Contents

IMBALANCED BINARY CLASSIFICATION ON US CENSUS BUREAU 1994/1995

Introduction

The project group is commissioned this POC by Sterling Cooper Advertising Agency willing to extend its data science capabilities. The motivation for the analysis is to (1) understand whether it is possible to predict individuals with a annual persona income above or below a certain threshold and (2) define a possible workflow as a starting point for a future implementation in the case results from (1) are promising. The scope of the project is to define a clear data science pipeline going from data to model predictions on a test set. The POC 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 this project includes:

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

The objective of the solution is to solve an imbalanced binary classification problem predicting total personal income level binned as a dichotomous variable at $50,000 threshold.

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 platform the tool will be deployed and how it will integrate with its current IT systems. Therefore, the second milestone consists of operationalizing the workflow – including monitoring tools – on the platform chosen by the Sterling Cooper Advertising Agency.

The dataset to be used is US Census Bureau which includes data for 1994 and 1995. The attached metadata file provides further information for each variable and the sampling technique used to perform the census which was based on stratified sampling. The training set and the test set are both composed of mixed data of 1994 and 1995. The project group can include any data source as long as it will be available for new predictions; this to avoid risk of data leakage and biasing analysis. The dataset provided to the project group is already partially prepared following the same procedure that will be applied to future datasets so they will be consistent with the data structure provided for US Census Bureau of 1994 and 1995. The project group can use any possible variable for the prediction, also obtained from external data sources (as long as it will be available for new predictions).

The report is focused on the accomplishment of the first milestone. The amount of effort to be involved is 5 man/days over an elapsed of two weeks. **Considering the very limited effort (only 5m/d), no advanced implementation is expected from the project group**. Instead, this POC will be use as starting point for future advancements based on feedbacks from Sterling Cooper Advertising Agency.

In some section is added a paragraph about possible enhancement containing proposed alternative approaches in the case more effort is to be spent.

Activities

This section includes a recap of daily activities performed to complete the POC.

Time slot	Day 1	Day 2	Day 3	Day 4	Day 5
9-10AM	Project planning	Plan daily activities	Plan daily activities	Plan daily activities	Plan daily activities
10-11AM	Setup environment	Exploratory Data Analysis (Part 1)	Data Preparation	Evaluation	Exploratory Data Analysis (Part 2)
11-12AM			Modelling	Data Preparation	Modelling
15-16PM	Analyze and prepare metadata		Evaluation	Modelling	Evaluation
16-17PM	Exploratory Data Analysis (Part 0)		Data Preparation	Evaluation	Summarize daily activities in diary
17-18PM		Data Preparation	Modelling	Data Preparation	Review document
18-19PM	Summarize daily activities in diary	Summarize daily activities in diary	Summarize daily activities in diary	Summarize daily activities in diary	Review document

Environment setup

In order to provide Sterling Cooper Advertising Agency the best from both Python and R, this project will be based on an interface allowing to collaborate using both the environment. The interface used is the enhanced reticulate package (v1.14). For this purpose, two virtual environments are setup in order to guarantee reproducibility of the work: one for Python (Conda) and one for R (RStudio). The package manager used is *renv* because supports both R and Python virtual environments.

Versioning is performed using a free license for private github repository allowing collaboration up to 3 users.

In order to fasten visual prototyping, the project group might take advantage of other open source data visualization tools (e.g. Power BI Desktop, Amazon QuickSight).

Possible enhancement

Even though the setup is performed on local machines, to be as portable as possible on other platform (e.g. Dataiku, GCP, AWS, Azure) and to allow easier reproducibility and collaboration, it should be preferred to develop on Docker Containers.

Data Source

The data sources are two csv files stored on an open AWS S3 bucket; one for training and one for testing. The integration process for deployment is not defined and out of scope for this POC.

For the purpose of this POC data are stored locally in the project folder without allowing that folder to be synched with git (using *.gitignore* command). Datasets will follow the immutability principle such data no dataset is overwritten, instead, a new dataset is created for each transformation.

