Next steps

Paint simple possible architecture for production. Batch or online prediction. Involve database

Model

Try simple model (e.g. naïve bayes, DT) but also complex (LR, SVM, RF).

Talk about testing a possible evolutionary algorithm for AutoML (TPOT)

Assess overfitting

Variable selection: domain knowledge or autoML

Regularization

Class imbalance: (1) assign larger penalty to the majority class. Or upsampling the minority class, or down-sampling the majority class, or generate synthetic training examples. There is no universally best solution or technique that works best across different problem domains. Thus, it is recommended to try out different strategies on a given problem, evaluate the results, and choose the technique that seems most appropriate.

Monitoring: talk about validation curve for monitoring

https://cran.r-project.org/doc/manuals/R-lang.html#Indexing

# Intro

I was commissioned this POC by Sterling Cooper Advertising Agency willing to extend its data science capabilities. The scope of the project is to define a clear data science pipeline going from data to model predictions on a test set. The project is to be developed on R or Python and eventually be translated on other platforms for deploy into production (e.g. Dataiku, AWS, GCP, or Azure).

The first milestone for the project must accomplish:

1. Knowledge Transfer: describe the steps took to accomplish the Data Science workflow
2. Predictive Modelling: Find clear insights on the profiles of the people that make more than $50,000 / year
3. Possible Enhancement: state how steps could be improved or possible alternative approaches

During the review of the first milestone, several possible approaches for production deployment will be discussed. The second part of the project will start once Sterling Cooper Advertising Agency will have decided on which Cloud platform the tool will be deployed and how it will integrate with its current systems.

Therefore, the second milestone will consists of operationalizing the workflow – including monitoring tools – on the platform chosen by the client.

The report is focused on the accomplishment of the first milestone. The project will follow some characteristics of Agile Approach such that scope and cost are rigid. The amount of cost will be 5 man/days over an elapsed of two weeks.

Below the recap of activity performed on each day:

Agenda

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Time slot | Day 2/1 | Day 2/2 | Day 2/9 | Day 2/10 | Day 2/11 |
| 9-10AM | Planning | Review daily activities | Review daily activities | Review daily activities | Review daily activities |
| 10-11AM | Setup environment | Exploratory Data Analysis (Part 1) | Data Preparation | Evaluation |  |
| 11-12AM | Modelling | Data Preparation |  |
| 15-16PM | Analyze and prepare metadata | Evaluation | Modelling |  |
| 16-17PM | Exploratory Data Analysis (Part 0) | Data Preparation | Evaluation |  |
| 17-18PM | Data Preparation | Modelling |  |  |
| 18-19PM | Summarize daily activities in diary | Summarize daily activities in diary | Summarize daily activities in diary | Summarize daily activities in diary | Summarize daily activities in diary |

# SETUP

I am a big fun of R’s elegance and grace. And I am aware of the effectiveness of Python. After the 2020 rstudio conf in San Francisco, I decide this is a good chance to give reticulate a try.

Therefore, I will setup two virtual environment: one for Python (Conda) and one for R (RStudio).

Next I will work on the project as a rmarkdown combining Python and R.

In order to simplify visual prototyping I might give Power BI a chance.

I was not able to use Anaconda because RStudio is available only up to version 1.1. Instead, RStudio 1.2 has enhancements for using Python. <https://blog.rstudio.com/2019/04/30/rstudio-1-2-release/>

Next step I needed a package manager in order to mange in RStudio virtual environments for R and Python. Seems like renv is the suitable candidate <https://blog.rstudio.com/2019/11/06/renv-project-environments-for-r/>

Therefore, I checked out the documentation for RStudio and reticulate <https://docs.rstudio.com/tutorials/user/using-python-with-rstudio-and-reticulate/>

Versioning  
Nevertheless I created a new repository on github and synched my local repository to work on RStudio. This question <https://stackoverflow.com/questions/34565238/where-does-github-desktop-install-command-line-version-of-git> helped me find the git.exe executable installed in the github desktop distribution.

Also this was useful to setup RStudio and git. https://cfss.uchicago.edu/setup/git-with-rstudio/

I decided to inspire the structure of the project according to this <https://drivendata.github.io/cookiecutter-data-science/>

Considering important aspects is reproducibility, I will create virtual environment. Ideally, at the end of the project I would create a Docker container ready-to-use.

# Data Source

Ideally, I would have preferred not to manually download the data from the zip file sent by Marie, instead download it directly from AWS S3. Additionally, I would have preferred to keep all my data in AWS S3.

However, I added a .gitignore in order not to version changes in data using github. Instead, data must be permanent such that once created, a dataset is never overwritten. Instead, a different dataset is created upon transformation.

Refers to 1994 data:

# Environment

The project is accomplished on my local computer. An alternative approach would have been to entirely build it using Docker container.

I created a conda virtual environment for Python.

I created a project library for R using renv.

