Prediction and Decision Making - Lesson Summary Notes

==

What You Learned:

Linear Regression Concepts

- Linear regression uses **one independent variable** to predict a **dependent variable (y)**.
- **Simple Linear Regression (SLR)**: Analyzes the relationship between one x and one y variable.
- **Multiple Linear Regression**: Involves two or more predictor (x) variables to model a continuous target (y).

Visualization Tools with Seaborn

- Use `regplot()` for regression visualization.
- Use `residplot()` to inspect residuals and assess model fit.
- A good residual plot:
  - Residuals centered around zero
  - Even distribution across x-axis
  - Constant variance (homoscedasticity)

Distribution Plots

- Compare predicted vs. actual values.
- Especially helpful when using multiple features in a regression model.

Polynomial Regression

- Polynomial regression fits a non-linear curve.
- The **order of the polynomial** affects model flexibility and fit.
- Use `np.polyfit()` for creating polynomial regression models.

Feature Transformation & Normalization

- Use `PolynomialFeatures` from `sklearn.preprocessing` to expand features.
- Use `StandardScaler` to normalize data (zero mean, unit variance).
- Proper transformation improves model accuracy and interpretability.

Pipeline in scikit-learn

- Pipelines automate the workflow: transformation training prediction.
- Simplifies code and prevents data leakage.
- Example tasks handled in a pipeline:
  - Normalization
  - Polynomial feature generation
  - Model training & prediction

Model Evaluation Techniques

- Use `mean_squared_error` to measure the **average squared difference** between predicted and actual values.
- Use `.score()` or `r2_score()` for the **R-squared (coefficient of determination)**:
  - Closer to **1.0** = better fit
  - Negative R^2 = poor model or overfitting

Interpreting Model Fit

- Good model:
  - High R^2 (e.g., > 0.8 depending on context)
  - Low MSE
- Poor model:
  - Low R^2

  - High MSE

  - Residual plot shows patterns (non-randomness)

Best Practices

- Use both **visual** (e.g., plots) and **numerical** (e.g., MSE, R^2) metrics to evaluate models.

- A distribution plot is ideal for **multiple linear regression** diagnostics.

- Always validate assumptions using residual plots:

  - Random residuals good model

  - Curved or patterned residuals non-linearity or model issues

Tip:

Understand the context of your data. An "acceptable" R^2 score depends on the problem domain.