Possible enhancement

Ideally, all data should be stored in an encrypted cloud storage (e.g. AWS S3) and partitioned. In this scenario, a possible partition should have been the time (i.e. year of the census).

Metadata

The metadata text file provided by the client is structured into a csv file in order to consistently keep track of:

- variables name and description
- data type

The structure of the file is shown in table below:

code	description	variable_type	possible_values
AAGE	age	continuous	continuous

This step is important for the workflow because:

1. Column headers are missing in the data files
2. The metadata text file does not provide a clear understanding of which column in the data source is associated to which variable. In fact, the metadata file refers to more than 50 columns, but the file has only 42. The process was by exclusion: first assigned the column name to the field (e.g. Married will be marital status), next proceed by exclusion.

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

Possible enhancement

Review with the client if the project group assigned the correct variable name to each column in the data source. Also, define a structure governance process to maintain and organize the metadata file.

Workflow

There are many data mining workflow, however, in the end they all express the same approach. The project group decides to represent its data mining workflow using below graph because (1) it explicitly includes business understanding in the picture and (2) it reflects the cyclical and iterative nature of data mining processes.

Considering Census Bureau are yearly snapshot, it does not make sense to combine and analyze together 1994 and 1995 census. Instead, they are longitudinal data. There are many approaches for dealing with longitudinal data, the one proposed for this POC consists of performing Data Understanding separately on 1994 and 1995 in order to identify both common and different patterns. Next, a model is trained on 1994 and used to predict on 1994. Finally, the generalization power for the model trained on 1994 will be tested on 1995 using the model trained on 1995 as benchmark.

When the most performant model is trained, the proposed approach for production will be to train a model on year T and use it to predict on year $T + 1$.
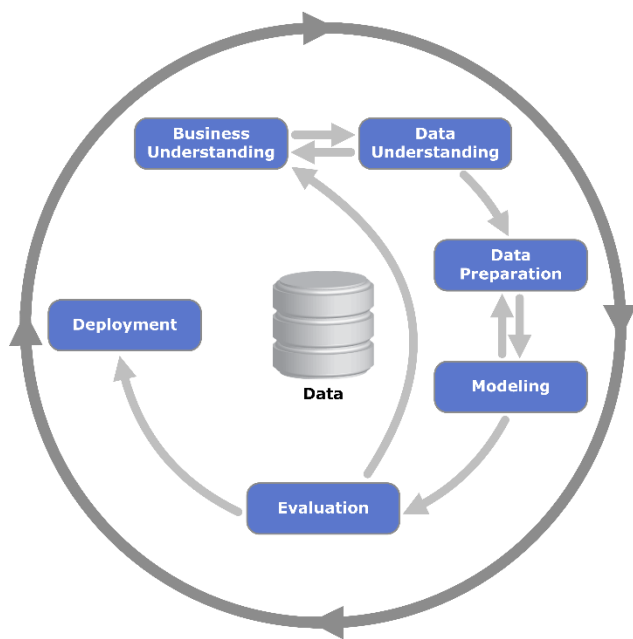
Business Understanding

From a business perspective, the scope of this project is to improve marketing effectiveness by identifying individuals above the $50,000/year personal income threshold with better balance between precision and recall. In fact, this would allow the client to reduce direct mail costs by better identifying individual with enough personal income available to be targeted by the client's advertising campaigns.

Another important asset for this project is the discovery of clear insights on the profiles and characteristics of individuals that earn more than $50,000/year.

Possible enhancements

One of the possible enhancements for the business understanding phase is interviews with business users. Some of the reasons why interviews with business are important:

1. Deepen understanding of their needs
2. Understand current implementation
3. Define a clear expectation for this project
4. Define metrics for success

Data Understanding (EDA)

In this section are presented descriptive statistics for categorical (nominal and ordinal) and for measurable (interval and ratio) variables. Considering time constraints and priorities, only an overview of all variables is proposed. A deeper individual analysis of each variable from univariate and multivariate perspective is referred in the possible enhancements. EDA is performed on 1994 US Census data because of time constraints. EDA on 1995 and comparison of common patterns is included in possible enhancements for this section.

