United States Census Income Data

Literature Review

Introduction

Census is the process of systematically recording statistical information about the nation's population. The data is comprised of various people, their distribution, their living conditions, education and other key factors, which are critical for the development of the nation. This data helps the policy makers construct decisions for the future and betterment of the country. Income is one of the primary concerns for the standard of living and economic status of an individual and thus, has a significant impact on determining the nation's growth and prosperity.

In this project, the aim is to explore the U.S census income data set, relating the earnings/income with demographic information such as; age, gender, marital status, race, education level, employment type etc… and to build a classifier which can predict whether the income of an individual is greater than or less than $50K/year. The problem is a binary classification problem, since the target variable income has binary values.

To make these predictions, I will explore the census income data and determine the most remarkable features in the census income data, for predicting the income class of an individual and use these features to build/train models on training data, using different machine learning classification algorithms. The models will be tested on the test data, and the one with the highest accuracy will be selected as the best predictive model.

Research Questions:

1. Is the income of an individual greater than or less than $50K a year?

Literature Review

To obtain a better understanding and insight of the given problem, I reviewed several publications which focus on dealing with binary classification problems and adult census income data. Ron Kohavi in 1996, used income dataset to build a hybrid algorithm NBTree [1], which combines the features of decision-tree and naïve bayes classifiers. In this algorithm, the decision-tree nodes contain univariate splits (segmentation), like decision tree but at the leaves classification is done using naïve Bayesian classifiers. The author is able to prove through his experiments that NBTree works better on real-world datasets,

like adult census data. It outperformed both decision tree and naïve bayes, but takes more time with regards to speed.

In another study, the author used random forest classifier as it gives better performance than decision tree and naïve Bayes. He achieved 85% accuracy with random forest considering marital status, capital gain, education, age and hours per week as the top features. In preprocessing author converted categorial variables into dummy variables, and merged capital gain and capital loss as one column. The author explores that all numerical attributes have a positive correlation with income. The groups with higher income are male, married people, self, or government employees, Whites and Asians, professionals, specialists, technology workers, and managers. The model has a good accuracy on low income but is weak when it comes to predicting high income [2].

In this paper [3], the author presented the effect of dimensionality reduction using; Principal Component analysis for dimensionality reduction on the performance of support vector machine on Adult income data set. The author also emphasized on the consequences in terms of accuracy and computation time, by reducing the input data vertically with regards to the number of training examples and horizontally in terms of the number of features and to find out the ideal ratio between them. They build four data sets: adult_13 dataset was built by removing education number, based on the strong correlation 0.8881 between education and education number. Next, adult_10 was built by removing four variables age, education number, marital status and sex. Third, adult_6 they keep components with eigenvalues greater than 1, and the last dataset adult_8, contains 13 principal components generated by the PCA method. As SVM method deals better with real numbers, they used the scaling method instead of discretization for continuous variables. They reported that the accuracy of adult_8, is the highest among all the datasets. This shows that PCA components capture maximum information and improved the performance of the Support Vector Machine.

In this study [5], the author compares the performance of different classification algorithms based on accuracy, ROC area and precision. They used eight supervised machine learning algorithms: SVMs, neural nets, decision trees, random forest, k-nearest neighbor, bagged trees, boosted trees and boosted stumps on different data sets including Adult income dataset and conducted empirical comparison. The Adult income data set is the only problem that has nominal attributes. The study concludes that bagged trees, random forest and neural nets have the best average performance over all the metrics and dataset. In terms of KNN, it gives phenomenal results, when attributes are weighted by gain ratio instead of unweighted Euclidean distance. Furthermore, for the selection of the best model they used the IK validation test, and then checked the performance on the test set.

After conducting a series of literature reviews, I have decided to use logistic regression, random forest and naïve bayes for this project. It is further supported by the fact that the project of predicting income from census data is a binary classification problem, as the target variable having binary output and data has mixed numerical and nominal attributes.

These are the benchmark algorithms, which are used for dealing with binary classification problems.

Logistic regression is also helpful as its gives the idea of important variables, which will be useful for predicting income, based on p values. Random forest gives better performance as compared to decision trees, in most of the studies. Feature selection will be done based on the correlation between attributes, exploratory analysis, literature review and in the model building step with forward/backward feature elimination techniques. Principle Component analysis (PCA) as used in [4] to reduce the dimensionality will not be used at this stage due to the scope and time limitation. The data used in this project will be the adult income census data extracted from complete census data, as used by most of the researchers.

Data Description

U.S Adult Census Income Data

The dataset used for the analysis is an extraction from the 1994 census data by Barry Becker and downloaded from UCI Machine Learning repository [6].

The data is comprised of 14 attributes and 32561 observations. 8 are categorical and 6 are numerical. Income is the dependence variable/class variable and rest are independent variables. The data attributes, their type, categories and description are explained as follow:

1. **Attribute Name** : Age
   Data Type : integer
   Description : Represents the age of an individual.

