Usecase 7 - (Lab)

By: eng. Esraa Madhi

So we defined data science as: It's the process of asking interesting questions, and then answering those questions using data.

For any **Data project** we will go through these steps:

1. Defining the Problem Statement
2. Collecting Data
3. Data Quality Checking and Remediation
4. Exploratory Data Analysis
5. Building Machine Learning Models
6. Model Evaluation
7. Communicating Results
8. Model Deployment
9. Model Performance Maintenance in Production

usecase 7

Step 1: Defining the Problem Statement

We are interested in developing a robust predictive model that can accurately estimate the transfer values of football players based on a comprehensive set of variables. This model aims to leverage both basic player information and detailed professional statistics to determine a player's market value when they are transferred between clubs.

Step 2: Collecting Data

Football players dataset was collected for this lab encompassing player demographics (age, height, playing position) and performance metrics (goal

scoring, assists, injury history) for the seasons 2021-2022 and 2022-2023.

Step 3: Data Quality Checking and Remediation

Step 4: Exploratory Data Analysis

- For these two steps, make sure to do:
  a. Data Profiling: apply the 7 types of data profiling
  b. Data Cleaning: handle missing values, correcting errors, and dealing with outliers.
  c. Univariate Analysis &Bivariate/Multivariate Analysis: to understand their distribution and look at the relationships between variables. For your visualizations make sure to:
     - Drive meaningful insights that would help you in building models (at least 3 different charts).

Step 5: Building Machine Learning Models

Step 6: Model Evaluation

The requirements for students during the model training phase, along with the types of evaluations required:

- Feature Engineering: Apply feature engineering techniques to create new features or modify existing ones such as :
  - Encode categorical variables
  - Normalize or standardize numerical features.
- Model Training: Train **the selected** models on the training dataset. Ensure that you have a separate validation set or employ cross-validation to assess model performance during training.
- Performance Metrics: Use appropriate performance metrics to evaluate the models like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared value.
- Hyperparameter Tuning: Fine-tune the hyperparameters of each model to optimize performance. Utilize techniques like grid search.
- Model Validation: Validate the model's performance using the test set to ensure that the model generalizes well to unseen data.

- Overfitting Check: Ensure that models are not overfitting by comparing training and validation performance.
- Underfitting Check: Ensure that models are not underfitting by assessing if the performance is significantly poor both on training and validation sets.

Step 7: Communicating Results

When conducting the analysis and building the predictive model, it is crucial to maintain clear and comprehensive documentation. This will not only facilitate understanding and reproducibility of the work but also allow others to follow and build upon your methodology. Use **markdown cells** in your notebook to provide detailed commentary on the following key aspects:

- Feature Engineering: Summarize the feature engineering process.
- Hyperparameter Optimization: Outline the approach taken for hyperparameter tuning.
- Performance Metric Visuals: Include visuals of model performance metrics.
- Feature and Prediction Insights:
  - Feature Importance: Note which features significantly impact predictions.
  - Prediction Interpretation: Explain how feature values influence predictions.
  - Limitations: Acknowledge any model limitations and suggest areas for improvement.

Step 8: Model Deployment

Not applicable

Step 9 : Model Performance Maintenance in Production

Not applicable