EDA considers entire training data for 1994 US Census.

The outcome expected for this phase is:

-   Identification of promising variables
-   Identification of problematic variables that requires outliers treatment, normalization, etc.
-   Identification of interesting patterns to validate

Duplicates

To perform 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).

In general, it does not make sense from analytical perspective to sum two snapshots (1994, 1995) which represents longitudinal data. 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.

Further investigation on duplicates and possible alternative approaches in dealing with duplicated are referred in possible enhancements.

Instance Weight

Before proceeding further, it is important to discuss the instance weight variable (MARSUPWT) present in the raw data. According to the metadata text file and this link: "*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.

As per request of metadata text file, instance_weight is included for data visualizations but excluded from modelling.

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 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 approach helps in identifying variables that are not relevant (e.g. unary) and also distinguish between nominal and ordinal. During the data preparation, this information should be taken into consideration. The difference between nominal and ordinal variables might be relevant to take into consideration in the case it contains predictive power. This dataset seems to present some variable ordinal by nature (e.g. education), however, further analyses are needed in order to understand whether the ranking 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).

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

1. Only 6.3% of the instances present the target class of interest.
2. Most of the instances are related to race "white" (82.7%).
3. Invalid or missing values seem to be remapped in "Not in universe" or "?" with exception for few variables (e.g. country_of_birth_father).
4. For some variables (e.g. region_of_previous_residence), instances are concentrated only in few classes out of the many.
5. full_or_part_time_employment_stat is unary in the 1994 training set.
6. Only education seems to be an ordinal variable but there are so many different levels that it is difficult to define a clear ranking.

Deeper univariate and multivariate analysis of categorical variables, measurable variables and interaction with target are referred to possible enhancement. The results of this analysis might suggest possible feature engineering approaches (e.g. transformation based on weight of evidence).

Variable	Stats / Values	Rel Freqs	Missing

target	1. - 50000. 2. 50000+.	92440 (94.1%) 5839 (5.9%)	0 (0%)
class_of_worker	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	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	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	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	1. Not in universe 2. High school 3. College or university	91984 (93.6%) 3387 (3.4%) 2908 (3.0%)	0 (0%)
marital_stat	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	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	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	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	1. All other 2. Mexican-American 3. Mexican (Mexicano) 4. Central or South American 5. Puerto Rican 6. Other Spanish 7. NA	84432 (85.9%) 4068 (4.1%) 3413 (3.5%) 1927 (2.0%) 1632 (1.7%) 1212 (1.2%) 687 (0.7%) 551 (0.6%)	0 (0%)

	8. Cuban 9. Do not know 10. Chicano	212 (0.2%) 145 (0.1%)	
sex	1. Female 2. Male	51211 (52.1%) 47068 (47.9%)	0 (0%)
member_of_a_labor_union	1. Not in universe 2. No 3. Yes	88718 (90.3%) 8019 (8.2%) 1542 (1.6%)	0 (0%)
reason_for_unemployment	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	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	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	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	1. Not in universe 2. California 3. Utah 4. Florida 5. North Carolina 6. Abroad 7. Oklahoma 8. Minnesota	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%)	707 (0.72%)

	9. Indiana 10. North Dakota [40 others]	497 (0.5%) 7707 (7.9%)	
detailed_household_and_family_stat	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	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	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	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	1. Nonmover 2. Same county 3. Different county same sta 4. Not in universe 5. Different state in South	81128 (82.5%) 9779 (10.0%) 2792 (2.8%) 1419 (1.4%) 972 (1.0%)	0 (0%)

	6. Different state in West 7. Different state in Midwes 8. Abroad 9. Different state in Northe	678 (0.7%) 551 (0.6%) 530 (0.5%) 430 (0.4%)	
live_in_this_house_1_year_ago	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	1. Not in universe 2. No 3. Yes	82547 (84.0%) 9959 (10.1%) 5773 (5.9%)	0 (0%)
family_members_under_18	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	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	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	1. Panama 2. Holand-Netherlands 3. United-States 4. Mexico 5. Puerto-Rico 6. Germany	0 (0.0%) 0 (0.0%) 87378 (90.4%) 2781 (2.9%) 708 (0.7%) 439 (0.5%) 420 (0.4%)	1606 (1.63%)

