Final project

Analyzing StackOverflow Developer Survey

Each year, [StackOverflow (https://stackoverflow.com/)](https://stackoverflow.com/) asks the developer community about everything from their favorite technologies to their job preferences, then they publish Survey results of over 100,000 developers.

You can see an example of the analysis they made for last year's survey [here (https://insights.stackoverflow.com/survey/2018/)](https://insights.stackoverflow.com/survey/2018/).

👉 Initial analysis and cleaning

- Extract and load the ***data/stack-overflow-2018-developer-survey.zip*** data into a `DataFrame` .
- Process the data and clean it according to your own criteria, for example:
  - Set the right type of the columns;
  - Remove outliers;
  - Remove unused columns;
  - Transform categorical into dummies.
- Make an EDA *(Exploratory Data Analysis)* to understand the data.
  - Find out variable relationships;
  - Correlations;
  - Actionable insights that can be drawn from the given data (scatterplots, boxplots, lineplots, etc.).

👉 Questions

Aside from the analysis performed in the previous point, please answer these questions extending your EDA:

1. Is there a correlation between the years of experience of a developer and their salary? What about age?

- Is there any relationship between bootcamp education and salaries? What about formal education?
- Is there any relationship between years of experience and the technologies/languages employed by developers?
- Which are the top-three the most common dev types per country? And which are the best paid ones?
- Which are the most popular programming languages? And which are the highest-paying ones?
- Are the best paid jobs the most satisfying?
- Does career satisfaction depends on company or organization size? Are there any other variables that are correlated to career satisfaction?

👉 Modeling

Machine Learning time! We want you to conduct the required analysis to make some predictions. It's not necessary that the model predicts correctly, we're more interested about your approach to select variables, make the analysis, cleaning and selecting the model. Keep in mind that this is a large dataset with many columns; not all of them will be relevant.

Predicting Salary

Create a model that predicts the salary of a developer. What are the most relevant variables when comes to predicting salary and why?

Predict job satisfaction

Create a model that predicts Job satisfaction for a developer. What variables are the most relevant for this? And what factors have a greater influence in higher job satisfaction?

👉 Optional

Finally, if you have extra time, here are two optional points to work on:

- Look up other datasets to enrich the data. For example, you could get a dataset with Country GDP and combine it with survey data, to see how Job Satisfaction is related to the country's GDP.
- Create an API endpoint to predict salaries. The endpoint should receive the appropriate variables and return the prediction.