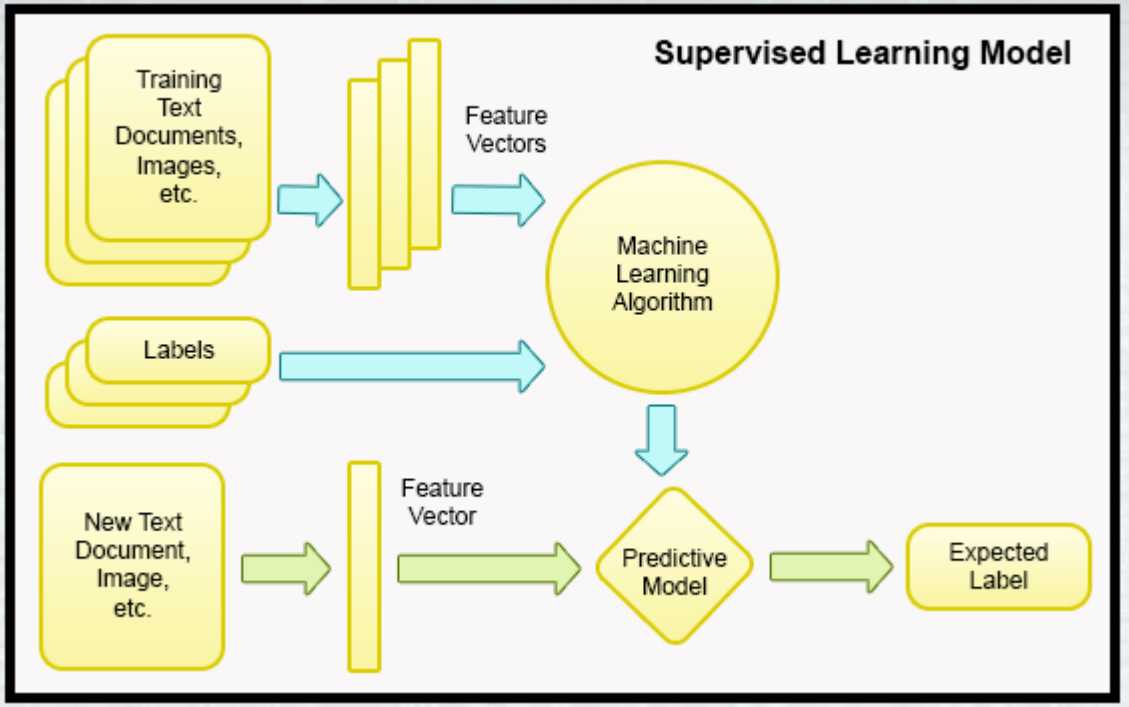
Classification Model

Supervised Learning Model

As you can derive from the description of conditional probability, Scott and Nathan will need an existing dataset. Therefore, they should try and find out a list of passengers whose survival statuses are known.

Once they have the dataset, they can use the Supervised Learning Model. The dataset will be used to construct features, and their survival statuses can be used as labels.

Refer to the image of a Supervised Learning Model provided here.



Classification Model

Dataset

Nathan realizes that his earlier research would be useful now. He hands over the data he found in the Internet to Scott.

However, Nathan wonders what details would be relevant for their study. Scott comes to his rescue.

Survived	Pclass	Name	Sex	Age	Cabin
0	3	Braund, Mr. Owen Harris	male	22	
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	C85
1	3	Heikkinen, Miss. Laina	female	26	

Classification Model

Dataset

It reflects the survival status of the passenger. This is also the response variable.	It reflects the passenger class that the passenger was traveling in. It is a proxy for their socio-economic status.	It reflects the name of the passenger. It does not hold any clue to the survival of the passenger; hence, this is not an important feature for the machine learning model.	It reflects the gender of the passenger.	It reflects the age of the passenger.	It reflects the cabin in which the passenger traveled.

Survived	Pclass	Name	Sex	Age	Cabin
0	3	Braund, Mr. Owen Harris	male	22	
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	C85
1	3	Heikkinen, Miss. Laina	female	26	

Classification Model

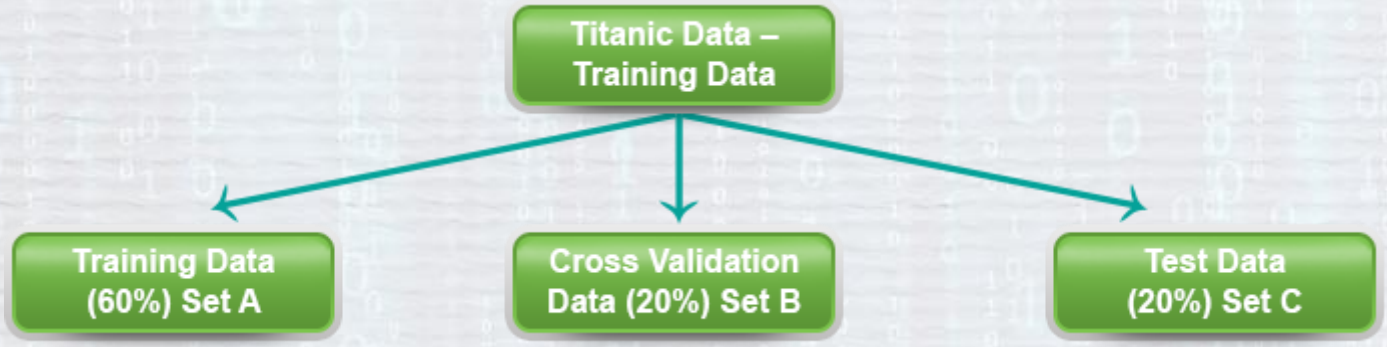
Data Selection Strategy for Training

Before they start analyzing the data, Scott decides to verify the data Nathan has got. He realizes that the list inside the letter would be the final test data since it does not have the status of the survival. He knows that the Training Data that Nathan shared has to be split across three different samples and he needs to adopt a proper sampling strategy.