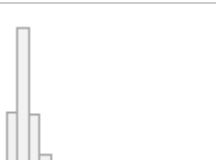|				
	7. Cuba 8. Philippines 9. El-Salvador 10. Canada [32 others]	373 (0.4%) 360 (0.4%) 341 (0.4%) 3873 (4.0%)		
citizenship	1. Native- Born in the Unite 2. Foreign born- Not a citiz 3. Foreign born- U S citizen 4. Native- Born abroad of Am 5. Native- Born in Puerto Ri	87378 (88.9%) 6432 (6.5%) 2824 (2.9%) 885 (0.9%) 760 (0.8%)		0 (0%)
own_business_or_self_employed	1. 0 2. 2 3. 1	89114 (90.7%) 7939 (8.1%) 1226 (1.2%)		0 (0%)
fill_inc_questionnaire_for_veteran.s_admin	1. Not in universe 2. No 3. Yes	97291 (99.0%) 792 (0.8%) 196 (0.2%)		0 (0%)
veterans_benefits	1. 2 2. 0 3. 1	74922 (76.2%) 22369 (22.8%) 988 (1.0%)		0 (0%)

Measurable variables
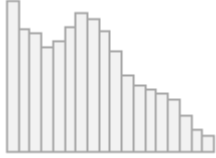
There are 8 measurable variables in the dataset. Measurable variables can be further distinguished in interval and ratio. The nature of the measurable variable determines the statistical techniques applicable. In ratio variables, the zero point define the absence of that variable, whereas in interval variable there is no fixed zero point. For example, the nature of age variable 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 zero value for age and also very right skewed distribution (e.g. wage_per_hour) which median is zero.
3. Considering only US Census for 1994, year variable becomes unary.

Deeper univariate and multivariate analysis of categorical variables, measurable variables and interaction with target are referred to possible enhancement. The results of this analysis might suggest possible feature engineering approaches (e.g. log transformation) and well as outliers' treatments.

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		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		0 (0%)
capital_gains [integer]	Mean (sd) : 416.6 (4566.3) min < med < max: 0 < 0 < 99999 IQR (CV) : 0 (11)	127 distinct values		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		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		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		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%)		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		0 (0%)

year [integer]	1 distinct value	94 : 98279 (100.0%		0 (0%)

Validation

It is always best practice to validate the data. In this project, it seems data was already validated in terms of guaranteeing a minimum viable data quality. For example:

- age is already capped to 90,
- there is no university student below 20 years old
- All children are below $50K

Further validation activities are referred in possible enhancements.

Possible enhancements

In the case the client wishes to invest additional effort on this POC, the proposed priority for additional activities in data preparation is:

1. Extend EDA on 1994 Census data: medium effort.
   Deeper individual analysis of each variable from univariate and multivariate perspective. Adding correlation matrix and measures of correlation between mixed type variables.

2. Trend comparison: high effort.
   Perform EDA on 1995 Census training data and compare results with the training data of 1994.

Data Preparation

Data preparation consists of two main components: (1) create and validate a pipeline to obtain the final dataset starting from training data and (2) extend the pipeline in order to automate the process on any new dataset (i.e. to be exactly replicated on test set). Additionally, it is best practice to include a report on checks to make sure that once in production the data ingested by the pipeline do not present new issues that requires adaptation of the data pipeline.

The data preparation includes but is not limited to:

1. Creation of new variables
    a. Data enrichment
    b. Features engineering
2. Transformation of variables
    a. Categorical: weight of evidence, one-hot-encoding, etc.
    b. Measurable variables: log transform, standardization, etc.
3. Treatment of
    a. Missing values
    b. Invalid data
    c. Outliers
    d. Duplicates

4. Label encoding of target variable

Creation of new variables

Variables can be created from at least two sources: (1) enrichment with additional data sources that add further context to the dataset under investigation and (2) use of domain knowledge to include new business concepts not available in the raw data.

