Classification of Income Statistics

Abstract

A detailed case study of census income, to predict if an individual's annual income exceeds $50k. The adult dataset used is from the Census database. In our project, we worked on this Census data set. In our initial stages, we pre-processed the data and develop understanding of the data and its useful features that explain the variances by doing various types of exploratory analysis. Later, we moved on to a classification task of predicting whether the income is >=50k/year from a person's attributes, by using important features. We grouped the data based on different attributes and came to conclusions. We have evaluated and compared various supervised machine learning methods such as Naïve Bayes, Decision Tree, MLP with Back propagation, KNN, Logistic Regression, AdaBoost Classifier, Random Forest Classifier, Extra Trees Classifier, Gradient Boosting Classifier, XGB Classifier, Support vector machines as well as LGBM Classifier. We classified the data based on their income and also provide the count of various people in each sector.

Keywords – Income, Classification, Decision tree, GB, Logistic Regression, SVC, KNN, MLP

I. INTRODUCTION

With the changing economic trends in the present generation, we can see a huge demand for Loans and Credit systems.

Nowadays, people prefer loaning from the bank rather than Mortgage loan. In this process, sometimes the banks face fraud by the Borrowers at the time of repayment.

So, to overcome these type of fraud cases, we have come up with a Machine learning project in which we can classify a borrower's annual income based on the attributes given by the borrowers and by doing so, we can flag the borrowers that if they'll be able to repay the amount or not.

The flagged borrowers will be kept under surveillance for a regular check of their payments.

II. BACKGROUND

The adult dataset used is from the Census database. It is also known as "Census Income" dataset. Details of this dataset can be found at UCI Machine Learning Repository.
Also in this project, we have used Python modules like Scikit learn and other libraries like XBG, LGBM, etc.
The code is written, implemented and was worked on Jupyter Notebooks.

In Machine Learning, **Classification** is a supervised learning concept which basically categorizes a set of data into classes. The most common classification problems are – speech recognition, face detection, handwriting recognition, document classification, etc. It can be either a binary classification problem or a multi-class problem too. There are a bunch of machine learning algorithms for classification in machine learning. There are mainly two types of learners in Classification:

Lazy Learners – Lazy learners simply store the training data and wait until a testing data appears. The classification is done using the most related data in the stored training data. They have more predicting time compared to eager learners. Examples – K-nearest neighbours, case-based reasoning.

Eager Learners – Eager learners construct a classification model based on the given training data before getting data for predictions. It must be able to commit to a single hypothesis that will work for the entire space. Due to this, they take a lot of time in training and less time for a prediction. Examples – Decision Tree, Naive Bayes, Neural Networks.

III. Literature Survey

[1]. This paper attempts to investigate the co-integration relationship between consumption, income and GDP per capita in time-series cross-section data. The authors have applied tests to verify if the time series are non-stationary and co-integrated. The study found that, in general, there is an association between consumption and income. Study confirms that the psychological law stated by Keynes, according to which as the level of income increases, the difference between income and consumption increases as well, is validated by the empirical evidences.

[2]. They have used linear discriminant, this classification rule is based on an index called the linear discriminant function which provides a method for such analysis. The linear discriminant function is used to identify characteristics that distinguish between communities in Arkansas in which per capita incomes are growing rapidly and those in which incomes are growing more slowly. The same set of variables used to account for differences between slow- and fast-growing cities in Arkansas is applied to Oklahoma to test the validity of the model. The discriminant function presented here suggests that percentage increase in per capita income is associated negatively with base year income and city size; it is associated positively with the proportion of the county population living on farms, the school dropout rate, and median educational levels.

[3]. By means of an unsupervised classification, such as the clustering data analysis, it is proposed to examine whether there are natural groupings of cities the basic aspects mentioned. For this, we used the hierarchical method WARD and the non-hierarchical k-means method, with the criteria of validation width silhouette (SWC) and sum of squared errors (SSE) to find groups of municipalities in three basic aspects of development.

[4]. Employing data from the China rural–urban mobility survey conducted in2010, this study investigates the influence of family demographic characteristics on the income, life satisfaction, and potential for rural–urban mobility at the rural household level of two provinces of China: Shaanxi and Henan. They conclude that a larger family size does not translate to more benefits for a rural household. Family size preference is found to be a reflection of parents concerns about elderly care and is deemed to be unfavorable for urbanization in P. R. China.

