Rosalie Day – Milestone Report, Capstone 1
Prediction of Income from Education and Sector Status
December, 2019

Problem statement: Why it's a useful question to answer and for whom (source this from your proposal)

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 analysis is intended to describe education levels and employment sector and status associated with incomes, specifically at some threshold of income presumably a "good job" provides a living wage.

Initial Findings Summary

This study explores the predictors of whether a person's income was less or greater than $50,000, a selected "good job" threshold corresponding to a 1994 dataset. The focus is on education levels and employment sector and status as predictors, a departure from current focus on gender and race, which have been widely considered.

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" was second largest, followed by bachelor's degrees across work class categories. The three largest work class categories were private sector at 69 percent and dropping far below public sector 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 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.

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 working data set included 10 predictor variables: age, workclass, education/educational-num, marital-status, occupation, relationship (to family), race, gender, and hours-per-week, native-country; and the income target variable.

There were few modifications 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 variables 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 income (previously explained) and work class. The work class variable includes sector information, public, private, self-employed and an 'other' level which may be non-profit. Race, gender, relationship (to family) and marital-information were not explored, alt hough these will be used for classification of income outcomes.

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.

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 = -52.408681619739234 and a p-value = 0.0. From this comparison of income outcomes, the probability is 0.0 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 variables. A cross-tabulation of education and workclass was created from crosstab function in pandas. The chi-squared statistic = 3906.1080264021853 with degrees of freedom = 120 and p-value = 0.0 Therefore, the null hypothesis that these two variables are independent can be rejected. Workclass, which is a combination of work-sector and -status, and level of education are correlated.

Third, if the hours worked were correlated positively, 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 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.

Presentation – Slide deck in GitHub folder

Go by snippets

```
#Answer - top 10 countries with most projects - grouping by country
wbproj_df.groupby(['countryshortname']).count()['_id'].sort_values(ascending=False)[:10]

#ALT - top 10 countries with most projects - by value counts
top10countries= wbproj_df['countryshortname'].value_counts()[:10]
print(top10countries)
```

ERRORS – key error 'educational-num'
Create dataframe of education (labels) and sorted educational-num variables

```
#create a names map by dropping missing values then dropping missing duplicates
names_map = adult1.education.drop_duplicates()
names_map
#create a dictionary from the zipped columns, key= code: value= name
themes_dict= dict(zip(names_map['educational-num'], names_map.education))
themes_dict


#flatten lists of dicts
theme_list = [dict for sublist in themes for dict in sublist]
```

Chunk of mapping names and filling in missing names

```
#create a names map by dropping missing values then dropping missing duplicates
names_map = themes_df[themes_df.name != ''].drop_duplicates()
names_map
#create a dictionary from the zipped columns, key= code: value= name
themes_dict= dict(zip(names_map.code, names_map.name))
themes_dict
#Assign dataframe for completing with missing names
complete_themes_df = themes_df
#Add column derived from custom function and filling in
complete_themes_df['clean_name'] = complete_themes_df.apply(lambda row: themes_dict[row.code]
if row['name'] == '' else row['name'], axis=1)
```