HyperionDev

Lecture - Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly.
  - Please review Code of Conduct (in Student Undertaking Agreement) if unsure
- No question is daft or silly - **ask them!**
- There are Q&A sessions midway and at the end of the session, should you wish to ask any follow-up questions.
- Should you have any questions after the lecture, please schedule a mentor session.
- For all non-academic questions, please submit a query: www.hyperiondev.com/support

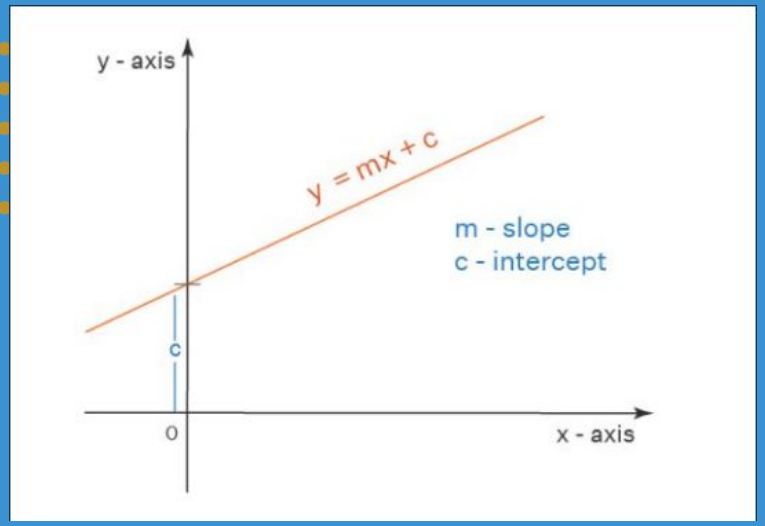
Lecture Objectives

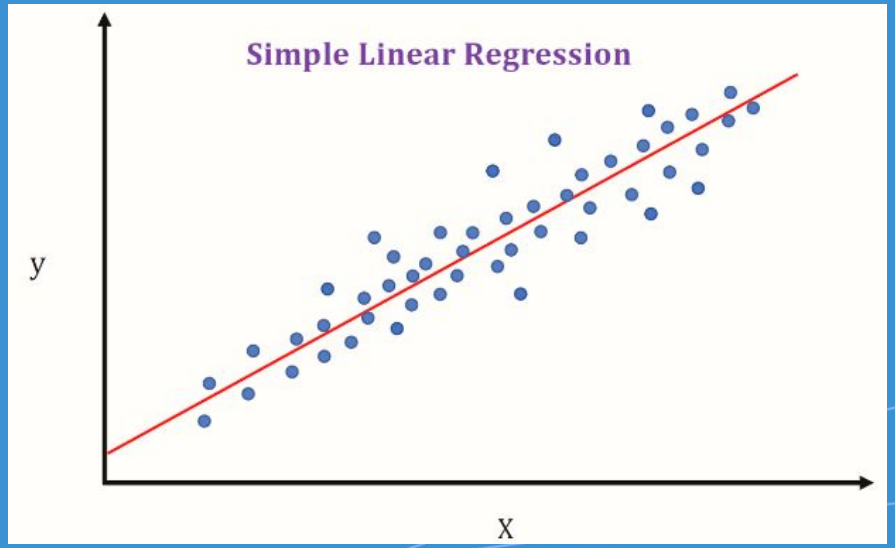
1. An overview of Regression and Classification models
2. An introduction to Decision Trees
3. Discuss Logistic Regression
4. Demonstrations of how we use regression models to make predictions

Regression

- Regression analysis is a statistical process used to estimate relationships between variables.
- There are two types of linear regression: simple linear regression and multiple linear regression.
- We also have logistic regression, which helps us classify observations.

