Data Science Case Study STVS 2019-12 013

At Nielsen we work with a wide variety of datasets from different sources. We would like to see how you approach working with a new dataset to create a model.

Scope and Expectations:

- We are interested in seeing how you use data and code to *begin to* solve a problem. We would like you to focus on data exploration and some *initial* model building.
- We are also interested in seeing your coding style. Please note we will be looking into how the code is organized and formatted as well as how you approach the modeling problem itself.
- We do not expect you to have a perfect model or code. We want to see your early steps to understanding the data and training a model and what coding strategies you implement to solve this.
- We know your time is limited. You will have several days to complete this task, but please do not spend more than **4 hours total time** on it.
- We would like to see all of your code, results, and documentation. This includes your data analysis, visualization, evaluation, written comments, etc.
- You can use whatever programming languages and development environment you like.

Data

The data for this exercise comes from the U.S. Census Bureau. It contains population statistics for counties in the U.S. from 2010 to 2015. Given several years of historical data, **your goal is to begin testing a model that will predict whether the population of a given county will increase or not from 2015 to 2016**. You may use the existing information to make the prediction and/or create additional features depending on your approach. At this stage, we are most interested in seeing how you use the data and structure this modeling problem over developing the most accurate model.

Data structure and contents:

All of the data is contained in the `census_labeled.csv` file. Each row of data contains information for a particular county and year. The year, county name, and state in which the county is located are given.

Additional columns show the total female population (`female_total_population`) and male population (`male_total_population`) and the percentages of females and males in different age groups, ranging from under 5 years old to over 85 years old. For example, the `female_age_30_to_34_pct` shows the percentage of females that are 30 to 34 years old out of the female total population. The `male_age_10_to_14_pct` shows the percentage of males that are 10 to 14 years old out of the male total population.

The labeled outcome you are trying to predict is whether the county's population increased from 2015-2016 (`county_population_increased_2015_2016`). A value of `TRUE` means the population increased, and `FALSE` means it did not. Note that the value will always be the same for a given county for each year it appears in the dataset, but your model should provide only one prediction per county in the end. Also note that this is the complete labeled data; we are not holding out a test data set.

Deliverables

Please email the following items to your recruiter or team member by the date specified:

1. **Documentation** - A summary of what you did and why, including an explanation of the model you chose and the analysis supporting your decisions.
2. **Code** - The code you used to explore the data and develop and test the model.

You may combine items 1 and 2 into a single document or notebook.

If you have questions, experience issues opening the dataset, or will not be able to complete the task in the time given, please contact your recruiter or team member as soon as possible.