## Dependency manager

# METADATA

I reorganized the metadata information in a simple table like the following using Excel:

Code| description|data\_type|possible\_values

Once organized the environment, I started preparing the metadata for two reasons:

1. Column headers are missing in the data files
2. Metadata file does not provide a clear mapping of column name and column position in the data file

I proceeded the metadata preparation in Excel because the columns were less than 50 so it was reasonable to proceed in Excel. The process was by exclusion: first assigned the column name to the field (e.g. Married will be marital status).

Then I remained with the field in position # 2 #3 not clear.

During this process I observed that in the 18 records I used for understanding the column names, nominal categorical variables might have required normalization through remapping.

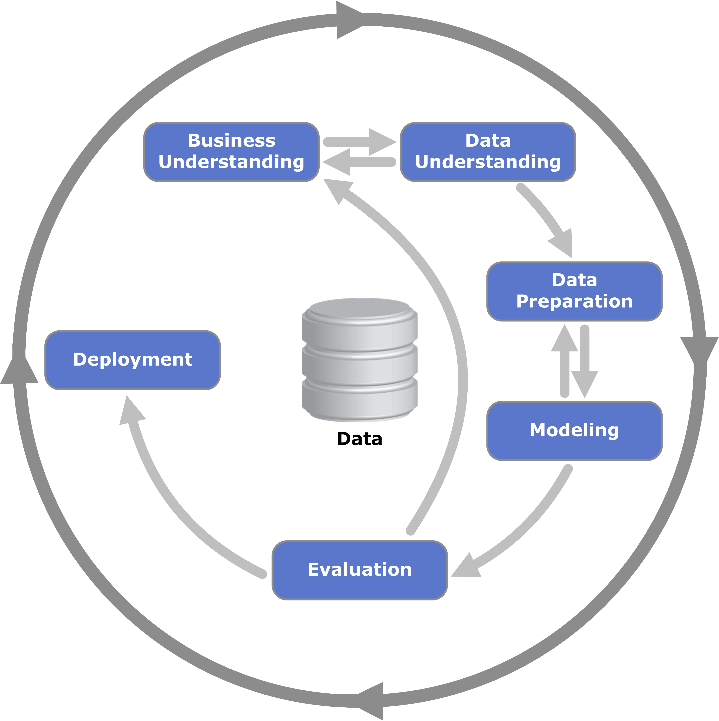
In the end I observed that all variables related to income – except the target – were removed from the training set therefore this allowed me to solve doubts about remaining column positions.

The output of the process is “/docs/metadata/census-income-metadata-prepared.v00.csv”

Considering the project is business oriented, I decide to use business names for variables instead of code.

# Workflow

There are many data mining workflow, however, in the end the all express the same approach. I like the below graph because I believe it includes the business in the picture and because it reflects the cyclical and iterative nature of data mining.



From a business perspective, the scope of this project is to improve marketing effectiveness by identifying which higher precision individuals above the $50,000/year threshold. In fact, this would allow the company to reduce direct mail costs by better identify individual with enough personal income available. Find clear insights on the profiles of the people that make more than $50,000 / year. For example, which variables seem to be the most correlated with this phenomenon?

I want to define a first iteration of EDA analyzing variables and their distribution split by target value.

The outcome I expect is:

* Identify promising variables
* Identify variables that might need treatment
  + Outliers
  + Remapping of nominal variables

Test set will remain not used.

All analysis are applied on the training set and defined replicable so they will be easily applied also to the test set.

# Exploratory Data Analysis

In this section we proceed on calculating descriptive statistics for categorical (nominal or interval) and for measurable (interval or ratio) variables. After a synthetic overview, each variable is individually analyzed to better understand it, discover how it relates with the target variable and find evidence for decisions to be taken in data preparation stage. This section concludes presenting main results of multivariate analysis involving correlation and multidimensional variables interactions.

As explained in the “Duplicates” section, EDA will be performed on 1994 US Census data.

### Duplicates

To perform the EDA, raw training data is loaded in the analysis environment and duplicated records are removed: 1,548 records accounting for 2,401,889.19 instance weight. In the dataset – excluding the target – there are 8 measurable variables (interval or ratio) and 33 categorical variables (nominal or ordinal).

Investigating duplicates a bit further and considering the data set is US Census data for 1994 and 1995, we might observe the same record appears once for each year. However, it is not observed any record that appears equally – with exception of year – in both 1994 and 1995.

In general, it does not make sense from analytical perspective to sum two snapshots (1994, 1995). Instead, it is more appropriate to analyze them separately and eventually study temporal trends. Considering time constraints for this project, this EDA will focus only on 1994 US Census data and in the next step it will be included a possible task of validating temporal trends. This decision is supported by evidence showing there is no significant different in distributions between 1994 and 1995, with exception for full\_or\_part\_time\_employment\_stat which seems to be introduced in 1995 since in 1994 all instances were assigned to the same level.

Please note that 1995 data are not dropped, they will simply not be included in the EDA. 1995 data will be used as data for either training or validating the model. Another possible approach to investigate is to build a new model every year.

