WorldQuant Predictive Research Technology Team
Machine Learning Engineer
Take-Home Challenge
2/15/2022

Candidate: Alex Salman
Email: aalsalma@ucsc.edu

In order to help the Intergovernmental Panel of Practitioners of Impractical Practices (IPPIP) to find out the average height of all current living humans, once and for all, I am proposing the following study/work to be conducted for the organization.

In this paragraph and for non Machine Learning (ML) technical readers, I am identifying the required elements to conduct the study. The first and most important is to identify the features that are believed to be critical in determining the height of all human beings. Geographical location, height, age, gender are features I will be considering when collecting samples from people. For example, from three African countries (South East, North, West), I will collect samples from 100 people each. That will be my 300 samples for Africa. In these samples, I am accurately taking four different ranges of people ages and the two genders. And the same will be applied to other geographical locations considering people's races in some geographical locations that are considered rich in racial diversity. Since I am sampling sets of people, then I am collecting additional historical data if available that could be used in the study when some outliers happen. For instance, I want to consider collecting migrations status from some geographical locations if any; like in Brazil, there is the largest Japanese population outside of Japan. That could drastically change my metrics if not considered.

For ML readers and now as I have the data collected from all geographical locations in the world, I will conduct my study using a Machine Learning model, the Logistic Regression model, to create estimation for peoples' heights through some representation of the dataset collected. As a starting point, I will load then analize the data collected. I will need some data manipulation way for example converting gender and geolocations to numbers using sklearn LabelEncoder. Then I will need to place the data into DataFrame. The data is ready to be splitted into training and test sets so I can fit the linear regression model. Now I can start predicting the test set values; I will compare the results and measure the model accuracy. For model accuracy, I am using Mean Squared Error. After that the model can be tested using some datasets to predict heights of people. Datasets should be checked for outliers as the model will give inaccuree results with having them.

Thank you for considring my proposal