

**Assignment 4: Part 1. Understanding how a model works and analyzing data**

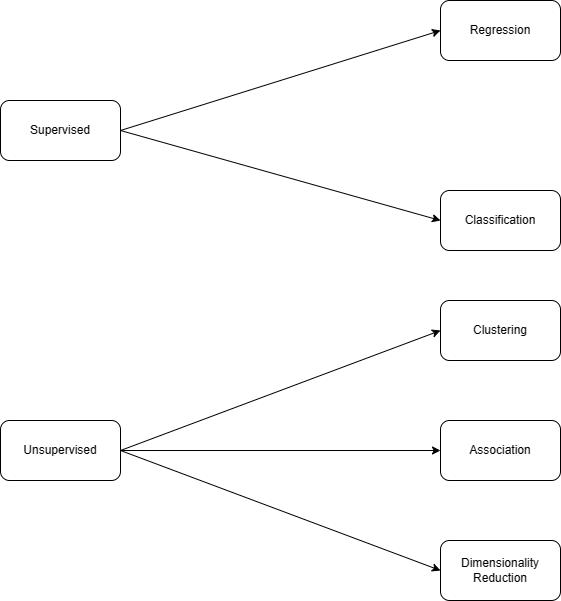
### Main Learning Goal

To understand how to train Machine Learning models to analyze financial data.

### How can we train ML models to analyze data?

Machine Lending

Hello, and welcome to the Bank of the Future™ weekly tour! Here we will show you the exciting models that we and other banks use, such as supervised and unsupervised models.

Below is a reading on the different models and examples of how these models are used in Industry. This is based on [blog posts by Terence Shin](https://terenceshin.medium.com/?source=post_page9fe30ff6776a) and [IBM's website](https://developer.ibm.com/articles/cc-models-machine-learning/%C2%A0)  
   
All machine learning models are categorized as either supervised or unsupervised.      
   
 

   
If the model is a supervised model, it’s then sub-categorized as either a regression or classification model. We’ll go over what these terms mean and the corresponding models that fall into each category below.    
   
**What is supervised learning?**  
Supervised learning is a machine learning approach that’s defined by its use of labeled datasets. These datasets are designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.    
   
**What is unsupervised learning?**  
Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are “unsupervised”).    
   
The main distinction between the two approaches is the use of labeled datasets: supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.    
In supervised learning, the algorithm “learns” from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. While supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately. For example, a supervised learning model can predict how long your commute will be based on the time of day, weather conditions and so on. But first, you’ll have to train it to know that rainy weather extends the driving time.    
Unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabeled data. Note that they still require some human intervention for validating output variables. For example, an unsupervised learning model can identify that online shoppers often purchase groups of products at the same time. However, a data analyst would need to validate that it makes sense for a recommendation engine to group baby clothes with an order of diapers, applesauce and sippy cups.  

### Example of a machine learning model in Finance

Now that you have learned what supervised and unsupervised models are, you are now going to learn how supervised models are used in financial technology. The folder for this activity contains the data set and the Matlab file you can run.

Here you will learn in Matlab how machine learning can be used to detect credit card fraud. You will use a data set you can download from this site. <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

We also have the data shared in the files.

1. **Look at the data set, what do the columns represent? What data is in these columns?**

They say the data has been PCA Transformed. Principal component analysis (PCA) is a dimensionality reduction and machine learning method used to simplify a large data set into a smaller set while still maintaining significant patterns and trends. Basically they fold lots of data columns into one larger column to try and reduce all the variables you are working with.

1. **Where does the data come from? How old is the data?**

Watch this video on how MatLab is used for a model. We are going to draw out the steps of what they are doing in the video and not actually going to program, but you can if you want to. Use the template below to log in the steps they went through to analyze the data. Some examples are filled in for you

|  |  |  |
| --- | --- | --- |
| Step number | What they did in Matlab | Results of what they did |
| 1 | Create variable T and readtable for the data in the csv file from Kaggle | All the data they wanted is read into a table for them to use |
| 2 | Convert the class value from the spreadsheet into categorical data (number) that they can read for fraud detected | Changes the values in the column to something the computer can read |
| 3 | Remove time |  |
| 4 |  | Splits the data set into two segments, one to use for the training of the model and one to use to test the model. 30% of the data is used for testing. |
| 5 | Apply machine learning model to fir data using classification trees |  |
| 6 |  | Tested the model out using the data held back for the testing. Here they visualized the classification trees to see how they wound up with a 1 or 0 for fraud. |
| 7 | Calculate actual fraud |  |
| 8 | Look at model performance/accuracy |  |
| 9 | Display Area under the precision recall curve (AUPRC) |  |
| 10 | Decide if your model training and model is good enough for the job |  |

[Machine Learning for Risk Management: Fraud Detection Using Machine Learning - YouTube](https://www.youtube.com/watch?v=ONNk9ypWzeU)

### What are classification Trees

Imagine you're a detective trying to figure out if a credit card transaction is legitimate or fraudulent. You can ask yourself a series of yes-or-no questions about the transaction, and based on the answers, you'll decide whether it's likely to be fraud or not.

This series of questions forms a "decision tree." At each step, you ask a question, and the answer leads you to either ask another question or make a final decision.

For example, let's say the first question you ask is: "Is the purchase amount more than $1,000?" If the answer is "yes," that could be a red flag for fraud, so you move to the next question. If the answer is "no," it's less likely to be fraud, but you still need to check other factors.

The next question could be: "Is the purchase from a different country than the cardholder's home address?" If yes, that's another potential sign of fraud, so you continue down that branch of the tree. If no, you move to a different question.

You keep asking questions like this, such as "Is the purchase for an unusual type of item for this cardholder?" or "Has the credit card been reported lost or stolen?" Each answer leads you down a different path until you reach a final decision: either "Likely Fraud" or "Likely Legitimate."

The decision tree helps you organize and prioritize the questions in a logical way, so you don't waste time asking unnecessary questions if earlier answers already point strongly toward fraud or a legitimate purchase.

Companies that issue credit cards can use decision trees like this, with many different questions and branches, to automatically flag suspicious transactions for further review by human analysts.

Watch this video from Matlab on how Classification Trees work

<https://www.youtube.com/watch?v=PbzjDIKWKSw>

1. **In the Matlab finance example you saw earlier, they used a classification tree model to identify if something is fraud or not fraud. How is this like a yes or no shown in this video? What did they use to show fraud? Was it a zero or a 1?**
2. **If you collected a lot of data on an artist or sports player and were trying to decide if you were going to sign them or not, could you use a classification tree model? Explain how that model might work.**

Resources

<https://builtin.com/data-science/step-step-explanation-principal-component-analysis>

Read about how MatLab uses Machine Learning and different models

[Machine Learning in MATLAB - MATLAB & Simulink (mathworks.com)](https://www.mathworks.com/help/stats/machine-learning-in-matlab.html)