Regression

Unleashing the **predictive power** of regression analysis

5 WEEKS **175** HOURS **9** LESSONS

In this module, we delve into the realm of **regression analysis**, covering **essential techniques and methodologies crucial for predictive modelling** and data-driven decision-making. We start by exploring fundamental concepts such as **linear regression** and **model performance evaluation**, gradually progressing to more advanced topics like **multiple linear regression**, **variable selection**, **regularisation**, **ensemble methods**, and **bootstrapping**.

Throughout the module, we'll engage in **examples** and **practical assessments** to reinforce our understanding and application of regression analysis techniques in **realistic scenarios**. By using **real-world datasets and examples**, we'll contextualise the concepts and demonstrate their practical relevance in solving complex problems.

Module objectives

Introduction to machine learning

Gain a fundamental understanding of **machine learning** and **artificial intelligence**, recognising their significance in diverse domains. Understand **predictive modelling**, distinguish between **regression** and **classification** tasks, and assess **model performance** using different evaluation metrics.

Linear regression models

Master the fundamentals of **linear regression**, including **simple** and **multiple** linear regression, and **model assessment techniques**. Become proficient in utilising **Python's sklearn library** to build, assess, and interpret linear regression models, ensuring reliability and making informed decisions based on the results.

Variable selection

Understand the significance of variable selection in increasing **model efficiency** and **interpretability** by using a variety of strategies to **choose key variables**.

Model performance evaluation

Understand the importance of **model accuracy** and its evaluation metrics, while recognising common challenges that may impact accuracy. Learn how to **implement the train-test split** in Python to **train models on training data** and **assess their performance using testing data**.

Decision trees

Understand the conceptual **structure and training process of decision trees**, implement them for classification and regression tasks, and **evaluate their performance**. Apply decision tree algorithms to real-world datasets, enabling data-driven decision-making to solve practical problems effectively.

Improving model performance

Grasp the importance of **regularisation**, **data scaling**, **ensemble methods**, and **bootstrapping** in preventing overfitting and enhancing the performance of machine learning models. Gain proficiency in applying techniques such as **ridge and LASSO regression**, along with random forests and ensembling methods, to build more **accurate** and **robust** predictive models capable of generalising well to unseen data.

Learning activities

Participating in various learning activities will **enhance our understanding** of machine learning and regression, fostering a diverse skill set for **real-world challenges** through **hands-on problem-solving** and **authentic projects**.

We learn by doing. We'll work on practical problem-solving and real-world projects.

Learn Watch animated videos and explore practical examples to learn regression concepts.

- 9 **Animated videos**
- 9 **Sketch videos**
- 2 **Slide decks**
- 17 **Examples**
- 2 **Walk-throughs**

Apply Practice regression modelling and assessment by following detailed step-by-step guides and applying these techniques to real-world scenarios.

- 9 **Exercises**

Assess Test and track your understanding of regression and its application.

- 31 **KQ assessments**
- 3 **Code Challenges**
- 2 **MCQ assessments**

Regression

Unleashing the **predictive power** of regression analysis.

Week 1

Lesson: An introduction to machine learning

In this lesson, we will look at the fundamental principles of **machine learning**, before delving into the **metrics** we can use to determine if a model is **performing** well or not.

- Differentiate between **machine learning and artificial intelligence**, understanding their interplay and significance in various domains.
- Define **predictive modelling** and explain its significance in data analysis and decision-making processes.
- Differentiate between **regression and classification** tasks.
- Explain the rationale behind the **train-test split** and its importance in assessing model performance.
- Be familiar with the different measures available to **evaluate the performance of predictive models**.

Lesson: Linear models

In this lesson, we delve into the basics of **simple linear regression** and its application in modelling the relationship between two variables to make predictions. We then look at the **least squares method** and how it is used to find the **line of best fit**. Finally, we learn how to **implement a linear regression model** using Python's sklearn library, **evaluate its performance**, and **interpret the results**.

- Understand the fundamentals of **simple linear regression** and how it uses the relationship between two variables to predict outcomes.
- Understand what **least squares regression** is and how this method is used to find the line of best fit.
- Know how to utilise Python's **sklearn library** to build and apply simple linear regression models, including **data preparation**, **model fitting**, and **making predictions**.
- Understand how to assess the performance of a linear regression model using metrics like **RSS**, **MSE**, and **R^2**, to ensure model reliability and improve prediction outcomes.

