Rosalie Day – Final Report, Capstone 1
Prediction of Income from Education and Sector Status
January, 2020

Problem statement

For the past decade, the conventional wisdom has been jobs that pay a living wage require a college degree. The claim that having a bachelor's degree is necessary for getting a "good" job has been around for decades. In the current discourse, the bachelor degree is the symbolic threshold of being able to adapt to workforce needs in the digital age going forward.

This study is intended to describe education levels and employment sector and status associated with incomes. The analysis explores predictors of whether a person's income was less or greater than $50,000, a selected threshold corresponding to a "good Job" for the 1994 dataset. The focus is on education levels and employment sector and status as predictors. It does not explore race, gender and marital information; although these will be used for classification of income outcomes for the full set of predictor attributes. The data was extracted from the 1994 US Census and appears as the Adult Census Income dataset on the Kaggle website.

This information on both education level and employment sector status used for predicting outcomes inform public policy. Understanding both are critical elements for projecting employment and workforce development needs. These techniques can be used on recent Census data to identify trends of both the relationship of education level to income threshold and the relationship of sector employment, education level and income.

Findings Summary

Counts revealed for education levels (zero grades completed through a doctorate degree), high school graduates comprised the largest level in almost all work class, employment sector and status, categories. The "some college" level was the second largest and the median, followed by bachelor's degrees across work class categories. The three largest work class categories were private sector at 69 percent, dropping far below, public sector at 13 percent, and 11 percent self-employed. The average age in the sample was 38 and hours worked was 40.4 with standard deviation of 12.4. The US accounted for 89.7 percent as country of origin for the sample. For the income variable, 76 percent of individuals were at or below $50,000.

Inferential statistics revealed three results. In a two-sample t-test, a four-year college degree, the bachelor's degree, significantly impacted the income outcome as compared with no bachelor's degree. Work class, which is a combination of work-sector and -status, and the level of education are significantly correlated resulting from chi-squared test. In contrast, hour-per-week worked was independent from education.

Machine learning unsupervised models, Gradient Boosting and Logistic Regression performed the best on the adult income sample based on the complete set of attributes. Both models had an accuracy score of .84. This is the harmonically weighted f1-score of precision (for all positive, the proportion that is correct) and recall (the probability of finding all positive incidences). From these models the area under the receiver operator characteristic ("ROC") curves ("AUC") round to .89. They both identified the attributes to classify as either less than or equal to $50,000, or greater than $50,000 annual income, regardless of classification threshold.

These preferred models on performed almost equally well on the subset focused on education and work class and attributes dropped for privacy. With five fewer predictor variables, the accuracy scores lost .04 for Gradient Boosting and .05 for Logistic Regression, resulting in scores .80 and .79. This performance differential could reflect the small edge in AUC scores that the Gradient Boosting model had.

The Data

The Adult Income data set is an extraction of the 1994 US Census. The data was downloaded in csv file format from the Kaggle website https://www.kaggle.com/uciml/adult-census-income into a Panda's dataframe.

The data set included 48842 samples, composed of mainly categorical variables and no missing values. Notable exceptions to the categorical class were age and hours-per-week. The variables, capital gains, capital loss and final weight variables were excluded as they did not contribute to the intended analysis. The full working data set included: 10 predictor variables, age, "workclass," education/"educational-num," marital status, occupation, relationship (to family), race, gender, hours-per-week, and native country; and the income target variable.

Few modifications were necessary for analysis. The income target variable, which was an object, was converted into the integers 0 and 1 for equal or less than and greater than $50, respectively. There were two education attributes containing the same information: explanatory labeling in string objects; and integers which corresponded to the string values. These levels did not correspond to education years completed. The choice of string or integer was made on a case-by-case basis depending on the use.

The variables were largely self-explanatory except for the previously explained income and work class. The work class variable includes conventional sector information, public, private, and "other" which may be non-profit, and self-employment with incorporation status. Unemployment was not captured in this data set. Race, gender, relationship (familial) and marital information were not explored, although used for classification in the full set of attributes.

Descriptive statistics, distribution and count plots for this sample, included education levels, ages, employment sector and status, hours worked per week and income. The sample ages

ranged from 17 to 99. Other variables, for example, the distribution of country of origin were highly skewed, where 43,832 were listed as from the US. *Please see presentation for highlight graphics and Jupyter notebook for detailed visual data exploration.*

Inferential statistics

This study focuses on exploring education and sector as predictors of whether a person's income is less or greater than $50,000.

The inferential statistical analysis includes hypothesis testing:

1. Outcomes, greater than $50k income, are more likely with bachelor's degrees than any education level short of a bachelor's degree;
2. Work class is associated with education level; and
3. Education is correlated with hours worked.