### Instance Weight

Before proceeding further, it is important to discuss the instance weight variable (MARSUPWT) present in the raw data. According to the data description and this link, <https://www.census.gov/programs-surveys/sipp/methodology/weighting.html>: “*the weight for a responding unit in a survey data set is an estimate of the number of units in the target population that the responding unit represents. In general, since population units may be sampled with different selection probabilities and since response rates and coverage rates may vary across subpopulations, different responding units represent different numbers of units in the population. The use of weights in survey analysis compensates for this differential representation, thus producing estimates that relate to the target population*”.

After performing some analyses at aggregated level, we can observe that sum of instance weight and count of records is perfectly correlated (Pearson) in every variable, with exception for univariate variables (e.g. migration\_code.change\_in\_msa). On the basis of this evidence, summary statistics tables will use the number of records as a proxy for the instance weight. This allows to use packages already available in order to generate such summary statistics tables, the alternative is either to customize a summary stats function or to build ad-hoc summary stats function to aggregate on instance weight instead of number of records.

Instead, for data visualizations the instance weights will be used.

### Categorical variables

There are 32 categorical variables – not counting the target. Also, five variables are manually redefined as categorical for example own\_business\_or\_self\_employed and detailed\_occupation\_recode. In below table a high-level summary about these categorical variables, including the target distribution.

For the purpose of this visualization, Category levels are ordered descending based on frequency in order to facilitate exploration. This will allow to identify variables that are useless and also distinguish between nominal and ordinal. During the data preparation, this information will be taken into consideration.

Considering the summary table extends for several pages, we can draw in this paragraph some of general remarks:

1. Only 6.3% of the records present the target class of interest
2. Most of the instances are related to race “white” (82.7%)
3. Missing values seem to be replaced by “Not in universe” or “?”
4. For some variables (e.g. region\_of\_previous\_residence), instances are concentrated only in few classes out of the many

Another aspect to take into consideration is the difference between nominal and ordinal variable. It seems some variable might be ordinal by nature (e.g. education), however, further analyses are needed in order to understand whether the order is relevant or not. Distinguishing between nominal and ordinal categorical variable is important because it influences decision about how further proceeding for pre-processing (e.g. type of categorical encoding) and modelling (e.g. use algorithm that can process ordinal variables).

