Autorank - a simple python package

CMSC 6950 - COMPUTER BASED RESEARCH TOOLS AND APPLICATIONS

TERM PROJECT

6TH AUGUST, 2020

SUBMITTED BY

ANNE ODEH

Memorial University of Newfoundland
St. John's, Canada.

Abstract

Reproducibility is seen as one of the pillars of the entire scientific method, a criterion on which to measure the efficacy of an experiment. This paper presents how autorank can be reproduced by understanding how it works and the necessary tools that are required to mimic its exact work.

1   Introduction

Autorank is a simple Python package with one task: simplify the comparison between (multiple) paired populations. This is, for example, required if the performance different machine learning algorithms or simulations should be compared on multiple data sets. The performance measures on each data set are then the paired samples, the difference in the central tendency (e.g., the mean or median) can be used to rank the different algorithms. This problem is not new and how such tests could be done was already described in 2006 in the well-known article Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7 (December 2006), 1–30.

Regardless, the correct use of Demšar guidelines is hard for non-experts in statistics. Correct use of the guidelines requires the decision of whether a paired t-test, a Wilcoxon's rank sum test, repeated measures ANOVA with Tukey's HSD as post-hoc test, or Friedman's tests and Nemenyi's post-hoc test to determine an answer to the question if there are differences. For this, the distribution of the populations must be analyzed with the Shapiro-Wilk test for normality and, depending on the normality with Levene's test or Bartlett's tests for homogeneity of the data. All this is already quite complex. This does not yet account for the adjustment of the significance level in case of repeated tests to achieve the desired family-wise significance. Additionally, not only the tests should be conducted, but good reporting of the results also include confidence intervals, effect sizes, and the decision of whether it is appropriate to report the mean value and standard deviation, or whether the median value and the median absolute deviation is more appropriate.

2 Reproducing Autorank

To better understand Autorank, I executed the python file (example.py) to fully comprehend how it works, what it does and how it can be of use to my project. The actual result was the same after doing the specified steps multiple times.

- First, import the functions from Autorank and create some data.

- Second, the result of the statistical analysis was displayed on the terminal.

- Then, Autorank generates a plot that visualizes the statistical analysis.

- Lastly, Autorank can also generate a latex table with relevant results.

3 Data Exploration

3.1 Dataset Summary

This dataset has 17 columns and 11163 rows. Below is a detailed description of each column:

- age: the age of group of ban customers

- job: type of job (admin., blue collar, entrepreneur, housemaid, management, retired, 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')

- marital: marital status (divorced, married, single, unknown; note: divorced means divorced or widowed)

- education: (primary, secondary, tertiary, and unknown)

- default: has credit in default? (no, yes, unknown)

- housing: has housing loan? (no, yes, unknown)

- loan: has personal loan? (no, yes, unknown)

- balance: Balance of the individual

- contact: contact communication type (cellular, telephone)

- month: last contact month of year

- day: last contact day of the week

- duration: last contact duration, in seconds (numeric).

- campaign: number of contacts performed during this campaign and for this client

- pdays: number of days that passed by after the client was last contacted from a previous campaign

- previous: number of contacts performed before this campaign and for this client

- poutcome: outcome of the previous marketing campaign (failure, non-existent, success)

- deposit: has the client subscribed a term deposit? (yes or no)

3.2   Attributes Types

Attributes	Types
age	numeric
job	categorical
marital	categorical
education	categorical
default	categorical
housing	categorical
loan	categorical
contact	categorical
balance	numeric
contact	unknown
month	categorical
day	numeric
duration	numeric
campaign	numeric
pdays	numeric
previous	numeric
poutcome	unknown
deposit	categorical

Table 1: Attributes Types

3.3 Implementing autorank in my dataset

Banks use marketing campaigns as tools to focus on customer needs and their overall satisfaction strategically. Improving customer experience requires truly understanding your customers and relating to them in ways that they understand. This includes taking a 360-degree view of your banking customer and leveraging the gold mine of data available to you today, including:

- Core customer information including contact and location data

- Additional experiential customer information gathered from all stages of the customer life-cycle.

- Transaction information including checking, savings and credit card transactions; loan draws and repayments; investment positions and balances.

Banks of every size are drenched with data, but harnessing and leveraging that organizational data for more effective banking operations has always been a challenge. In the current market, it's more important than ever that you understand customers, products, channels and pricing – all to ensure it is tailored towards product offerings to customers while maximizing the potential revenue.

Using transaction and core customer information, you can determine the life stages and family dynamics that allow for better product bundling and targeted marketing for your customers.

