Week – III
Exploratory Data Analysis & Feature Engineering

ML Bootcamp 2021

GDG Cloud EDMONTON

AlbertaAI

Careera Analytics Lab

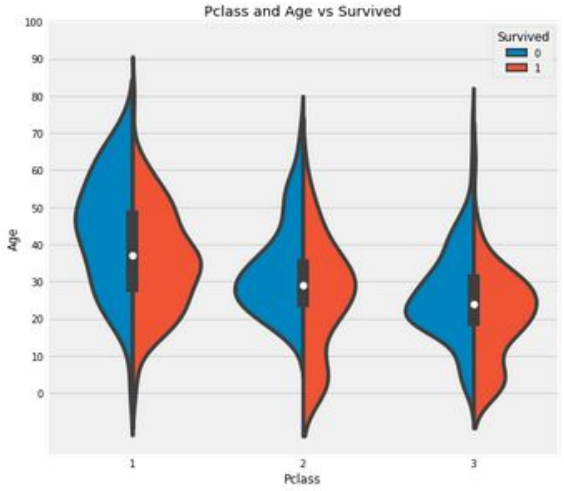
Slido: #gdg_ml_bootcamp_w3

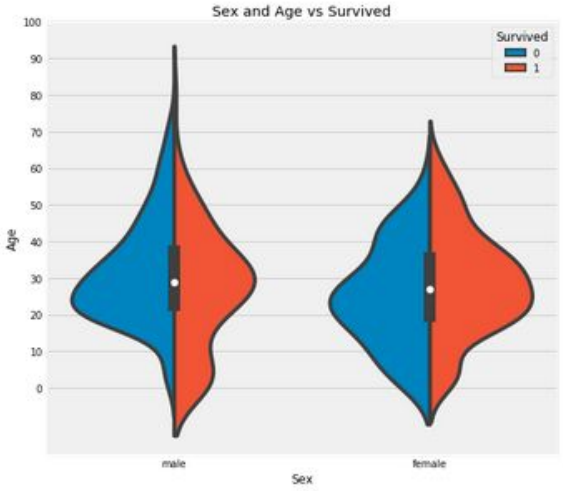
Table of contents

Slido: #gdg_ml_bootcamp_w3

Exploratory Data Analysis

- Just as the title says, it is all about exploring your data
- You will have lots of data: numerical, categorical, text etc.
- Each data needs to be explored in a different way based on the type of the data



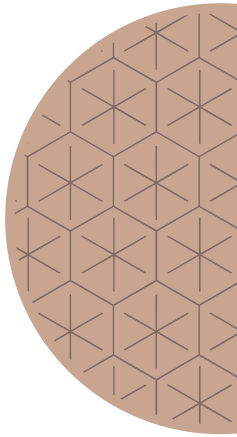


Slido: #gdg_ml_bootcamp_w3

EDA is about creating a story...

- Exploring the data is all about asking questions
- You need to investigate and “Explore” the data
- For Example:
  - How many people survived/didn’t survive?
  - What is Pclass?
  - Is there a relationship between Pclass and Survived?
  - How is age related to the survival rate?

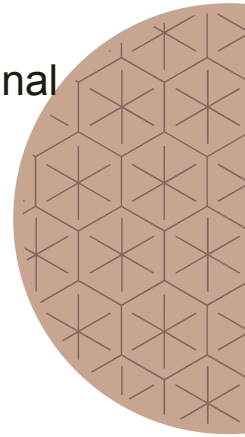
- These questions help us move on to the task of Feature Engineering



Slido: #gdg_ml_bootcamp_w3

Dimensionality Reduction Techniques

- The data that we collect usually has large number of columns. Each column is called a feature
- Visualizing beyond 3D is beyond the scope of any human being as of today. Hence, we resort to visualizing data in 1,2 or 3D
- Dimensionality reduction allows us to better understand very high dimensional data and helps us deal with the issue of “Curse of Dimensionality”



Slido: #gdg_ml_bootcamp_w3

Feature Engineering

- ”Features” are nothing but the characteristics that define your data.
- For example
  - Height, weight, age are three features that describe a human being or an animal.
  - Number of pages, color, paperback or hardcover are features that describe a book
- Curse of dimensionality is one reason for Feature Engineering or Dimensionality Reduction

Sample Image

Features: ?