Out of the 32 categorical variables, only education is ordinal. However, there are so many different levels that it might be difficult to define a clear order.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Stats / Values** | **Rel Freqs** | **Missing** |
| target  [factor] | 1. - 50000.  2. 50000+. | |  |  |  |  | | --- | --- | --- | --- | | 92440 | ( | 94.1% | ) | | 5839 | ( | 5.9% | ) | | 0  (0%) |
| class\_of\_worker  [factor] | 1. Not in universe  2. Private  3. Self-employed-not incorpo  4. Local government  5. State government  6. Self-employed-incorporate  7. Federal government  8. Never worked  9. Without pay | |  |  |  |  | | --- | --- | --- | --- | | 48685 | ( | 49.5% | ) | | 35919 | ( | 36.5% | ) | | 4286 | ( | 4.4% | ) | | 3903 | ( | 4.0% | ) | | 2092 | ( | 2.1% | ) | | 1662 | ( | 1.7% | ) | | 1424 | ( | 1.4% | ) | | 224 | ( | 0.2% | ) | | 84 | ( | 0.1% | ) | | 0  (0%) |
| detailed\_industry\_recode  [factor] | 1. 0  2. 33  3. 43  4. 4  5. 42  6. 45  7. 29  8. 41  9. 37  10. 32  [ 42 others ] | |  |  |  |  | | --- | --- | --- | --- | | 48909 | ( | 49.8% | ) | | 8650 | ( | 8.8% | ) | | 4076 | ( | 4.1% | ) | | 2949 | ( | 3.0% | ) | | 2326 | ( | 2.4% | ) | | 2153 | ( | 2.2% | ) | | 2110 | ( | 2.1% | ) | | 2006 | ( | 2.0% | ) | | 1976 | ( | 2.0% | ) | | 1775 | ( | 1.8% | ) | | 21349 | ( | 21.7% | ) | | 0  (0%) |
| detailed\_occupation\_recode  [factor] | 1. 0  2. 2  3. 26  4. 19  5. 29  6. 36  7. 34  8. 10  9. 16  10. 23  [ 37 others ] | |  |  |  |  | | --- | --- | --- | --- | | 48909 | ( | 49.8% | ) | | 4297 | ( | 4.4% | ) | | 3976 | ( | 4.0% | ) | | 2696 | ( | 2.7% | ) | | 2595 | ( | 2.6% | ) | | 2068 | ( | 2.1% | ) | | 2002 | ( | 2.0% | ) | | 1831 | ( | 1.9% | ) | | 1758 | ( | 1.8% | ) | | 1675 | ( | 1.7% | ) | | 26472 | ( | 26.9% | ) | | 0  (0%) |
| education  [factor] | 1. High school graduate  2. Children  3. Some college but no degre  4. Bachelors degree(BA AB BS  5. 7th and 8th grade  6. 10th grade  7. 11th grade  8. Masters degree(MA MS MEng  9. 9th grade  10. Associates degree-occup /  [ 7 others ] | |  |  |  |  | | --- | --- | --- | --- | | 24343 | ( | 24.8% | ) | | 22382 | ( | 22.8% | ) | | 13950 | ( | 14.2% | ) | | 9767 | ( | 9.9% | ) | | 3991 | ( | 4.1% | ) | | 3836 | ( | 3.9% | ) | | 3442 | ( | 3.5% | ) | | 3159 | ( | 3.2% | ) | | 3014 | ( | 3.1% | ) | | 2712 | ( | 2.8% | ) | | 7683 | ( | 7.8% | ) | | 0  (0%) |
| enroll\_in\_edu\_inst\_last\_wk  [factor] | 1. Not in universe  2. High school  3. College or university | |  |  |  |  | | --- | --- | --- | --- | | 91984 | ( | 93.6% | ) | | 3387 | ( | 3.4% | ) | | 2908 | ( | 3.0% | ) | | 0  (0%) |
| marital\_stat  [factor] | 1. Married-civilian spouse p  2. Never married  3. Divorced  4. Widowed  5. Separated  6. Married-spouse absent  7. Married-A F spouse presen | |  |  |  |  | | --- | --- | --- | --- | | 42079 | ( | 42.8% | ) | | 41798 | ( | 42.5% | ) | | 6372 | ( | 6.5% | ) | | 5262 | ( | 5.3% | ) | | 1694 | ( | 1.7% | ) | | 741 | ( | 0.8% | ) | | 333 | ( | 0.3% | ) | | 0  (0%) |
| major\_industry\_code  [factor] | 1. Not in universe or childr  2. Retail trade  3. Manufacturing-durable goo  4. Education  5. Manufacturing-nondurable 6. Finance insurance and rea  7. Construction  8. Business and repair servi  9. Medical except hospital  10. Public administration  [ 14 others ] | |  |  |  |  | | --- | --- | --- | --- | | 48909 | ( | 49.8% | ) | | 8650 | ( | 8.8% | ) | | 4613 | ( | 4.7% | ) | | 4076 | ( | 4.1% | ) | | 3470 | ( | 3.5% | ) | | 3065 | ( | 3.1% | ) | | 2949 | ( | 3.0% | ) | | 2782 | ( | 2.8% | ) | | 2326 | ( | 2.4% | ) | | 2224 | ( | 2.3% | ) | | 15215 | ( | 15.5% | ) | | 0  (0%) |
| major\_occupation\_code  [factor] | 1. Not in universe  2. Adm support including cle  3. Professional specialty  4. Executive admin and manag  5. Other service  6. Sales  7. Precision production craf  8. Machine operators assmblr  9. Handlers equip cleaners e  10. Transportation and materi  [ 5 others ] | |  |  |  |  | | --- | --- | --- | --- | | 48909 | ( | 49.8% | ) | | 7446 | ( | 7.6% | ) | | 6843 | ( | 7.0% | ) | | 6148 | ( | 6.3% | ) | | 6095 | ( | 6.2% | ) | | 5854 | ( | 6.0% | ) | | 5264 | ( | 5.4% | ) | | 3177 | ( | 3.2% | ) | | 2102 | ( | 2.1% | ) | | 2049 | ( | 2.1% | ) | | 4392 | ( | 4.5% | ) | | 0  (0%) |
| race  [factor] | 1. White  2. Black  3. Asian or Pacific Islander  4. Other  5. Amer Indian Aleut or Eski | |  |  |  |  | | --- | --- | --- | --- | | 83295 | ( | 84.8% | ) | | 10099 | ( | 10.3% | ) | | 2535 | ( | 2.6% | ) | | 1251 | ( | 1.3% | ) | | 1099 | ( | 1.1% | ) | | 0  (0%) |