[5]. The US Census Bureau conducts the American Community Survey generating a massive dataset with millions of data points. The rich dataset contains detailed information of approximately 3.5 million households in regard to who they are and how they live including

ancestry, education, work, transportation, internet use and so on. This enormous data encourages the need to know more about the population and to derive insights.

[6]. The ever-demanding requirement in exposing the subtlety in case of economic issues is the motivation behind to construe meaningful conclusions in income domain. Hence the focus is to concentrate on bringing out unique insights into the financial status of the people living in the country. These conclusions delineated might aid in delivering wiser decisions in regard to economic growth of the country.

[7]. Using relevant attributes, demographic graphs are plotted aiding the conclusions drawn. Also, classifications into various economic classes are done using well known classifiers.

[8]. In this work they extend deep learning into a new application domain – namely classification on mobile phone datasets. We implement a simple deep learning architecture and compare it with traditional data mining models as our benchmarks. Using only a single dimension of the data in its raw form, achieves a 7% better performance compared to the best traditional data mining approach based on custom engineered features from multiple data dimensions.

[9]. This paper shows how the reduction of income inequality through tax policy affects economic growth. Using US state-level data and micro-level household tax returns over the last three decades, they find that reducing income inequality between low- and median-income households improves economic growth. reducing income inequality through taxation between median and high-income households reduces economic growth. These asymmetric economic growth effects are attributable both to supply-side factors and to consumption demand.

[10]. This article has analyzed the full set of labor market VSL studies to measure the income elasticity of the VSL. Their results demonstrate that the income elasticity ranges for the United States are not consistent with the income elasticities in non-U.S. countries. That international estimates of the income elasticity may be greater is borneout in our quantile regression analysis, which demonstrated that the VSL incomeelasticity falls as the VSL level increases.

[11]. In this study Random Forest Classifier machine learning algorithm is applied to predict income levels of individuals based on attributes including education, marital status, gender, occupation, country and others. Income levels are defined as a binary variable 0 for income <=50K/year and 1 for higher levels. .The data is acquired from UCI Machine Learning Repository and includes 32,561 individual's data on 13 attributes based on 1994 census

database. Random forest classifier is used since it gave better accuracy compared to decision tree classifier and naïve bayes classifier. Important features prediction shows marital status, capital gain, education, age and hours per week are the top features which account for larger shares of the model accuracy.

[12]. In this paper, the dataset is used to analyze and categorize the customer based on their purchase behavior. The classification is performed by SVM algorithm. The inventory data set and sales data set which is available in the internet is used in this work and the performance is evaluated by using the algorithms. The experimental results are analyzed and it shows that the proposed methodology analyze a customer behavior in a better way.

[13]. In this paper, they propose a new way to calculate the similarity matrix used by a Laplacian score in order to perform the selection. Laplacian score used to select the most relevant income (input) indicators for Middle East countries, has shown good classification performances of those countries, while reducing their input indicator space. Results show the interest of the proposed approach for indicator selection to perform classification of those Middle East countries

[14]. The present paper attempts to examine the discrepancies in results that may arise due to differential classification of income levels. For the purpose of the study, analysis of a fieldwork-based data set of 708 individuals (20-70 years) from 470 households based on different income classifications and various socio-demographic, behavioural and nutritional indicators were used. It is observed that varying classification of income levels may not influence the interpretation of results in case of discrete or continuous variables, however, it is found to skew the direction of interpretation of results to a considerable extent in categorical variables.

[13]. It has been demonstrated that socioeconomic factors affect health through material and psychosocial pathways. The income inequality was calculated on the basis of self-reported income. The special requirements for complex survey data analysis were considered in the bivariate analysis and linear regression models. Income inequality has damaging effects on HRQL in Shaanxi, China, especially for people with low income. In addition, people living in rural regions were more vulnerable to economic factors.