Sample Review

The earphones that I purchased were working very well at first but later stopped working completely

Features: ?

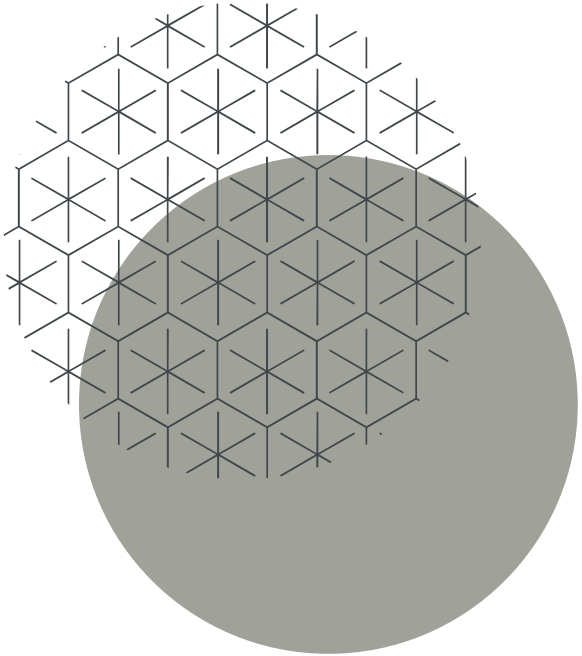
Different Methods of Feature Engineering

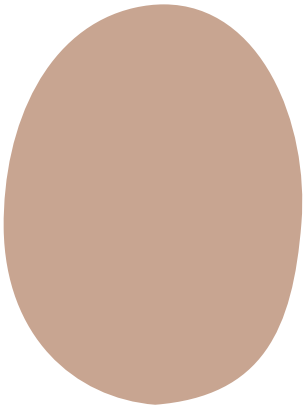
- **Feature Extraction**
  - Derive/create new features from existing features
  - These features are either derived from existing features based on calculations or dimensionality reduction techniques such as PCA(although originally intended to reduce the dimensions for better visualization) may be used
  - One hot encoding is another example of generating new features based on existing ones.
- **Feature Selection**
  - This process involves select a subset of the features from your existing set of features
  - Various statistical techniques, regularization and feature importance techniques can be used here
  - SelectKBest and Recursive Feature Elimination(RFE) provided by scikit-learn are a good start
- **Domain knowledge specific features**
  - This category of feature engineering comes from domain knowledge
  - A Subject-Matter Expert(SME) is a person who is the expert in that field and will be

Dealing with Data

- A Machine Learning algorithm can only accept number, and hence we need to convert all our data into numbers that the machine can understand
- **Numerical data:**
  - Since this data is already in the form of numbers some of the work is done
  - However, looking across different numerical features, sometimes you will notice an imbalance in the scale of the data
  - Example: Age and Salary.
- **Categorical data:**
  - This data consists of categories. Eg: Gender, T-shirt size, Weather Condition etc.
  - All this data is in the form of strings.
  - You would need to convert it into numbers using techniques such as encoding the categories as numbers(One hot encoding, label encoding etc.)
- **Text data:**
  - Free text data is also very commonly collected. Eg: Emails, Survey responses, Chats from chatbots etc.
  - All this text needs to be converted into numbers for the machine to be able to understand it. Vectorizing these texts using techniques such as Word Embeddings, TF-IDF and CountVectorizer

Lets start Practicing!





Slido: #gdg_ml_bootcamp_w3

Thanks

Do you have any questions?


slidesgo