| hispanic\_origin  [factor] | 1. All other  2. Mexican-American  3. Mexican (Mexicano)  4. Central or South American  5. Puerto Rican  6. Other Spanish  7. NA  8. Cuban  9. Do not know  10. Chicano | |  |  |  |  | | --- | --- | --- | --- | | 84432 | ( | 85.9% | ) | | 4068 | ( | 4.1% | ) | | 3413 | ( | 3.5% | ) | | 1927 | ( | 2.0% | ) | | 1632 | ( | 1.7% | ) | | 1212 | ( | 1.2% | ) | | 687 | ( | 0.7% | ) | | 551 | ( | 0.6% | ) | | 212 | ( | 0.2% | ) | | 145 | ( | 0.1% | ) | | 0  (0%) |
| sex  [factor] | 1. Female  2. Male | |  |  |  |  | | --- | --- | --- | --- | | 51211 | ( | 52.1% | ) | | 47068 | ( | 47.9% | ) | | 0  (0%) |
| member\_of\_a\_labor\_union  [factor] | 1. Not in universe  2. No  3. Yes | |  |  |  |  | | --- | --- | --- | --- | | 88718 | ( | 90.3% | ) | | 8019 | ( | 8.2% | ) | | 1542 | ( | 1.6% | ) | | 0  (0%) |
| reason\_for\_unemployment  [factor] | 1. Not in universe  2. Other job loser  3. Re-entrant  4. Job loser - on layoff  5. Job leaver  6. New entrant | |  |  |  |  | | --- | --- | --- | --- | | 95014 | ( | 96.7% | ) | | 1148 | ( | 1.2% | ) | | 1091 | ( | 1.1% | ) | | 501 | ( | 0.5% | ) | | 301 | ( | 0.3% | ) | | 224 | ( | 0.2% | ) | | 0  (0%) |
| full\_or\_part\_time\_employment\_stat  [factor] | 1. Unemployed part- time  2. Unemployed full-time  3. PT for non-econ reasons u  4. PT for econ reasons usual  5. PT for econ reasons usual  6. Not in labor force  7. Full-time schedules  8. Children or Armed Forces | |  |  |  |  | | --- | --- | --- | --- | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 98279 | ( | 100.0% | ) | | 0  (0%) |
| tax\_filer\_stat  [factor] | 1. Nonfiler  2. Joint both under 65  3. Single  4. Joint both 65+  5. Head of household  6. Joint one under 65 & one | |  |  |  |  | | --- | --- | --- | --- | | 36092 | ( | 36.7% | ) | | 33652 | ( | 34.2% | ) | | 18695 | ( | 19.0% | ) | | 4247 | ( | 4.3% | ) | | 3695 | ( | 3.8% | ) | | 1898 | ( | 1.9% | ) | | 0  (0%) |
| region\_of\_previous\_residence  [factor] | 1. Not in universe  2. South  3. West  4. Midwest  5. Northeast  6. Abroad | |  |  |  |  | | --- | --- | --- | --- | | 82547 | ( | 84.0% | ) | | 4875 | ( | 5.0% | ) | | 4068 | ( | 4.1% | ) | | 3559 | ( | 3.6% | ) | | 2700 | ( | 2.8% | ) | | 530 | ( | 0.5% | ) | | 0  (0%) |
| state\_of\_previous\_residence  [factor] | 1. Not in universe  2. California  3. Utah  4. Florida  5. North Carolina  6. Abroad  7. Oklahoma  8. Minnesota  9. Indiana  10. North Dakota  [ 40 others ] | |  |  |  |  | | --- | --- | --- | --- | | 82547 | ( | 84.6% | ) | | 1710 | ( | 1.8% | ) | | 1061 | ( | 1.1% | ) | | 847 | ( | 0.9% | ) | | 810 | ( | 0.8% | ) | | 671 | ( | 0.7% | ) | | 622 | ( | 0.6% | ) | | 572 | ( | 0.6% | ) | | 528 | ( | 0.5% | ) | | 497 | ( | 0.5% | ) | | 7707 | ( | 7.9% | ) | | 707  (0.72%) |
| detailed\_household\_and\_family\_stat  [factor] | 1. Other Rel <18 ever marr n  2. Householder  3. Child <18 never marr not 4. Spouse of householder  5. Nonfamily householder  6. Child 18+ never marr Not 7. Secondary individual  8. Other Rel 18+ ever marr n  9. Grandchild <18 never marr  10. Other Rel 18+ never marr [ 28 others ] | |  |  |  |  | | --- | --- | --- | --- | | 0 | ( | 0.0% | ) | | 26698 | ( | 27.2% | ) | | 23761 | ( | 24.2% | ) | | 20784 | ( | 21.1% | ) | | 10953 | ( | 11.1% | ) | | 6116 | ( | 6.2% | ) | | 2964 | ( | 3.0% | ) | | 923 | ( | 0.9% | ) | | 904 | ( | 0.9% | ) | | 864 | ( | 0.9% | ) | | 4312 | ( | 4.4% | ) | | 0  (0%) |
| detailed\_household\_summary\_in\_household  [factor] | 1. Householder  2. Child under 18 never marr  3. Spouse of householder  4. Child 18 or older  5. Other relative of househo  6. Nonrelative of householde  7. Group Quarters- Secondary  8. Child under 18 ever marri | |  |  |  |  | | --- | --- | --- | --- | | 37660 | ( | 38.3% | ) | | 23809 | ( | 24.2% | ) | | 20791 | ( | 21.2% | ) | | 7328 | ( | 7.5% | ) | | 4764 | ( | 4.8% | ) | | 3824 | ( | 3.9% | ) | | 80 | ( | 0.1% | ) | | 23 | ( | 0.0% | ) | | 0  (0%) |
| migration\_code.change\_in\_msa  [factor] | 1. Nonmover  2. MSA to MSA  3. NonMSA to nonMSA  4. Not in universe  5. MSA to nonMSA  6. NonMSA to MSA  7. Abroad to MSA  8. Not identifiable  9. Abroad to nonMSA | |  |  |  |  | | --- | --- | --- | --- | | 81128 | ( | 82.5% | ) | | 10572 | ( | 10.8% | ) | | 2802 | ( | 2.8% | ) | | 1419 | ( | 1.4% | ) | | 787 | ( | 0.8% | ) | | 615 | ( | 0.6% | ) | | 453 | ( | 0.5% | ) | | 430 | ( | 0.4% | ) | | 73 | ( | 0.1% | ) | | 0  (0%) |