2. **Attribute Name** : Workclass
   Data Type : Categorical data (9 Distinct categories)
   Description : It describes the work class of an individual.
   Distinct Categories : 9 Distinct categories are as follow:
   Federal-gov, Local-gov, Never-worked, Private, Self-em, p-inc, Self-emp-not-inc, State-gov, Without-pay

3. **Attribute Name** : Fnlwgt
   Data Type : integer
   Description : Represents final weight. It is the number of units in the target population that the responding unit represents.

4. **Attribute Name** : Education
   Data Type : Categorical (16 Distinct categories)

Distinct Categories : 16 Distinct categories are as follow:
1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate , HS-grad, Masters, Preschool, Prof-school, Some-college.

Description : It describes the education levels and consist of 16 different categories.

5. **Attribute Name :** Education number
   Data Type : integer
   Description : It specifies number of year of educations in total.

6. **Attribute Name :** Marital status
   Data Type : Categorical (7 Distinct categories)
   Distinct Categories : 7 Distinct categories are as follow:
   Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed
   Description : Describes the marital status of an individual.

7. **Attribute Name :** Occupation
   Data Type : Categorical (15 Distinct categories)
   Distinct Categories : 15 Distinct categories are as follow:
   ?, Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, Transport-moving.
   Description : Describes the occupation of an individual.

8. **Attribute Name :** Relationship
   Data Type : Categorical with 5 distinct categories.
   Distinct Categories : 5 Distinct categories are as follow:
   Husband, Not-in-family, Other-relative, Own-child, Unmarried, Wife.
   Description : It represents the role of an individual in a family.

9. **Attribute Name :** race
   Data Type : Categorical with 5 distinct levels.

Distinct Categories: Distinct categories are as follow:
White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

10. **Attribute Name :** Sex
Data Type : factor 2 levels Female, Male

11. **Attribute Name :** Capital gain
Data Type : integer
Description : It represents information gain during a year.

12. **Attribute Name :** Capital loss
Data Type : integer
Description : It represents information losses during a year.

13. **Attribute Name :** Hours.per.week
Data Type : integer
Description : It represents the number of hours worked in a week.

14. **Attribute Name :** Native.country
Data Type : factor (42 levels) Cambodia
Description : It represents the native country of an individual.
Categories/Levels : Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Outlying-US(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad & Tobago, United-States, Vietnam, Yugoslavia.

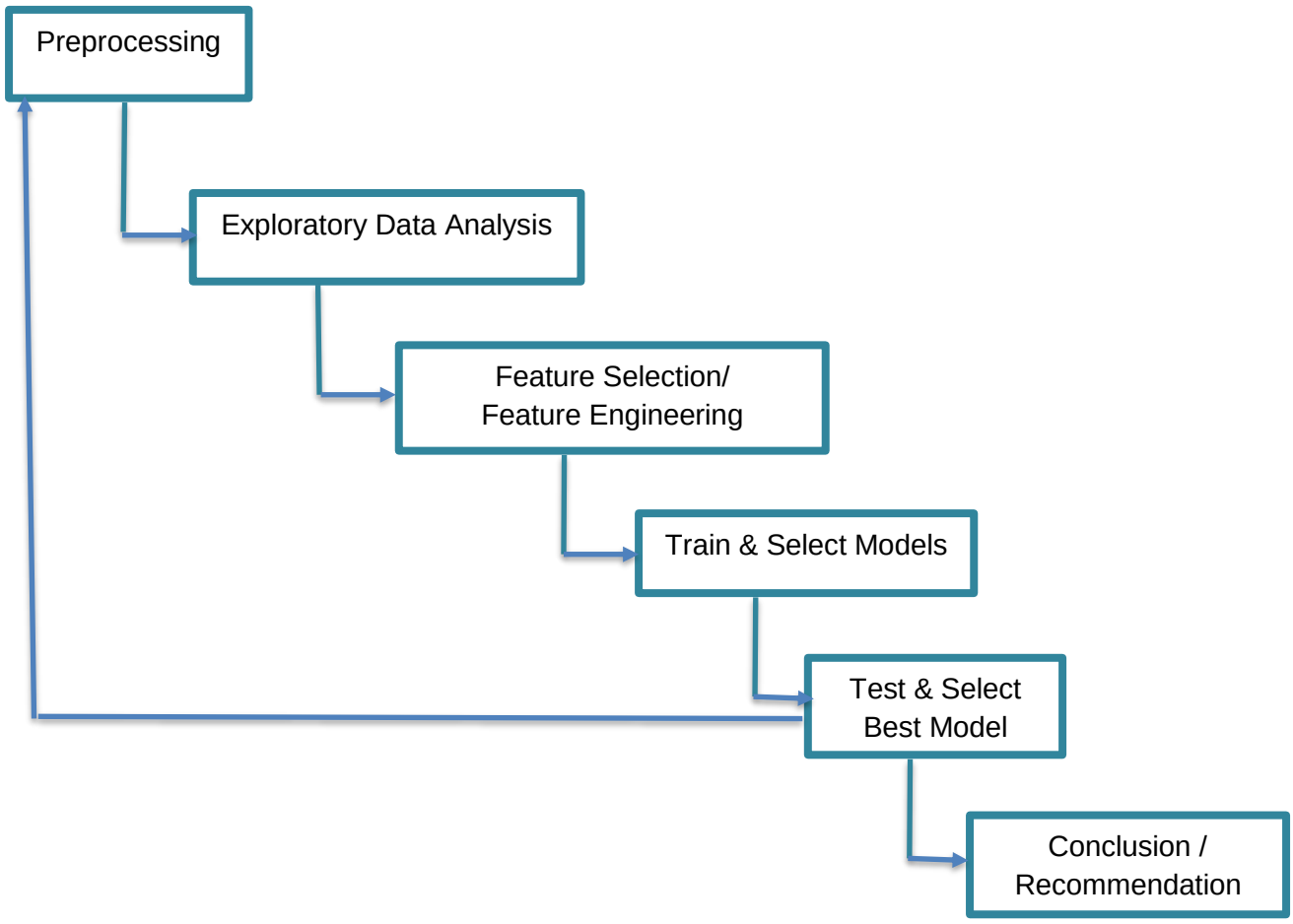
15. **Attribute Name :** Income
Data Type : factor (2 levels)
Categories/Levels : <= 50K, > 50 K
Description : Class Label

