DAT565/DIT407
Introduction to Data Science and AI

Assignment 4

The file `life_expectancy.csv`, available on Canvas, contains data gathered from different UN sources [2, 3]. The data includes statistical indicators about demographics and human development of different countries over the years. Our task will be to construct a linear model to predict the *life expectancy at birth (LEB)*[1], from the other observed variables.

Before you start: Make sure you have read the chapters in Skiena for this week! This will make it a lot easier.

Problem 1: Splitting the data

Perform an appropriate train-test split on the data. In your report, justify why you chose to do the split in the way you did. In further problems, only ever perform exploratory data analysis on the *training data*, as testing data should never be used to draw hypotheses.

Problem 2: Single-variable model

a) Identify the variable that has the strongest linear relationship with the target variable. Do so by calculating and reporting the Pearson product-moment correlation coefficient between this variable and the target variable (LEB) in your report. Hint: use functions for doing this from the Pandas dataframe package.

b) Construct a single-variable linear model using this variable. In your report, report the coefficient of determination, and the coefficients of your model, including the intercept. Also, produce a scatter plot showing the LEB as the function of the variable you identified. Plot the regression line over the scatter plot.

c) Finally, use the model to predict the LEB for the test set, and report the correlation between the predicted values and the target variable, and the Mean Squared Error (MSE) between the predictions and the true target variable values in your report.

d) Look up information about the meaning of the variable, and describe *why* you think it has such a strong relationship with LEB. Remember to cite your sources.

[1] LEB is the hypothetial length of the human life span in a population, calculated from the observed mortality rates at a given year; for more information, see [1].

Problem 3: Non-linear relationship

Recall sections 9.2.2 and 9.2.3 in Skiena. Now, explore your data and try to find a variable which has a non-linear monotonic relationship with the target variable. **Hints:** such variables could for instance be power-law distributed, resulting in a large difference in size between the min and max values they take on. You may also want to investigate the Pearson and Spearman correlations between candidate variables and the target. Finally, you may want to plot the candidate variable against the target to visualise the relationship, does it remind you of the shape of some non-linear function?

Perform an appropriate transformation for the selected values, such that the transformed value have an approximately linear relationship with the target variable. In your report, report the variable that you found, the function that you used for transformation, and the Pearson correlation coefficient before and after transformation.

Problem 4: Multiple linear regression

Let us exclude the variable we identified in Problem 2 from explanatory variables. Perform a systematic search to find a subset of variables that you expect to produce a better linear model than the model produced in Problem 2.

Describe how you discovered the subset of variables in your report. Also include the following pieces of information:

- Names of the variables included in your model,
- The coefficients of your model (including the intercept),
- The coefficient of determination,
- Pearson correlation coefficient between the values your model predicts for the test set vs. the actual values in the test set,
- MSE between your predictions and the actual values in the test set.

Your model should outperform the single-variable model, but try not to include more variables than is needed.

Hints

- There are multiple ways to conduct the train test split, but you should be able to justify how you did it.
- If there are missing values, handle them appropriately.
- Some variables may, in fact, measure more or less the same thing; this can be detrimental.
- Even if the variable has an underlying nonlinear relationship, the linear relationship can be nonnegligible.
- You can include the transformed variable from Problem 3 in your model of Problem 4.

- Use the `LinearRegression` class from the `sklearn.linear_model` module.
- When choosing the variables for your model, you *must not* look at the test set; the test set is only for evaluating the generalization properties of your model.
- There are multiple ways to conduct such a systematic search; a statistically grounded way would be to evaluate the coefficient of determination for a subset of variables; a more machine learning oriented approach could be to split the training data into two sets and use the other set as a validation set, and then evaluate some numerical statistic (such as MSE) against that set.
- It is probably not necessary (or even feasible) to try all possible subsets of variables; you should use an appropriate metric to select which variables to try to include in the model (such as correlation coefficients).

Returning your report

Write a report, typeset in LaTeX, that answers *all* questions above. Include all your Python code in your report as an appendix, preferably using the `listings` package. Your report should be legible even without having a look at your code. The report, excluding appendices, should not be much more than five pages long, but can be shorter. Clear and concise answers are preferred, don't pad answers with irrelevant text.

If you refer to outside sources, remember to add an appropriate literature reference (including websites) in references by `\cite`ing the references. It is recommended that you use the package `biblatex` to manage citations.

Place your figures in numbered `figure` environments, with descriptive captions and `\ref` to the figures in your discussion. Likewise, place your tables in numbered `table` environments with desciprive captions and `\ref` to the tables in your discussion.

After grading, you will be given another attempt to revise your report according to TA comments if it is not considered acceptable.

The deadline is *hard*. Late submissions will not be read at all and are considered failed. This means you will not get any feedback for the first round and the submission is considered a revision; there will be no third attempt, so if a late submission is failed, you will need to participate in a later iteration of the course for a re-attempt.

References

[1] Esteban Ortiz-Ospina. *"Life Expectancy" – What does this actually mean?* Retrieved 2023-11-24. 2017. URL: `https://ourworldindata.org/life-expectancy-how-is-it-calculated-and-how-should-it-be-interpreted`.

[2] UNDP (United Nations Development Programme). *Human Development Report 2021/2022: Uncertain Times, Unsettled Lives: Shaping our Future in a Transforming World.* Retrieved 2023-11-24. 2022. URL: `https://hdr.undp.org/data-center/documentation-and-downloads`.

[3] Department of Economic United Nations and Population Division Social Affairs. *World Population Prospects 2022, Online Edition*. Retrieved 2023-11-24. 2022. URL: `https://population.un.org/wpp/Download/Standard/CSV/`.