The first tests the difference in income outcomes between having specifically a bachelor's degree (BA) and no graduate degrees, with not having a bachelor's degree (noBA). Appropriate filters were set to select only the rows of interest: the noBA had a sample size of 36372; and the BA, 8025. This hypothesis test uses a two-sample t-test from the scipy stats library with parameters set for equal variances.

The null hypothesis posits that there is no differential effect on the income outcome with respect to having a bachelor's degree; it is the same without a BA and with only a BA. The results were a significant t-statistic, -52.4087 and a p-value of 0.0. From this comparison of income outcomes, the probability is zero that the 2 samples, BA and noBA, are alike, therefore rejecting the null hypothesis. The alternative hypothesis that a bachelor's degree impacts the income outcome holds.

Second, testing the association of the categorical variables, education and workclass, will use the chi-squared test. The null hypothesis is that education and workclass are independent attributes. A cross-tabulation of education and workclass was created from crosstab function in pandas. The chi-squared statistic is 3906.1080 with 120 degrees of freedom and p-value of 0.0. Therefore, the null hypothesis that these two variables are independent can be rejected. Work class, which is a combination of work-sector and -status, and level of education are correlated.

Third, if the hours worked were correlated positively with education level, that relationship could affect the income outcome. Testing for correlation of education level and hours worked per week is implemented with correlation pairs of the researchpy library. The null hypothesis, that education (specifically 'educ_num') and 'hours-per-week' are independent variables, cannot be rejected with the very small r value of 0.1437. There appears to be no significant correlation between education levels and hours worked per week.

Classification Prediction

The predictor variables were selected and data was preprocessed with the Pandas get_dummies function turning the level of all values into binary pairs. The sample was then split into training and testing data, using 70 percent to train. The training data was fitted on the following models, with the default parameters, from Scitkit Learn:
K-Nearest Neighbors (KNN); Logistic Regression; Decision Trees, Random Forest; and Gradient Boosting.

The 30 percent of sample test data was used for prediction. A performance evaluation for each model was composed of a confusion matrix ([true positive, false positive], [false negative, true negative]) and classification report, both functions were from sklearn.metrics.

KNN -max true neg

```
Matrix [[10057  1090]
        [ 1512  1994]]
              precision    recall  f1-score   support
           0       0.87      0.90      0.89     11147
           1       0.65      0.57      0.61      3506
    accuracy                           0.82     14653
```

Logistic Regression **AUC = .8863**

```
Matrix [[10328   819]
        [ 1562  1944]]
              precision    recall  f1-score   support
           0       0.87      0.93      0.90     11147
           1       0.70      0.55      0.62      3506
    accuracy                           0.84     14653
```

Decision tree

```
Matrix [[9676 1471]
        [1659 1847]]
              precision    recall  f1-score   support
           0       0.85      0.87      0.86     11147
           1       0.56      0.53      0.54      3506
    accuracy                           0.79     14653
```

Random Forest

```
Matrix [[10096  1051]
        [ 1674  1832]]
              precision    recall  f1-score   support
           0       0.86      0.91      0.88     11147
           1       0.64      0.52      0.57      3506
    accuracy                           0.81     14653
```

Gradient Boost -max true pos **AUC = .8936**

```
Matrix [[10413   734]
        [ 1561  1945]]
              precision    recall  f1-score   support
           0       0.87      0.93      0.90     11147
           1       0.73      0.55      0.63      3506
    accuracy                           0.84     14653
```

The Area Under (the ROC) Curve (AUC) for the two top performers are .886 for the Logistic Regression model and .894 for the Gradient Boosting model.

Prediction on Limited Attributes

Classifications were repeated for selected features on the two preferred models, top performing Gradient Boosting and second, Logistic Regression. In keeping with privacy considerations and emphasis on education, the subset of four predictors include: education; hours per week; work class; and age. The same train-test split was created as on the full set of attributes.

For visual simplicity's sake, a Pandas scattermatrix was created to include age, education and hours per week. The scattermatrix confirms the lack of correlation between education and hours per week. The contribution of the age attribute, as expected, was highly overlapping when ages and education levels are low and sparse when ages and hours worked levels are high.

These two models, Gradient Boosting and Logistic Regression, perform roughly the same at .84 f1-scores on the full set of attributes. The AUC scores only gave a narrow edge to Gradient Boosting. Logistic Regression on the subset of attributes has an accuracy score of .79 as compared to Gradient Boosting with .80, perhaps reflecting that slight edge in performance on the whole set of attributes.

More research

The parameters for Logistic Regression and Gradient Boosting could be tuned for better performance. Also, classification modeling on data at 10- to 12-year increments to explore whether the education levels and sector profiles changed for the income threshold similar to this one could provide more information for workforce development policy.

Presentation – Slide deck in GitHub repository – Income_Outcomes