Lesson: Model performance

In this lesson, we will explore the **various aspects of evaluating model performance**, such as the **metrics** and **challenges** that underpin this evaluation. We will also look at how to assess a model's **ability to generalise to new data** and why this is an important indicator of a model's real-world performance.

- Understand the significance of **model accuracy** and the metrics used to evaluate it.
- Identify **common challenges** that can affect model accuracy and the need to know about them.
- Understand why it's necessary to **split a dataset** and the techniques we can use.
- **Implement the train-test split in Python** to create training and testing sets.
- **Train a model on the training data** only and **assess** its performance **using the testing set**.

Week 2

Lesson: Multiple linear regression

In this lesson, we will cover the **fundamentals of multiple linear regression**, including its assumptions, implementation in Python using libraries like **sklearn** and **statsmodels**, and evaluation techniques. We'll explore how to check for **linearity**, **multicollinearity**, **independence**, **homoscedasticity**, **normality**, and **outliers** in regression models.

- Implement multiple linear regression models using **sklearn** and **statsmodels**.
- Check for **violations of** multiple linear regression **assumptions** and **interpret their implications**.
- **Visualise** relationships between **variables and residuals** to identify **patterns**.
- Use **diagnostic plots** and statistical tests to **assess the validity of regression models**.
- Apply techniques to handle **multicollinearity** and **influential outliers** in regression analysis.
- **Interpret regression model outputs** and make informed decisions based on the results.

Lesson: Variable selection and model persistence

In this lesson, we will examine **variable selection techniques**, and how to apply them to select the **most informative features** for our models. We will also look at how to effectively **save** and **load** a **trained machine learning model** for future use, ensuring it can be embedded into real-world systems, shared, or deployed without the need to retrain.

- Understand the concept of **variable selection** and **why it is important**.
- Know and **apply variable selection techniques**, and observe their impact on model performance.
- Know how to **save a trained model** and its parameters for future use.
- Know how to **load a saved model** and use it.

Lesson: Regularisation

In this lesson, we'll unravel the mechanics and **benefits of regularisation methods**, including **ridge** and **LASSO** regression. We'll delve into the intricacies of **data scaling**, **overfitting**, and the strategic application of **regularisation** to **enhance model performance** in real-world data science scenarios.

- Explain the concept of **overfitting** and the **importance of regularisation** in machine learning.
- Implement **data scaling techniques** to improve model performance.
- Apply **ridge and LASSO regression methods** to prevent overfitting.
- Differentiate between **L1 and L2 regularisation** and understand their applications.
- **Utilise regularisation techniques** to build more accurate and robust machine learning models.

Week 3

Lesson: Decision trees

In this lesson, we will delve into the **fundamentals of decision trees**, exploring **how they work**, **how to train them effectively**, and **how to implement them** using Python libraries like sklearn. Through a combination of theoretical explanations, practical examples, and hands-on coding exercises, we will gain a comprehensive understanding of decision trees and their application in real-world scenarios.

- Understand the **conceptual structure** and workings of **decision trees**.
- Explain the process of training a decision tree model, including **partitioning** and **recursive binary splitting**.
- **Implement decision trees** for both classification and regression tasks using sklearn.
- **Evaluate the performance of decision tree models** using appropriate metrics such as **accuracy** and **mean squared error**.
- Recognise the **advantages and disadvantages of using decision trees** in machine learning applications.

Lesson: Ensemble methods and bootstrapping

In this lesson, we delve into ensemble methods and bootstrapping to significantly improve the **robustness**, **accuracy**, and **generalisability** of predictive models.

- Comprehend the **foundational principles and applications of ensemble methods** and bootstrapping in data science.
- Analyse the effectiveness of various **ensemble strategies**, such as **bagging**, **boosting**, and **stacking**, in enhancing model performance.
- Apply **bootstrapping techniques** to estimate model accuracy and stability.
- Integrate ensemble methods and bootstrapping into data science projects to **address complex predictive modelling challenges**.

Week 4

Lesson: Random forests and applying our knowledge

We have previously looked at ensemble learning where we **combine multiple models** to improve overall performance. In this lesson, we look at a well-known application of this concept, the random forest, which leverages the **collective strength of multiple decision trees** to produce more accurate results. We will examine how random forests work and how to apply them in Python.

In this lesson, we also apply the knowledge we have acquired in the earlier weeks by going through the **data science process** of evaluating a dataset from scratch and testing various modelling methods **to solve a regression problem**.

