Identifying Characteristics Associated with Income

On US Census Bureau data

Valentin TASSEL

TABLE OF CONTENTS

A collection of economic and demographic data

The sample dataset contains a detailed extract of the American population

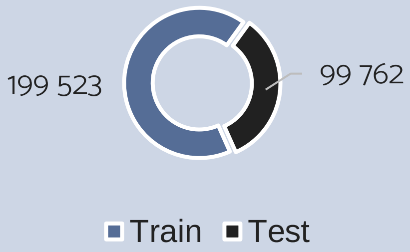
42 different attributes about:

- Social situation of the individual
- Employment situation
- Demographic situation
- Financial situation

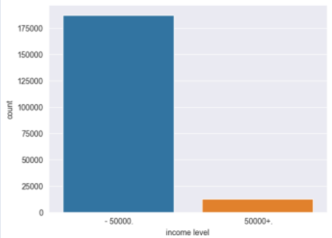
A mix of categorical and numeric features record those different attributes

Total number of instances: 299 285

Set Separation:



The Income level is defined by **two classes** with an **imbalance** with the major class (93.6 %) « Income level < 50K USD »



Personal information obtained from a survey

Each instance represent a person

The income level is the target feature for our analysis

Leverage Machine Learning methods

Using scientific process to uncover characteristics of income level

Goal: Identify characteristics that are associated with a person making more or less than 50K USD per year

From a business knowledge perspective, these following aspects might be important

- Individual personal characteristics
  Age, Gender, Race
- Work situation
  Wage per hour, Occupation, Industry
- Other revenues
  Capital gains, Dividends from stocks, Capital losses
- Other unknown characteristics

Machine learning will be used to discover the unknown characteristics having an impact and rank by importance all the mentioned characteristics

The computational capacity is a challenge when using machine learning, we will have to transform and select the right data before constructing a model

Engineering of the data modeling pipeline to obtain the results

Data exploration

We validates hypothesis and observe visually the insight coming from the features via multiple type of graph according the data type during the exploration phase.

Feature engineering & transformation pipeline

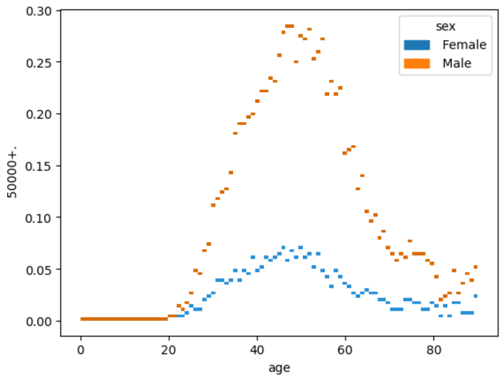
A critical part of the success of a Machine Learning project is coming up with a good set of feature to train on.
For this we apply multiple modification in the features (Modification, Selection, Extraction) coming from observation during the exploratory phase.
We automate the process via SciKit Learn Pipeline and Column transformer package

Modelling for feature importance

Using machine learning model able to learn hidden pattern from the data we will be able to assess and rank the feature importance via a machine point of view completing the first hypothesis observed from a business point of view.
By having a precise and accurate model for the prediction of the target variable we will leverage the importance of the features in this prediction.

First exploration to validate our hypothesis

To organize our exploratory phase, we began to observe combined features



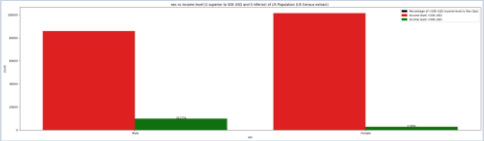
Histogram of age and sex compared to the proportion of high income level (>50K US$)

Combined features observation

The combination of features allows to validate our hypothesis and bring new observations

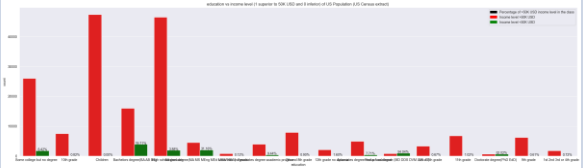
- Male population has an higher income level than female
- Children and young adults are not part of high income level group

