Practical Data Science: Assignment 3

Setup

Welcome to the third assignment for this course. At this point you've written some Python, you've worked with the web and you are familiar with Pandas and data frames. The next step is to fit some models to data that you collect. For this assignment, you **must** use the `statsmodels` Python module that we looked at together in class. Please review the material on this module before tackling this assignment. We've looked at ANOVA, linear regression and logistic regression using this module so read up on these 3 approaches.

Questions

There are **three** questions in this assignment and you must complete all three. The style of this assignment is different from the previous ones in that it is slightly more open ended. You will be given three data sets and asked to investigate a particular problem. You must choose the model or approach to use, fit it using the `statsmodels` module, and then using the model's output or an any auxiliary checks necessary, answer the question you have been given. Thus, you will be doing complete, albeit relatively simple, end to end analyses of the data. For this assignment we want you to always use the entire dataset, without doing any splits (e.g. for training / testing).

Question 1 (10 points)

Diversity in the workplace has become a hotly debated topic in recent years, and one that strives to ensure that workplaces provide equal opportunities to people of different races, genders and backgrounds. In the USA, companies often file special reports to indicate their hiring practices by showing the race and gender composition of their workforce.

Visit the following link to obtain a data set comprised of filings of different companies in Silicon Valley:

https://github.com/cirlabs/Silicon-Valley-Diversity-Data/blob/master/Reveal_EEO1_for_2016.csv

Using this data set, and by running an appropriate analysis with the `statsmodels` module, investigate the following questions:

- Is the proportion of males in senior positions greater than that in non-senior positions?

- Is the proportion of whites in senior positions greater than that in non-senior positions?

Question 2 (10 points)

The computer game industry can often be a very interesting source of data for experimenting with different models. Massively multiplayer online role-playing (MMORPG) games like World of Warcraft often have their own in-game virtual economies based on currencies (in this case gold pieces) or trade goods that players collect. In 2015, World of Warcraft initiated a system whereby players could purchase their monthly subscription using its in game currency, gold. The price of this monthly subscription fluctuates according to the demand and supply of gold in the economy and thus, we can study how it has changed over time.

Visit the following link in order to obtain a data set with the price of the monthly subscription measured at various days and times since 2015 (you will need to create a free account at data.world if you do not already have one):

https://data.world/helithumper/prices-of-world-of-warcraft-token/workspace/file?filename=wowcointotal.csv

- Using this data set and the `statsmodels` module, fit a model that predicts the price of the monthly subscription as a function of time and discuss, citing approrpiate metrics, how well it fits the data
- Every time World of Warcraft releases an expansion, there is usually an inflation in the economy as monsters, quests and other in-game activities reward more gold and the supply of gold goes up. Based on this fact, and by looking at the data, can you see roughly when an expansion was released during this time period?
- Split the data into two time periods, one before the expansion and one after and fit two separate models that predict the price of the monthly subscriptions as a function of time, and discuss whether this results in an overall better fit to the data again using appropriate metrics.

Question 3 (10 points)

An excellent site for practicing data science is of course Kaggle. Kaggle hosts competitions by uploading datasets and allowing individuals and teams around the globe to compete for a specific period of time in order to build models that perform the best on these datasets. You should absolutely spend time on this site to get practicing with the techniques that you learn in your MSc. For this assignment, we will use the data from an old competition on credit card fraud.

Visit the following link in order to obtain the data and to read a description of what the fields are:

https://www.kaggle.com/dalpozz/creditcardfraud

- Using this data set and the `statsmodels` module, fit a model that predicts whether a particular transaction was fraudulent
- Compute the accuracy of your model on the data it was trained on. Discuss whether this is a good result or not, especially with regard to class inbalance.
- Look up the definition of Precion and Recall (e.g. on Wikipedia) and compute the precision and recall (do not use built in functions or another module for this) that your model outputs
- How can you make your model become more sensitive to the rare (fraudulent) class? Implement this change and compute the precision and recall values on the result and comment.

HINT: The `fittedvalues` attribute of your model outpus the model's prediction on its training datas on a logarithmic scale. Be sure to reverse this (e.g. by applying np.exp() on them) in order to obtain values in the range that you expect.