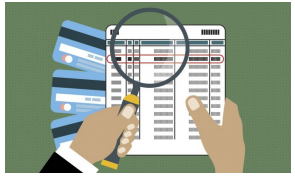
EXPLORE | DIGITAL SKILLS

Advanced Regression

Overview

What is Machine Learning?

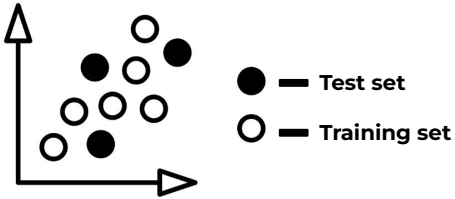
Machine learning is the study of **software that automatically learns** from experience.

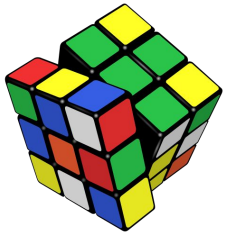
Types of machine learning		Where is it used?	
1 Supervised	Given sets of input-output pairs (x's and y's), the algorithm finds **hidden relationships** in the data (function approximation)	Prediction / Classification Credit & Insurance underwriting Cancer detection Speech recognition	
2 Unsupervised	Given input data only, the algorithm identifies **clusters**, **patterns** and **anomalies** in the data.	Clustering / Grouping Recommendation systems Fraud Detection	
3 Reinforcement	Given an environment and a reward function, the algorithm optimises its actions to **maximise its reward**.	Self-driving Cars Autonomous Robots Utilities (traffic management) Game Playing	

EXPLORE | DIGITAL SKILLS

Machine Learning in Practice

Preprocessing → **Model Building** → **Model Evaluation** → **Deployment**





- Data Cleaning
  - Impute
  - Normalise/Standardise
  - Label/Dummy encode
- Train-Test split / Kfold

- Model Selection
- Model Training
- Hyperparameter Tuning

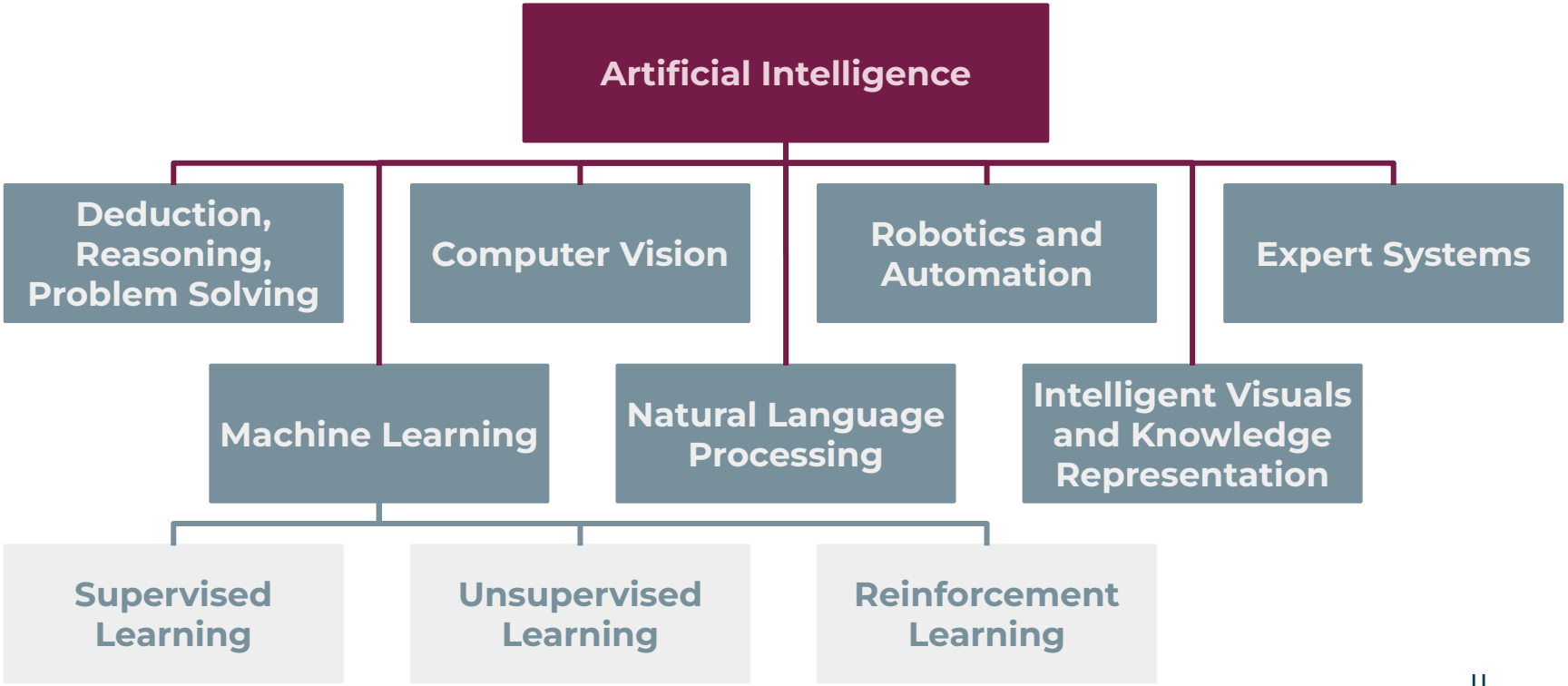
- Evaluate Model on Test set
- Report Performance metrics