An example of data enrichment consists of integrating training data with socio-demographic context (e.g. population and income) for each U.S. countries. This would allow to update prior probabilities of individuals based on their country of residence. For example, if in Utah the average salary is $10,000 with a small variance, then it is very unlikely to observe an individual above $50,000/year.

Creation of new variables is refereed in possible enhancements.

Variables transformation

The scope of variable transformation is to (1) better expose the signal contained in variables and (2) change the format in order to be the one required for the algorithm. For example, if there is no specific reason to give measurable variables different weights it is best practice to standard scale them. Another example is related to the implementation of algorithms in Machine Learning library, in fact, Python Scikit-Learn requires all training data to be in numeric format regardless of the algorithm, whereas R Gradient Boosting Trees in gbm implementation accepts categorical variables.

Hypothesis on variables transformation are identified during EDA phase and validated in modelling. Considering EDA was limited to an overview, the only variables transformation applied is standardization whereas further investigations are referred in possible enhancements. Additional investigation for categorical variables are mostly related to the normalization of values because some variables are very sparse over many different possible values (e.g. state_of_previous_residence has 50 possible values).

Data Treatment

Missing values

We observed a minority of missing values (i.e. less than 3%) in some categorical variables (i.e. country_of_birth_father) which are imputed with a constant string "Missing". Other and more advanced approaches exist for imputing missing values, but the effort is not justified but the very small number of missing.
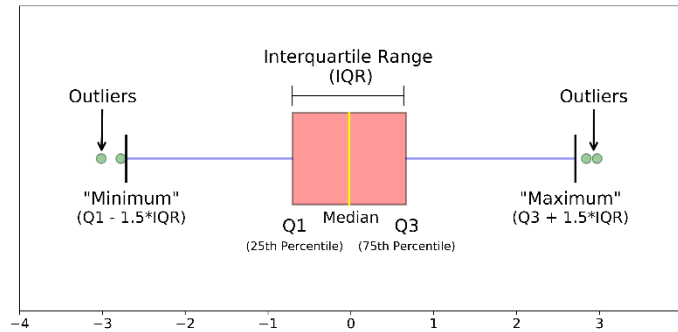
Invalid data

During a rapid data validation in EDA phase no striking invalid data is observed. However, further effort in data validation might highlight the need to deal with some invalid data.

Outliers

Outliers are observed in the following measurable variables: wage_per_hour, capital_gains, capital_losses, dividends_from_stocks and instance_weight.

There are many approaches to deal with outliers from either univariate or multivariate perspective. Approaches differ on the metric used to identify outliers. For small dataset with a limited number of variables, it is preferred to not apply a general rule to remove all outliers (e.g. remove all data above 1.5 IQR from maximum or minimum).

Interquartile Range
(IQR)

Outliers

Outliers

"Minimum"
(Q1 - 1.5*IQR)

Q1
(25th Percentile)

Median

Q3
(75th Percentile)

"Maximum"
(Q3 + 1.5*IQR)

−4   −3   −2   −1   0   1   2   3   4

Considering time constraints, it was not possible to investigate the outlier treatment. This is a known limitation of the project. Outliers treatment is referred as possible enhancement; it will be interesting to compare the performance of models without and with outliers treatment.

Duplicates

Duplicates were identified during EDA phase and it was decided to exclude them considering they represented a minority portion. Duplicates deletion is included in the data preparation pipeline.

Label encoding

In the raw data, target variable is categorical with possible values " - 50000." and " 50000+.". Label encoding is the process through which the target variable is remapped into different values. In this case, value " 50000+." is remapped to "1" because it is the class of interest, and " - 50000." is remapped as 0. Depending on the algorithm use, it might require different remapping; e.g. SVM requires binary classes to be -1 and +1.

Possible enhancements

In the case the client wishes to invest additional effort on this POC, the proposed priority for additional activities in data preparation is:

1. Data Treatment: medium effort.
   Proceed in outliers' treatment investigation.

2. Variables transformation: high effort.
   Review variables transformation especially for categorical variables which represent the majority of the dataset.