Autorank was used to achieve a quick statistical analysis of the duration of calls made to different clients with different marital statuses. Banks will find this helpful as it will aid them to design different deposit package that benefits everyone.
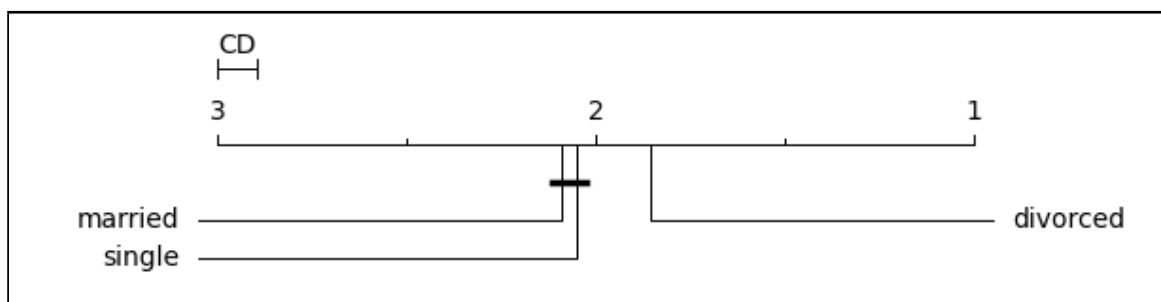

Figure 1: Autorank Plot

The plot shows the number of contacts with different subset of customers during the campaign.
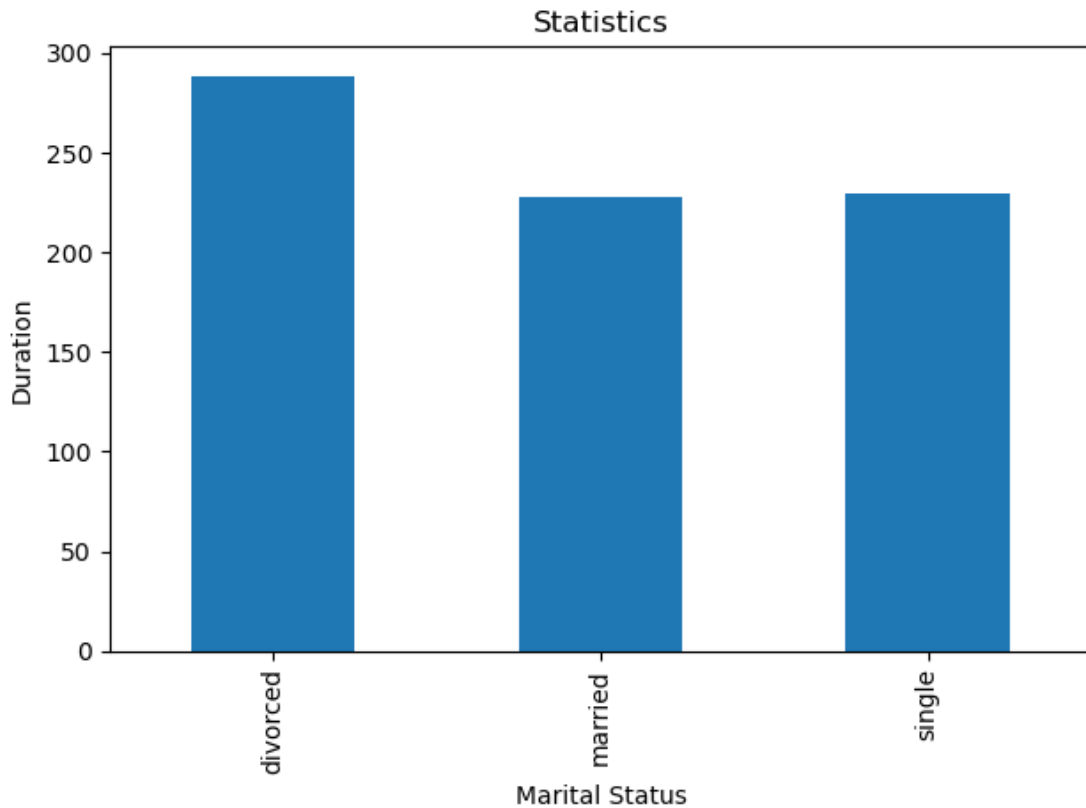

Figure 2: Bar Plot of the

4   Conclusion

The goal of Autorank is to simplify the statistical analysis for non-experts. Autorank takes care of all of the above with a single function call. Additional functions allow the generation of appropriate plots, result tables, and even of a complete latex document. All that is required is the data about the populations is in a Pandas dataframe.Overall, Autorank statistical analysis involves evaluating and then summarizing the data into a mathematical form that was easy to implement in my dataset.

References

Martinez, J. (2017). Bank Marketing Dataset. Retrieved July 31, 2020, from `https://www.kaggle.com/janiobachmann/bank-marketing-dataset`

S. Herbold (2020). Autorank: A Python package for automated ranking of classifiers. Journal of Open Source Software. https://doi.org/10.21105/joss.02173