Scott's findings are as follows:

- He needs to create three different samples of the training data having the following proportions.
  - 60% for training the model
  - 20% for cross validating the model
  - Last 20% for finalizing the model

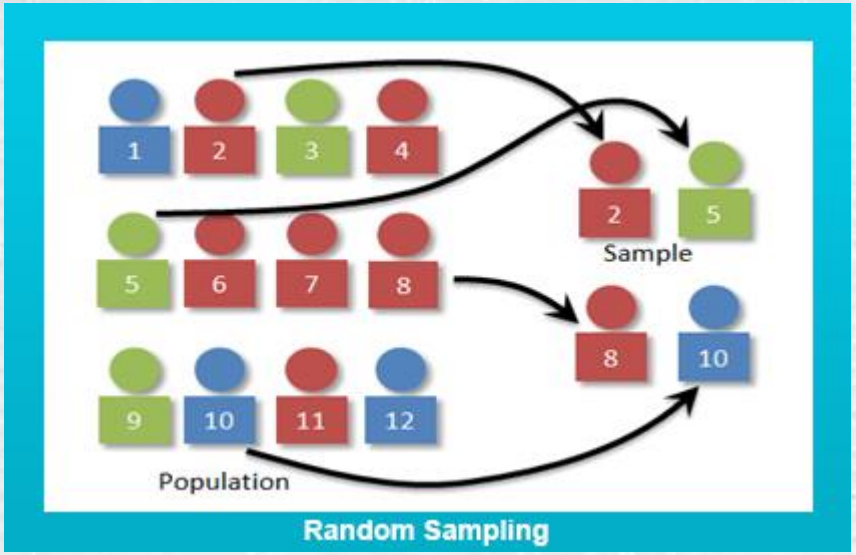
Refer to the diagram to view a visual representation of the Titanic dataset.

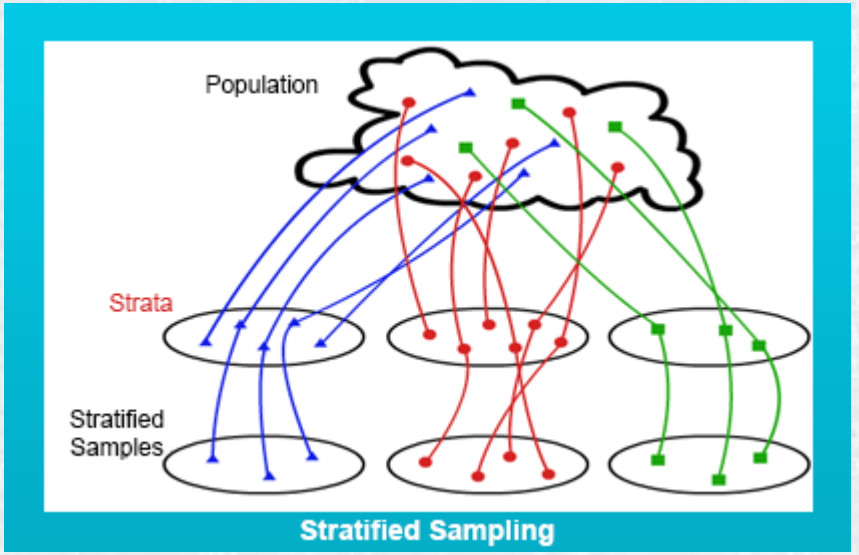


Classification Model

Random and Stratified Sampling

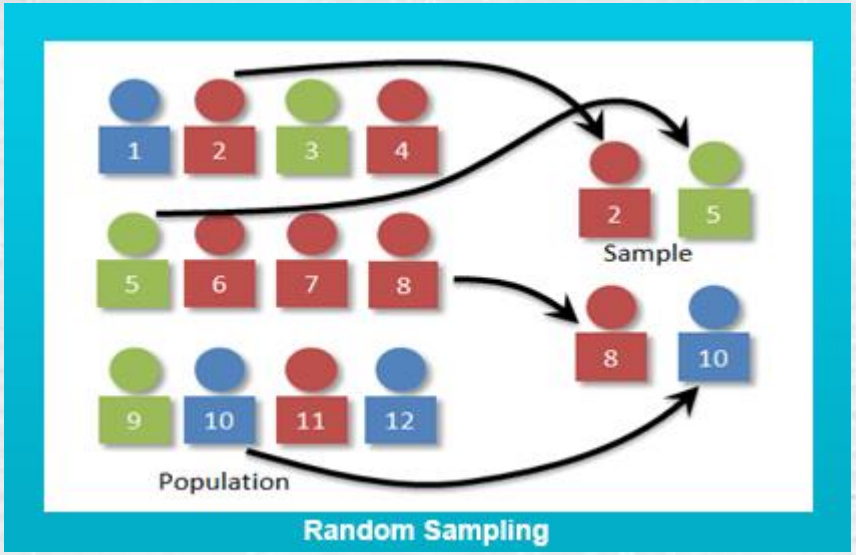
Nathan is not sure about data sampling. He asks Scott to elaborate. Scott explains that there are two types of data sampling: Random and Stratified.