- Hosting and Versioning
- Dashboards
- Containerization
  - Docker
  - Kubernetes

EXPLORE || DIGITAL SKILLS

Things to consider - the larger AI spectrum

Artificial Intelligence is the broad term used to **describe machines with cognitive ability**.



EXPLORE | DIGITAL SKILLS

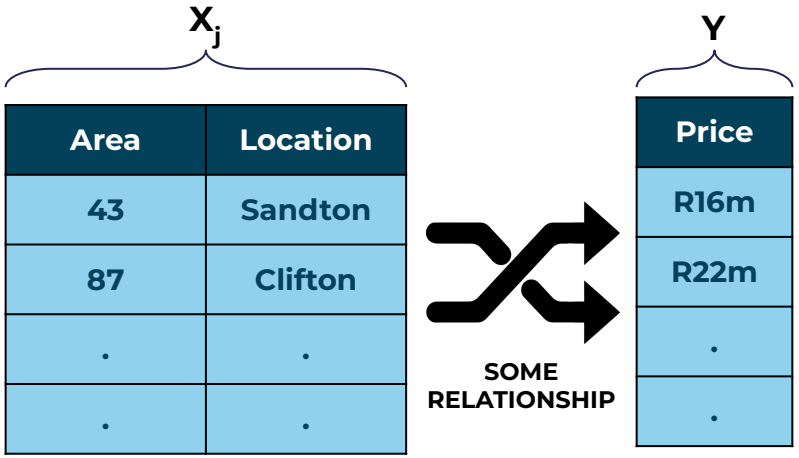
Introduction to regression

Regression is finding relationships between:

- A numeric, dependent variable *Y*
  - Like the price of a house, size of a loan, etc.
- And, a number of independent predictor variables, each known as an X_j
  - Like the size and location of the house, type of tree, etc.

It's a type of supervised learning:

- Supervised learning is a subset of machine learning - simply the process of working out how some inputs relate to some outputs.
  - *Inputs*: independent X_j's described above;
  - *Outputs*: dependent *Y*, as above.