[14]. The author has tried to predict each adult's income potential and classify them according to the quantitative attributes. Deleting incomplete data, normalize and re-arranging both can help improve the effect and effectiveness but re-arranging need to depend on specific

dataset and attribute's meaning. Two detailed categorical attributes are re-arranged but there are actually more

[15]. It is concluded from the given analysis that an average surveyed household earns annually Rs. 96199 in rural Uttar Pradesh. Farm business income is the most important component of household income. An average sampled farm household earns per capita income of Rs. 17547 annually. The study reveals a positive relationship between farm-size and income levels, i.e., as the farm-size increases, the average income of the households will also increase

[16]. Their study adds to the growing body of literature on the relationship between SES and smoking behaviors, and found discrepancies in the relationship between multiple measures of SES and smoking behaviors. Although we detected a small or insignificant relationship between income and smoking behaviors, a significant gradient in multiple measures of smoking behaviors by occupation and education were observed among the middle aged and elderly Chinese population.

[17]. This study compared the use of the income- and asset-based measures to determine the poverty status of households in a South African Township.The income-based poverty was measured using the Household Subsistence Level (HSL); while Principal Component Analysis (PCA) was applied to determine the asset-based poverty status. The Analysis of Variance (ANOVA) was used to assess whether there is a significant difference between the results of these two measures of poverty. This study concludes that, in the absence of the income, the asset index can be used as measurement of poverty in low-income areas.

[18]. This paper provides estimates of Gini coefficients for each of the 31 Chinese provinces for the years 2000-2012. The estimates bring out the extremely high-income inequality to be found in nearly half of the provinces in sharp contrast to the extremely low-income inequality to be found in the other half. The country seems to be sharply divided into two extremes when it comes to considering the extent of income inequality within its provinces. Policy implications of the findings are considered.

[19]. Here they used "Random Forest" classifier to predict "Income Level" based on various attributes. they tested their Random Forest model on two types of datasets which are without dimensionality reduced and dimensionality reduced respectively. We observed that we got higher accuracy on the dataset which is not dimensionality reduced than the dataset which is dimensionality reduced.

[20]. The government promulgated the compliance policy of net contract vehicles, which raised the cost of drivers. In order to preserve the capacity of the platform, it is necessary to

ensure the income of drivers. In order to determine the reasons for the low driver's income, this paper analyzes the driver's attributes, driver's behavior, passenger's behavior and platform factors, screens the characteristic indicators, and uses decision tree and logistic regression machine learning algorithm to judge the important influencing factors and their influence areas, so as to provide support for the follow-up strategy of enterprises.

[21]. This paper aims to empirically test the hypothesis for ten middle income Sub-Saharan African countries over the period of 1971–2012. The study utilizes panel unit root, panel cointegration and panel dynamic ordinary least squares technique for empirical analysis. Empirical findings suggest that it is possible for Sub-Saharan African countries to continue their efforts to spur economic growth while minimizing environmental damage at the same time.

[22]. This study was conducted to fill in the gap of income inequality and economic growth on two different samples for the period from 1960 to 2014: (i) a full sample of 158 countries; and (ii) a sample of 86 middle-income countries. The Granger causality test and a system generalized method of moments (GMM) are utilized in this study. The findings from this study indicate that causality is found from economic growth to income inequality and vice versa in both samples of countries. In addition, this study also finds that income inequality contributes negatively to the economic growth in the middle-income countries in the research period.

[23]. The aim of the paper is to compare personal income distributions in selected countries of the European Union, taking into account gender differences. They examined the income inequalities between men and women in each country using the Oaxaca-Blinder decomposition procedure. They extended the decomposition procedure to different quantile points along the whole income distribution. They found that there exists an important diversity in the size of the gender pay gap across members of the European Union. The results obtained for these countries allowed us to group them into clusters.

[24]. This study aims to show the usage of machine learning and data mining techniques in providing a solution to the income equality problem. The Gradient Boosting Classifier Model was deployed which clocked the highest accuracy of 88.16%, eventually breaking the benchmark accuracy of existing works.

[25]. This paper analyzes the relationship between the taxi driving mode and the driver's income based on the GPS trajectory data of 10,000 taxis in Chengdu. We first extract and clean the GPS positioning data to obtain the data set of effective trips. Based on the data analysis, the revenue data are identified by high/low-income groups, and the indicators with obvious differences between the two groups are analyzed. Next, a decision tree model is established

based on these indicators to classify drivers. The accuracy of the classification rules is then verified and operational advices for improving drivers' incomes are provided.