3. Creation of new variables: medium effort.
   Integrate relevant data sources to enrich context and use domain knowledge to create new relevant features.

Modelling

The training data used for the modelling phase are 1994 Census obtained as output of the data preparation pipeline described in previous phase. Considering the target variable is binary, the project group decides to frame this as a binary classification problem.

Considering no treatment was applied to categorical variables – and they count for 32 out of 41 explanatory variable – the decision is to use algorithms and implementation that accepts categorical variables. The decision a priori goes toward Gradient Boosting Trees implemented in R.

We refer in possible enhancement the exploration and comparison of additional algorithms and implementations, which heavily depends on the effort spent on data preparation. For example, more effort should be invested on outliers' treatment and variables transformation (e.g. one-hot-encoding for categorical).

Variables selection

Variables selection should include at least two different perspectives: (1) a priori multicollinearity between variables identified during EDA and (2) variable importance calculated using multiple approaches.

For exceptional time constraints of this POC, variables are selected based on a trade-off between domain knowledge and speeding up training time. In fact, considering there are 32 categorical variables and many of them have above 40 levels, it is better to exclude them in order to avoid extremely long training time. We refer to more advanced variables selection only after extensive EDA is performed along with more robust data preparation for categorical variables.

Therefore, the variables selected for this iteration are:

- Age: measurable;
- Education: categorical
- Marital status: categorical
- Race: categorical
- Sex: categorical

An additional topic for possible enhancement is ethic about the use of certain variables. For example, in insurance industry gender cannot be used as discriminant for pricing the risk.

Class imbalance

In the 1994 US Census training data only 5.9% of individuals present the class of interest. On the one hand, this is an imbalanced dataset, on the other hand, this data is representative of the entire population from statistical perspective.

There are three main approaches in dealing with class imbalance:

1) Rework the dataset: e.g. oversampling, undersampling, generating synthetic data.
2) Getting additional features: i.e. to make classes more separable by the classifier.
3) Rework the problem: e.g. cost-based classification, probability threshold, classes reweight.

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. For this POC it is decided to compare performance of models based on:

- No class rebalance technique applied
- Classes reweight

Model

In order to speed-up performance we defined (1) 5-fold cross validation with stratified sampling to be repeated 2 times on each training and (2) maximum of 60 iterations for gradient boosting algorithm.

Here the results for the model trained using only 1994 training data.

No class re-balancing

Below the performance of the model with highest accuracy:

Confusion matrix:

```
              Reference
Prediction above below
     above    863    464
     below   4976  91976

               Accuracy : 0.9446
                 95% CI : (0.9432, 0.9461)
    No Information Rate : 0.9406
    P-Value [Acc > NIR] : 2.751e-08

                  Kappa : 0.2238

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.147799
            Specificity : 0.994981
         Pos Pred Value : 0.650339
         Neg Pred Value : 0.948676
             Prevalence : 0.059412
         Detection Rate : 0.008781
   Detection Prevalence : 0.013502
      Balanced Accuracy : 0.571390

       'Positive' Class : above
```
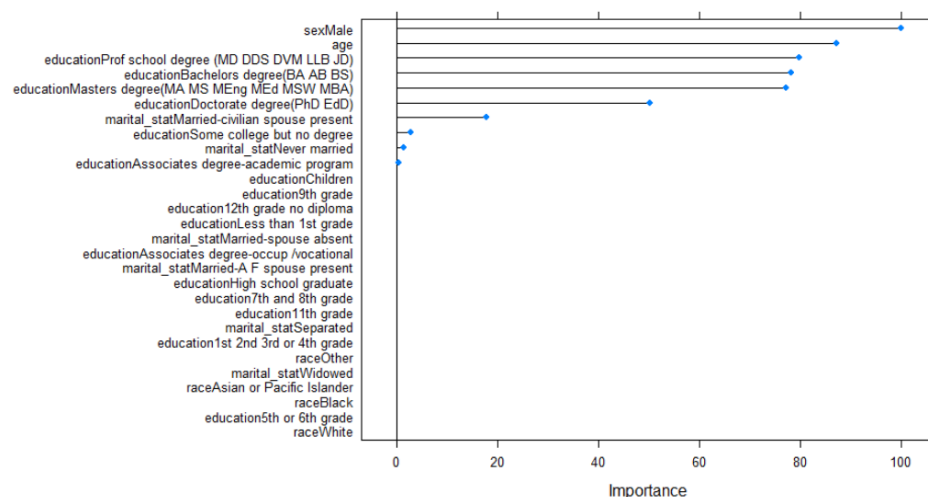
➢ The model presents 94% accuracy because it predicts correctly 99.49% of the negative class
➢ The model predicted as "above" only those individuals with very high probability
➢ Only 14% of the positive class was correctly identified

