COMP 333 — Week 2 Example 3: Titanic

Titanic Survival Dataset

There is an ongoing challenge on Kaggle on the Titanic disaster:
`https://www.kaggle.com/c/titanic`.

As a result there is plenty of information on how to analyse the data.

None of these has provided a conclusive "analysis", so the challenge is still ongoing.

Maybe you can do better?

> The sinking of the Titanic is one of the most infamous shipwrecks in history.
>
> On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there werent enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.
>
> While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.
>
> In this challenge, we ask you to build a predictive model that answers the question: *"what sorts of people were more likely to survive?"* using passenger data (ie name, age, gender, socio-economic class, etc). [from Kaggle]

Our Example 3 follows the article *A Gentle Introduction to Exploratory Data Analysis* by Daniel Bourke that is included in the supplementary material. The supplementary material also includes his Jupyter notebook, and a very long video of him working through the notebook.

Overview

The article is indeed a gentle introduction.

He discusses

- an EDA checklist
- types of data
- how to deal with missing values
- outliers
- feature engineering: add, remove, change features
- contribution of a feature to a model
- building a model

And he summarises:

We covered a non-exhaustive EDA checklist with the Titanic Kaggle dataset as an example.

1. What question are you trying to solve (or prove wrong)?

   Start with the simplest hypothesis possible. Add complexity as needed.

2. What kind of data do you have?

   Is your data numerical, categorical or something else? How do you deal with each kind?

3. What's missing from the data and how do you deal with?

   Why is the data missing? Missing data can be a sign in itself. You'll never be able to replace it with anything as good as the original but you can try.

4. Where are the outliers and why should pay attention to them?

   Distribution. Distribution. Distribution. Three times is enough for the summary. Where are the outliers in your data? Do you need them or are they damaging your model?

5. How can you add, change or remove features to get more out of your data?

   The default rule of thumb is more data = good. And following this works well quite often. But is there anything you can remove get the same results? Start simple. Less but better.

Next

You can see the details of this example of EDA in the supplementary material:

- the article by Daniel Bourke
- the corresponding Jupyter notebook
- the corresponding video working through the Jupyter notebook (over 2 hours long!)

The article is highly recommended. READ the article.

Remember to use this example as an introduction.

It gives you an overview of EDA and gently helps clarify many concepts about data, data analysis, and EDA.

You do not need to understand everything just yet.

Come back again and again to the example as you meet the details of concepts, steps, and techniques in future lectures.

We will return in future lectures to

Descriptive Data Analysis

Missing Values

Exploratory Data Analysis

Feature Engineering

Modeling