[26]. The authors examine reasons for declining gender inequalities, and most notably concentrate on explanations for the closing gender gap in low-wage employment risks. Based on regression techniques and decomposition analyses (1996-2016), the authors find significantly decreasing labor market risks for the female workforce. Detailed analysis reveals that the concrete positioning in the labor market shows greater importance in explaining declining gender differences compared to personal characteristics. The changed composition of the labor markets has prevented the low-wage sector from increasing even more in general and works in favor of the female workforce and their low-wage employment risks in particular.

[27]. They study the income distribution for India from 2014 to 2019 and find that while income inequality remains generally consistent through this period, income shares at the very bottom of the distribution decline substantially. this fragility of incomes lower in the distribution and decline in real incomes at the bottom is reflective of broader economic trends including informalization of the formal workforce, and agrarian distress. The design of specific policies bearing upon incomes of marginal farmers and wage labourers is therefore an area that requires immediate attention.

[28]. Based on influence function regression methods, this paper explores the potential consequences in the labour income distribution related to a long-lasting increase in WFH feasibility among Italian employees. Results show that a positive shift in WFH feasibility would be associated with an increase in average labour income.

[29]. They have investigated the key contributors (electricity consumption, foreign direct investment, carbon dioxide emissions, and population) of economic growth in Africa, and clustered the selected countries into their income levels spanning from 1990 to 2018. These findings imply that electricity usage and economic growth are highly corrected.

[30]. In this paper, the American Time Use Survey (ATUS) Eating & Health Module Files from 2014 survey is used to predict the BMI of people based on their income and through it. To do this analysis different machine learning algorithms were used and finally a comparison of all the algorithms are done with the help of ROC curve. They have tried to establish a relationship between overweight/obese likely based on the income of the family.

IV. DATASET DESCRIPTION & SAMPLE DATA

The adult dataset used is from the Census database. It is also known as "Census Income" dataset. Details of this dataset can be found at UCI Machine Learning Repository.

Source:

Donor:

Ronny Kohavi and Barry Becker - Data Mining and Visualization Silicon Graphics.

Attribute Information:

Listing of attributes*:* >50K, <=50K.

age*:* continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: Built with a certain formula which takes different parameters. Made for governmental use.

education*:* Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num*:* continuous.

marital-status*:* Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation*:* Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship*:* Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race*:* White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex*:* Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Non United States

age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

V. PROPOSED ALGORITHM WITH FLOWCHART

Reading Dataset

- We will import all the necessary python modules
- Then we will use .read_csv() from pandas to read our dataset

Exploratory Data Analysis

- Here we have plotted various graphs for all attributes of our dataset to identify data imbalance and to obtain rough margins.

Finding and Handling Missing Data

- During the analysis, we have checked if there is any missing value which is not 'nan' and then we have converted them to 'nan'.
- Then to handle the null values, we have used backward fill and mode of the features.

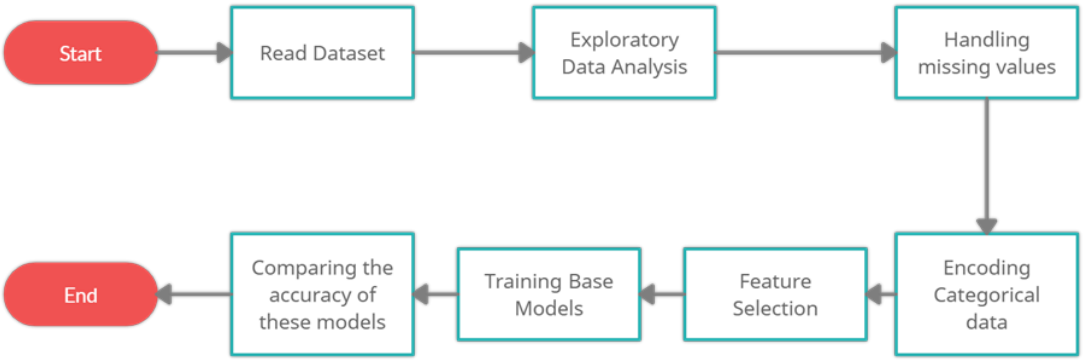
Feature Selection:

- Here, we have used Pearson Correlation to check correlation of features with the target column i.e. Income.
- And then we have removed the columns that were not very useful.
- We have also removed redundant columns which showed multicollinearity.