| migration\_code.change\_in\_reg  [factor] | 1. Nonmover  2. Same county  3. Different county same sta  4. Not in universe  5. Different region  6. Different state same divi  7. Abroad  8. Different division same r | |  |  |  |  | | --- | --- | --- | --- | | 81128 | ( | 82.5% | ) | | 9779 | ( | 10.0% | ) | | 2792 | ( | 2.8% | ) | | 1419 | ( | 1.4% | ) | | 1178 | ( | 1.2% | ) | | 990 | ( | 1.0% | ) | | 530 | ( | 0.5% | ) | | 463 | ( | 0.5% | ) | | 0  (0%) |
| migration\_code.move\_within\_reg  [factor] | 1. Nonmover  2. Same county  3. Different county same sta  4. Not in universe  5. Different state in South  6. Different state in West  7. Different state in Midwes  8. Abroad  9. Different state in Northe | |  |  |  |  | | --- | --- | --- | --- | | 81128 | ( | 82.5% | ) | | 9779 | ( | 10.0% | ) | | 2792 | ( | 2.8% | ) | | 1419 | ( | 1.4% | ) | | 972 | ( | 1.0% | ) | | 678 | ( | 0.7% | ) | | 551 | ( | 0.6% | ) | | 530 | ( | 0.5% | ) | | 430 | ( | 0.4% | ) | | 0  (0%) |
| live\_in\_this\_house\_1\_year\_ago  [factor] | 1. Yes  2. No  3. Not in universe under 1 y | |  |  |  |  | | --- | --- | --- | --- | | 81128 | ( | 82.5% | ) | | 15732 | ( | 16.0% | ) | | 1419 | ( | 1.4% | ) | | 0  (0%) |
| migration\_prev\_res\_in\_sunbelt  [factor] | 1. Not in universe  2. No  3. Yes | |  |  |  |  | | --- | --- | --- | --- | | 82547 | ( | 84.0% | ) | | 9959 | ( | 10.1% | ) | | 5773 | ( | 5.9% | ) | | 0  (0%) |
| family\_members\_under\_18  [factor] | 1. Not in universe  2. Both parents present  3. Mother only present  4. Father only present  5. Neither parent present | |  |  |  |  | | --- | --- | --- | --- | | 72038 | ( | 73.3% | ) | | 18275 | ( | 18.6% | ) | | 6282 | ( | 6.4% | ) | | 901 | ( | 0.9% | ) | | 783 | ( | 0.8% | ) | | 0  (0%) |
| country\_of\_birth\_father  [factor] | 1. Panama  2. Holand-Netherlands  3. United-States  4. Mexico  5. Puerto-Rico  6. Italy  7. Germany  8. Canada  9. Dominican-Republic  10. Poland  [ 32 others ] | |  |  |  |  | | --- | --- | --- | --- | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 78665 | ( | 82.8% | ) | | 4783 | ( | 5.0% | ) | | 1347 | ( | 1.4% | ) | | 1152 | ( | 1.2% | ) | | 685 | ( | 0.7% | ) | | 685 | ( | 0.7% | ) | | 654 | ( | 0.7% | ) | | 572 | ( | 0.6% | ) | | 6484 | ( | 6.8% | ) | | 3252  (3.31%) |
| country\_of\_birth\_mother  [factor] | 1. Panama  2. Holand-Netherlands  3. United-States  4. Mexico  5. Puerto-Rico  6. Italy  7. Canada  8. Germany  9. El-Salvador  10. Cuba  [ 32 others ] | |  |  |  |  | | --- | --- | --- | --- | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 79354 | ( | 83.3% | ) | | 4666 | ( | 4.9% | ) | | 1240 | ( | 1.3% | ) | | 940 | ( | 1.0% | ) | | 714 | ( | 0.7% | ) | | 708 | ( | 0.7% | ) | | 564 | ( | 0.6% | ) | | 554 | ( | 0.6% | ) | | 6578 | ( | 6.9% | ) | | 2961  (3.01%) |
| country\_of\_birth\_self  [factor] | 1. Panama  2. Holand-Netherlands  3. United-States  4. Mexico  5. Puerto-Rico  6. Germany  7. Cuba  8. Philippines  9. El-Salvador  10. Canada  [ 32 others ] | |  |  |  |  | | --- | --- | --- | --- | | 0 | ( | 0.0% | ) | | 0 | ( | 0.0% | ) | | 87378 | ( | 90.4% | ) | | 2781 | ( | 2.9% | ) | | 708 | ( | 0.7% | ) | | 439 | ( | 0.5% | ) | | 420 | ( | 0.4% | ) | | 373 | ( | 0.4% | ) | | 360 | ( | 0.4% | ) | | 341 | ( | 0.4% | ) | | 3873 | ( | 4.0% | ) | | 1606  (1.63%) |
| citizenship  [factor] | 1. Native- Born in the Unite  2. Foreign born- Not a citiz  3. Foreign born- U S citizen  4. Native- Born abroad of Am  5. Native- Born in Puerto Ri | |  |  |  |  | | --- | --- | --- | --- | | 87378 | ( | 88.9% | ) | | 6432 | ( | 6.5% | ) | | 2824 | ( | 2.9% | ) | | 885 | ( | 0.9% | ) | | 760 | ( | 0.8% | ) | | 0  (0%) |
| own\_business\_or\_self\_employed  [factor] | 1. 0  2. 2  3. 1 | |  |  |  |  | | --- | --- | --- | --- | | 89114 | ( | 90.7% | ) | | 7939 | ( | 8.1% | ) | | 1226 | ( | 1.2% | ) | | 0  (0%) |
| fill\_inc\_questionnaire\_for\_veteran.s\_admin  [factor] | 1. Not in universe  2. No  3. Yes | |  |  |  |  | | --- | --- | --- | --- | | 97291 | ( | 99.0% | ) | | 792 | ( | 0.8% | ) | | 196 | ( | 0.2% | ) | | 0  (0%) |
| veterans\_benefits  [factor] | 1. 2  2. 0  3. 1 | |  |  |  |  | | --- | --- | --- | --- | | 74922 | ( | 76.2% | ) | | 22369 | ( | 22.8% | ) | | 988 | ( | 1.0% | ) | | 0  (0%) |

