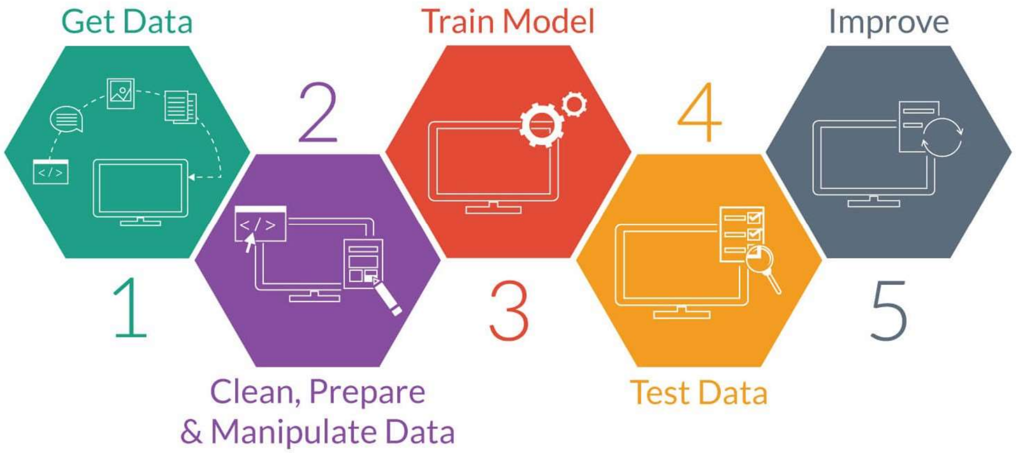
IU 4.6.3
Exploratory Data Analysis (EDA) for Classification

Topics

- Machine Learning Process
- Classification
- Data Exploration
- Visualizing for Classification

Machine Learning Process

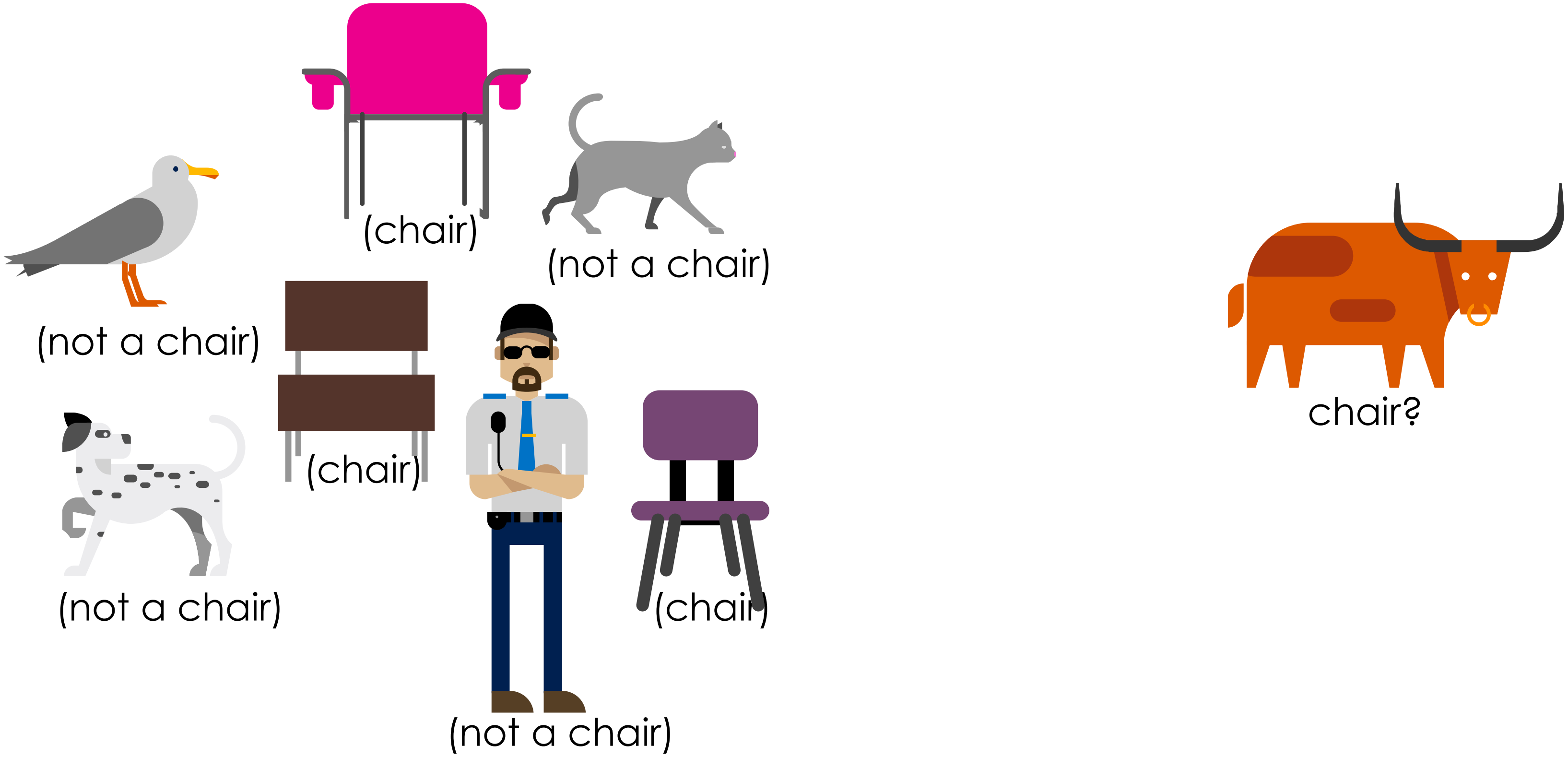
- 1. Get Data
- 2. Clean, Prepare & Manipulate Data
- 3. Train Model
- 4. Test Data
- 5. Improve (Iterate)