EXPLORE | DIGITAL SKILLS

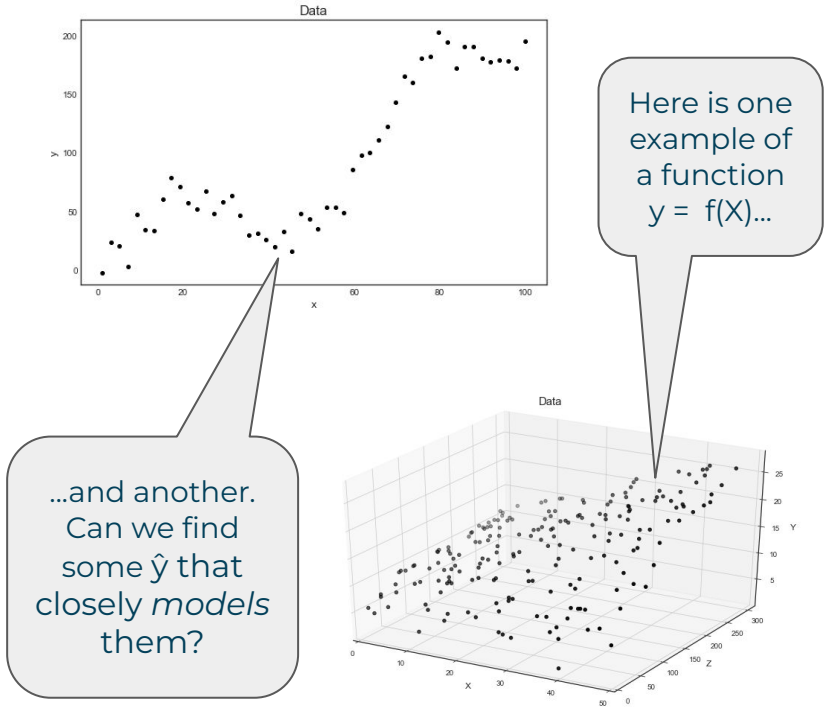
Introduction to regression (continued)

Our data is some mystery function y = *f*(X):

- We will never know what the true *f* is.
- Machine learning is the process of getting as close as we can to that *f*.
- The gap between the true *f* and our approximation of it is known as *error*.
- The model we produce is $\hat{y} = f(X) + \text{error}$.
- We have means of determining how close our $\hat{y}$ is to *f - called goodness of fit* measures.

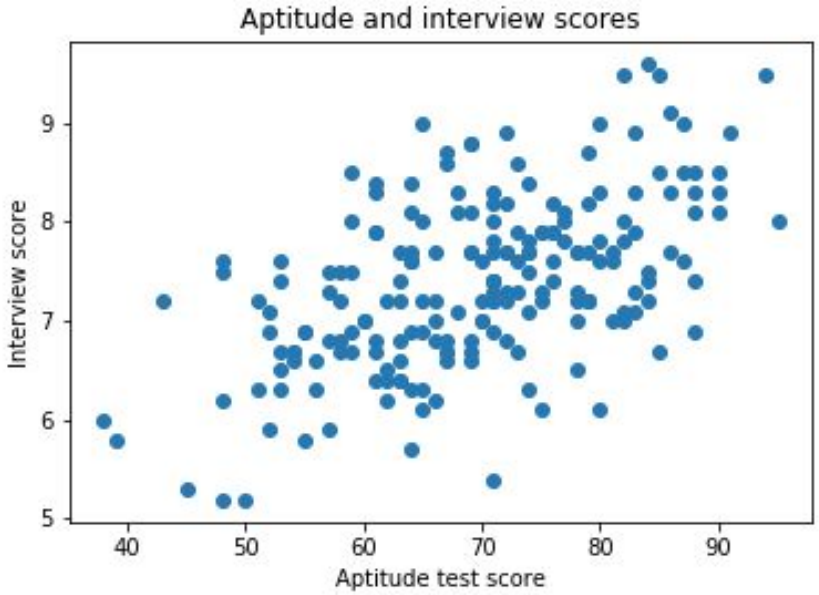
Regression is useful for:

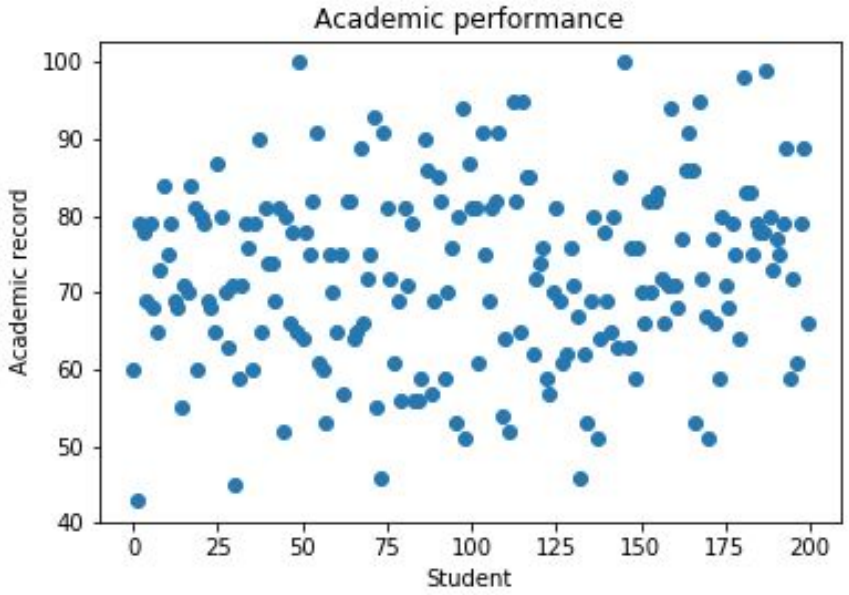
- Modelling and analysing data;
- Explaining relationships between variables (the X_js);
- Predicting numerical values.



