Data Science Honors Course: Assignment 1: Titanic Data Preprocessing

Contents:

Plagiarism
1. Students may discuss with each other for information sharing regarding solutions, tools, and techniques to approach the assignment. However, they should implement their own code from scratch. Students found violating general code of conduct might be penalized by reducing their scores.

2. These assignments are for learning and should be the mutual goal of everybody, teachers, and students alike.

Problem Definition
Access an open source dataset “Titanic”.
Apply pre-processing techniques on the raw dataset.

Data analysis

1. Data Acquisition
2. Data cleaning/transforms
3. Data Visualization
4. Feature Engineering
5. Dimensionality Reduction/Feature Selection

Question Answers

1. What is Data?
Ans: Collection of facts, numbers, words, measurements, observations ,etc
2. What should be the data format?
Ans: Structured, Tabular, Well-defined
2. How do we acquire it?
Ans: Sensors,People counter, Cookies data, Surveys, records,etc.

3. What is data cleaning?
Ans: Processing to more readable format, Identifying outliers, removing errors and missing values.

4. Why to do data cleaning?
Ans:Formatting, Data-type - validating data, making reliable dataset.
5. What is feature engineering?
Ans: to extract features from raw data.
6. Why to do feature engineering?
7. What is visualization?

Ans: It is graphical representation of data using elements like charts, graphs ,etc to see and understand trends, patterns and outliers.

8. What are the various types of data?

Ans 1. Textual- written , printed data

2. Categorical-

- Stored in groups with labels
- Groups are made on characteristics
- Nominal data
- Ordinal data

3. Numerical-

- Discrete data- countable elements
- Continuous data- height,temperature ,etc

9. What are the techniques to handle different kinds of data?

1. Textualq
    1. Removing stop-words, punctuation
    2. Stemming
    3. Embedding
    4. pronoun-noun- entity recognition
2. Categorical
    1. Nominal
    2. Ordinal
    3. Boolean
3. Numerical
    1. Change scale
        1. Normalize
        2. Standardize
        3. Robust
    2. Change distribution
        1. Power
        2. Quantile
        3. Discretize
    3. Engineer
        1. Polynomial
4. Data imputation
    1. Fill Nan's or nulls
        1. Textual
        2. Categorical
        3. Numerical

10. Once data is enriched, what to do next?

11. What are the different varieties of plots?

1. Line
2. Bar
    1. Stacked
    2. Non-stacked
3. Histogram
4. Box and whisker
5. Scatter
6. Pie chart
7. Wind-rose
8. Correlation and partial correlation
9. Many more - (list what interests you with examples)

12. What are Feature Selection techniques?

1. Matrix Factorization
    1. PCA
    2. SVD

References

1. [Data Preparation techniques](https://machinelearningmastery.com/data-preparation-techniques-for-machine-learning/)
2. [List of books for data processing](https://machinelearningmastery.com/books-on-data-cleaning-data-preparation-and-feature-engineering/)
3. [Data Visualization - python matplotlib](https://machinelearningmastery.com/data-visualization-methods-in-python/)
4. [Data preparation](https://machinelearningmastery.com/what-is-data-preparation-in-machine-learning/)
5. [List of Visualization plots](https://datavizcatalogue.com/)
6. [RandomForestClassifier model](https://machinelearningmastery.com/random-forest-ensemble-in-python/)

Assignment ToDo List

Notes:
1. Take it as a challenge to go beyond boudaries of the assignment
2. Apply all that you can.
3. Lookout for improvisations.
4. Students will be scored based on efforts taken. (Copying is strictly prohibited and will be treated severely.)
5. As a study material for further projects and your understanding, elaborately add more points to this file.
6. You can also make a GitHub project and use this first draft for continuous updates.
7. Ideally use Python3.7.x, Pandas, Matplotlib, Seaborn, Sklearn, etc. Optionally use R, Matlab, etc.

ToDo List:
1. Download the dataset.
2. Apply the relevant data processing techniques
    1. Remember to do the analysis separately for test and train data.
    2. Ideally prepare a pipeline for data processing
        1. Input - single data instance
        2. Output - Transformed instances
3. Plot various visualizations
    1. Make it generic - Reusable for the last assignment
4. Think and answer a few data scientific questions (interesting and insightful) using data analysis?
    1. How many of the survived were male, female? Within this, how many were children in each gender category?
    2. What does the SibSp, Parch, Cabin, Embark column signify? Can we attach external datasets to enrich the information?
    3. Is Fare just a number or is correlated with other columns? If yes, which ones?
    4. Students can add more such questions.

5. Advanced (assignment 2): Fit a model like LogisticRegression and RandomForestRegressor using sklearn.

6. Send your doubts to shreedhar.kodate.ta@gmail.com with email subject line of the format DYPCOE_DSHC_{firstName_lastName}_{snake_case_short_title}

[top](#assignment-1-titanic-data-preprocessing)