Classification Model

Random and Stratified Sampling



Random Sampling

In random sampling, a sample is selected so that each item or person in the population has the same chance of being included. Thus, it refers to the process rather than the outcome of the process. This method assembles samples easily.

For example, if you want to draw a sample of five items for a group of 50 using random sampling, you can place them in a hat and draw five of them randomly.

However, random sampling may not present the proportion present in the original database. So, if the population of a city can be separated into smaller groups based on one or more distinguishing characteristics, random sampling may not be the best solution.

Classification Model

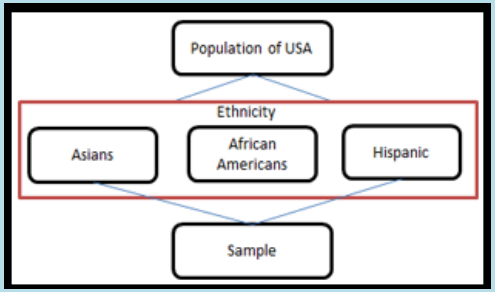
Random and Stratified Sampling

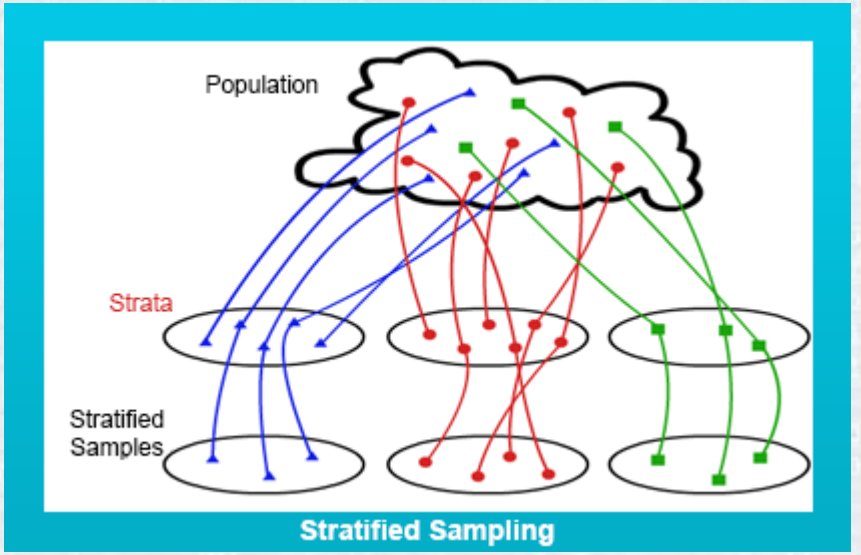
Stratified Sampling

In stratified sampling, elements in the dataset are first divided into various strata. Random sample is then taken from each stratum. A strata column is a categorical column with discrete values. For example, you can divide the population of USA into various ethnic groups. These groups will serve as strata.

Stratification ensures that the ratios of the selected values are preserved. Therefore, stratified techniques are generally used when population can be separated into smaller groups based on one or more distinguishing characteristics.

In the context of the Titanic, stratified sampling can be effective as there is a need to maintain the original proportion of males and females in the individual sub population.

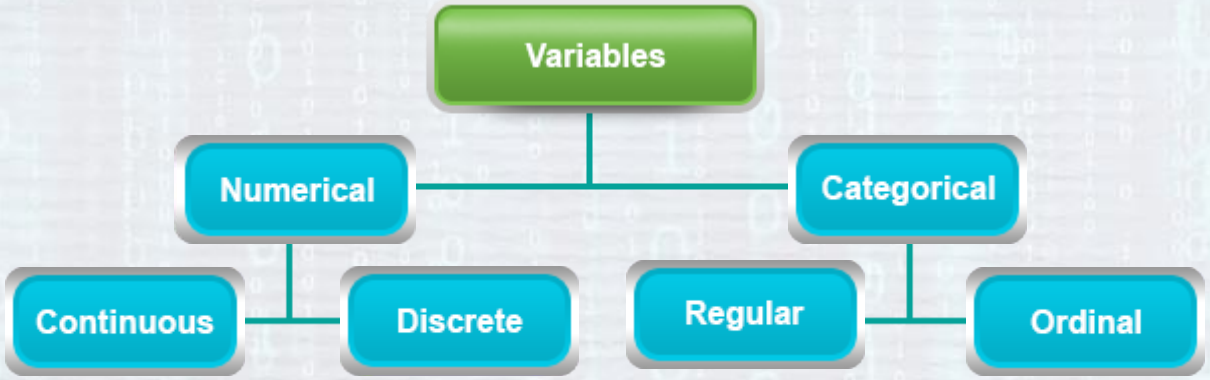




Classification Model

Understanding the Dataset: Variables

Nathan now understands data sampling. He can see that the dataset are selected through stratified sampling. However, Scott decides to explain variables to Nathan. After all, the Titanic dataset has several variables. Understanding them will help Nathan deal with the dataset better.



County	State	Population	Federal Spending	Poverty	Smoking Ban
Autauga	AL	58687	6.068	10.6	None
Barbour	AL	100536	8.752	25	None
Baldwin	AL	256874	6.14	12.2	None
Blount	AL	97856	5.131	13.4	None

Classification Model

Understanding Variables

Numerical Variable

It can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. Example: Federal Spending.