Training Base Models:

- We have trained the following models using the training set:
    - Logistic Regression
    - Naïve Bayes
    - Decision Tree
    - K-Nearest Neighbours
    - Multi-Layer Perceptron with Back Propagation
    - Support Vector Classifier
    - AdaBoost classifier
    - ExtraTrees classifier

- Gradient Boosting classifier
- Random Forest Classifier
- eXtreme Gradient Boosting Classifier
- Light Gradient Boosting Machine



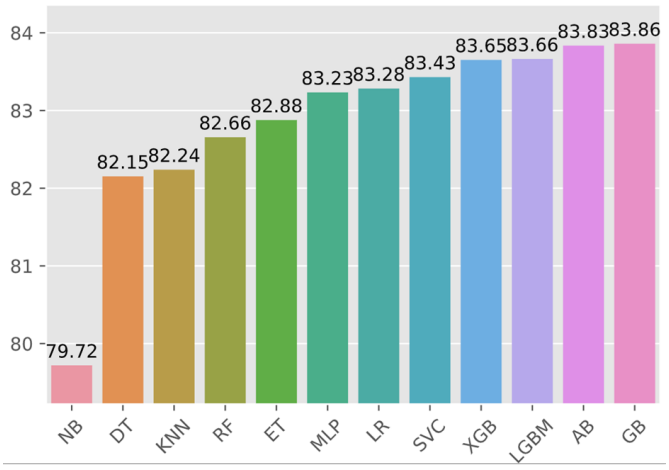
VI. EXPERIMENT RESULTS

We have tested our models using the test sets and based on the results we have compared which model better suits our needs for our purpose. So the results are as followed:

1. Naïve Bayes: 79.72%
2. Decision Tree: 82.15%
3. K-Nearest Neighbors: 82.24%
4. Random Forest Classifier: 82.66%
5. ExtraTrees Classifier: 82.88%
6. MLP: 83.23%
7. LR: 83.28%
8. SVC: 83.43%
9. XGB: 83.65%
10. LGBM: 83.66%
11. AB: 83.83%
12. GB: 83.86%

We can see that **Gradient Boosting** is best suited for our purposes and hence we can use it to implement a fraud detection system for various banks and financial institutions.

VII. COMPARATIVE STUDY / RESULTS AND DISCUSSION



Model	Accuracy
Naïve Bayes	79.72
Decision Tree	82.15
K Nearest Neighbors	82.24
Random Forest	82.66
ExtraTrees	82.88
Multi-Layer Perceptron	83.23
Logistic Regression	83.28
Support Vector Classifier	83.43
eXtreme Gradient Boosting	83.65
Light Gradient Boosting Machine	83.66
Adaptive Boosting	83.83
Gradient Boosting	83.86

Here, Naïve Bayes model shows the least accuracy. This is due to the the fact that Naïve Bayes works by taking the attributes as an individual – which in real life, is not always a suitable case.

Decision trees can create over-complex trees causing overfitting. That is why accuracy isn't that high; while this issue is fixed in random forest due to generation of multiple trees. Hence it's accuracy is higher than decision tree. Then we have extra randomized tree which provides much greater accuracy than random forest when dataset is noisy, also it is extremely fast and that is why its the best among its peers.

Lastly, Boosting Algorithms showed the best outcome due to the fact that they make a large number of trees which have their own weight – showing their importance in the final outcome. Moreover, it tests different attributes considering different thresholds and thus takes out the best form of its own.

VIII. CONCLUSION AND FUTURE WORK

At last, we can conclude that among all these classifiers, gradient boosting is best suited for our purpose with an accuracy of 83.86, hence it can be successfully used in prediction of fraud cases in financial background, future works may include implementation of similar systems in finding the relevance of education with respect to income, so that we can spread awareness about the education system and its importance. We can classify the different income groups on basis of education backgrounds.

Also the future works may include creation of newer enhanced algorithms which may further improve Gradient Boosting, SVC, Tree based models and may also combine their advantages to make a better appropriate model.