EXPLORE DIGITAL SKILLS

Simulated aptitude and interview scores, and academic performance





EXPLORE | DIGITAL SKILLS

Linear regression mechanics

Linear regression:

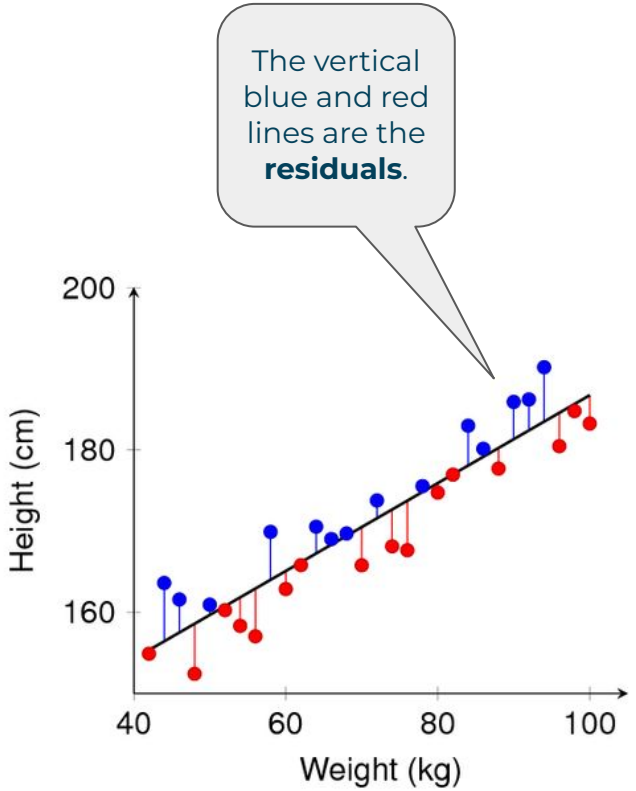
- Is the process of fitting a straight line through the data.
- Known as the line of **best fit.**

How is the line fit?

- Method known as **ordinary least squares** (OLS);
- Residuals: vertical distances between the line we fit and the data points we have (see right).
- OLS attempts to minimise the residuals to fit the line.

Common measures of fit:

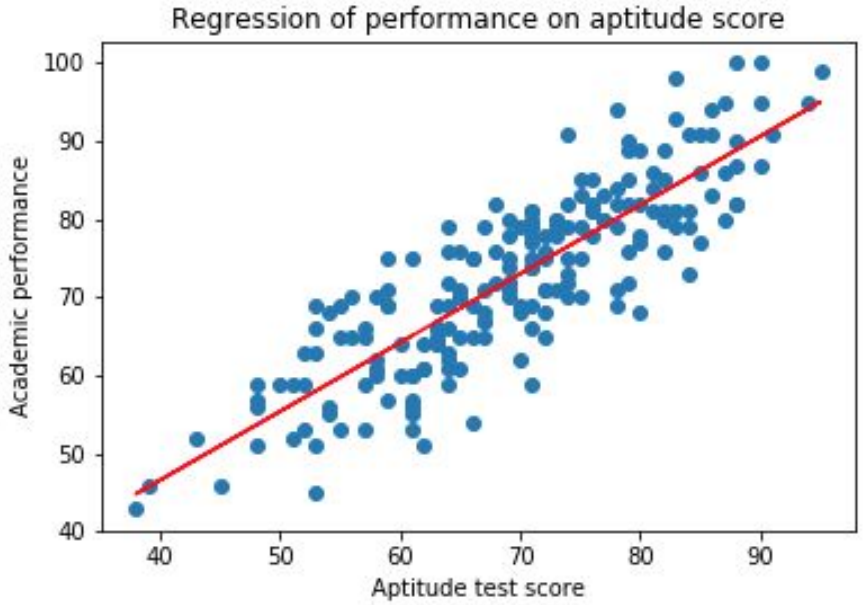
- Mean squared error (MSE);
- R^2 (quality of fit).