Simple Linear Regression





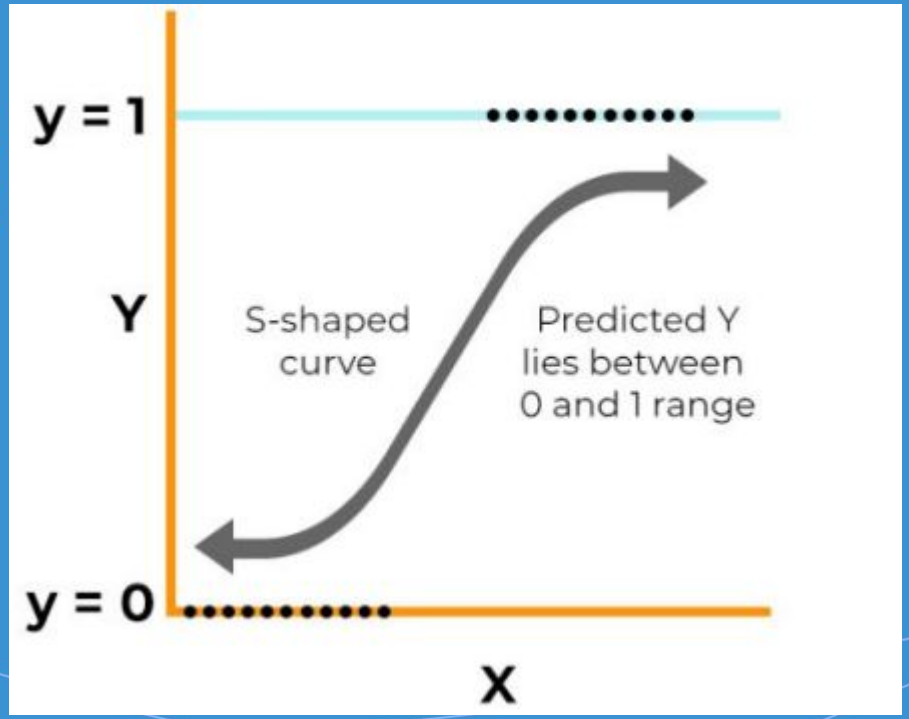
Logistic Regression

- When linear regression models are used, we must have a dependent variable, Y, which is a continuous numerical variable. However it very common in machine learning problems instead to be dealing with categorical variables. These variables take on distinct non-continuous values which will correspond to a specific set of categories.
- Predicting categorical variables is called classification. Classification problems are very common, perhaps even more so than for problems suited for regression.

Logistic Regression

- One approach to classification is logistic regression, which is a common way to do binary logistic regression, which is classifying into two categories.
- It works by using the logistic function, also known as the sigmoid function. This is an S-shaped curve that maps input values to x output values y.
- Logistic regression is similar to linear regression, however the output is not continuous along a line, but a value between 0 and 1.
- That value can then be interpreted as the probability of that the instance belongs to a certain category.

Logistic Regression



$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Intro to Decision Trees

- Decision tree work by formulating simple rules that partition data into even smaller regions. Each partition can be thought of as a fork in the road, where a decision will be made.

- The decision is made based on rules which are derived from learning experiences. Decision trees are among the most interpretable machine learning techniques based on how they resemble how humans make decisions.

- For example, to make dinner, one would first think if there are enough ingredients in the fridge to make a meal. If there aren't enough ingredients, we'll need to consider other options. Based on the time of day, you may decide to go to the grocery store or just decide on take-out instead.

Variables in Credit Risk Dataset

Volume: volume of bank transactions per month (Frequent, Seldom)

Value: income per transaction (High, Low)