IX. REFERENCES

1. The Relationship between Income, Consumption and GDP: A Time Series, Cross-Country Analysis by Paula-Elena Diacona, Liviu-George Mahab, April 2015
2. An Analysis of Rates of Change in Community Per Capita Income by Discriminant Analysis by Steve Murray, July 2015
3. Socioeconomic Class Of Brazilian Cities For Health, Education And Employment & Income IFDM: A Clustering Data Analysis by N. J. Martarelli and M. S. Nagano, March 2016
4. Geographical Mobility, Income, Life Satisfaction and Family Size Preferences: An Empirical Study on Rural Households in Shaanxi and Henan Provinces in China by Jiangsheng Chen, Hong Yang, October 2016
5. The Data Analytics to predict the Income and Economic Hierarchy on Census Data by Sharath R, Krishna Chaitanya S, Nirupam K. N, Sowmya B J, October 2016
6. Deep Learning Applied to Mobile Phone Data for Individual Income Classification by Pål Sundsøy, Johannes Bjelland, Bjørn-Atle Reme, Asif M.Iqbal and Eaman Jahani, January 2016
7. Income Inequality, Tax Policy, and Economic Growth by Siddhartha Biswas, Indraneel Chakraborty, May 2017
8. Income Elasticities and Global Values of a Statistical Life, by W. Kip Viscusi, Clayton J. Masterman, July 2017
9. Using decision tree classifier to predict income levels by Bekena, Sisay Menji, July 2017
10. Predicting customer behavior in online shopping using SVM classifier by Dr.K.Maheswari, P.Packia Amutha Priya, March 2017
11. Selection of Income Indicators for Middle East Country Classification by - Ali Kalakech and Mariam Kalakech, May 2017
12. The Dilemma of Classification of Income Levels in Social Research, by Jenny Jami, August 2018
13. Household income, income inequality, and health-related quality of life measured by the EQ-5D in Shaanxi, China: a cross-sectional study by Zhijun Tan, Fuyan Shi, Haiyue Zhang, Ning Li, Yongyong Xu, and Ying Liang, March 2018
14. Classifying Income Potential From The Adult Dataset: Comparing Different Understanding And Pre-processing Data, by Hongzheng He, December 2018
15. Rural household income patterns in uttar pradesh: primary data by Manoj Kumar Sharma, KS Rao, BVS Sisodia, Sandhya Verma, June 2018

16. Income, occupation and education: Are they related to smoking behaviours in China? by Qing Wang, Jay J.Shen, Michelle Sotero, Casey A. Li, Zhiyuan Hou, February 2018

17. A comparative analysis of income- and asset-based poverty measures of households in a township in South Africa by - Paul-Francois Muzindutsi, December 2018

18. On Income Inequality within China's Provinces: by - Prabir Bhattacharya, Javier Palacio-Torralba, Xinrong Li, May 2018

19. Predict the Level of Income using Random Forest Classifier by Tejas Phase, Suhas Patil, December 2019

20. Analysis of Income Attribution of Ride-hailing Drivers Based on Machine Learning by Shuaisa Sun, November 2019

21. Income heterogeneity and the Environmental Kuznets Curve hypothesis in Sub-Saharan African countries by - Muhammad Maladoh Bah, Muhammad Mansur Abdulwakil, Muhammad Azam, February 2019

22. What Factors Affect Income Inequality and Economic Growth in Middle-Income Countries? by Duc Hong Vo, Thang Cong Nguyen, Ngoc Phu Tran and Anh The Vo, February, 2019

23. Differences in Income Distributions for Men and Women in the European Union Countries by Joanna Małgorzata Landmesser, February 2019

24. A Statistical Approach to Adult Census Income Level Prediction by Navoneel Chakrabarty, Sanket Biswas, July 2019

25. based analysis on the relationship between taxi travelling patterns and taxi drivers' incomes by Guangxin Ou; Youkai Wu, Gangqing Wang, Zhaoxia Guo, October 2019

26. Declining Gender Differences in Low-Wage Employment in Germany, Austria and Switzerland, by Nina-Sophie Fritsch, Roland Verwiebe, Bernd Liedl, October 2019

27. Income distribution and inequality in India: 2014-19 by Anand Sahasranman, Nishanth Kumar, October 2020

28. Working from home and income inequality: risks of a 'new normal' with COVID-19 by Luca Bonacini, Giovanni Gallo & Sergio Scicchitano, September 2020