EXPLORE | DIGITAL SKILLS

Univariate regression 1: On aptitude test score

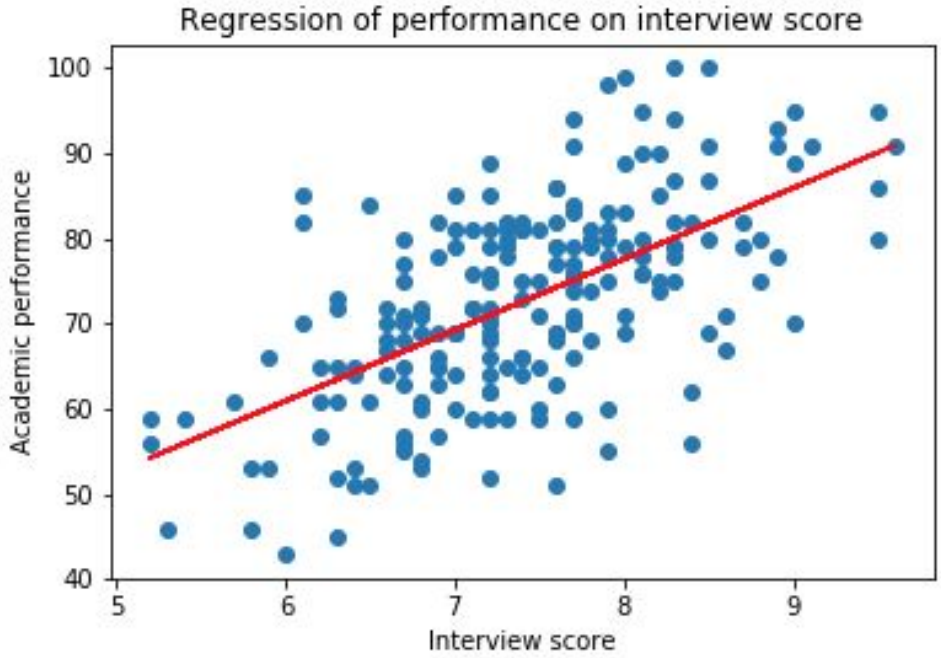
Univariate means that we only have **one predictor variable** (one X_j).



