STEM

LUND UNIVERSITY FINANCE SOCIETY EST 1991

Lund University Finance Society

Advanced Python Workshops Project

VT 24

Created by:

Amanda Ramirez, Michal Nowak & Yasser Mahfoud

Description

You can work individually or in groups of up to three to deliver the project. Your code should all be your own although you can of course reference the internet and other sources, but you should not just copy and paste an entire python file.

The data set used for this project is available on canvas as `credit card customers.xlsx`.

This data set contains data on 10 000 customers containing 19 different variables regarding each customer such as age, gender, education level, marital status, months on book, etc. The manager of the bank has asked you, the in-house data scientist, to dig into this data and create a model to predict which customers are likely to leave the bank, and provide insight as to the biggest reasons for leaving. Your output should be your code for the model and a dashboard which displays the major factors you identified, what your model predicts, and the accuracy of your model.

The "Attrition Flag" columns tells you if a customer currently has an open or closed account. You can find the description of the data in each column in the second tab of the excel file named "variable description".

Project Steps

1. To begin, you should do some exploratory data analysis. The first thing we want to know is:
   (a) Proportion of clients who have left the bank
   (b) Within the clients that left, what is the distribution of, for example:
      i. Gender
      ii. Marital status
      iii. Income category
      iv. Months on book
      v. Education level
2. Then, you want to select the features that could be used in your model. To do this, you need to clean your data, and try to remove the variables which have low predictive power.
   (a) Clean data
   (b) Which variables seem to have a low predictive power? Depending on which model you choose, you may want to carry out different tests between the variables, e.g.:
      i. Check the correlation between variables
      ii. ANOVA, chi squared, Pearsons correlation
      iii. (Advanced) PCA
3. Choose and create your prediction model:
   (a) Logistic regression
   (b) Clustering
   (c) Decision tree
   (d) Random forest
   (e) Any other model, e.g. K-nearest neighbours or Neural network

4. Evaluation of your model. Depending on your model, you can carry out several tests, usually within the package you use to create the models (scikitlearn, etc). e.g.

5. Present key findings in suitable visualizations on a dashboard made in Plotly and/or Dash. You could present things like:

   (a) Key statistics (computed in first step)

   (b) Most relevant variables

   (c) Correlation matrix

   (d) Pie charts

   (e) Score of model

   (f) Confusion matrix

   (g) ROC curve

   (h) Elbow plot

   (i) If you are using neural networks it would be nice to see the structure of your network (plotted)

Submission

Your (group's) submission should contain:

1. source code for your model,

2. html file of dashboard, source code (if used Dash) and screenshot (png).

Your submission should be all the files required to run your model and dashboard, i.e. the python scripts, any data sets used, and any visual assets like pngs (if any). You submit the files through canvas by 23:59 June 13th, 2024.

Examples

You can reference dashboard examples on the Dash website, on the link below:

`https://dash.gallery/Portal/`

Here is a specific example with both the front end app and the back end on github:

`https://github.com/plotly/dash-sample-apps/tree/main/apps/dash-financial-report`