29. Investigation on the main contributors of economic growth in a dynamic heterogeneous panel data (DHPD) in Africa: evidence from their income classification by Olivier Joseph Abban & Yao Hongxing, January 2021

30. Impact of Income Level of an Individual on his BMI and Performance Analysis using Various Machine Learning Approaches on ATUS Survey 2014–16 by Neha Singh, Sinu Mathew, Neha Kunte, January 2021

Appendix

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings

warnings.filterwarnings("ignore")
plt.style.use('ggplot')

from imblearn.over_sampling import RandomOverSampler
from sklearn.model_selection import train_test_split
from numpy import mean, std

df = pd.read_csv("adult.csv")
df.head()

sns.countplot(df.income)
sns.distplot(df[df.income==">50K"].age, color='g')
sns.distplot(df[df.income=="<=50K"].age, color='r')
plt.xticks(rotation=90)
sns.countplot(df.workclass, hue=df.income, palette='tab10')
sns.distplot(df[df.income=='<=50K'].fnlwgt, color='r')
sns.distplot(df[df.income=='>50K'].fnlwgt, color='g')
plt.xticks(rotation=90)
sns.countplot(df.education, hue=df.income, palette='muted')
sns.countplot(df["education.num"], hue=df.income)
plt.xticks(rotation=90)
sns.countplot(df['marital.status'], hue=df.income)
plt.xticks(rotation=90)
sns.countplot(df.occupation, hue=df.income, palette='rocket')
```

```
plt.xticks(rotation=90)
sns.countplot(df.relationship, hue=df.income, palette='muted')
plt.xticks(rotation=90)
sns.countplot(df.race, hue=df.income, palette='Set2')
plt.xticks(rotation=90)
sns.countplot(df.sex, hue=df.income)
df['capital.gain'].value_counts()
df['capital.loss'].value_counts()
sns.distplot(df[df.income=='<=50K']['hours.per.week'], color='b')
sns.distplot(df[df.income=='>50K']['hours.per.week'], color='r')
df['native.country'].value_counts()

df[df.select_dtypes("object") =="?"] = np.nan
nans = df.isnull().sum()
if len(nans[nans>0]):
    print("Missing values detected.\n")
    print(nans[nans>0])
else:
    print("No missing values. You are good to go.")



df.workclass.fillna("Private", inplace=True)
df.occupation.fillna(method='bfill', inplace=True)
df['native.country'].fillna("United-States", inplace=True)
print("Handled missing values successfully.")

from sklearn.preprocessing import LabelEncoder
from sklearn.utils import column_or_1d
class MyLabelEncoder(LabelEncoder):
    def fit(self, y, arr=[]):
        y = column_or_1d(y, warn=True)
        if arr == []:
            arr=y
        self.classes_ = pd.Series(arr).unique()
```

```
        return self
le = MyLabelEncoder()


df['age_enc'] = df.age.apply(lambda x: 1 if x > 30 else 0)
def prep_workclass(x):
    if x == 'Never-worked' or x == 'Without-pay':
        return 0
    elif x == 'Private':
        return 1
    elif x == 'State-gov' or x == 'Local-gov' or x == 'Federal-gov':
        return 2
    elif x == 'Self-emp-not-inc':
        return 3
    else:
        return 4
df['workclass_enc'] = df.workclass.apply(prep_workclass)
df['fnlwgt_enc'] = df.fnlwgt.apply(lambda x: 0 if x>200000 else 1)
le.fit(df.education, arr=['Preschool', '1st-4th', '5th-6th', '7th-8th', '9th','10th', '11th',
'12th',
                                'HS-grad', 'Prof-school', 'Assoc-acdm', 'Assoc-voc',
'Some-college', 'Bachelors', 'Masters', 'Doctorate'])
df['education_enc'] = le.transform(df.education)
df['education.num_enc'] = df['education.num'].apply(lambda x: 1 if x>=9 else 0)
df['marital.status_enc'] = df['marital.status'].apply(lambda x: 1 if x=='Married-civ-
spouse' or x == 'Married-AF-spouse' else 0)
def prep_occupation(x):
    if x in ['Prof-specialty', 'Exec-managerial']:
        return 2
    elif x in ['Sales', 'Craft-repair']:
        return 1
    else:
        return 0
df['occupation_enc'] = df.occupation.apply(prep_occupation)
```