Data Preparation

- Sometimes can take up > 80% of time
- GI-GO : Garbage In Garbage Out
- Your model/prediction depends on how good the data used for training the model

Classification (Supervised Learning)



Features Type

- Numeric
  - Discrete
  - Continuous
- Category
  - Nominal: country, gender, race, hair color, blood type
  - Ordinal: Shirt size, age group,

Features Type (Quiz)

- What Type are these features (Numeric or Category?)

  If numeric (Continuous or Discrefe), If Category (Nominal or Ordinal?)

1. Customer Experience ?
2. Mile Per gallon (MPG)
3. Car Engine Location
4. Car number of doors
5. Origin Airport Code
6. Flight Departure Time

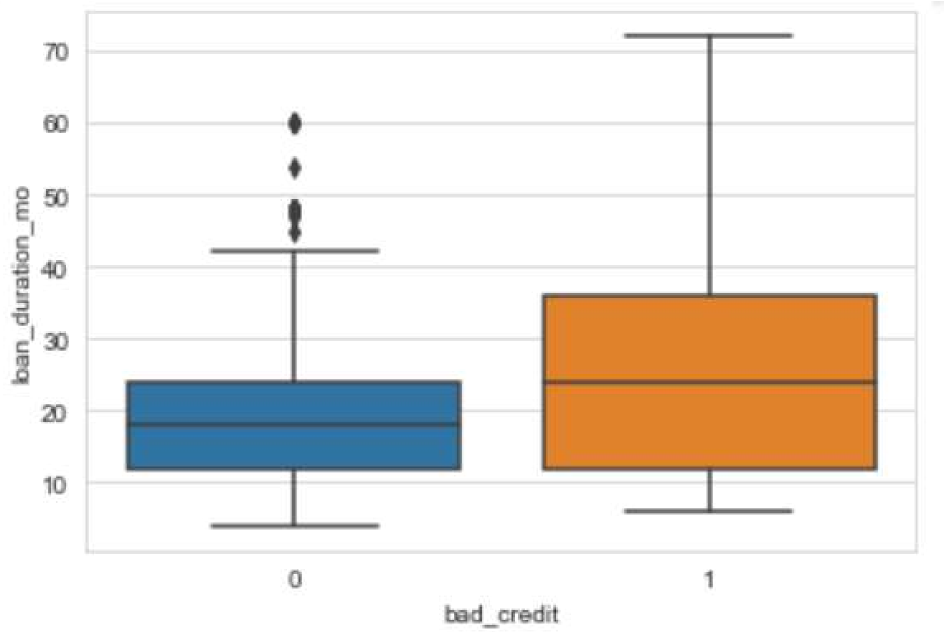
Visualizing for Classification

- Visualizing Numeric features
  - Using Box Plot
  - Using Violin Plotn (1 or 2 dimensions)
- Visualizing Categorical features
  - Using Bar chart or histogram