Categorical Variable

It consists of categories. Possible values are called the variable's levels. The State variable can take up to 51 values and they are categories. This is an example of categorical variable.

Continuous Variable

It takes any numerical value including negatives and decimals. Examples include Federal Spending and Poverty.

Discrete Variable

It takes only whole, non-negative numbers such as 0,1, and 2. The Population variable, which takes numeric values with jumps, is an example of discrete variable.

Regular Variable

It includes all categorical variables that are non-ordinal. Example includes the State variable.

Ordinal Variable

It contains categorical variables but its levels have a natural ordering. The Smoking Ban variable, which describes the type of county-wide smoking ban and takes values None, Partial, or Comprehensive in each county, is an example of the ordinal variable.

Classification Model

Age

With the understanding of variables, Nathan now looks at the train dataset to understand and analyze the data. The first thing he notices is the Age column.

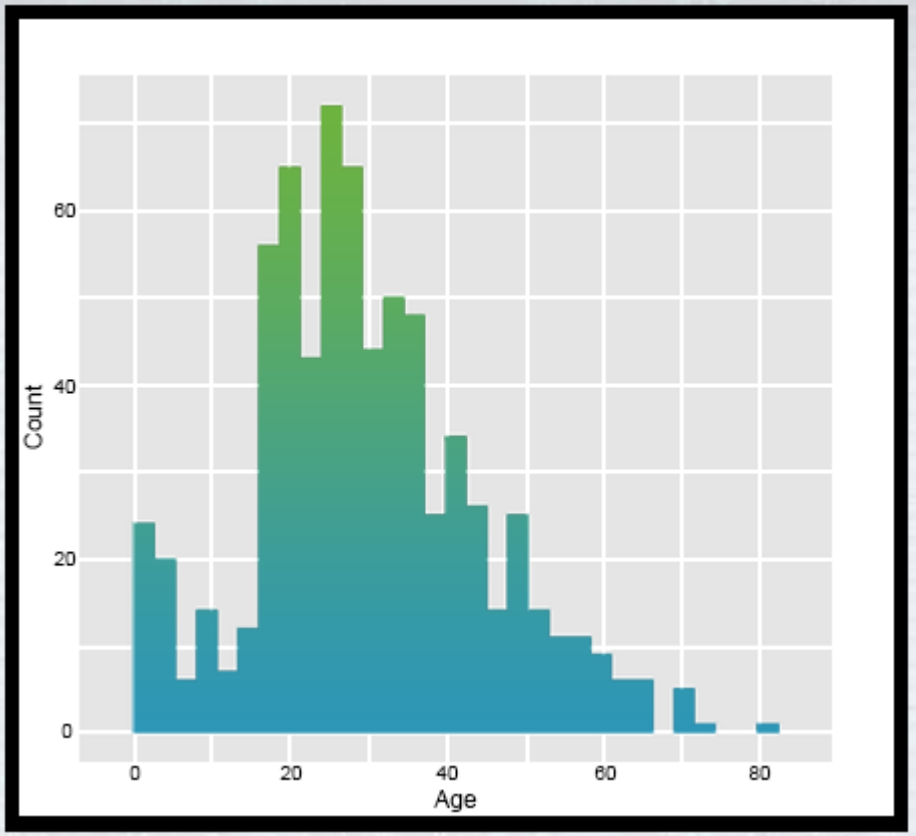
He identifies that the Age is a continuous variable here. At the same time, it seems to be a significant factor in determining who survived the shipwreck and who did not.

However, while collating the data, he comes across a few challenges. First of all, he observes that there are 177 missing values in the train dataset. He understands that removing so many rows can adversely affect the outcome.

On closer analysis of the data through a histogram plot—a frequency distribution—he finds that most of the passengers onboard the Titanic were people between age 0 and 65.

There were very few people beyond the age of 65 and there was only one instance of a survivor beyond the age of 65. Hence, passengers above the age of 65 can be considered as outliers.

How can Nathan solve the two challenges?



Classification Model

Age: Distribution Analysis

Nathan raises his concerns to Scott. Once again, he has a solution for the challenges.

Resolving the Missing Value

Resolving the Outliers

Classification Model

Age: Distribution Analysis

Resolving the Missing Value

The best strategy for the missing values would be to replace them with an appropriate measure. Some of the appropriate options are:

- Mean of the Age
- Median of the Age
- Mode of the Age
- Age derived using a probabilistic distribution

Please note that there isn't a single best way to select the replacement strategy of the missing values. You have to evaluate the performance of the model to identify the right replacement.

However, it's pertinent to note that mean is sensitive to outliers as opposed to the median. If the platform offers a function or a routine through which the best value can be replaced, it's the preferred choice. For example, in Azure ML platform, there are proprietary algorithms, which assist in replacing the missing values.

Resolving the Outliers

The best way to offset the influence of the outliers in the current context of Titanic is by replacing the age of outliers with normalized Age values. Since the objective is to create a predictive model for the normal data, one of the replacement strategies is to replace any data points that falls beyond the 99% percentile data point using a median, or a mode, or a system defined value.

Sample Representation Only	
Age	**Transformed Age**
75	43
70	42
85	42
66	45
65	65
60	60

Classification Model

Cabin

After resolving the queries on the Age, Nathan now looks at the Cabin column. He sees that the column contains alphanumeric values, and reflects the cabin in which the passengers traveled.

