SESSION PLAN

Session Name	EDA & Data Preprocessing

Learning Outcomes

- Reinforce your understanding of the concepts through a mentor-led onboarding session
- Revisit the topics and understand
    - Why preprocess and clean data
    - Outlier detection and treatment
    - Data Transformation and standardization
    - Data Encoding

Prerequisites for the Student

- Complete all the learning tasks & assessments in EDA and Data Preprocessing concept
- In case, the tasks are difficult, at least go through the relevant content and come prepared for the session.

Student Activities

- Key topic onboarding in this concept (2hrs)
    - Why preprocess and explore data - learn the importance of EDA and preprocessing for Data Science. Note that this is the phase where a lot of time is actually spent. There are no hard and fast rules here and everything depends on the data and use case we are working with.
    - Outlier detection and treatment - learn that boxplots are the best way to detect outliers. Sometimes removing outliers have adverse effects and for cases like fraud detection, outliers are the important points that must be retained. Talk about these points in detail.
    - Detecting and handling missing data - while handling missing data - focus more on what techniques you are using as per your hands-on experience. When do we impute missing data and when do we drop the data? Practical scenarios would improve learner understanding.
    - Data Transformation - Talk a little bit on the previous linear regression exercise where we applied the data transformation - and how the data transformation ensured that the data followed the assumptions closely.
    - Data encoding - explain in detail the encoding process. Learners generally get confused regarding the fit() and transform() which is very similar to training in sklearn. Clarify that upfront. Also talk in detail about the phenomenon of data leakage - and how it can introduce bias in your results.
- Code Along Notes (3 hrs)
    - Applying skills to solve a problem
        - Quiz learners on how to solve the problem posed given the concept that they have already learned. Let them come up with the approach.
    - Adapting to something new
        - How to look for help in documentation and quickly solve problems.

        - For error debugging how to quickly look at stackoverflow/google
    - The code along talks about the automobile price prediction.
        - Note that there are two files - one for EDA and other for Data Preprocessing.
        - For every visualization talk about the business insights you can infer and ask learners questions on the same.
        - For data preprocessing, talk about how missing values not only denote NaN but can come in any form like '?' or other junk values.
        - The details of the code along can be seen in the platform. Access the solution within the platform and keep in handy outside - printout / another screen different from where you are coding.

Next Session

- Concept - Logistic Regression (20-30 mins)
    - Talk about how this is a classification algorithm - which is different from the regression algorithm. So the metrics are quite different and it is going to take some time.
    - Highlight the topics that are going to be difficult for learners.
    - Some of the topics might take time to understand and that is ok. Everyone finds these topics quite difficult.
    - If some of the assignments are difficult, tell the learners to take the help of hints and solutions. Set the expectation to the learners to atleast go through the material end to end before the session (attempt the tasks but if you are unable to do them, that is ok. They can attempt post the session)