Data Statistics:

Statistics of the numerical variables are as follow:

No.	Attribute Name	Minimum	Ist Qu.	Median	Mean	3rd Qu.	Max.	Sd
1.	Age	17.00	28.00	37.00	38.58	48.00	90.00	13.64
2.	Education	1.00	9.00	10.00	10.08	12.00	16.00	2.57
3.	Capital Gain	0	0	0	1078	0	999	7385.29
4.	Capital Loss	0.0	0.0	0.0	87.3	0.0	4356.0	402.96
5.	Hours per week	1.00	40.00	40.00	40.44	45.00	99.00	12.35

Approach



Step 1: Preprocessing

In this step, data is loaded, and the initial cleaning of the raw data is performed. It involves analysis of attributes.

In our dataset we have a total of 16 attributes, in which 6 are numerical. Their statistical information's minimum, maximum value, mean, median and standard deviation was calculated and shown in Table 1. Some of the categorical variables i.e., education, native country and occupation have more than 10 levels with many levels having no or very few values, so I intend to combine those levels and refactorize them.

The data has missing values in three attributes, as stated; work class, occupation and native country (90% values correspond to the United States), which are categorical and represent a very small fraction of whole data. I removed the rows with missing values.

Additionally, in this step I will check for any inconsistence in data, distribution of data (normal or not), low variance attributes, and for outliers in data.

Step 2: Exploratory Data Analysis

This stage involves a detailed analysis of attributes and their relationships with each other, with target variable income. Some of the example queries are as follows:

- Does higher education facilitate in earning more money?
- Does high skill set a guarantee to high income?
- Do education and education number have a strong correlation?
- Does age have any effect on income?
- Does occupation contribute to earning a high income?

I will explore the above stated relations, using different plots for both numerical and categorical attributes. I will also use the correlation matrix for numerical variables to find how closely they are related with the income attribute.

Step 3: Feature Selection/ Feature Engineering

This step involves feature selection and feature engineering, based on the exploratory analysis of data done in previous steps and the classifiers to be used in next step. The attributes that are not important and don't provide any useful information, will be dropped. If needed, some attributes like age will be discretized into bins.

Step 4: Trian and Select Models

In this step, a different feature set selected in previous steps will be used to train models on the train data. The classification algorithms I will use are; Logistic Regression, Random Forest and Naïve Bayes. Different models will be created using significant features based on previous steps. The train data consists of 70% of the original data, and the test data constitutes 30%, of the dataset.

In this stage I will also analyze the most significant attributes for each algorithm. Moreover, it requires quick iterations of the previous steps. Finally, I intend to select 3-4 best models.

Step 5: Test and Select Model

The models built in Step 4 will be used on the test dataset (30%), and their performance will be measured. Performance will be measured based on accuracy, TPR, FNR and ROC. The model with the highest accuracy will be selected as the final model. It sometimes also involves the fine tuning of data and going back to previous steps.

Step 6: Prediction/Conclusion

In this step, we will draw inferences and final conclusions based on the model created and generate feedback from the previous step.

Github Link:

https://github.com/HumairaAsim/CKME136

Bibliography

[1] R. Kohavi, "Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996.

[2] S. M. Bekena, "Using decision tree classifier to predict income levels," MPRA Archive, 2017.

[3] A. Lazar, "Income prediction via support vector machine," in *2004 International Conference on Machine Learning and Applications, 2004. Proceedings.*, Louisville, Kentucky, USA, 2004.

[4] N. K. A. Y. Rich Caruana, "An Empirical Evaluation of Supervised Learning for ROC Area," in *Proceedings of the 25th international conference on Machine learning*, Helsinki, Finland, 2008.

[5] B. Becker, "UCI Machine Learning Repository," [Online]. Available: http://mlr.cs.umass.edu/ml/datasets/Census+Income. [Accessed 2018].