### Measurable variables

As mentioned before there are 8 measurable variables. Measurable variables can be distinguished in interval and ratio. The nature of the measurable variable determines the statistical techniques applicable. In ratio variable the zero point define the absence of that variable, whereas in interval variable there is no fixed zero point. For example, age is ratio.

Some of the main observations are:

1. No missing values are observed, presumably replaced with a constant value (e.g. 0)
2. This might explain why we observe age zero and also median of wage\_per\_hour is zero
3. Considering only US Census for 1994, year variable becomes unary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Stats / Values** | **Freqs (% of Valid)** | **Graph** | **Missing** |
| age  [integer] | Mean (sd) : 34.8 (22.3)  min < med < max:  0 < 33 < 90  IQR (CV) : 34 (0.6) | 91 distinct values | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\D74DD4F1.tmp | 0  (0%) |
| wage\_per\_hour  [integer] | Mean (sd) : 55.7 (279.3)  min < med < max:  0 < 0 < 9999  IQR (CV) : 0 (5) | 881 distinct values | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\39BD7963.tmp | 0  (0%) |
| capital\_gains  [integer] | Mean (sd) : 416.6 (4566.3)  min < med < max:  0 < 0 < 99999  IQR (CV) : 0 (11) | 127 distinct values | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\A8761C69.tmp | 0  (0%) |
| capital\_losses  [integer] | Mean (sd) : 37.9 (272.6)  min < med < max:  0 < 0 < 4356  IQR (CV) : 0 (7.2) | 107 distinct values | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\4E3FE35F.tmp | 0  (0%) |
| dividends\_from\_stocks  [integer] | Mean (sd) : 196.8 (1964.3)  min < med < max:  0 < 0 < 99999  IQR (CV) : 0 (10) | 950 distinct values | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\E3C4A3C5.tmp | 0  (0%) |
| instance\_weight  [numeric] | Mean (sd) : 1731.5 (973.9)  min < med < max:  81.3 < 1628.8 < 18656.3  IQR (CV) : 1093 (0.6) | 54010 distinct values | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\8DDC4F1B.tmp | 0  (0%) |
| num\_persons\_worked\_for\_employer  [integer] | Mean (sd) : 1.9 (2.3)  min < med < max:  0 < 1 < 6  IQR (CV) : 4 (1.2) | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 0 | : | 46615 | ( | 47.4% | ) | | 1 | : | 11911 | ( | 12.1% | ) | | 2 | : | 5311 | ( | 5.4% | ) | | 3 | : | 6848 | ( | 7.0% | ) | | 4 | : | 7374 | ( | 7.5% | ) | | 5 | : | 3059 | ( | 3.1% | ) | | 6 | : | 17161 | ( | 17.5% | ) | | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\8DDF52E1.tmp | 0  (0%) |
| weeks\_worked\_in\_year  [integer] | Mean (sd) : 23.4 (24.4)  min < med < max:  0 < 10 < 52  IQR (CV) : 52 (1) | 53 distinct values | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\F54DDBD3.tmp | 0  (0%) |
| year  [integer] | 1 distinct value | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 94 | : | 98279 | ( | 100.0% | ) | | C:\Users\espad\AppData\Local\Packages\Microsoft.Office.Desktop_8wekyb3d8bbwe\AC\INetCache\Content.MSO\F6A75859.tmp | 0  (0%) |

To boost my productivity in Exploratory Data Analysis, I loaded the deduplicated training data into Power BI. This tool allow quick data visualization and cross-filtering of several visualization in order to start grasping insights from the data.

At the end of the EDA, most relevant data visualizations identified using Power BI will be translated into code and loaded into the script in order to allow reproducibility. Some of the other possible alternatives to this approach would have been creating dashboards using Dataiku, creating reports using AWS QuickSight, or publishing the Power BI report on Power BI services.