However, while analyzing the data, Nathan realizes that as many as 687 values in the Cabin column are missing. That is almost 77% of the entire data. Since learning models are sensitive to missing data points, he decides to ignore the Cabin column.

Do you think his decision to ignore the Cabin details is correct?

- Correct
- Incorrect

Classification Model

Cabin

Correct Answer:

Correct

Point to Remember

Nathan is correct in his decision to omit the Cabin data from the Machine Learning Model. Otherwise, the missing value would have negative impact on the model. Moreover, Cabin, in this context, does not hold a clue to the survival of the passenger.

Classification Model

Sex

Although the Cabin details turn out to be insignificant, Nathan understands that the Sex data is important. It seems natural that kids and ladies have a greater chance to survive than gentlemen. The analysis of the data shows similar results:

- 26% of the population survived.
- 68% of the survivors are females.

The Sex column is a string containing two unique values: male and female. However, it needs to be transformed into an indicator form before it can be used in the learning model. Below is the representation of dataset where the column has been converted to the indicator values.

Name	Sex
Braund, Mr. Owen Harris	male
Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female

Name	Sex Male	Sex Female
Braund, Mr. Owen Harris	1	0
Cumings, Mrs. John Bradley (Florence Briggs Thayer)	0	1

Classification Model

Pclass

Next to the Sex column is the Pclass. It reflects the passenger class that the passenger was traveling in, and thereby suggests their socio-economic status. From the earlier discussion with Scott, Nathan can see that the column is an example of discrete variable. Nathan assumes that the Pclass played an important role in the survival of the passengers.

While analyzing the Pclass data, Nathan ferrets out some interesting details:

- If the passenger was a female, and was traveling in Class 1, the upper class, the survival rate is 97%.
- If the passenger was a female and, was traveling in Class 2, the second class, the survival rate is 94%.
- If the passenger was a female, and was traveling in Class 3, the lowest class, the survival rate is 50%.

Refer to the images to see how the data has been analyzed based on the Pclass variables.

		Female			
		Pclass			
		1	2	3	
Survived	0	3	6	72	81
	1	91	70	72	233
		94	76	144	
	Survival Rate	97%	92%	50%	

Survival Status of Female Passengers

		Male			
		Pclass			
		1	2	3	
Survived	0	77	91	300	468
	1	45	17	47	109
		122	108	347	
	Survival Rate	37%	16%	14%	

Survival Status of Male Passengers

Classification Model

Input

Scott is impressed by the way Nathan analyzes the data. He noticed that all the relevant data has been analyzed efficiently by Nathan. He now decides to use the data for the machine learning model.

Scott needs the following details for the input of the algorithm model.

Column	Description
Survived	The survival status of the passenger.
Pclass	The passenger class.
Sex	The sex of the passenger.
Age	The age of the passenger.

Note: The Survived column represents the response variable.

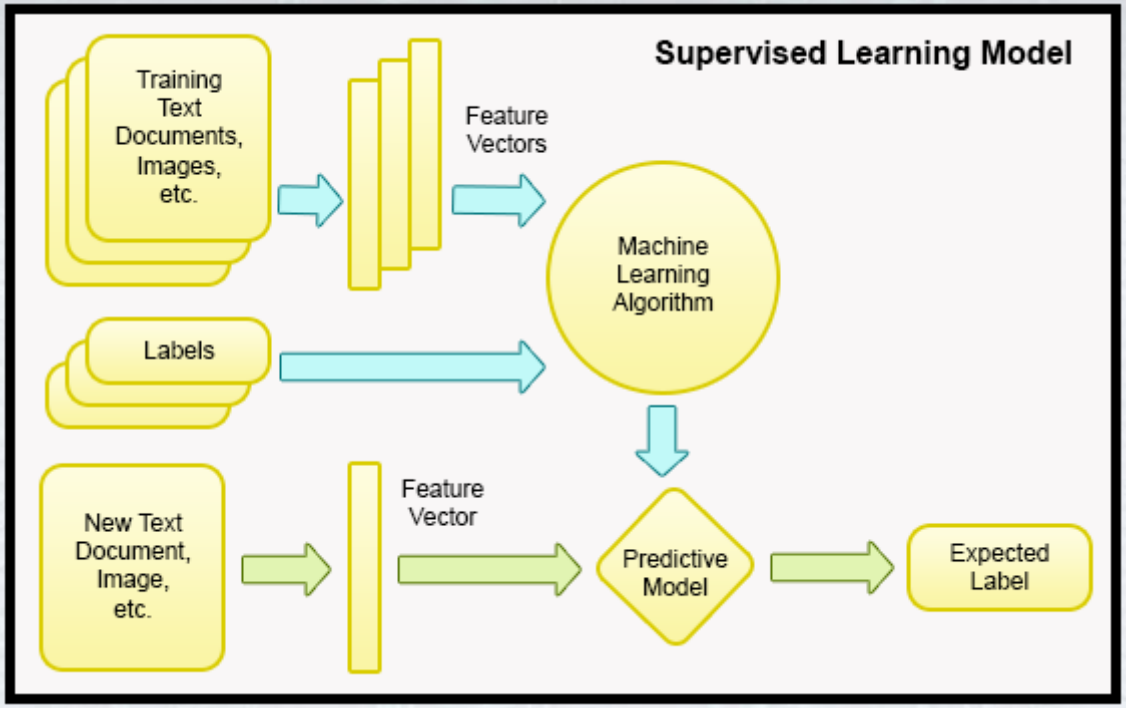
Classification Model

Learning Model

Scott shows the Learning Model to Nathan. The model is identical with the one Scott has shown earlier. However, Scott explains that he is going to enter the data and customize the model as per the Titanic case.