Variables importance:

➢ In this very naïve approach to variable importance, we can observe attributes with highest importance are Male (sex), age, high level of education, and Married (marital_stat)

Class re-weighted

Below the performance of the model with highest accuracy:

Confusion matrix:

```
          Reference
Prediction above below
     above  5088 20715
     below   751 71725

              Accuracy : 0.7816
                95% CI : (0.779, 0.7842)
   No Information Rate : 0.9406
   P-Value [Acc > NIR] : 1

                 Kappa : 0.2488

 Mcnemar's Test P-Value : <2e-16

           Sensitivity : 0.87138
           Specificity : 0.77591
        Pos Pred Value : 0.19719
        Neg Pred Value : 0.98964
            Prevalence : 0.05941
        Detection Rate : 0.05177
  Detection Prevalence : 0.26255
     Balanced Accuracy : 0.82365

      'Positive' Class : above
```
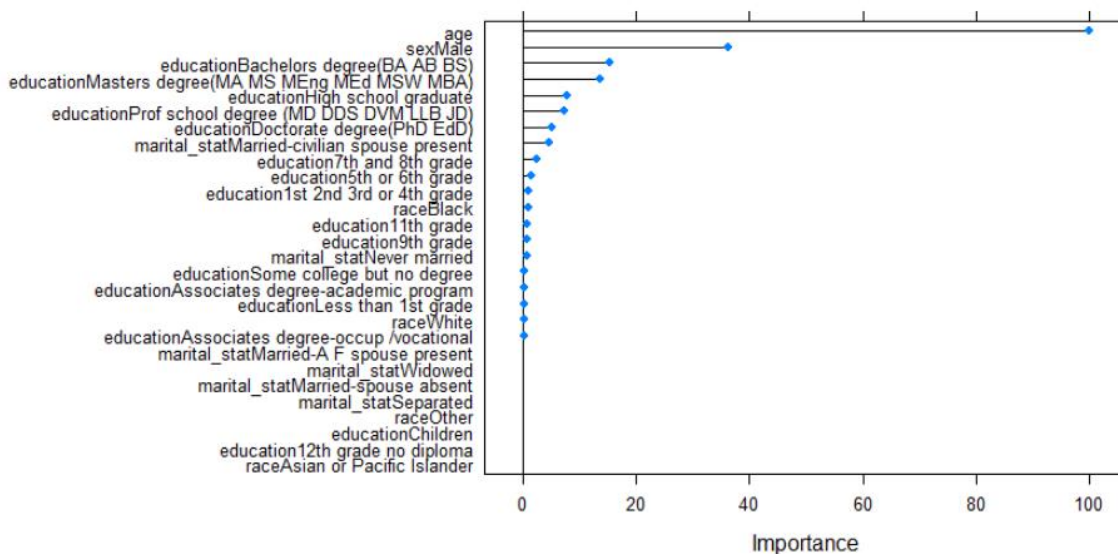
➢ Accuracy is reduced to 78% because more precision and recall costed a significant decrease in specificity
➢ The model predicted as "above" only those individuals with very high probability
➢ 87% of the positive class is correctly identified, but the precision is only 19%

Variables importance:

➢ age becomes by far the major discriminant

➢ other top attributes remain Male (sex) and high education

Hyperparameter tuning

Considering time constraints only a naïve approach to hyperparameter tuning is proposed using grid search over a limited list of parameter values. Possible enhancements include more advanced approaches to hyperparameter optimization. Also, regularization approaches and parameter value are referred in possible enhancements.