Considering the data are already sampled, the entire training set is used for EDA.

The output for the EDA will be (1) descriptive statistics, (2) data visualizations and (3) summary of insights.

Make a quick statistic based and univariate audit of the different columns’ content and produce the results in visual / graphic format. This audit should describe the variable distribution, the % of missing values, the extreme values, and so on

CATEGORICAL

target (TARGET)

class\_of\_worker (ACLSWKR)

detailed\_industry\_recode (ADTIND)

detailed\_occupation\_recode (ADTOCC)

education (AHGA)

enroll\_in\_edu\_inst\_last\_wk (AHSCOL)

marital\_stat (AMARITL)

major\_industry\_code (AMJIND)

major\_occupation\_code (AMJOCC)

race (ARACE)

hispanic\_origin (AREORGN)

sex (ASEX)

member\_of\_a\_labor\_union (AUNMEM)

reason\_for\_unemployment (AUNTYPE)

full\_or\_part\_time\_employment\_stat (AWKSTAT)

tax\_filer\_stat (FILESTAT)

region\_of\_previous\_residence (GRINREG)

state\_of\_previous\_residence (GRINST)

detailed\_household\_and\_family\_stat (HHDFMX)

detailed\_household\_summary\_in\_household (HHDREL)

migration\_code-change\_in\_msa (MIGMTR1)

migration\_code-change\_in\_reg (MIGMTR3)

migration\_code-move\_within\_reg (MIGMTR4)

live\_in\_this\_house\_1\_year\_ago (MIGSAME)

migration\_prev\_res\_in\_sunbelt (MIGSUN)

family\_members\_under\_18 (PARENT)

country\_of\_birth\_father (PEFNTVTY)

country\_of\_birth\_mother (PEMNTVTY)

country\_of\_birth\_self (PENATVTY)

citizenship (PRCITSHP)

own\_business\_or\_self\_employed (SEOTR)

fill\_inc\_questionnaire\_for\_veteran's\_admin (VETQVA)

veterans\_benefits (VETYN)

year (YEAR)

MEASURABLE

instance\_weight (MARSUPWT)

age (AAGE)

wage\_per\_hour (AHRSPAY)

capital\_gains (CAPGAIN)

capital\_losses (CAPLOSS)

dividends\_from\_stocks (DIVVAL)

num\_persons\_worked\_for\_employer (NOEMP)

weeks\_worked\_in\_year (WKSWORK)

## Validation

It seems data were already validated in terms of guaranteeing a minimum data quality:

* Age is already capped to 90
* There is no university student below 20 years old
* All children are below $50K

# Remapping Categorical Variables

Considering the high level of remapping for normalizing the data, it will be important to deploy a mechanism to monitor when appears data outside the distribution.

# External data:

Pull external data for income about the different countries. However, many have missing values so it was not very useful.

Instead, it is useful to understand the median salary (if not available then also the average) for 94 and 95.

Ideally, we would retrieve this information for each state such that we can assign a priori probability to each individual.

<https://catalog.data.gov/dataset?organization_type=State+Government&tags=income>

http://kdd.ics.uci.edu

https://www.google.it/search?client=opera&hs=EeI&ei=55g1XsamK4WYkwX8raOgDw&q=median+personal+income+1995+U.s.&oq=median+personal+income+1995+U.s.&gs\_l=psy-ab.3...17404.18861..18898...0.3..0.105.1275.14j1......0....1..gws-wiz.......0i71.KR1dA2xoyY0&ved=0ahUKEwiGtpaj1bDnAhUFzKQKHfzWCPQQ4dUDCAo&uact=5

If the variables representing region or state of residence were available – and with a small number of missing values – it would have been interesting to generate features based on socio-demographic data specific for that region. For example, if in Utah the average salary is $10,000 with a small variance, then it would have been very unlikely for an individual in Utah to be . Instead, if New York median salary is $100,000 then it would have been normal to observe.

# Objective:

<https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29>

https://kdd.ics.uci.edu/databases/census-income/census-income.data.html

Predict total personal income level as a dichotomous variable binned at the $50K

# Next steps

* Review code for production
  + E.g. Build in Dataiku
  + Connect to AWS data source
* Test on 2019 Census
  + This might require understanding whether attributes are the same
* Understand changes over time of different variables
* Database
* Use correctly predicted tested label to further improve model learning

# Complexities

## Environment Setup

One of the complexities of the project surely is the setup of the environment. For operationalization, probably a good approach would be to build a Docker container containing the setup of the environment in order to support results reproducibility. In addition, it is also required for environments to be aligned. I performed this using renv which is a tool that keep track of packages versions used in both R and Python.

In the case of multiple data scientist working on the same project, it would have been even more complicated to keep aligned all the different environments.

Nevertheless, this is a time consuming and error prone task. It is required for deploying on AWS, Azure or GCP. Whereas if deployed on a Data Science platform like Dataiku’s DSS, environment setup and operationalization are simplified to the root.

## Metadata

Instance weight