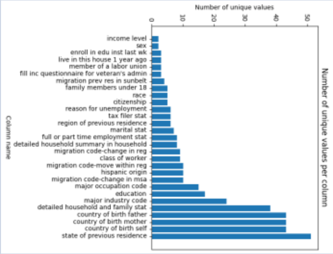
- Peak of high proportion of high income level is reach around age 50
- Peak seems to be the same for men and women

Categorical features	Numerical features	Geographical features
		
Grouped bar chart	Correlation heat map	Colored world map

An exploration to construct our first model

Exploration gives insight to which strategy to operate to reshape features





Categorical data challenge

By regrouping grade categories into degree categories, we keep the insight from the values and we have the possibility to have a better visualisation and transform it into an ordered value via ordinal encoding
One-Hot Encoding becomes a big problem in such a case since we have a separate column for each unique value.
Problem of space consumption, curse of dimensionality and visualisation

Handling high cardinality

- Regrouping categories and order groups
- Regrouping rare categories into on category
- Regrouping non-relevant categories and simplify the categories

An exploration to construct our first model

Fix and select features allows to unclutter the input for the model

Feature scaling

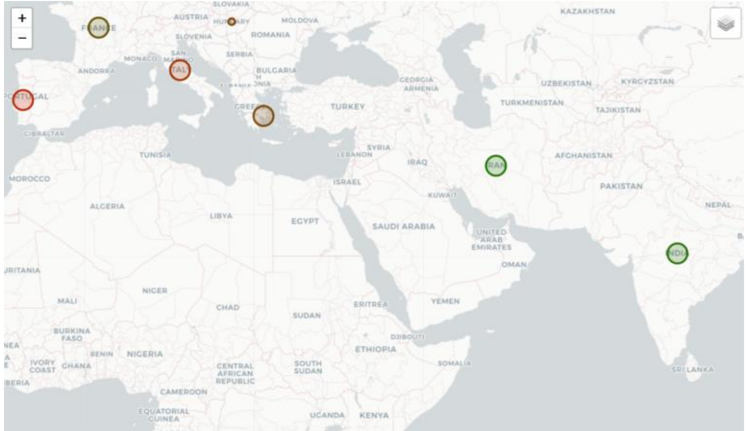
Standardisation of numerical features to avoid the lack of performance due to different scale.
It does not bound values to a specific range.

Useless features detection

By a manual observation of the graph we can identify pointless features

Geographical features handling

Some features require an other way to be explored and transformed



Plotting geographical features on a map can be useful for visualizing and analyzing spatial patterns and relationships in the data.

This interactive map show the country of birth vs the income level (greener the circle is higher is the income level)

We can observe different regions with some patterns in the income level underneath the quantity of data

Defining the right metrics to assess the model

By focusing on the right metrics, comparison between models will be possible

Cross validation: Split the training set into smaller training set – 3 distinct subsets, picking a different one for evaluation every time

Precision: Determine when the cost of wrongly identified low income level person is high

Recall: when there is a high cost associated with wrongly identified high income level person

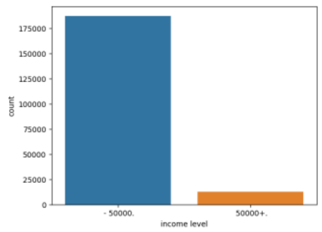
F1 Score
Favors classifiers that have **similar** precision and recall

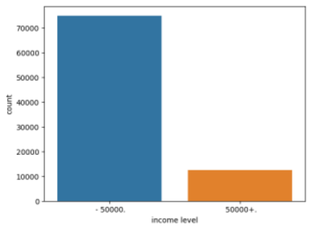
Accuracy fails on classification problems with a skewed class distribution

Modelling to determine the best model

It is important to have the best performing model in order to accurately analyse the feature importance.

