Practical Business Python

Lecture 6: Exploratory Data Analysis in Python.

Iegor Vyshnevskyi

Woosong University

October 19, 2023

Agenda

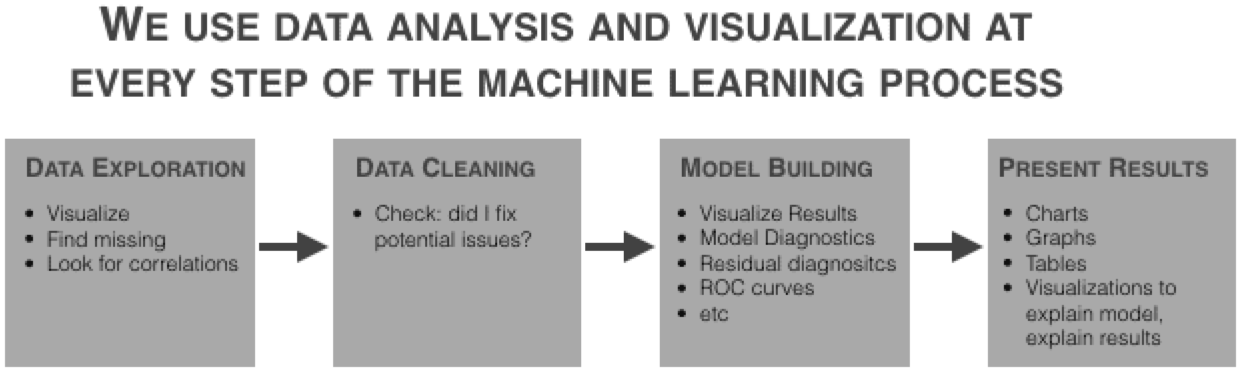
1. Intro to Exploratory Data Analysis (EDA)
2. Importance of EDA
3. Python packages for EDA
4. Common EDA Techniques
5. Summarizing and Computing Descriptive Statistics
6. In-class Assignment

1. Intro to Exploratory Data Analysis (EDA)

Intro

Exploratory Data Analysis (EDA) is the process of visually and quantitatively examining datasets to summarize their main characteristics and patterns, typically using statistical graphics, plots, and other data visualization techniques.

The **primary goal of EDA** is to understand the underlying structure, patterns, and relationships in the data to guide further analysis or model building.



WE USE DATA ANALYSIS AND VISUALIZATION AT EVERY STEP OF THE MACHINE LEARNING PROCESS

DATA EXPLORATION
- Visualize
- Find missing
- Look for correlations

DATA CLEANING
- Check: did I fix potential issues?

MODEL BUILDING
- Visualize Results
- Model Diagnostics
- Residual diagnositcs
- ROC curves
- etc

PRESENT RESULTS
- Charts
- Graphs
- Tables
- Visualizations to explain model, explain results

Source: Picture from Joshua Ebner. How to use data analysis for machine learning (example, part 1)

Intro

- You can either explore data using graphs or through some *python* functions.
- There are two type of analysis. *Univariate and Bivariate*.
  - In the univariate, you analyzing a single attribute.
  - But in the bivariate, you analyzing an attribute with the target attribute.
- In the *non-graphical approach*, you will be using functions such as shape, summary, describe, isnull, info, datatypes and more.
- In the *graphical approach*, you will be using plots such as scatter, box, bar, density and correlation plots.

- **Remember**:
  - EDA is often an iterative process where initial findings might lead to further data cleaning, visualization, or modeling efforts, refining the analysis and insights.
  - Documenting EDA findings, insights, and the visualizations generated is crucial for sharing results and collaborating with other stakeholders.

Intro

Key Steps in EDA:

- Importing a dataset
- Understanding the big picture
- Preparation
- Understanding of variables
- Study of the relationships between variables
- Brainstorming

Intro

Key Steps of EDA in more details:

- *Data Loading*: Load the dataset into Python using libraries like Pandas, ensuring it is in a suitable format for analysis.
- *Data Cleaning*: Identify and handle missing values, outliers, and anomalies that could affect the analysis and conclusions.
- *Summary Statistics*: Compute basic statistics (mean, median, standard deviation, etc.) to summarize the central tendency and dispersion of the data.
- *Data Visualization*: Create various plots (histograms, scatter plots, box plots, etc.) to visualize the distribution, relationships, and trends in the data.
- *Exploratory Data Visualization*: Use advanced visualization techniques to delve deeper into relationships and patterns within the data.
- *Correlation Analysis*: Explore correlations between variables to identify potential predictive relationships.
- *Feature Engineering*: Derive new features or modify existing ones to improve the quality of input data for machine learning models.
- *Dimensionality Reduction*: Reduce the number of features while preserving essential information, aiding in model efficiency and interpretability.

2. Importance of EDA

Why EDA

- EDA helps in detecting errors, outliers, and anomalies.
- It provides a comprehensive understanding of the data's structure, improving modeling decisions.
- EDA guides feature selection and engineering, optimizing model performance.
- EDA aids in choosing appropriate machine learning algorithms based on the data's characteristics.

- EDA is *crucial* in various domains like finance (analyzing stock market trends), healthcare (clinical data analysis), marketing (customer segmentation), and more.

3. Python packages for EDA

What we use…

- *Pandas*: For data loading, cleaning, and initial data manipulation.
- *NumPy* and *SciPy*: For statistical analysis and mathematical computations.
- *Matplotlib* and *Seaborn*: For creating various visualizations.
- *Plotly*: For interactive and advanced visualizations.

4. Common EDA Techniques

What we use…

- *Histograms and Distributions*: To understand data distribution.
- *Scatter Plots*: To observe relationships between two variables.
- *Box Plots*: To detect outliers and distribution characteristics.
- *Heatmaps and Correlation Plots*: To visualize correlations between variables.
- *Pair Plots*: To visualize relationships across multiple variables.

5. EDA Example

What can be done…

- Please open the file “L6_work”.

5. EDA Practicum

What you will be doing…

- It's group work. You will be assigned to groups.
- Perform EDA for a given dataset in file “titanic.csv”, and finish by 2:40 pm.
- Present your findings briefly (3 min. per group).

You can get the file from LMS or running code

“**titanic = sns.load_dataset('titanic')**”

5. In-class Assignment