Box Plot

- X-axis: categorical label

  Y-axis: numeric features value

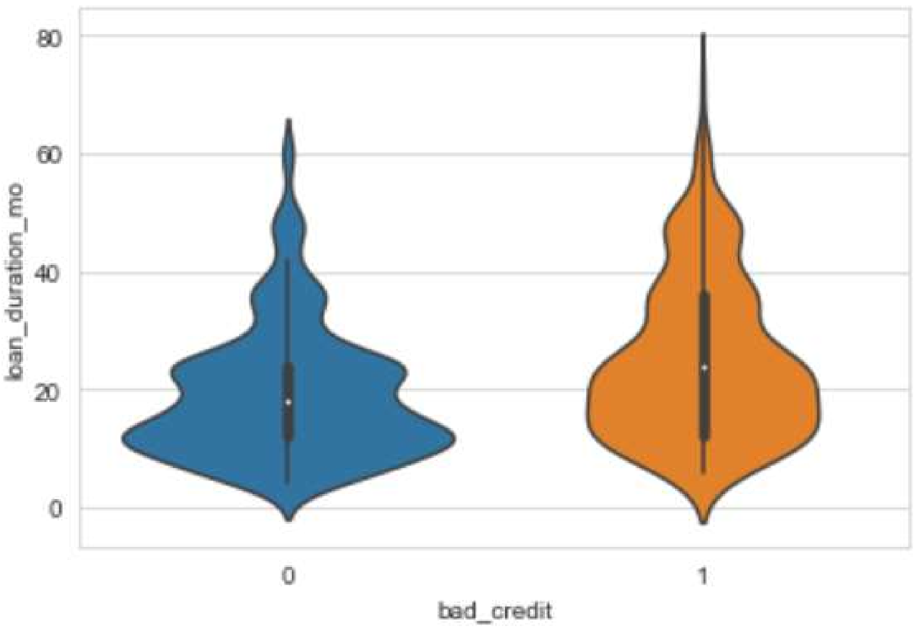


Violin Plot (1 dimension)

- X-axis: categorical label

  Y-axis: numeric features value

Similar to Box Plot, but violin plot also visualize the distribution of the numeric features



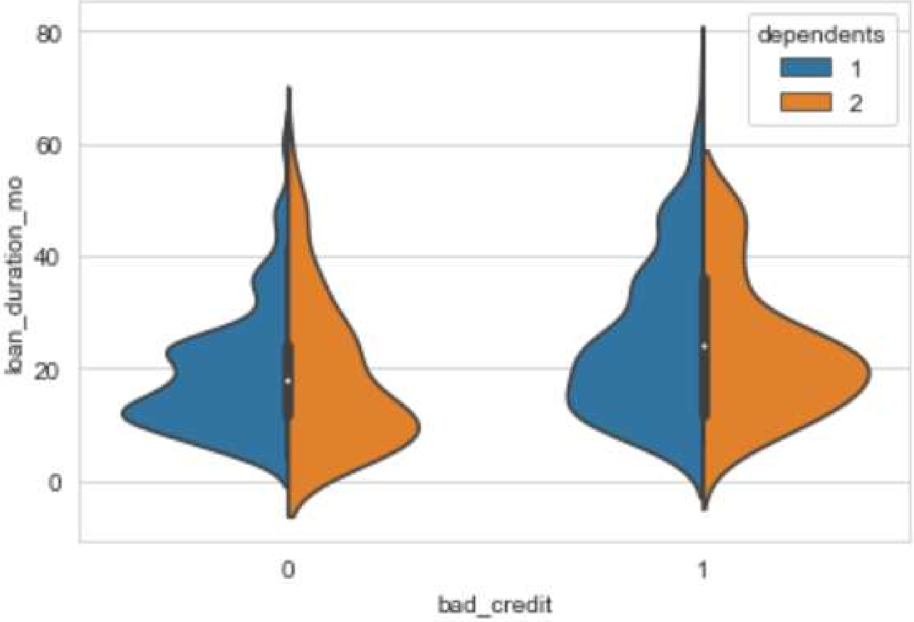
Violin Plot (2 dimension)

- X-axis: categorical label

  Y-axis: numeric features value

Use hue to split the violin chart to 2 dimensions (left & right)

sns.violinplot(x=col_x, y=col, data=credit,hue="dependents",split =True)



Frequency Tables

- Used to visualize categorical features
- X-axis: category name ; Y-axis: count (numeric)
- Normally presented as Bar/Column Chart or histrogram
- Can be one dimension or two dimensions
- Can be used to visualize the distribution of each category (how balance/imbalance of your data)