To resolve the imbalanced set challenge, we resampled the dataset by under sampling the majority class by 0.4 ratio





We trained a SGD Classifier and a Random forest on both sets

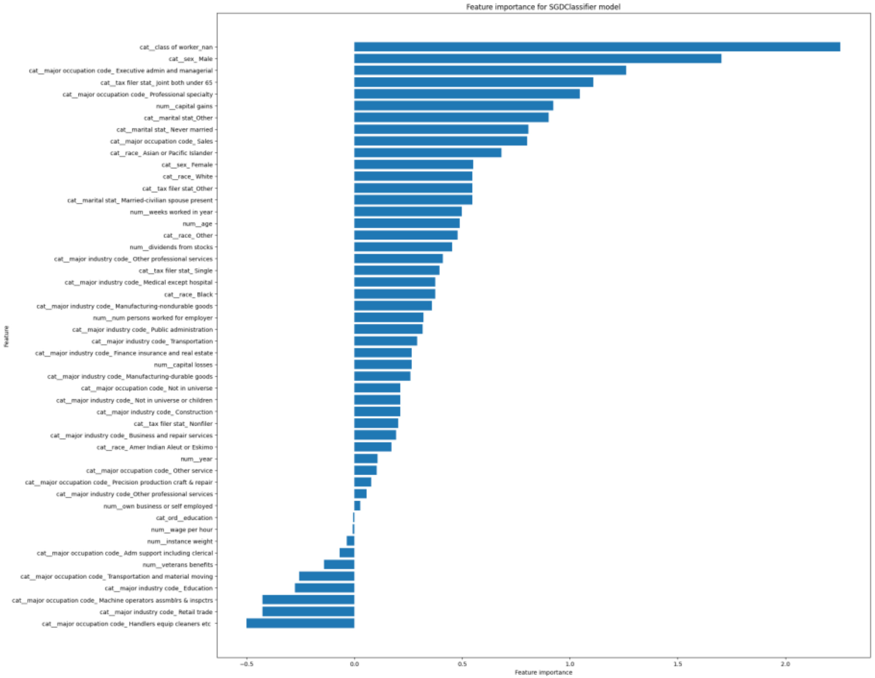
Our benchmark:

	Name	Accuracy	Precision	Recall	F1
0	Random forest trained on 0.4 balanced set	0.945249	0.882826	0.132072	0.227313
1	SGD Classifier trained on 0.4 balanced set	0.945991	0.770869	0.194633	0.303289
2	SGD Classifier trained on imbalanced set	0.945821	0.800649	0.170385	0.279745
3	Random forest trained on imbalanced set	0.945460	0.883404	0.132396	0.226812

After an observation of the results based on the F1 Score, the SGD Classifier trained on 0.4 balanced dataset is the most performing model

Analysis of the feature importance

The graph resulting from the model shows the impact of each class and numerical feature in the prediction



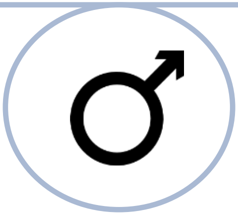
Top 5 features for **high** income level
Category: **Male**
Category: **Executive administration and managerial positions**
Category: **Joint and under 65 year old box in the tax filing**
Category: **Professional specialty**
Numerical: **High Capital gains**

Top 5 features for **low** income level
Numerical: **Low Capital gains**
Numerical: **Low Age**
Category: **Handlers / Cleaners**
Category: **Retail trade**
Category: **Machine assembler/Inspector**

The feature importance helps to define typical profile

The model allows to have a the typical profile according the income

Typical profiling of high income earner



Male

Executives

High Capital gains

Asian ethnicity

Typical profiling of low income earner

Working in Retail

Cleaners/Handlers

Low Capital gains

Young adult

Conclusion

Characteristics identified could be used for business application however risks need to be assesed

Business application

- Consumer profiling in Marketing
- Sociological studies
- Tax and government plans

Risks

- Etnicity data
- Trustfulness in the model
- Bias
- Survey data source

Next steps

- Dive deeper into the features
- Combine dataset
- Error analysis of the model
- Business application

THANKS

www.valentintassel.com