The current tools in the market such as Azure ML, SPSS, R offers an out of the box function to apply logistic regression model on the following set of parameters:

- Features in the Dataset
- Response Variable



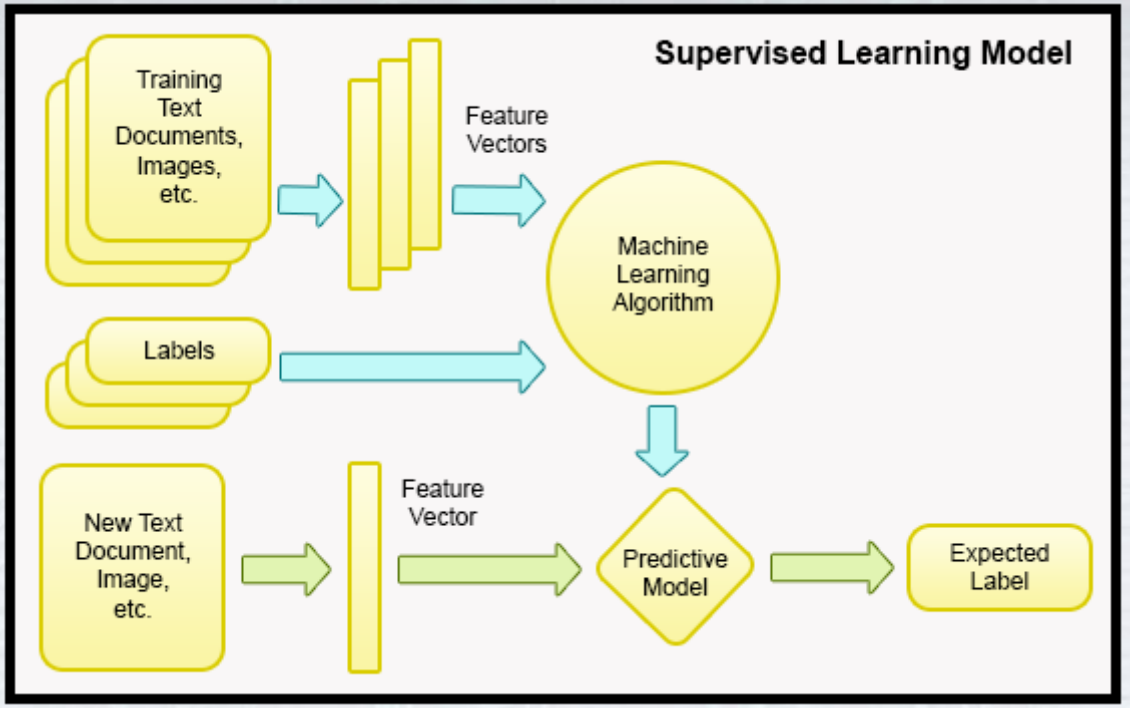
Classification Model

Learning Model

Feature Vectors

This is where the features derived from the train dataset should go. It should include the following:

Age	SexF	SexM	Pclass
20	1	0	1
10	1	0	2
30	0	1	3
40	0	1	1
50	1	0	2
0	0	1	3
4	1	0	3
5	0	1	3
6	1	0	3
7	0	1	3
8	1	0	3



