Rosalie Day – In-Depth Analysis Capstone 1
Prediction of Income from Education and Sector Status
Jaunary, 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 Adult Census Income on the Kaggle website at https://www.kaggle.com/uciml/adult-census-income.

Prior Data Exploration

Relevant data exploration includes counts for education levels, ranging from zero formal education to doctorate degree, compared to income outcomes and work-class. High school graduates comprised the largest level in almost all work class, employment sector and status, categories. The "some college" was second largest followed by bachelor's degrees.

The inferential statistical analysis included hypothesis testing that determined: there was a positive outcome in income from having only a bachelor's degree in contrast with have no bachelor's degree; education level and workclass are correlated; and education and hours worked per week were independent.

Classification Prediction

The predictor variables were selected and data was preprocessed with the Pandas get_dummies function. 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 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.

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

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 is as expected, highly overlapping when ages and education levels are low and sparse when ages and hours worked level 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 the 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 the 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.