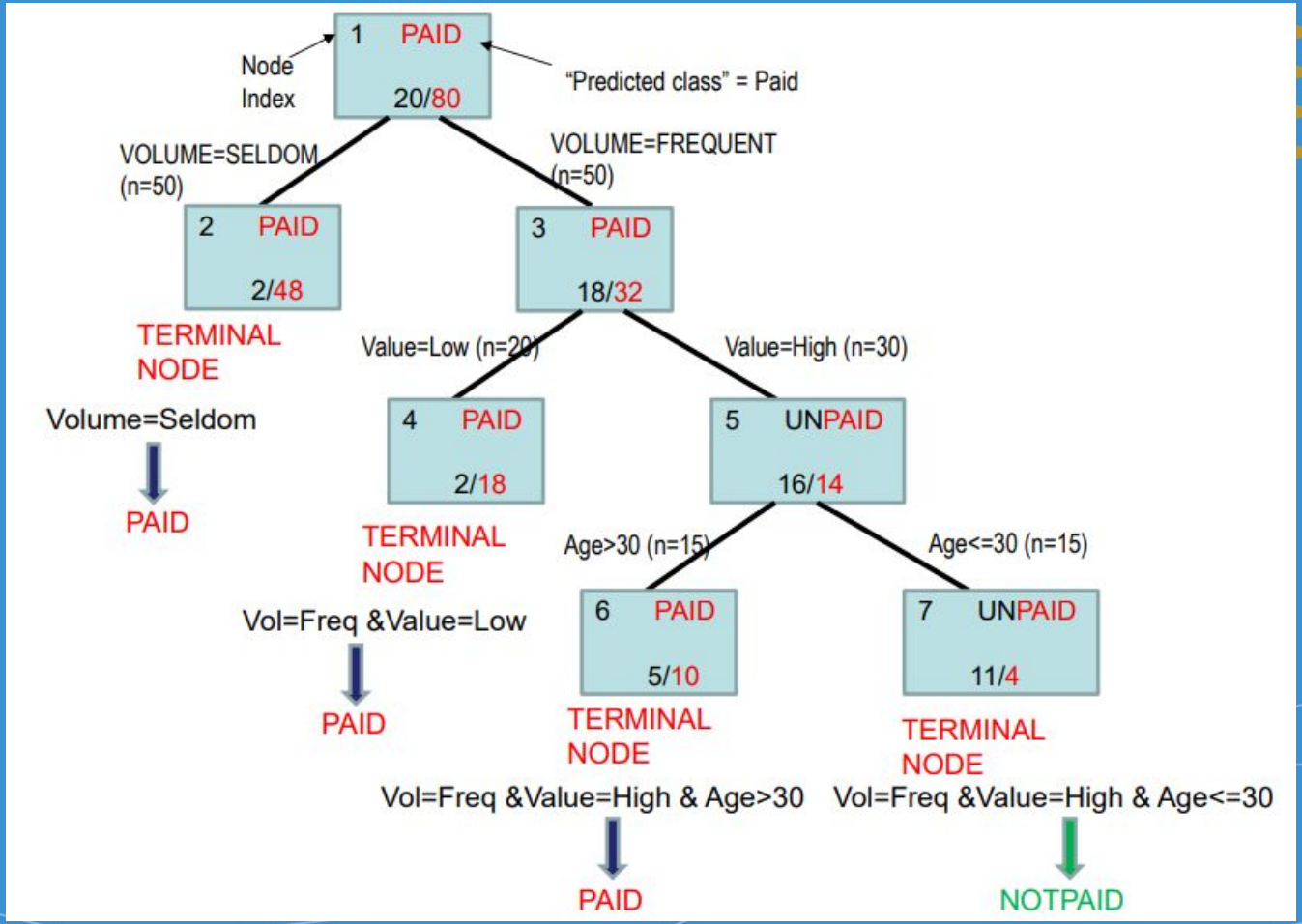
Age: age of customer (Younger than 25, 25 to 30, Older than 30)

Status: whether the loan was repaid (Not Paid/Paid)

Credit Risk Dataset

	A	B	C	D	E
1	ID	Volume	Value	Age	Status
2	1	Seldom	Low	>30	Paid
3	2	Seldom	Low	>30	Paid
4	3	Seldom	Low	>30	Paid
5	4	Seldom	Low	>30	Paid
6	5	Seldom	Low	<25	Paid
7	6	Seldom	Low	<25	Paid
8	7	Seldom	Low	<25	Paid
9	8	Seldom	Low	<25	Paid
10	9	Seldom	Low	<25	Paid
11	10	Seldom	Low	<25	Paid
12	11	Seldom	Low	<25	Paid

91	90	Frequent	High	<25	Notpaid
92	91	Frequent	High	<25	Notpaid
93	92	Frequent	High	<25	Notpaid
94	93	Frequent	High	25-30	Notpaid
95	94	Frequent	High	25-30	Notpaid
96	95	Frequent	High	25-30	Notpaid
97	96	Frequent	High	>30	Notpaid
98	97	Frequent	High	>30	Notpaid
99	98	Frequent	High	>30	Notpaid
100	99	Frequent	High	>30	Notpaid
101	100	Frequent	High	>30	Notpaid


1 PAID
20/80
Node Index
"Predicted class" = Paid
VOLUME=SELDOM (n=50)
VOLUME=FREQUENT (n=50)
2 PAID
2/48
3 PAID
18/32
TERMINAL NODE
Value=Low (n=20)
Value=High (n=30)
Volume=Seldom
PAID
4 PAID
2/18
5 UNPAID
16/14
TERMINAL NODE
Age>30 (n=15)
Age<=30 (n=15)
Vol=Freq &Value=Low
PAID
6 PAID
5/10
7 UNPAID
11/4
TERMINAL NODE
TERMINAL NODE
Vol=Freq &Value=High & Age>30
Vol=Freq &Value=High & Age<=30
PAID
NOTPAID

Project Idea

There is the advertising data set available on Kaggle:

https://www.kaggle.com/datasets/ashydv/advertising-dataset

Download the dataset and plot a scatterplot to analyse the relationship between ‘TV advertising’ and ‘Sales’

1. Why is a Simple Linear Regression model appropriate?
2. Fit the SLR model and make some predictions. Interpret these values.
3. Search for other datasets on Kaggle. Identify which ones would be suitable for Linear Regression

Questions and Answers

Questions around Regression and Classification