Classification Model

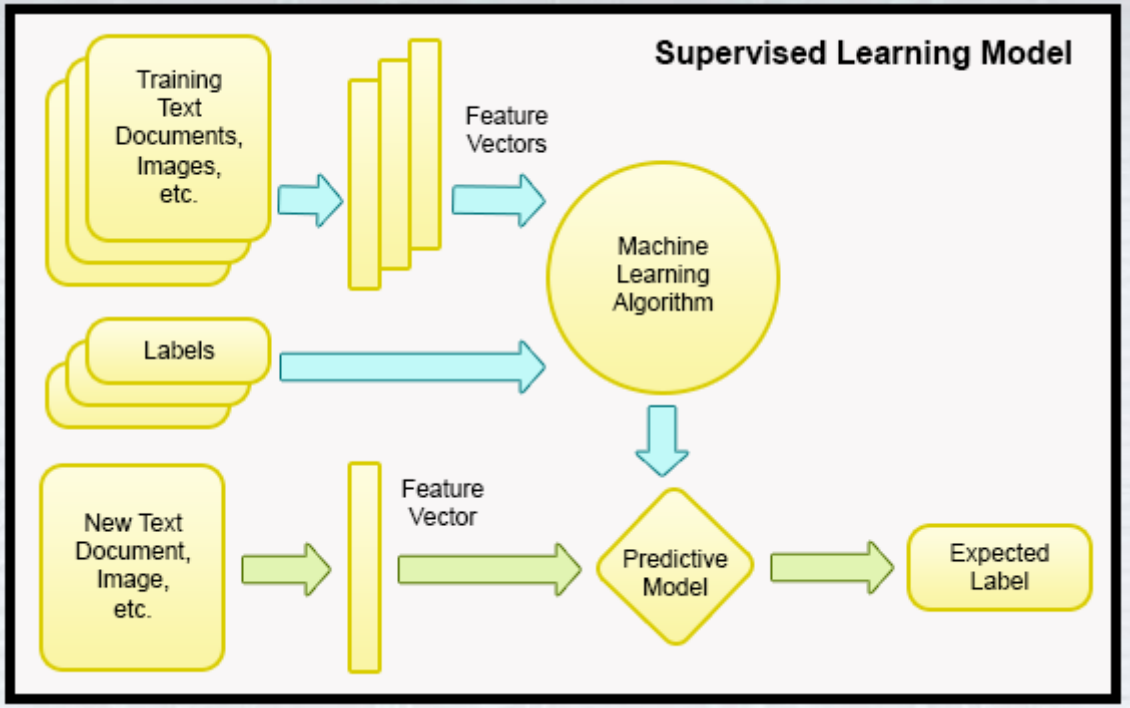
Learning Model

Labels

This is where the response variable should go. It should look like the following:

Survived

1
1
0
0
1
0
1
0
1
0
1



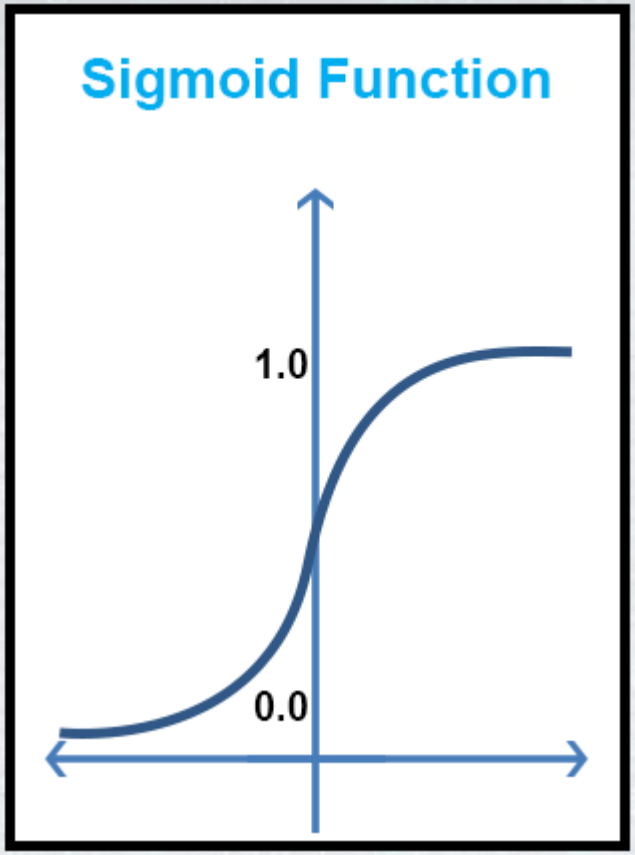
Classification Model

Output

Before Nathan could see the outcome of the learning model, Scott gives him an overview of the output. The output of the model would show the probability of survival of the passenger represented by the Passenger ID.

As Scott said earlier, if the probability is greater than 0.5 Threshold, the outcome is 1, which represents survived; else the outcome is 0, which represents not survived.

From a mathematical perspective, the logistic regression is represented by a Sigmoid function or logistic function where any value to the right represents positive and any value to the left represents negative.



Classification Model

Model Selection in Error Handling

Scott advises Nathan to use the error handling process to cross check his findings and reduce misclassification. Nathan wants to know more about error handling.

Error handling in machine learning for a Logistic Regression Model can be done using the confusion matrix. It involves calculating the metrics that reflect misclassification rate. Refer to the diagram to see how a confusion matrix looks like.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Classification Model

Model Selection in Error Handling

True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)
True Positive means the actual survivors are correctly classified as survivors.	False Positive means the non-survivors are incorrectly classified as survivors.	False Negative means the actual survivors are incorrectly classified as non-survivors.	True Negative means the actual non-survivors are correctly classified as non-survivors.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Classification Model

Precision, Recall, and F1 Score

Precision and Recall are two important factors that determine the F1 score, which in turn, determines the effectiveness of the model.

Precision: It is the fraction that actually survived among all the passengers they predicted as Survived.

Formula: Precision = TP/(TP+FP)

Recall: It is the fraction that they correctly detected as having survived among all passengers who actually survived.

Formula: Recall = TP/(TP+FN)

F1 Score: The F1 score can be interpreted as a weighted average of the Precision and Recall.

Formula: F1 Score = (2*Precision*Recall)/(Precision + Recall)

The higher the F1 Score, the better the model.

Based on the confusion matrix, the prediction is calculated. While predicting, you have to keep the following things in mind.

Classification Model

Errors in a Predictive Model

In a Predictive Model, there can be two types of error: Bias problem and Variance problem.

Bias Problem

Variance Problem

Cognizant

Classification Model

Errors in a Predictive Model

Bias Problem

Bias problem, or an under fitting problem, is a model that does not respond well to the training dataset itself.

Bias error can be solved by:

- Adding more features or deriving features based on the existing features.

Variance Problem

Variance problem, or an over fitting problem, is a model that fits well to the training data but does not fit to the cross validation or test dataset.