- Understand the fundamentals and mechanisms underlying **random forests**.
- Understand how random forests **mitigate the overfitting problem** in single decision trees.
- Know how to **build**, **evaluate**, and **apply a random forest** model in Python.
- Apply the **data science workflow** and **regression techniques** to solve a regression problem.

Week 5

Exam: **Regression**

It is time to review all the work that's been covered up to now and **test our knowledge**. Make sure to **cover each week's section in detail** before attempting this exam! It will be a combination of **theoretical** and **practical questions**, aimed at testing the **general understanding of concepts**, as well as the **application** and **interpretation of our new skills**.

Module summary

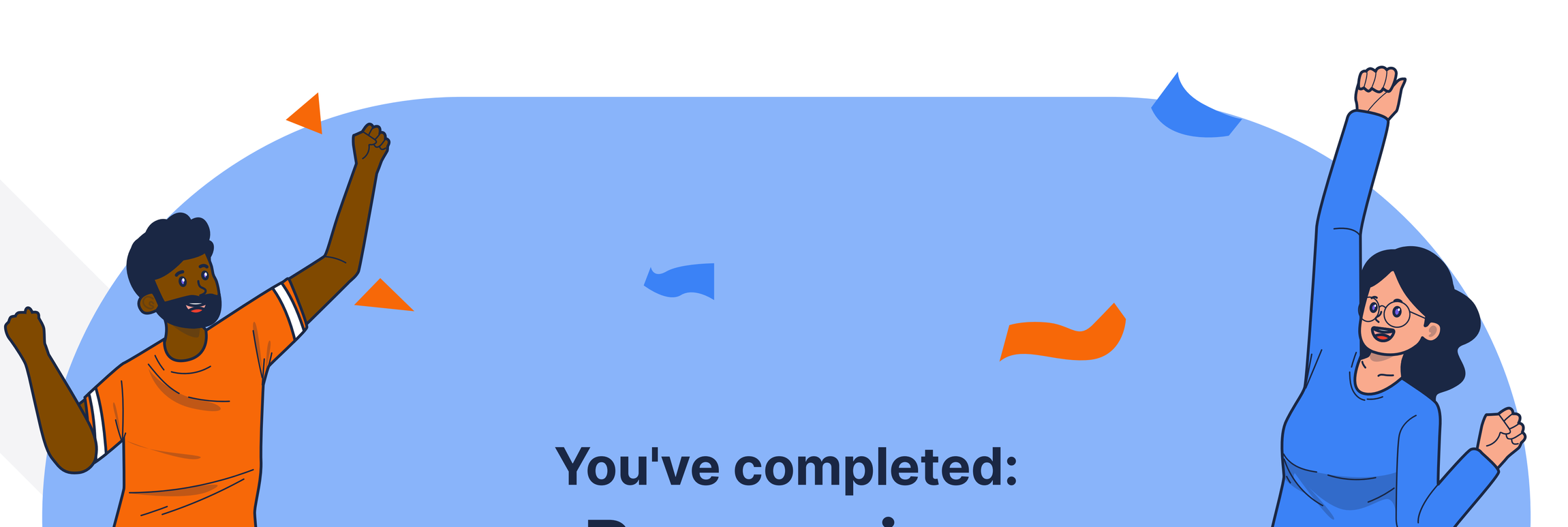
Throughout this module, we've **explored regression analysis** comprehensively, covering fundamental concepts and some more advanced techniques to model relationships between variables effectively. From understanding **linear and multiple linear regression** to exploring **variable selection**, **regularisation**, **ensemble methods**, and **bootstrapping**, we've gained valuable insights and practical skills essential for predictive modelling in various real-world scenarios.

By completing this module, we've equipped ourselves with a robust understanding of regression analysis and its applications, empowering us to **tackle complex data challenges** and make informed decisions using machine-learning techniques.

What's next?

As we conclude this module, we're ready to **apply our newfound knowledge and skills** to real-world data projects and problem-solving scenarios. In the next phase of our journey, we'll delve deeper into other **advanced topics** such as classification methods, natural language processing, and unsupervised learning, building additional skills and expertise to drive actionable insights and optimise decision-making processes.

Additionally, we'll continue to **refine our analytical abilities** and **expand our toolkit** by exploring other data science methodologies and technologies. Our commitment to lifelong learning and continuous improvement will propel us towards success in the dynamic and ever-evolving field of data science.



EXPLORE AI ACADEMY