```
df['relationship_enc'] = df.relationship.apply(lambda x: 1 if x in ['Husband', 'Wife'] else 0)
df['race_enc'] = df.race.apply(lambda x: 1 if x=='White' else 0)
df['sex_enc'] = df.sex.apply(lambda x: 1 if x=='Male' else 0)
df['capital.gain_enc'] = pd.cut(df["capital.gain"], bins=[-1,0,df[df["capital.gain"]>0]["capital.gain"].median(), df["capital.gain"].max()], labels=(0,1,2)).astype('int64')
df['capital.loss_enc'] = pd.cut(df["capital.loss"], bins=[-1,0,df[df["capital.loss"]>0]["capital.loss"].median(), df["capital.loss"].max()], labels=(0,1,2)).astype('int64')
df['hours.per.week_enc'] = pd.qcut(df['hours.per.week'], q=5, labels=(0,1,2,3), duplicates='drop').astype('int64')
df['native.country_enc'] = df['native.country'].apply(lambda x: 1 if x=='United-States' else 0)
df['income_enc'] = df.income.apply(lambda x: 1 if x==">50K" else 0)
print("Encoding complete.")

df.select_dtypes("object").info()

df.drop(['education', 'sex', 'income'], 1, inplace=True)

for feature in df.select_dtypes("object").columns:
   df[feature]=le.fit_transform(df[feature])
df.info()

pcorr = df.drop('income_enc',1).corrwith(df.income_enc)
plt.figure(figsize=(10,6))
plt.title("Pearson Correlation of Features with Income")
plt.xlabel("Features")
plt.ylabel("Correlation Coeff")
plt.xticks(rotation=90)
plt.bar(pcorr.index, list(map(abs,pcorr.values)))
```

```
df.drop(['workclass', 'fnlwgt','occupation', 'race', 'native.country', 'fnlwgt_enc', 'race_enc', 'native.country_enc'], 1, inplace=True)

sns.heatmap(df.corr().apply(abs))

df.drop(['age', 'education.num_enc', 'education_enc', 'marital.status_enc', 'capital.gain', 'capital.loss', 'hours.per.week'], 1, inplace = True)

df.info()
X = df.drop('income_enc', 1)
y = df.income_enc

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42, stratify=y)
print("No. of rows in training data:",X_train.shape[0])
print("No. of rows in testing data:",X_test.shape[0])

oversample = RandomOverSampler(sampling_strategy=0.5)
X_over, y_over = oversample.fit_resample(X_train, y_train)
y_over.value_counts()

from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import AdaBoostClassifier, ExtraTreesClassifier, GradientBoostingClassifier, RandomForestClassifier
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from sklearn.neural_network import MLPClassifier

seed = 2**32 -1
```

```
models = [
    ('LR',LogisticRegression(random_state=seed)),
    ('SVC',SVC(random_state=seed)),
    ('AB',AdaBoostClassifier(random_state=seed)),
    ('ET',ExtraTreesClassifier(random_state=seed)),
    ('GB',GradientBoostingClassifier(random_state=seed)),
    ('RF',RandomForestClassifier(random_state=seed)),
    ('XGB',XGBClassifier(random_state=seed,eval_metric='logloss')),
    ('LGBM',LGBMClassifier(random_state=seed)),
    ('KNN',KNeighborsClassifier()),
    ('NB',GaussianNB()),
    ('DT',DecisionTreeClassifier(random_state=seed)),
    ('MLP',MLPClassifier(random_state=seed))
]

name_scores=[]
for model in models:
    model[1].fit(X_over,y_over)
    print(model[0]," trained.")
    name_scores.append((model[0],100*model[1].score(X_test,y_test)))
name_scores.sort(key=lambda x:x[1])
print("Results are ready.")

def plot_scores(name_scores):
    names = [x[0] for x in name_scores]
    scores = [x[1] for x in name_scores]
    plt.ylim(ymax = max(scores)+0.5, ymin = min(scores)-0.5)
    plt.xticks(rotation=45)
    s = sns.barplot(names,scores)
    for x,y in enumerate(scores):
        s.text(x,y+0.1,round(y,2),ha="center")

plot_scores(name_scores)
```