Here are some measures of fit: **Mean squared error** and **R-squared**

MSE: 37.12
R^2: 0.731

Univariate regression 2: On interview score



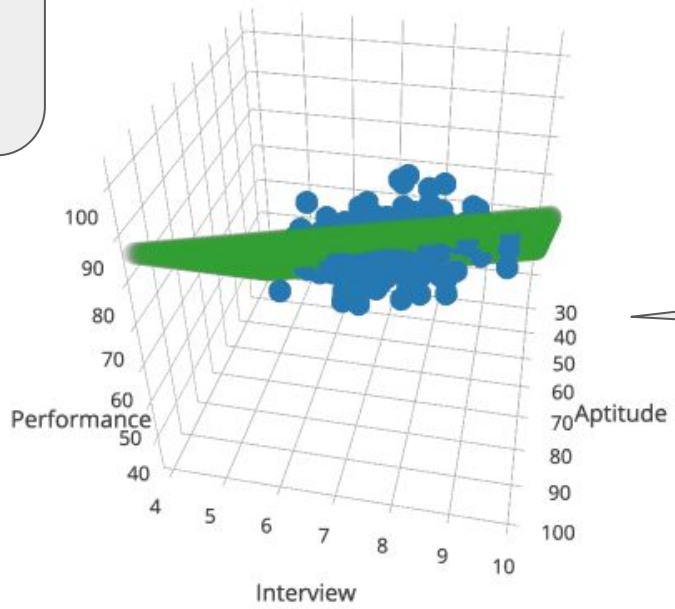
MSE: 87.42
R^2: 0.366

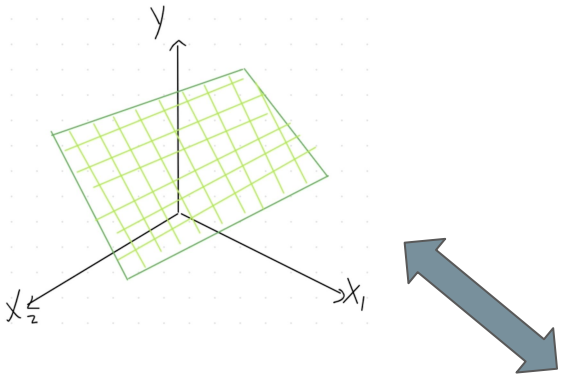
EXPLORE DIGITAL SKILLS

Multiple linear regression: On aptitude & interview scores

Multivariate means we have **multiple predictor variables** (multiple X_js).

MSE: 33.165
R^2: 0.756





With one variable, we had a **straight line**.

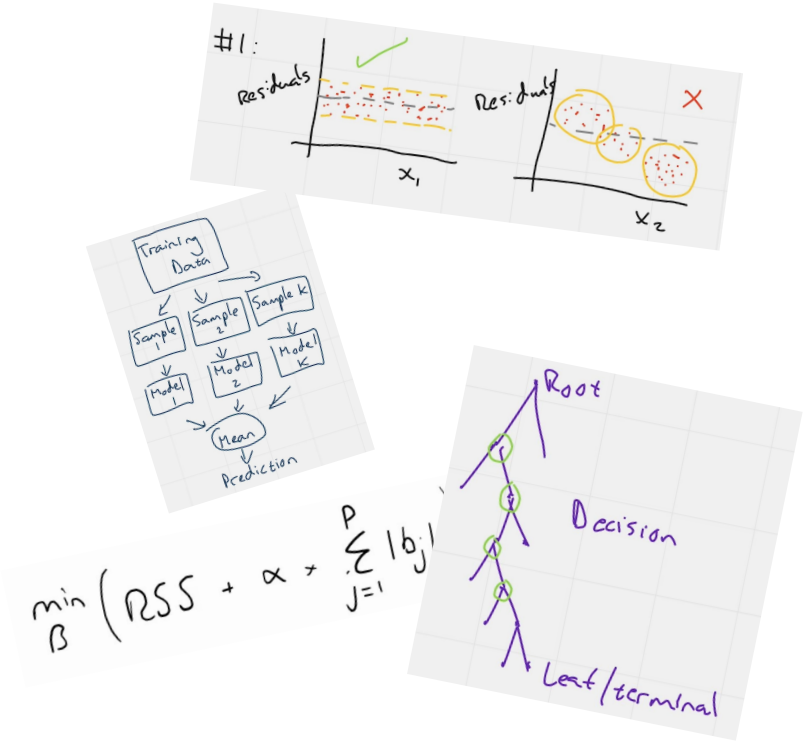
With 2+ variables, we get a **flat plane** (known as a hyperplane).

EXPLORE DIGITAL SKILLS

Model selection problems

We'll learn about ways to solve the following problems:

- Model overfitting - occurs when we have a large number of predictor variables (X_js) compared to the number of data points.
- Collinearity - when the predictor variables (X_js) are correlated to one another.
- Model interpretability - can we actually explain the model to someone, given that it has so many parameters?
- Assuming the wrong shape of *f* - there are many alternatives to linear regression, including random forests, support vector machines, decision trees, among others.



EXPLORE | DIGITAL SKILLS

Your EGAD Regression Sprint Heatmap

	Explain	Gather	Analyse	Deploy
DRAFT	Problem Statement	Problem Landscape	Equation of Value	Project Management
DO	Story Telling	Databases	Programming	Version Control
DELIVER	Communication	Data Engineering	Solution Governance	Production
DECOMPRESS	Feedback	Insights	Performance Metrics	Maintenance

EXPLORE DIGITAL SKILLS