Variance error can be resolved by:

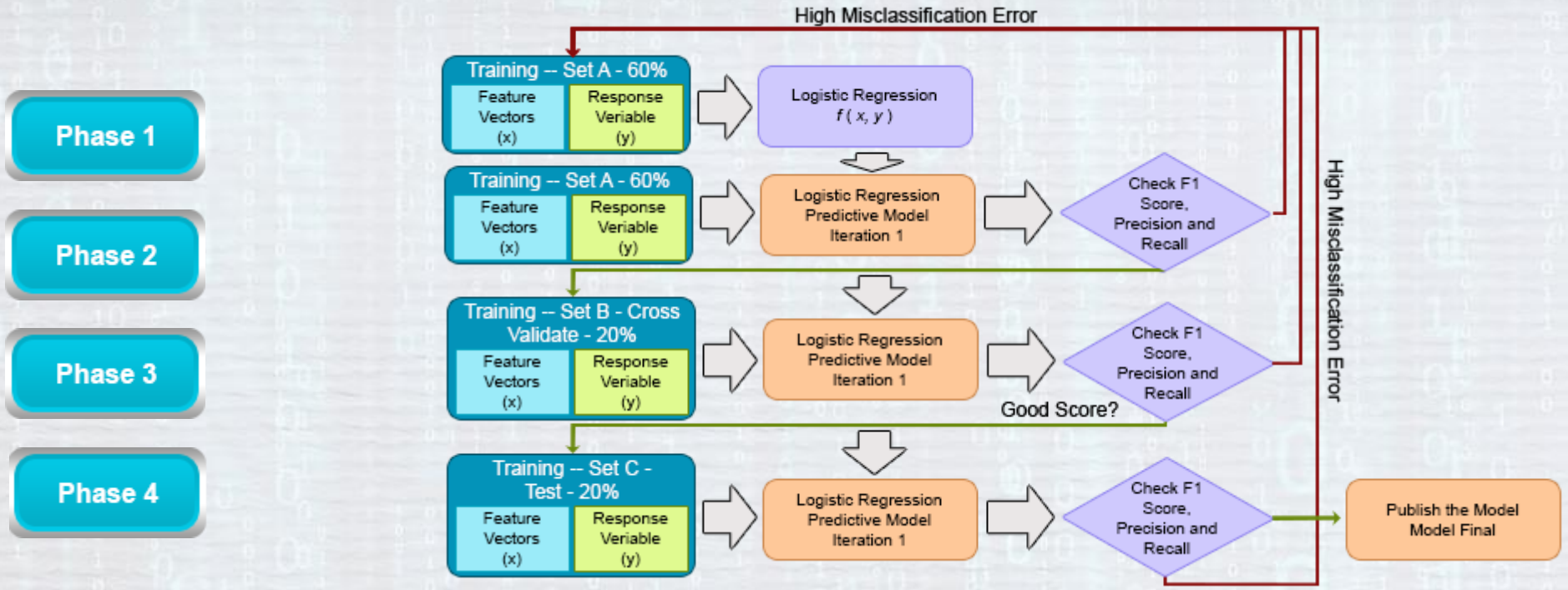
- Adding more dataset to the training data
- Reducing the number of features

Classification Model

Machine Learning Process

Since Nathan is now aware of all the important concepts of logistic regression, Scott explains how he uses the machine learning model to reach to the conclusion.

Take a look at how the machine learning process takes place.



Classification Model

Machine Learning Process

Phase 1

Phase 1:
The Training Data, Set A, having 60 % of the overall Training Data (Feature Vectors and the Response Variable) is passed as an input to the Logistic Regression Function.

Phase 2

Phase 2:
The resulting Model from the Step 1 is tested against the original 60 % data. The output of the Model is the predicted Survivors. The misclassification rate is calculated using the confusion matrix by comparing the actual and the predicted Survivors.

Phase 3

Phase 3:
If the metrics in the confusion matrix holds good then the model is tested against the cross validation data or the Set B of the Training Data. The output of the Model is the predicted Survivors. The misclassification rate is calculated using the confusion matrix by comparing the actual and the predicted Survivors.

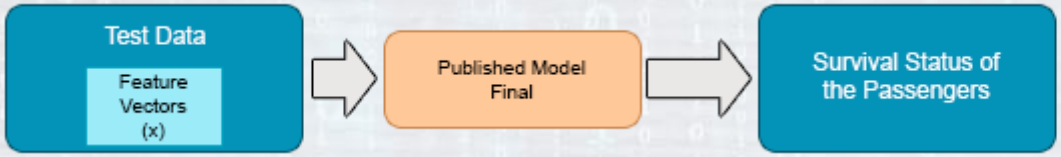
Phase 4

Phase 4:
If the metrics in the confusion matrix holds good then the model is tested against the test data or the Set C of the Training Data. This is the litmus test which conforms the model and its adaptability to the diverse input. If the model performs well then it is published for running it against the actual test data.

Classification Model

Machine Learning Process

However, in the Titanic case, the Test Dataset does not have a Response Variable, which is Survivor in this context. In fact, we need to determine the Survivor for the dataset. Therefore, the final phase of the process should be modified as below:



This step predicts the Survival of the Passengers whose actual Survival status is unknown.

Classification Model

Summary

Some of the key points of the course are mentioned below:

- Logistic Regression is based on conditional probability. It is effective in solving problems that have two outcomes: true or false.
- To predict an outcome using a supervised learning model, you need to have a pre- existing dataset (train dataset).
- While using the dataset, you have to analyze the data, identify the relevant features, replace missing values, and convert them into indicator form, if applicable.
- In a Predictive Model, there can be two types of errors: bias problem and variance problem. You should follow the proper strategies to overcome these errors. You should also use the confusion matrix for error handling.

Cognizant

Thank You