Possible enhancements

Some of the possible enhancements to extend the modelling phase are:

- Model selection based on more advanced metric rather than accuracy
- Comparison of several different algorithms and variables combinations
- Improving variables selection
- Deeper hyperparameter tuning
    - Including loss function and metrics used for model selection
- Investigation of other class re-balance techniques
- Outliers treatment
- Better preparation for categorical features

Evaluation

In this phase it is performed the evaluation of trained models on test sets. Considering only the model on 1994 is trained, it will be tested only on 1994 data.

No class re-balancing

Below the performance of the model with highest accuracy:

Confusion matrix:

```
          Reference
Prediction above below
     above   424    243
     below  2435  46307

              Accuracy : 0.9458
                95% CI : (0.9438, 0.9478)
    No Information Rate : 0.9421
    P-Value [Acc > NIR] : 0.0002238

                 Kappa : 0.2235

 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.148304
           Specificity : 0.994780
        Pos Pred Value : 0.635682
        Neg Pred Value : 0.950043
            Prevalence : 0.057864
        Detection Rate : 0.008581
   Detection Prevalence : 0.013500
      Balanced Accuracy : 0.571542

       'Positive' Class : above
```

➢ The performance of the model remains consistent in training and test set

Class re-weighted

Confusion matrix:

```
            Reference
Prediction above below
     above  2496 10370
     below   363 36180

               Accuracy : 0.7828
                 95% CI : (0.7791, 0.7864)
    No Information Rate : 0.9421
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2461

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.87303
            Specificity : 0.77723
         Pos Pred Value : 0.19400
         Neg Pred Value : 0.99007
             Prevalence : 0.05786
         Detection Rate : 0.05052
   Detection Prevalence : 0.26040
      Balanced Accuracy : 0.82513

       'Positive' Class : above
```

➢ The performance of the model remains consistent in training and test set

Model selection

It is not possible to perform a model selection without involving the business. In fact, before proceeding it is necessary to convert the performance of these two models into expected monetary value.

Possible enhancements

It is important to extend model evaluation phase to include additional metrics, better visualization and deeper variables importance analyses. Additionally, model evaluation should include 1995 data.

Deployment

The deployment of the product is out of scope for this POC. However, possible enhancement is:

1. Entirely review code in order to provide a data science product
2. Propose a possible architecture for the product to be deployed

Next steps

Some of the possible next steps are:

- Review code for production
  - o E.g. Build in Dataiku
  - o Connect to cloud data source and databases

- Test model on 2019 Census
    - This might require understanding whether attributes are the same
- Understand changes over time of different variables

Complexities

Time constraint

Given the short period of time available, it is important to be able to prioritize the essential and omit unnecessary. Following Andrea Ng suggestion, it is important to build first and end-to-end process and, next invest additional effort prioritizing actions with highest trade-off between impact and effort.

The report was considered as important as the code written because the methodology and documentation is a fundamental component of every project.

Census: snapshot data

It was not an easy decision the one to consider only 1994. But it would be analytically wrong to consider both censuses. US Census Bureau for 1994 and 1995 represents longitudinal data – snapshot at a certain point in time. Therefore, it is analytically and conceptually wrong to merge the two censuses. The decision was to treat the year separately.

Dealing with imbalanced dataset

The dataset is imbalanced therefore it required investigation of possible approaches for rebalancing the dataset. As a first iteration of the process, it was compared a model with no rebalance and one with classes reweight. There are many other possible approaches and hyperparameters to investigate.

Taking advantage from categorical features

Dealing with a dataset in which the majority of variables is categorical and most of them have many possible levels is complex. In fact, it limits the choices for possible implementation (i.e. scikit-learn only numerical) and if blindly using one-hot encoding the result would be an extremely sparse matrix.